

Bus Transport

Final Project Summary

Abstract: The project is on Indiana University Bus data and is intended to find the factors that affect the efficiency or the schedule of Indiana University Campus Bus system. IU campus shuttle is like a life line connecting the students and staff to all the corners of the campus and thus identifying any attributes that can improve the efficiency of campus shuttle is pivotal.

Data: The primary source of data was provided by IU Bus service. We also used data from Double Maps and Google Maps to calculate the distance between two stops and other characteristics of a particular route. To check the impact of external factors such as weather we also mapped the transport data with the weather data. The weather data was collected by using R package “weatherData” as it contained weather data in a more refined manner.

Final data is structured as persistent objects in such a way that it gives the details for each stops including dwell and transit time, average weather conditions per day, number of trips per route (a trip is an instance of a route), whether any particular stop was skipped or not, etc.

A summary of the steps in data processing is as follows:

- Division of the raw data into routes
- Division of routes into trips
- Creating internal reference index for each bus stop, in order to map sequence of stops
- Summary level data from R package “weatherData” using Monroe County Airport as station ID
- Creating customized variables like day of week, average transit time for each stop, label determining whether a bus is late or not based on calculated average time per stop

As a part of the requirement from the client we needed to understand not only the factors affecting the bus schedule, but also the average time at each bus stop. The data was preprocessed so that all factors could be incorporated in one file or table.

Process data for average time per stop:

The idea of average time was used to determine the average time required by a bus to travel from one stop to another bus stop and later comparing this average time with the actual time we can label whether that bus was on time or not i.e. time difference in seconds.

This label can later be used to find what are the characteristics that cause the bus to meet its schedule.

The start and stop time for each stop was derived from the time stamp provided in the data. The difference between the start and stop time gives the transit time for each stop. In order to determine genuine average time per stop there was a need to identify outliers, for this we clustered the data based on the time difference

and made 4 clusters. For each of the clusters average time was calculated. Then the Top 2 clusters, example cluster 1 and cluster 2 were selected which have highest and second highest frequency or cluster members, with a condition that the difference of frequency between these two clusters is not more than 25% of the largest cluster i.e. cluster 1.

If the difference is more than 25% we take only the mean of the highest frequency cluster, else the mean of the two cluster is taken, this helps us in removing the outliers in the data and helps us get a more refined version of average time for each stop.

Visualization:

The data from the transit was visualized for routes and in order to give the client more control or real time monitoring we included a metric which gives the efficiency or hit rate for each route (provided as a python script). Hit rate is measures how many time the bus was not late at a particular stop in reference to the average time for that bus stop.

Efficiency or Hit rate = Number of times the bus was not late for each stop/ Total number of stops covered

