The Anatomy of the Daily Usage of Bike Sharing Systems: Elevation, Distance and Seasonality

Injung Kim
University of Pittsburgh
Pittsburgh, Pennsylvania
injungkim@cs.pitt.edu

Konstantinos Pelechrinis
University of Pittsburgh
Pittsburgh, Pennsylvania
kpele@pitt.edu

Adam J. Lee
University of Pittsburgh
Pittsburgh, Pennsylvania
adamlee@cs.pitt.edu

ABSTRACT

Bike sharing systems have been in place for several years in many urban areas as alternative and sustainable means of transportation. Bicycle usage heavily depends on the available infrastructure (e.g., protected bike lanes), but other—mutable or immutable—environmental characteristics of a city can influence the adoption of the system from its dwellers. Hence, it is important to understand how these factors influence people's decisions of whether to use a bike system or not. In this this paper, we first investigate how altitude variation influences the usage of the bike sharing system in Pittsburgh. Using trip data from the system, and controlling for a number of other potential confounding factors, we formulate the problem as a classification problem, develop a framework to enable prediction using Poisson regression, and find that there is a negative correlation between the altitude difference and the number of trips between two stations (fewer trips between stations with larger altitude difference). We further, discuss how the results of our analysis can be used to inform decision making during the design and operation of bike sharing systems.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; **Model development and analysis**.

KEYWORDS

bike sharing systems, altitude difference, pairwise station trips, interactive analysis, Poisson regression

ACM Reference Format:
Injung Kim, Konstantinos Pelechrinis, and Adam J. Lee. 2020. The Anatomy of the Daily Usage of Bike Sharing Systems: Elevation, Distance and Seasonality. In *ACM SIGKDD urbcomp 2020, August 24, 2020, SanDiego, CA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

As cities are trying to build a sustainable and resilient environment for the future, bike sharing systems have become an important part of the urban transportation landscape during the past decade. The growth of bike sharing systems aims at encouraging sustainable transportation and reducing congestion in cities. However, the design of the system itself can attract or push away potential riders. Therefore, it is crucial to obtain actionable insights that can help design and operate *better* bike sharing systems.

More specifically, the presence of bicycling infrastructure (e.g., protected bike lanes, cycling-friendly street-scape, etc.) is crucial for riders to use the system [2, 3]. However, other—potentially immutable—environmental and topographic characteristics of the city can impact the number of bicyclists. For example, in a sprawling, city riders will inevitably have to ride longer distances to reach their destination. Adding on top of that things like weather effects, these factors can hinder the growth of cycling in a city. Recent studies have shown that the weather, capacity of bike stations, walkability, and job accessibility have a significant effect on the bike usage (e.g., [4–7]).

In this study, we focus on a particular environmental feature that can impact ridership, namely, elevation. It makes intuitive sense that riders might not be willing to ride (steep) uphill. For instance, Seattle's bike sharing system had to deal with the problem of everyone riding downhill, but no one returning bikes uphill [1]. An analysis of Montreal's "Bixi" system [6] examined how the elevation of the station correlates with the usage of the station (in terms of bike drop-offs), and identified that stations at higher elevations see lower usage. While this result points to elevation being negatively correlated with bike usage, these results focus on a specific station. In contrast, we are interested in pairwise interactions between stations i and j, and how the **elevation difference** between them relates with the number of trips from i to j and vice versa. This allows us to examine the relation between elevation and ridership in a finer granularity.

More specifically, we build a Poisson regression model for the number of daily (directional) trips between two stations of the bike share system in the city of Pittsburgh ("Healthy Ride" - HR for short) and an interactive analysis among factors such as elevation difference, distance, and seasonality. One of the benefits of using data from the system in Pittsburgh is that the city exhibits a fairly significant elevation variation, as compared to other cities that have a fairly flat surface such as New York, Chicago, Philadelphia, Washington D.C., Boston, or Minneapolis. For example, from an analysis of Chicago's bikeshare system usage [7], while initially considered the elevation as an independent variable, the final model excluded the variable due to small variation across the system's station. Figure 1 depicts the HR stations and the relative altitude map, where we can see there are non-trivial changes across different areas of the city.

ACM SIGKDD urbcomp, 2020, August 24, 2020, SanDiego, CA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

