INDIANA UNIVERSITY BLOOMINGTON

CSCI B 565
DATA MINING

IU Bus Route Optimization –
Group Name : Encoders

Author:

ISHTIAK ZAMAN

RASHMI RAYALA

ALKA NELAPATI

Supervisor:

DR. DALKILIC

December 16, 2015

Table of Contents

Introduction

The project is about optimizing the current schedule of IU bus system, a transport system organized by Indiana University Bloomington. The buses run in 4 routes namely A, B, E and X. We have proposed solutions and insights regarding the variance in schedule and factors affecting the variance that will be useful in optimizing the current schedule. Based on the analysis of results from our defined objectives, we have also proposed some changes to optimize the current schedule.

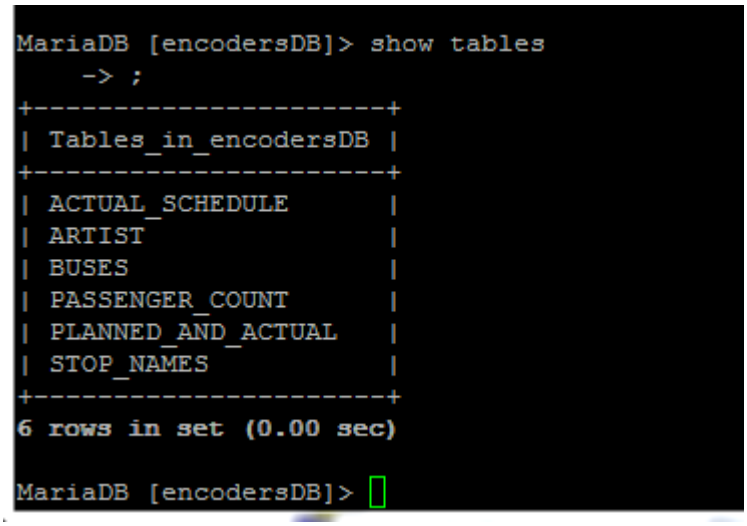
Objectives:

1. To identify a pattern in bus delays over weekdays (M-R) during 9AM to 5PM for all routes over Fall semester
2. Compare frequencies of overall delays and on time arrivals of buses over Spring semester
3. Analyze factors affecting delays – effect of Passenger In and Out bound on delays

Data Preprocessing

The given data is processed to extract planned and actual schedules of all routes, geocodes of each stop, passenger in and out bound count. The delay time is calculated as variance between planned and actual schedule. Geocodes are extracted using some optimized techniques to be close to accurate and used in generating the maps. The data is stored in MySql database and csv files for reference and processing.

The format of extracted information and process of extracting is as explained below.



1. Geocode generation
   To get the geocode of each stop, Geocode and interval data are divided into multiple files based on busid. From interval data, the when timestamp of each stop ("to" field) is used to match and filter geocode timestamps. This process is repeated for all the stops and buses. An average of 10 geocodes (latitude, longitude) is used to calculate the representative geocode of each stop.
   StopCode – Stop code of bus stop
   Latitude - latitude of the bus stop
   Longitude – longitude of the bus stop
2. Stop_Names

The purpose of this table is to link Stop code with its corresponding Stop name and geolocation (Latitude and Longitude)

The schema of this table is as follows:

Stop code – stop code of the bus stop

Stop name – name of the bus stop

Latitude - latitude of the bus stop

Longitude – longitude of the bus stop

Stop code and Stop name are extracted from Stop ID table.

3. Planned_Schedule

   The purpose of this table is to know the planned schedule of each bus (as scheduled by IU transit) i.e. the major stops at which the bus stops and the time at which a bus is scheduled to enter and leave a bus stop.

   The schema of this table is as follows:

   Schedule ID – schedule id corresponding to Bus

   From stop – stop from which the bus starts

   To stop - next destination stop

   Time out – time at which bus leaves from the from stop

   Time in - time at which bus reaches the to stop

   Transit time – Time taken by the bus to reach from the from stop to the to stop

   Data required for this table is being extracted from the following resources:

   From table's qry_Trips_A, qry_Trips_B, qry_Trips_E, qry_Trips_X in Fall Ridership 2014 database start times of the buses in all the four routes is being extracted.

   Time taken to travel between each stop is being extracted from Schedule data table in Double map database.

   Java code is being written to combine the above data into a single table.

4. Actual_Schedule

   The actual_data.csv file was generated from the given intervaldata2014-2015.tsv file. The purpose of the actual_data.csv to provide the actual timing when a bus start and end in a stop.