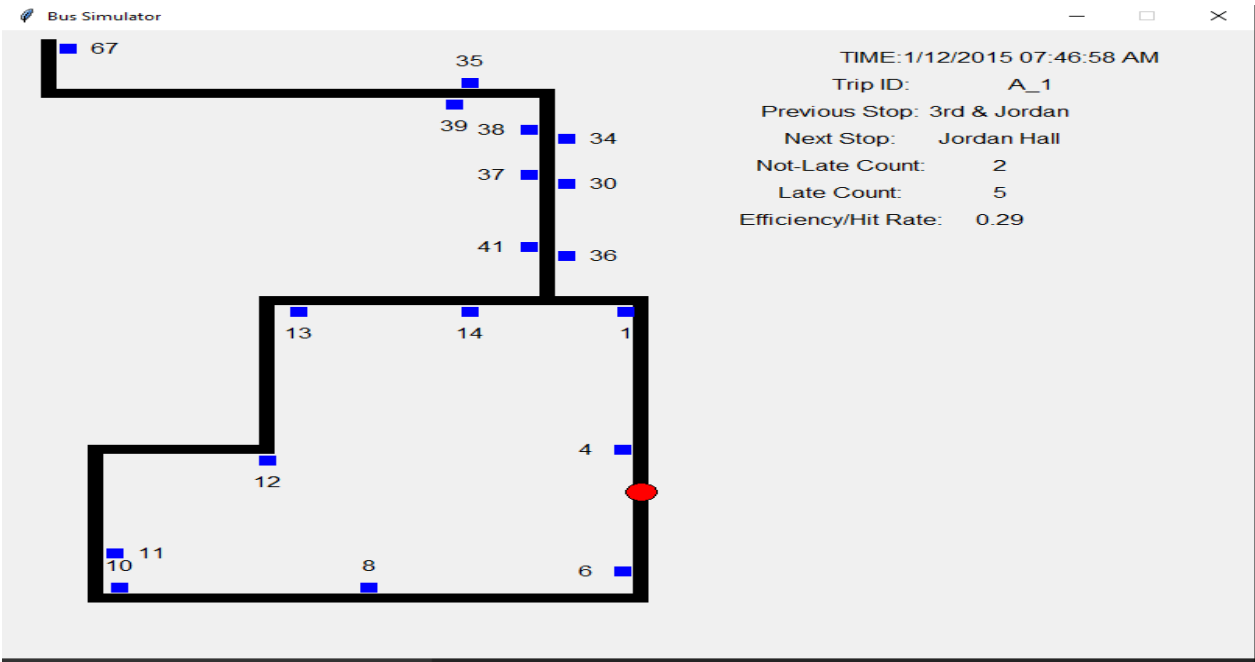


Figure 1.0

This metric can help the client to know what is the efficiency in terms of meeting the bus schedule for each stop and just not the major ones. This will also help the client identify issues in a particular bus stop.

Examining the data:

For the purpose of conciseness, we will discuss the results route “A” and then provide the same for other routes as a supplementary section.

- a) **Based on Hour of the day:** It can be noted that Hour of the day (HOD) is the primary factor that is common in identifying or predicting whether a given bus will be late or not. What we can notice is that, if the bus is operating in late night then the possibility of the bus being on time is high and the possibility reduces as the HOD is between morning 7 am to evening 8 pm (Note 1: This tendency is similar for all the routes). The below figure 2.0 gives a graphical overview of the above mentioned dependency.

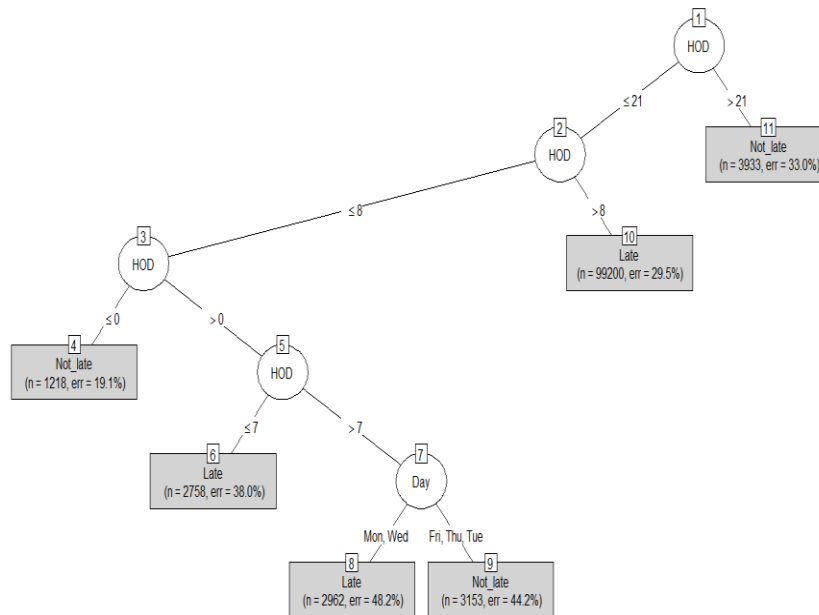


Figure 2.0

- b) **Based on weather:** it can also be noted that weather conditions also have an impact on the bus times. We took the number of times the bus was late and not late for a particular route and did a mapping this frequency on weather conditions and we observed the following (for base line we took the average percentage of the late and not late for the total number of records for a particular route).
- The combination of Fog-Rain-Snow affects the punctuality of the bus the most, when compared to the base line this combination give most numbers trips when the bus was late.
 - The weather factor combination when the bus punctuality was better than the average are: Rain-Thunderstorm, Fog-Snow, Snow, and Fog-Rain-Thunderstorm.