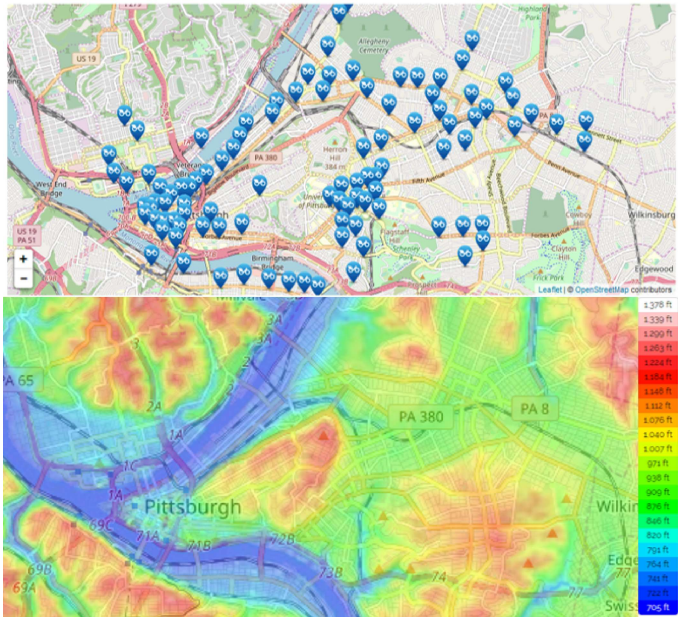

Figure 1: Pittsburgh "Healthy Ride" stations in June 2019 and its relative altitude map.

Our analysis clearly indicates that a higher altitude difference between the origin and destination stations, is associated with lower daily trips. The contributions of our work are thus, threefold:

(i) We provide additional evidence to the bikesharing literature for the negative correlation between altitude and bike sharing usage, by focusing on altitude *differences* in pairwise station trips. This is an important aspect of the scientific process with regards to result reproducibility and generalizability.

(ii) We build and evaluate predictive models, in contrast to existing literature in the area that is mainly focused on descriptive models [6, 7]. This is important, since predictive models can facilitate planning operations and examination of "what-if" scenarios. For example, identifying stations that might be *suffering* from lack of trips due to elevation, or pairs of stations that are expected to see highly asymmetric trips, will help to make decisions on a number of fronts including the decision of whether to invest on e-bikes, or informing the design of rebalancing schemes.

(iii) Given the type of model we choose (i.e., Poisson regression), we examine the probabilistic interpretation of the predictions, something absent from the existing literature.

The rest of this paper is organized as follows: in Section 2, we discuss related to our study literature, while we describe the dataset and experimental setup used in this study in Section 3. We also present some descriptive analysis of the bike sharing system trips, while Section 4 describes our Poisson regression model for the daily trips between two stations. Finally, Section 5 concludes our work and future work.

2 RELATED WORK

A large volume of research on bike sharing systems deals with the problem of rebalancing. Rebalancing refers to the operation of transferring bikes from docks with large numbers of bikes, to docks with fewer bikes available. This is a large operational cost for the system (approximately 20-25% of the total OPEX) and hence, any improvement is beneficial. There are several approaches that have been proposed to study and solve this problem, e.g., [8–10, 17] (with the list of course being non-exhaustive). More recently, operators have started moving away from dock-based systems and building dockless bikesharing that does not require docking stations. These systems provide more *flexibility* in terms of returning the bike to a specific location, but at the same time can create problems when it comes to finding a bike. Hence, part of the literature on dockless systems deals with the detection of *parking hotspots* for bikes [11, 12].

Closer to our work, a number of studies have explored the connection between the system's usage and other external factors. For example, in [5, 9] the authors show that the weather, walkability, and job accessibility have a significant effect on bike usage. The authors in [4] compare different regression models for predicting bike availability at a dock, while the authors in [22] take a clustering approach in predicting usage. A number of *urban environment* features, such as station density, capacity, points of interest, population, and housing units are correlated with demand [6, 7, 10, 18, 25]. Some of these studies also examine the elevation of the stations, and identify a negative correlation with the station's usage. However, as previously mentioned, the majority of these studies is focused on individual stations (rather than interactions between pairwise stations), and are performed on cities with small elevation variation to begin. Our work supports prior research by uncovering similar negative relationships between elevation and ridership, but extends the literature by focusing on altitude *differences* between stations comprising a trip rather than focusing on the absolute altitude of individual stations.

3 DATA AND EXPERIMENTAL SETUP

In order to perform our analysis, we collected and analyzed data from various sources. More specifically, we used the following data:

HR Trips: HR provides information about every trip taken on the system. Every trip is represented by a tuple with the following format: <date, start time, end time, origin station ID, destination station ID>. Our data spans a period of about 4 years, between 05-31-2015 and 06-29-2019, and includes a total of 76 million trips recorded over 112 stations. We also have additional information for the stations including their location (latitude, longitude) as well as their capacity (number of docks). Table 1 explains the details.

MapQuest bike paths: In order to calculate the distance between two bike stations, we do not use the Haversine distance between the two stations, but we rather take into consideration the (bike) street network. We use MapQuest's API and identify the biking distance between every pair of stations in the system.

Elevation: We get the station's altitude information from http://freemaptools.com by using station's latitude and longitude information.

Weather data: We obtained weather data from https://www.usclimatedata.com, which provides information about the high and low temperature recorded and precipitation levels for each day.