   Route ID - Integer route ID of the bus

   From - Bus stop ID of the last bus stop

   To - Bus stop ID of the bus where it just reached

   TimeOut - Time of leaving the "From" bus stop

   TimeIn - Time it reached the "To" bus stop

   DwellTime - Amount of time in seconds it waited in the "From" bus stop before starting for "To" bus stop.

5. Delay Time

   The 8 files: difference_Fall/Spring_A/B/X/E_Route_M-R.csv file were generated by joining the actual_data.csv, Fall/Spring_Scheduled.csv and FinalFallSpringRoutes.csv files together and comparing the actual time and scheduled time at each stops. The purpose of this file is to show how much delay a bus made in each major stops.

   Route ID - Integer route ID of the bus

   StopID - Stop ID of the bus

   ActualTime - Actual time it reached the stop. This value was fetched from the actual_data.csv file.

ScheduledTime - Time when it was supposed to reach the stop. This value was fetched from the Fall/Spring_Scheduled.csv files.
Delay - Delay in seconds the bus made in that particular stop. We got the value by subtracting ScheduledTime from ActualTime. Negative delay tells us that the bus reached there early.

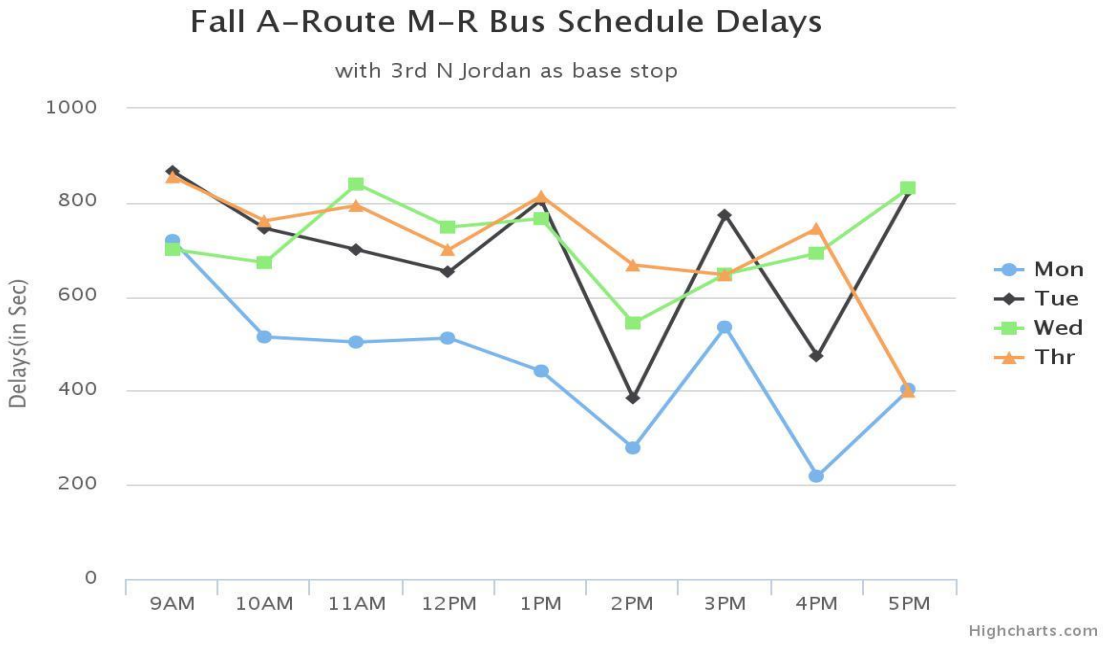
6. Passenger_count
   Contains passenger in and out bound information