- We also noted a counter intuitive insight that Rain to an extent helps improve the bus Punctuality. We hypothesize that this is because in Rain, student or pedestrians would not be walking around or there could be less traffic to deal with in the transit route.
- c) **Day of the week (DOW):** It can be noted that for route “A”, Wednesday is the day when the bus has the highest number of delays for this route and Friday being the least.
 - d) **Based on the stops or stop combos:** We can notice Foster to Kelley School (), Jordan Hall to Maurer School of Law, IMU to 10th & Woodlawn, Kelley School to Wells Library have the highest number of instances when the bus is late. And the stops with least number of late instances for route “A” as per order, are McNutt to Kelley School, Stadium (A) to Alumni Center and Assembly Hall () to Stadium () .
 - e) **Based on Hopping:** We noticed that most of the hopping instance are small in frequency, but still we noticed one stop to be predominant that is Well's Library to 3rd & Jordan.

Similarly, inference for other routes are as follows:

For Route B:

- a) **Based on HOD:** As mentioned in Note 1.
- b) **Based on weather:** We notice that the weather combination consisting of Fog-Rain-Snow, Fog-Rain, Fog, Rain-Snow have more adverse impact on the punctuality of the bus, where as a combination of weather conditions like Rain-Thunderstorm, Rain, Fog-Rain-Thunderstorm we see that the performance of the Bus was better than the average rate.
- c) **Based on DOW:** We notice that on Monday the instance of bus being late is high and it is less on a Friday.
- d) **Based on the stops or stop combos:** We notice that the stops Jordan Hall to Maurer School of Law, Lingelbach to Kappa Delta, Chi Omega () to Balfour () have the highest instance of bus being late, whereas the stops 3rd & Jordan to Jordan Hall, SRSC () to Lingelbach (), Fisher Court () to ZBT have less number of instance when the bus is late when compared to the average when the bus is late for route B.
- e) **Based on Hopping:** No hopping data noticed for this route.

For Route E:

- a) **Based on HOD:** As mentioned in Note 1.
- b) **Based on weather:** We notice that the weather combination consisting of Fog-Rain-Snow, Fog-Snow, Rain-Snow have more adverse impact on the punctuality of the bus, where as a combination of weather conditions like Fog-Rain-Thunderstorm and Rain we see that the performance of the Bus was better than the average rate for the Route E.

- c) **Based on DOW:** We notice that on Wednesday the instance of bus being late is high and it is less on a Friday.
- d) **Based on the stops or stop combos:** We notice that the stops IMU to 10th & Woodlawn, 10th & Woodlawn to Psychology, Well's Library to School of Education have the highest instance of bus being late, whereas the stops Nutt Apartment () to Bicknell (), 3rd & Jordan to Jordan Hall and Union & 10th to Union & 7th have less number of instance when the bus is late when compared to the average for route E.
- e) **Based on Hopping:** No hopping data noticed for this route.

For Route X:

- a) **Based on HOD:** As mentioned in Note 1.
- b) **Based on weather:** We notice that the weather combination consisting of Fog-Rain-Snow, Rain-Snow, Rain-Thunderstorm have more adverse impact on the punctuality of the bus, where as a combination of weather conditions like Fog-Rain-Thunderstorm, Fog-Rain we see that the performance of the Bus was better than the average rate for the Route X.
- c) **Based on DOW:** We notice that on Thursday the instances of bus being late are high.
- d) **Based on the stops or stop combos:** We notice that stops Stadium () to Stadium (X), Auditorium (X) to Stadium () have the highest instance of bus being late instances, whereas the stop 7th & Woodlawn to IMU (X) has less number of instance when the bus is late when compared to the average instance for route E.
- e) **Based on Hopping:** No hopping data noticed for this route.

Suggestions

1. In Areas like Wells Library where there is delay due to heavy traffic, maybe we can construct a skywalk or create underground pass way. This will prevent the delay caused due to students crossing the road.
2. Creating awareness about fast lanes. Most students don't know shortcuts to different parts of the campus. We need to create this awareness and may be highlight these paths in some particular color, so that they students can try it. This will reduce the dependency on the IU campus transport and will encourage (to certain extent) a healthy lifestyle among students.

Further Work

3. We are planning to map the Dwell time with the number of passengers travelling in the bus. We hypothesize that there has to be a correlation between the number of people getting in and out the bus and the dwell time of the bus. We want to quantify this relation.
4. Exploring avenues for dynamic routing.

Appendix

- Please refer to the attached files to get the average time for each bus stop
- The code and analysis file are also provided

Note: The analysis is done for spring data 2015.