HR trip dataset	date	trip date
	start time	bike rental time
	end time	bike return time
	origin station ID	bike rental station
	destination station ID	bike return station
HR station dataset	station ID	station unique ID
	station name	station street name
	latitude	station's latitude
	longitude	station's longitude
	capacity	number of docks

Table 1: Healthy Ride dataset

Our analysis is split into two parts. We begin by performing a descriptive analysis of the trips during the span of the four years that our data covers. This allows us to better understand the patterns in the trips and choose appropriate variables for the modeling of trips between two stations. In the second part of our analysis, we build a Poisson regression model for estimating the number of trips between specific pairs of stations. For training and evaluating this model, we only use information from the last year covered from our data for a variety of reasons. For instance, riding patterns from four years ago might not be representative of the way HR riders use the system today (e.g., they might not be as hesitant taking longer trips today as they were in the beginning of the system). Furthermore, the system has been updated during these four years (new stations, stations moved etc.) and hence, we also wanted to focus on a period of time without significant changes to the network.

3.1 Descriptive Analysis of Daily Bike Trips

We begin by performing some basic, exploratory analysis to understand the trip patterns during the first four years of HR operations better. In particular, we explore the time-series of daily trips, as well as trip characteristics like altitude difference and distance covered.

We first examine the basic temporal dynamics of the number of daily trips observed in the system. Figure 2 (top) depicts the time-series, where we can see, as one might have expected, that there is a seasonality with the system's usage, with more trips happening during warm weather seasons. The middle part of the same figure depicts the average altitude difference for all the trips that happened in a day. As we can see most days exhibit a negative value, which means that users tend to ride the bikes downhill more than uphill. It seems that there are some uphill trips in every winter season, however, it is actually shown due to the average daily trips' altitude difference over the small number of daily trips. We omit the uphill trips analysis, but in uphill trips over the year, there are more uphill trips in the summer season than the winter. The bottom part of the figure is the average of trip distance covered per day. As we can see, this average is stable over time, and less than 2 miles, supporting the idea that biking is a good alternative for short trips [20].

Central to our study is the relationship between altitude differences between the origin and destination of a trip and the volume of such trips. Figure 3 (top) presents the distribution of the station altitude difference of the trips. As we can see, the majority of trips happen between stations at about the same altitude, with smaller

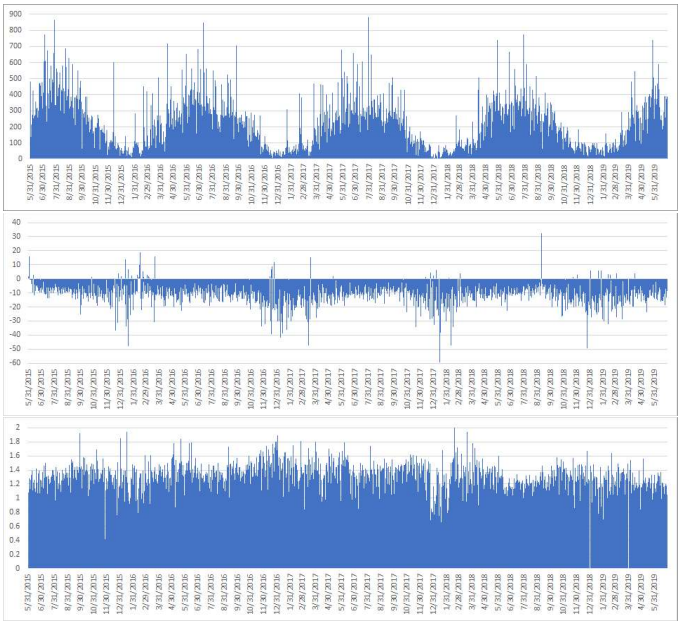

Figure 2: The number of daily trips (top), average daily trips' altitude difference in feet (middle), and average daily trips' distance in miles (bottom) to show the bicycle patterns for four years

numbers of downhill (left) and uphill (right) trips. This points to the absence of a linear correlation between the altitude difference and the number of trips at that altitude difference. As we will elaborate on in the following sections, discretizing this feature leads to better performance for our prediction models.

We also examine the distance of trips. Figure 3 (bottom) presents the distribution of the trip distances, which shows—as one might have expected—that the majority of the trips (77.9%) are less than 2 miles in distance. It also worth noting that for these distributions we have removed the trips that start and end at the same station, since even though the nominal distance is 0, this does not mean that the trip was of actual length 0.

4 PREDICTIVE MODEL FOR DAILY BIKE TRIPS

In order to better understand how the various variables relate with the number of daily trips observed between two stations, we build an (interpretable) prediction model. We will use a Poisson regression model, where the dependent variable is the number of daily trips between a (directional) pair of stations, and we perform feature selection through a validation set.

4.1 Prediction Features

As our predictors, we use three different types of features:

- **Time-related**: These are features that capture dependencies from *temporal* variables, and in particular, the day of the week and the season of the year.

