Statistical Analysis of Bike Sharing Data

Abstract

In this report we develop statistical analysis that can be applied to bike sharing data. The goal of this analysis is to 1: understand the nature of bike usage in a bike sharing system, from angles of users, the providers of the bike sharing system, and the city where the bike sharing system is situated, 2: define and quantify the performance of a bike sharing system, 3: suggest improvements to the system's performance, and 4. to develop our analysis into visualizations for the end user who can be the bike sharing system providers, the city, or the users of the bike sharing system. To stengthen our analysis we have used example data form online resources, in addition to the data provided by the bike sharing company.

Introduction

Several bike sharing systems have started recently in different parts of the world. Because it is geographically distributed, and used both as well as use by the bikers, it is hard to define a performance metric for a bike sharing system. We study the usage patterns across the users (bikers), and across the stations to define several consistent, quantitative metrics that measures the performance of the system. Such metrics will allow us to study the performance when some changes are made to the system, such as the addition or removal of a bike station.

We consider several viewpoints of the bike sharing system (BSS), which will be addressed as independent analysis of the data,

1. **Bird's eye perspective**: an analysis of the overall usage of the system
2. **User perspective**: an analysis of the user behavior
3. **Station perspective**: an analysis of the usage of bike stations
4. **Network flow perspective**: an analysis of the flow of the bicycles between stations.
5. **Geospatial perspective**: an analysis of the geospatial features of the BSS network
6. **Multimodal transport perspective**: an analysis of how the BSS is situated in the larger multi-modal transportation network of the city
7. **Operational perspective**: an analysis of the operations involved in running a BSS network
8. **Business perspective**: an analysis of how the usage translates into revenue

Below we summarize what we intend to do in each of these perspectives.

Bird's eye Perspective

Without going into the details of each trip, a bird's eye perspective should address the overall usage pattern of the BSS. This analysis will consist of the statistics of the aggregated usage. Questions that can be addressed include,

```
1. How has the usage grown over time?
2. What is the distribution of usage over time? We can consider the average and variability of usage  fo

     2a. the hour over a day,
     2b. by the weekday over a week,
     2c. by month over a year,
```

```
    2d. by season over a year.

3. How do registered users differ in their usage compared to casual users?
4. Is usage different on holidays? Or if their is a special event in the city?
5. How does usage depend on the season, and the weather?
6. How does usage vary between different bike sharing systems?
```

After an exploratory analysis of an example data sets, we will build a model that predicts the usage of a BSS by a specified time resolution. The model should predict the usage for a day given the season, working day or a weekend or a holiday or a special event, the weather – rainy or hot. A more detailed model should do the same by the hour. If it rains in the morning, should we expect more or less usage later in the day when it is dry?

User Perspective

Under user perspective we will analyze the features of the data that have a relevance on characterizing the users. Some features may be available directly in the data,

- user's susbcription type,
- age,
- user's start date
- if no longer a user, their end date

Others can be extracted from the data, * frequency of trips by season/weather/day of week/period of day * duration of a user's trips * intermittency period between a user's trips * length of period since the user's last trip, from a given time cycle's end * distribution of the user's start and stop stations.

We will discuss the distributions, and dependences of each derived statistics. Based on the directly available and derived features for each user, we will segment the users into clusters. This analyses will allow us to have an insight into the diversity of a BSS's users.

Station Perspective

Similar to the user perspective, under station perspective we will deal with the features of each station. Features available in the data are

- date when a station was commissioned
- if a station is still operational, and if not, the date it was decommissioned.
- the number of docks available at a station
- the number of bikes available at a staion, by the time of day

Features that can be extracted from the trips data are,

- frequency of trips for which the station is a trip's start point, by season/weather/day of week/period of day
- how unbalanced is the traffic at a station: the mismatch between outgoing and incoming trips to a station, by the hour, day, week, and month.
- mean and variability of intermittency of trips that start at a station
- mean and variability of intermittency of trips that end at a station