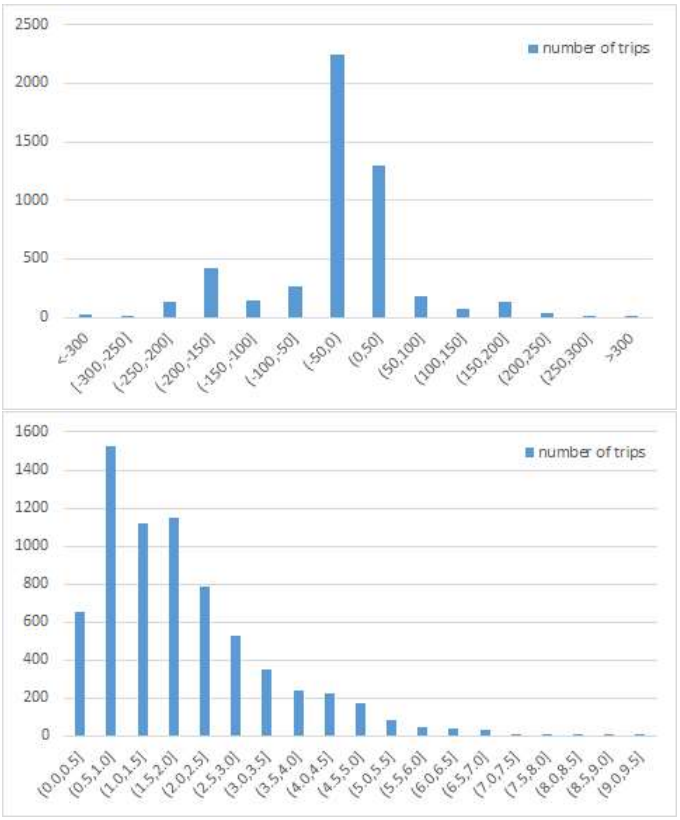

Figure 3: The number of bicycle trips of various altitude difference range in feet (top) and trip distance range in mile (bottom) in June 2019.

- **Weather-related**: These are features that capture information relevant to the weather conditions during a day. These include, highest daily temperature, lowest daily temperature, and precipitation during the day.
- **Urban environment-related**: These are features that capture information about the built environment and the urban landscape. The features we use here are the number of racks at the origin and destination station, the trip altitude difference, and the trip distance;

For building our models we rely on data from the last year of operation of the system. We make this decision to avoid issues with changing patterns over the 4 years of operations of the system[1], which are further pronounced by changes in the system itself. Focusing only on the most recent year of operation eliminates these issues. In particular, we use data from February 1, 2019 to February 28, 2019 (winter data) and May 30, 2019 to June 29, 2019 (summer data). These two periods correspond to distinct usage patterns for the system, which should allow us to understand the relationships between the daily usage and the independent variables better.

For the altitude difference, we will explore two different approaches; (i) a continuous variable for the difference between the origin and destination station, and (ii) a categorical variable that indicates the range of the altitude difference. We use the following ranges: < -200, [-200,-100), [-100,0), [0,100), [100,200), ≥ 200 feet, marked as -300, -200, -100, 100, 200, 300 in the model. As mentioned in the previous section, a range variable might be better at capturing non-linearities between altitude difference and daily usage. Part of the feature selection is then related to choosing which representation is better.

The trip distance is a continuous variable obtained through MapQuest's suggested bicycle route (i.e., we do not simply use the distance as the crow flies). We should note here that the (bike) distance between station A to B is usually different from the distance from station B to station A due to the structure of the street network. Even though based on the results from Fig. 3 there seems to be a linear relationship between distance and daily usage among two stations, we examine a binned variable for distance as well. We use the following ranges: [0,2), [2,4), [4,6), ≥ 6 mile in the feature selection in Table 3. Our model also includes an interaction term between the trip distance and the altitude difference variable.

We also include in our features a binary variable that indicates whether the origin and destination stations are the same. The rationale behind this is that for trips with the same origin and destination stations, while the nominal distance and altitude difference is 0, these are plasmatic differences (e.g., the actual distance covered from the rider is certainly > 0). Inclusion of this variable or not will again be driven from the feature selection. Table 2 summarizes the features considered when building our model.

Variable	Description
DOW	seven variables for day of the week
Separate Season	each summer and winter seasonal dataset
Combined Season	combined summer and winter seasonal dataset
Same Origin-Destination	one variable for marking same origin and destination station trips
Dis * Alti_diff (both range)	interaction variable between distance and altitude difference in range value
Dis * Alti_diff (both real)	interaction variable between distance and altitude difference in real value
Dis (real) * Alti_diff (range)	interaction variable between distance in real value and altitude difference in range value
Dis (range) * Alti_diff (real)	interaction variable between distance in range value and altitude difference in real value
High Temp	one variable for highest temperature in Fahrenheit per day
Low Temp	one variable for lowest temperature in Fahrenheit per day
Precipitation	one variable for precipitation inch per day
O_Station_Rack	one variable for the number of racks at origin station
D_Station_Rack	one variable for the number of racks at destination station

Table 2: Set of independent variables considered.

[1]Even though identifying the temporal differences in these patterns is themselves interesting.

Features	FS1	FS2	FS3	FS4	FS5	FS6	FS7	FS8	FS9	FS10
DOW	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Separate Season	✓		✓	✓	✓	✓		✓	✓	✓
Combined Season		✓					✓			
Same Origin-Destination						✓	✓	✓	✓	✓
Dis * Alti_diff (both range)	✓	✓				✓	✓			
Dis * Alti_diff (both real)			✓					✓		
Dis (real) * Alti_diff (range)				✓					✓	
Dis (range) * Alti_diff (real)					✓					✓
High Temp	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Low Temp	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Precipitation	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
O_Station_Rack	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
D_Station_Rack	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Train set	0.18173	0.2004	0.17839	0.17725	0.1827	0.17588	0.2004	0.17626	0.17588	0.17636
Validation set	0.17503	0.1951	0.17177	0.17062	0.17591	0.16931	0.1951	0.16963	**0.1693**	0.16969
Test set									0.16736	

$x1 * x2$ is an interaction term in regression.

Table 3: Feature selection results.

4.2 Poisson Regression

Our goal is to model the number of daily trips Y_{ij} from station i to station j. Given that our dependent variable Y_{ij} is a non-negative integer, a linear regression is not an appropriate model. Hence, we choose to use a Poisson regression, where essentially the data is assumed to follow a Poisson distribution, i.e., $Y_{ij} \sim Pois(\lambda_Y)$. A Poisson regression essentially models the average rate of the dependent variable through a linear combination of a set of independent variables $\mathbf{X}$ as:

$$\lambda_Y = e^{\alpha + (\mathbf{b} \cdot \mathbf{X})} \tag{1}$$

The parameters α and $\mathbf{b}$ are obtained through maximum likelihood estimation and we can thus, estimate the distribution for Y_{ij} as:

$$p(Y_{ij} = k | \mathbf{X}, \mathbf{b}, \alpha) = \frac{e^{k \cdot (\alpha + (\mathbf{b} \cdot \mathbf{X}))}}{k!} \cdot e^{-e^{\alpha + (\mathbf{b} \cdot \mathbf{X})}} \tag{2}$$

4.3 Feature Selection

In order to perform feature selection, we randomly split the dataset into three sets: train, validation, and test, at a 7:2:1 ratio.

We examine ten different models, and Table 3 presents the features used in each model. Note that the second and third features do not correspond to a specific feature but rather on the way the model is trained. In particular, "separate season" corresponds to two different models being trained, one on summer data only and one on winter data only. A "combined season" model uses all the data for training, and includes a categorical variable to indicate the season. For the cases where separate models for summer and winter are trained, we calculate the average of root mean square error (RMSE) weighted by the number of data points from the corresponding season. This setting ensures that all models are trained, validated and tested on exactly the same data.

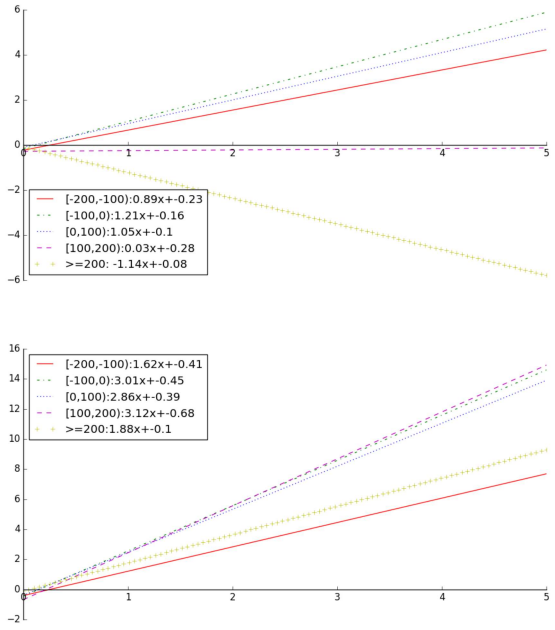

Figure 4: Figure of the total coefficient/effect for each range of altitude difference in summer (top) and winter (bottom). $a + bx$ where x is the distance, a and b are given coefficients

Based on the results on the validation set, FS 9 exhibits the lowest RMSE, and hence, our analysis focuses on this model. FS 9 uses a range variable for altitude (as opposed to a continuous variable) and trains separate models for summer and winter. Table 4 shows the models obtained for the two seasons.