- mean and variability of durations of trips that start at a station
- mean and variability durations of trips that end at a station
- number of distinct users that use a station to start their trips at
- number of distinct users that use a station to end their trips at
- diversity (measured as entropy) of the number of a user's trips that start at a station
- diversity (measured as entropy) of the number of a user's trips that end at a station
- diversity (measured as entropy) of the end stations of trips that start at a station
- diversity (measured as entropy) of the start stations of trips that end at a station
- diversity of users (as observed in our user perspective analysis) that use a station

None of the station features listed are geo-spatial in nature. The purpose of station analyses without geospatial information is to focus only on understating how diverse the stations are given only their usage statistics. We will discuss the geo-spatial features of the stations in a separate analysis, and combine them with the features discussed in this section.

As for the user perspective, we will discuss and analyze the distributions, and dependences fo each derived statistics, and use the station features to cluster the stations into correlated groups.

Performance of a station

Performance of a station can be simply measured as the amount of usage traffic it draws. Similar to a predictive model for the usage of the entire BSS, we will make a model that predicts a station's performance by season, day of week, workday/weekend/holiday, or period of day.

How does the closing or opening of a station influence the performance of other stations?

How do special events influence a station's performance?

Network flow perspective

The station perspective considers single stations, either as start or end points of a trip. Statistics related to the pairs of start and end stations can be studied using a network of flows among the stations. We will construct a network, with each station as a node. A weighted directed edge station **A** to station **B** will indicate the number of trips that were made from **A** to **B**. To obtain a network that captures the overall usage of the BSS, we will use the number of trips over the entirety of time period available in the data. We can analyse properties of tis network,

- degree distribution will tell us the important nodes (stations) by amount of traffic
- assortativity between nodes will tell us the correlation of traffic between stations
- betweenness centrality will tell us the central nodes
- community structure will allow us to see the subnetworks within the bigger network.

In addition to the static network built using the data for all the entirety of the existence of a BSS, we can make time dependent networks and consider how its structure varies over time, between weekdays and weekends, between seasons and weather, or over time as the BSS has evolved.

Other ways of defining networks

Besides using the amount of traffic between two stations, we can also use the reciprocity of trips to define the strength of a link between two stations. For stations **A** and **B** if more traffic flows from **A** to **B** then the link $A \to B$ is +1 and the reverse link is -1. This network will allow us to study the stock balancing properties of the BSS as a whole. If we use the duration of trips between pairs of stations to define link strengths . . .

Contagion, and perturbations.

After studying the structure and time evolution of the networks, we can simulate the perturbations caused by a contagion of the form of a close down of a station. If the closing was because of an executive decision, we can see how that influences the yearly traffic at other stations. This analyses can inform executive decision about which stations to close, which one to split, or whether to open a new one. Along with geo-spatial information (discussed in a separate analyses and section) we can use network analyses to decide where to open new stations.

If the closing was because of an accident, or failure of the computer system at a station, we can see how the user's distribute over other stations in the system, and if the failure will choke the system.

Geospatial perspective

Where are the stations? Population density Catchment area, people would walk 500m to rent a bike. Accessibility of a station by its location Nature of the location of a station. Is it in residential area, commercial area, or close to a railway station? How do the geospatial properties of a station's location influence its usage, for example the features discussed under station perspective? NetKDE analysis

Multimodal transport perspective

Travel times b/w stations, correlate with geo-data walking bus car How does the multimodal nature of a city's transport affect the performance of a BSS's performance?

Operational Perspective

- Operational diagnostics and planning.
- An algorithm that gives the optimal end of day bike balancing route

Business Perspective

- Quantify selling channels.
- Accessibility to selling points: where can the user buy BSS tickets?
- How far are these from bike stations?
- Is there more traffic at stations that are close to BSS ticket sale points?
- How many BSS tickets sold at a given selling point?
- Category of sale by selling point.
- Can we suggest where selling points should be located?