Dependent variable: daily bike usage	coef_summer	coef_winter
Intercept	-6.0234***	-9.0476***
C(dayofweek)[T.Monday]	-0.3787***	-0.0154
C(dayofweek)[T.Saturday]	0.1470***	-0.3716***
C(dayofweek)[T.Sunday]	-0.1828***	-0.2233*
C(dayofweek)[T.Thursday]	-0.3819***	0.2581***
C(dayofweek)[T.Tuesday]	-0.0228	0.1237
C(dayofweek)[T.Wednesday]	-0.2748***	0.0924
C(alti_range)[T.-200]	0.8861***	1.6210
C(alti_range)[T.-100]	1.2130***	3.0086***
C(alti_range)[T.100]	1.0517***	2.8639***
C(alti_range)[T.200]	0.0321	3.1160***
C(alti_range)[T.300]	-1.1421*	1.8787*
C(Same Origin-Destination)[T.1]	1.8198***	1.0229***
high_temp	0.0059**	0.0225***
low_temp	0.0013	0.0045
precipitation	-0.1179***	-0.1664***
o_station_rack	0.0724***	0.0516***
d_station_rack	0.0893***	0.0678***
distance	-0.3859***	-0.2652
distance:C(alti_range)[T.-200]	-0.2326***	-0.4117
distance:C(alti_range)[T.-100]	-0.1596**	-0.4466*
distance:C(alti_range)[T.100]	-0.0986	-0.3863
distance:C(alti_range)[T.200]	-0.2766***	-0.6793**
distance:C(alti_range)[T.300]	-0.0814	-0.1017

Significance codes: $*p < 0.05, **p < 0.01, ***p < 0.001$

Table 4: The FS 9 Poisson regression models for the daily bike usage in summer and winter.

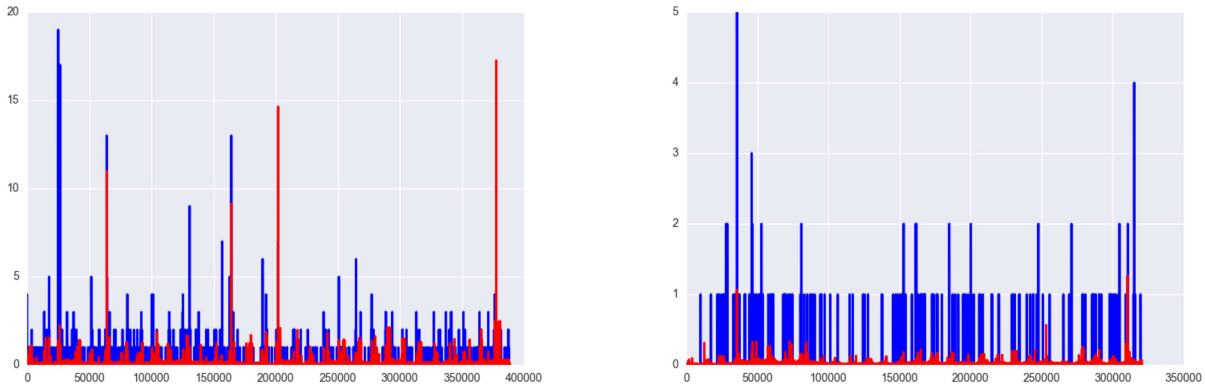

Figure 5: FS 9: actual (blue) and predicted (red) values for the test data points for summer (left) and winter (right) where X is a daily station pair trip indexed by smaller station id and earlier date (i.e., 135,213 is a trip from station id 49,401 to 1,028 on 6/9/2019) and Y is the number of trips.

4.4 Results

Table 4 presents the results from our selected model. As we can see the majority of the features have the same direction for their effect (i.e., positive or negative) in both the summer and winter seasons. For example, precipitation is negatively correlated with usage in both seasons. However, the effect size is different. Furthermore, large distances between origin and destination stations are associated with smaller number of trips.

Given the interaction term between distance and altitude difference, in order to interpret the relationship between distance and

daily usage we need to know the altitude difference as well. For example, if altitude difference is between -200 and -100 feet, then the total coefficient for distance is -0.39-0.23=-0.62. However, for most differences in altitude the corresponding total coefficient are going to be positive. Similar interpretation should be made for the altitude difference effect. From the "pure" coefficients for the altitude difference we can see that as the altitude difference approaches 0 we have the largest positive impact on daily usage, while the effect reduces as we move to downhill or uphill trips. Furthermore, uphill trips do not only have lower coefficients, but this is also negative (similar to the shape of the distribution at Figure 3). Again, in order to quantify the total effect of the altitude difference we would need to know the distance of the trip. For example, for an average distance trip of 1.8 miles, the total coefficient/effect for the altitude difference would be: $0.89+1.8\cdot(-0.23) = 0.47$ for trips with altitude difference between -200 and -100 feet, $1.21 + 1.8 \cdot (-0.16) = 0.92$ for trips with altitude difference between -100 and 0 feet, and so on. Figure. 4 shows the total coefficient/effect for each range of altitude difference in summer and winter, and the total coefficient is essentially a line $a+bx$, where x is the distance, a and b are given coefficients.

As mentioned earlier, we evaluate the selected model using the RMSE achieved over the test set, which is approximately 0.16. Figure 5 depicts the performance on each data point on the test set. In particular, every point on the x-axis corresponds to a test point, with the blue "bar" corresponding to the actual number of trips observed for this data point, and the red "bar" representing the predicted value from our model. As we can see, during the summer the daily number of trips is larger on average as compared to the winter and so are the predictions from our model. Furthermore, when comparing the performance of our model during the different seasons, we find that the RMSE of the model on the winter test set is smaller (0.0881), as compared to that of the summer (0.2327). This can be attributed to the fact that there is less variability for the trips over the winter, since the majority of the winter trips will not be recreational (which are typically less predictable).

One benefit of our analysis is the ability to obtain an actual distribution for the predicted number of trips between two stations. In particular, if our prediction for the number of trips is y_n, then the actual distribution for the number of trips is a Poisson distribution with mean $\lambda = y_n$. Hence, one can also examine the probability π_{pred} assigned by our model to the value of the actual trips observed. For example, if our prediction is 2.4 trips, while the observed number of trips was 4, then our prediction provided a 0.125 probability to this outcome ($pmf_{Poisson}(k = 4, \lambda = 2.4) = 0.125$). While this might seem low, the same prediction provides its maximum probability to the value of 2, which is equal to 0.26 - i.e., $\max_k pmf_{Poisson}(k, \lambda = 2.4) = 0.26$, for $k = 2$.

4.5 Discussion and Limitations

Discussion. To the best of our knowledge, our analysis is the first to consider the *relative* altitude difference between the origin and destination stations in bike sharing systems as a factor when predicting trip likelihood. As shown in Section 4.4, this was a significant factor in predicting trips between stations. Prior work considering only absolute station altitude could easily overcompensate for the effect of altitude when looking at trips in areas of the city with little altitude variation, whereas our work accounts for this explicitly. In terms of future work, it would be interesting to investigate whether health-related 'microtasks' could exploit this negative correlation to help with bicycle rebalancing in small local areas. Phrased another way, can individuals be incentivized to take short breaks (e.g., 15 minutes) to ride a bike uphill from one station to another nearby (but higher altitude) station and walk back to their home or workplace? This would have positive health benefits for individuals, positive benefits for the bike sharing system in terms of bike availability, and is highly applicable in cities like Pittsburgh with high variance in altitude over small distances.

Our daily bike usage predictive models are built and evaluated by narrowing down the features and analyzing in an interactive way among the features. Unlike other descriptive model works [6, 7], our predictive models can facilitate planning operations and examination. Not just improving urban bike sharing system, but also resolving one transportation problem with other transportation methods is also possible through our prediction. For example, city planner can build more bus lines at uphill, reduce the lines at the short distance downhill, and expand the bike station docks at downhill stations, then it can ultimately reduce the bicycle rebalancing cost.

Limitations. Given that we are interested in the elevation changes during a trip, it would be ideal if we were able to identify all sources of altitude variation along the trip, and not simply the altitude difference between the origin and destination stations. Similarly, the slope of these changes would be an interesting feature to consider in an analysis such as ours. Unfortunately, this is not possible for two reasons. First, we do not know which path the riders followed between the origin and destination stations. Further, even if we assume that riders took the shortest bike route between the two stations, the services that we have used to identify bike paths (e.g., MapQuest) do not provide path altitude information. While Google Maps visualizes altitude variation along a route, the available APIs do not provide access to this data.

In addition, our analysis considered only one city. It would be interesting to carry out our analysis using bike sharing data from other hilly cities (e.g., San Francisco) as well as flatter cities (e.g., Chicago). Furthermore, extending our analysis with a wider feature set inclusive of factors such as job density, population, nearby bus stations, likely improves the performance of our models, as well as allows us to more fully understand the importance of relative altitude as compared to, say, access to other forms of public transportation.

5 CONCLUSION

In this paper, we explore the daily usage in terms of number of trips between a (directional) pair of stations for the bike share system in Pittsburgh. We use a Poisson regression model and perform feature selection through a validation set. Our results indicate that the altitude difference, station distance, as well as, weather features impact the usage of the system. This adds evidence to an existing body of literature on the impact of elevation on bike usage.

As part of our future work, we intend to explore several directions in improving the predictive power of our model. For example, we have used Poisson model, but a negative binomial might be a better fit for the trip data. Furthermore, regularization can further help improve the predictive performance. We also plan to extend our study to cover other cities (both in the US and outside the states) in order to make direct comparisons and obtain more generalizable conclusions.

ACKNOWLEDGMENTS

This work is part of the PittSmartLiving project which is supported by NSF award CNS-1739413.

REFERENCES

[1] "People Don't Like To Bike Uphill — And That's Where Pronto's 'Rebalance' Comes In", https://www.knkx.org/post/people-dont-bike-uphill-and-thats-where-prontos-rebalance-comes (last accessed, 18 May, 2020)
[2] "Here's what bike-sharing programs need to succeed", The Conversation, https://theconversation.com/heres-what-bike-sharing-programs-need-to-succeed-85969 (last accessed, 18 May, 2020)
[3] C. M. de Chardon, Geoffrey Caruso, and Isabelle Thomas. "Bicycle sharing system 'success' determinants". Transportation research part A: policy and practice 100 (2017): 202-214.
[4] J. Holmgren, S. Aspegren, and J. Dahlströma, "Prediction of bicycle counter data using regression", Procedia computer science, 113 (2017) 502-507.
[5] A. Nasri, H. Younes, and L. Zhang, "Multi-level urban form and bikesharing: insights from five bikeshare programs across the United States", Annual Meeting of the Transportation Research Board (TRB 2019)
[6] C. Morency, M. Trepanier, A. Paez, H. Verreault, and J. Faucher, "Modelling bikeshareing usage in Montreal over 6 years", CIRRELT-2017-33, 2017
[7] A. Faghih-Imani, and N. Eluru, "Analysing destination choice preferences in bicycle sharing systems: an investigation of chicago's Divvy system", 94th Annual Meeting of the Transportation Research Board (TRB 2015)
[8] P. McMullen, "A markov simulation approach to balancing bike-sharing systems", American Journal of Operational Research, 2019, 9(1): 12-17
[9] P. Hulot, D. Aloise, and S. Jena, "Towards station-level demand prediction for effective rebalancing in bike-sharing systems", KDD '18: The 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 19-23, 2018
[10] A. Faghih-Imani, R. Hampshire, L. Marla, and N. Eluru, "An empirical analysis of bike sharing usage and rebalancing: evidence from Barcelona and Seville", Transportation Research Part A: Policy and Practice, Volume 97, March 2017, Pages 177-191
[11] K. Zamir, I. Bondarenko, A. Nasri, S. Brodie, and K. Lucas, "Comparative analysis of user behavior of dock-based vs. dockless bikeshare and scootershare in Washington, D.C.", Annual Meeting of the Transportation Research Board (TRB 2019)
[12] Z. Liu, Y. shen, and Y. Zhu, "Where will dockless shared bikes be stacked? - parking hotspots detection in a New City", KDD '18: The 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 19-23, 2018
[13] A. Faghih-Imani and N. Eluru, "Role of Bicycle Sharing System Infrastructure on Usage: Evidence from Montreal", 5th Innovations in Travel Modeling Conference, 2014
[14] M. Oppermann, T. Moller, M. Sedlmair, "Bike sharing atlas: visual analysis of bike-sharing networks", International Journal of Transportation, 2017
[15] S. Graham, B. Dean, O. Rawashdeh, N. Dahl, and A. Simenauer, "Low cost bicycle share security solution for universities", Proceedings of the 2014 ASEE North Central Section Conference, 2014
[16] K. Pelechrinis, C. Zacharias, M. Kokkodis, and T. Lappas, "Economic impact and policy implications from urban shared transportation: The case of Pittsburgh's shared bike system", PLoS ONE 12(8): e0184092 (2017)
[17] R. Saltzman, and R. Bradford, "Simulating a more efficient bike sharing system", Journal of Supply Chain and Operations Management, Volume 14, Number 2, December 2016
[18] D. Shinghvi, S. Singhvi, P. Frazier, S. Henderson, E. O'mahony, D. Shmoys, and D. Woodard, "Predicting bike usage for New York city's bike sharing system", Computational Sustainability: the 2015 AAAI Workshop
[19] G. Wu, Y. Li, J. Bao, Y. Zheng, J. Ye, and J. Luo, "Human-centric urban transit evaluation and planning", 2018 IEEE International Conference on Data Mining (ICDM), Singapore, 2018, pp. 547-556
[20] McClintock, Hugh, ed. Planning for cycling: principles, practice and solutions for urban planners. Elsevier, 2002.
[21] T. Arabghalizi, and A. Labrinidis (2019), "How full will my next bus be? A Framework to Predict Bus Crowding Levels", 10.13140/RG.2.2.12969.75368, KDD urbcomp, 2019
[22] M. Yang, C. Chen, L. Wang, X. Yan, and L. Zhou, "Bus arrival time prediction using support vector machine with genetic algorithm.", Neural Network World. 26. 205-217. 10.14311/NNW.2016.26.011., 2016
[23] X. Ling, Z. Huang, C. Wang, F. Zhang, and P. Wang, "Prediction subway passenger flows under different traffic conditions", PLoS ONE 13(8):e0202707, 2018
[24] C. Ding, D. Wang, X. Ma, and H. Li, "Predicting short-term subway ridership and prioritizing its influential factors using gradient boosting decision trees", Sustainability, MDPI, Open Access Journal, vol. 8(11), pages 1-16, October 2016
[25] I. Mateo-Babiano, R. Bean, J. Corcoran, and D. Pojani, "How does our natural and built environment affect the use of bicycle sharing?", Transportation Research Part A: Policy and Practice, Volume 94, 2016, Pages 295-307