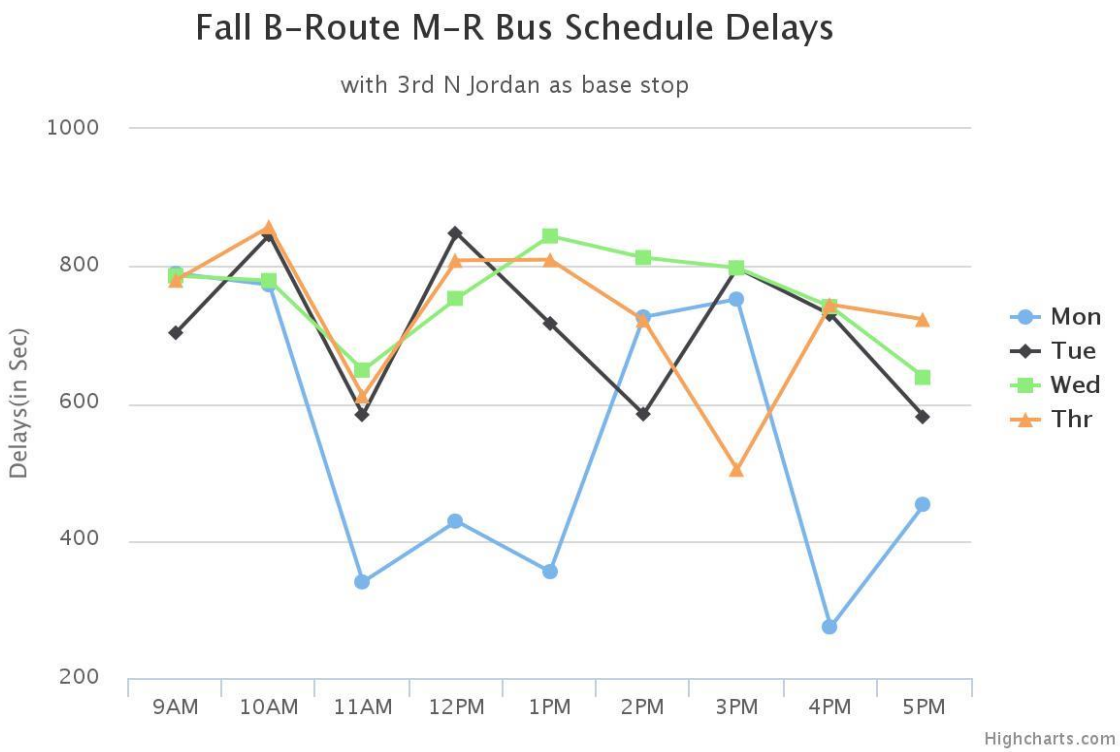
Objectives Implementation

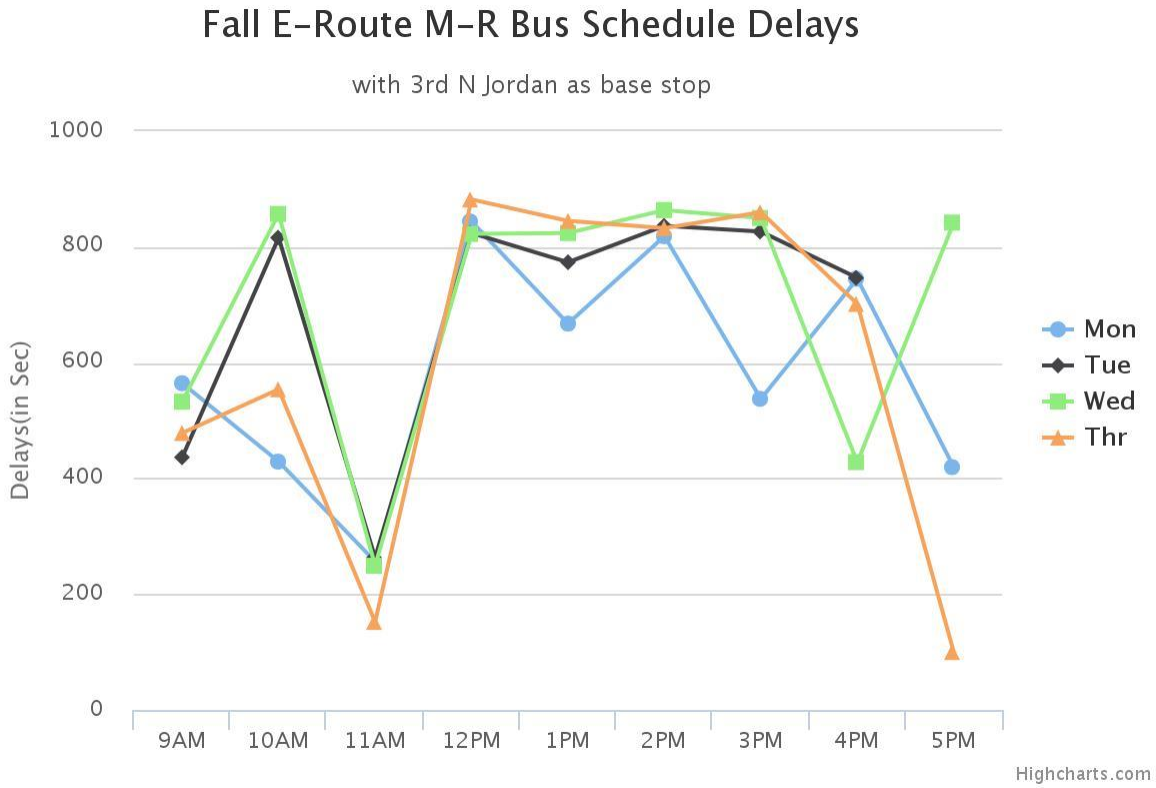
1. Identifying pattern in bus delays:

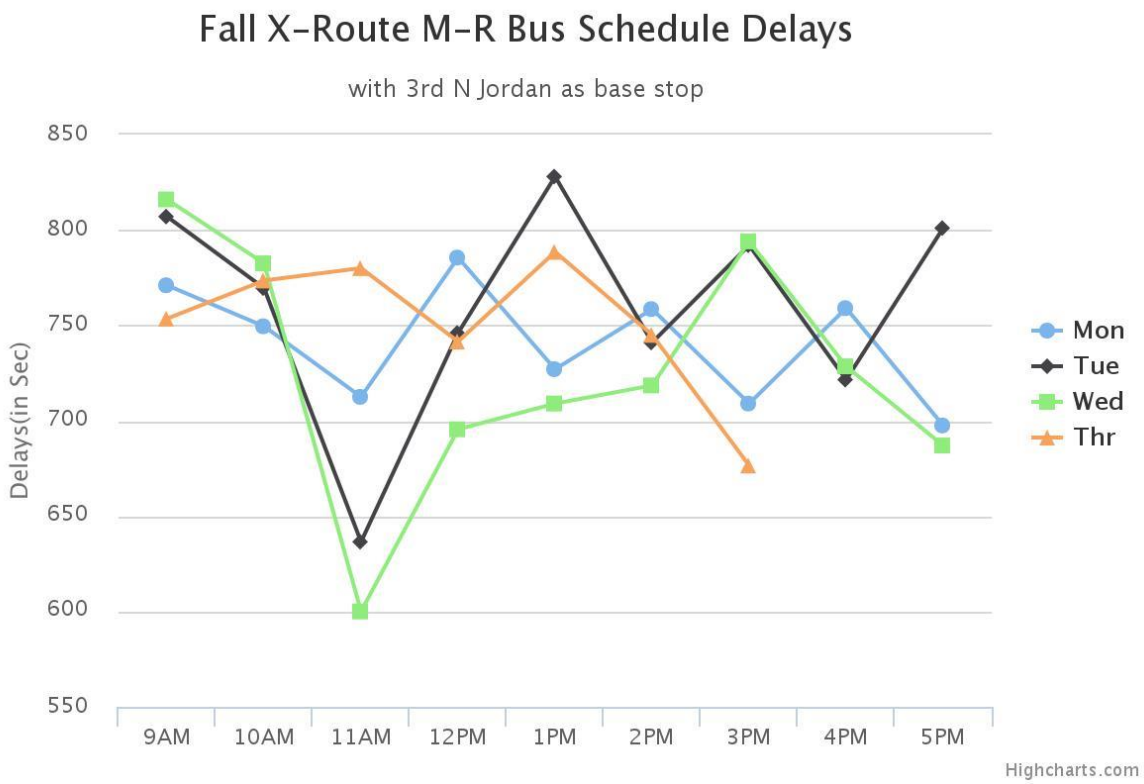
For Fall semester, the average of highest delays of each hour from 9AM to 5PM for each day categorized according to day of week (M-R) are computed. We used 3rd N Jordan (Stop ID 6) as a base stop for calculating delays since it is the mid stop for most of the routes and is identified to be a good measure.

The graphs are plotted with each hour from 9AM to 5PM on x-axis and delays (in secs) on y-axis for A,B,E and X routes. The analysis is focused on the business hours of working days.




Fall B-Route M-R Bus Schedule Delays
with 3rd N Jordan as base stop
Delays(in Sec)
1000
800
600
400
200
9AM
10AM
11AM
12PM
1PM
2PM
3PM
4PM
5PM
Mon
Tue
Wed
Thr
Highcharts.com


Fall E-Route M-R Bus Schedule Delays
with 3rd N Jordan as base stop
Delays(in Sec)
1000
800
600
400
200
0
9AM
10AM
11AM
12PM
1PM
2PM
3PM
4PM
5PM
Mon
Tue
Wed
Thr
Highcharts.com



Results and Analysis:

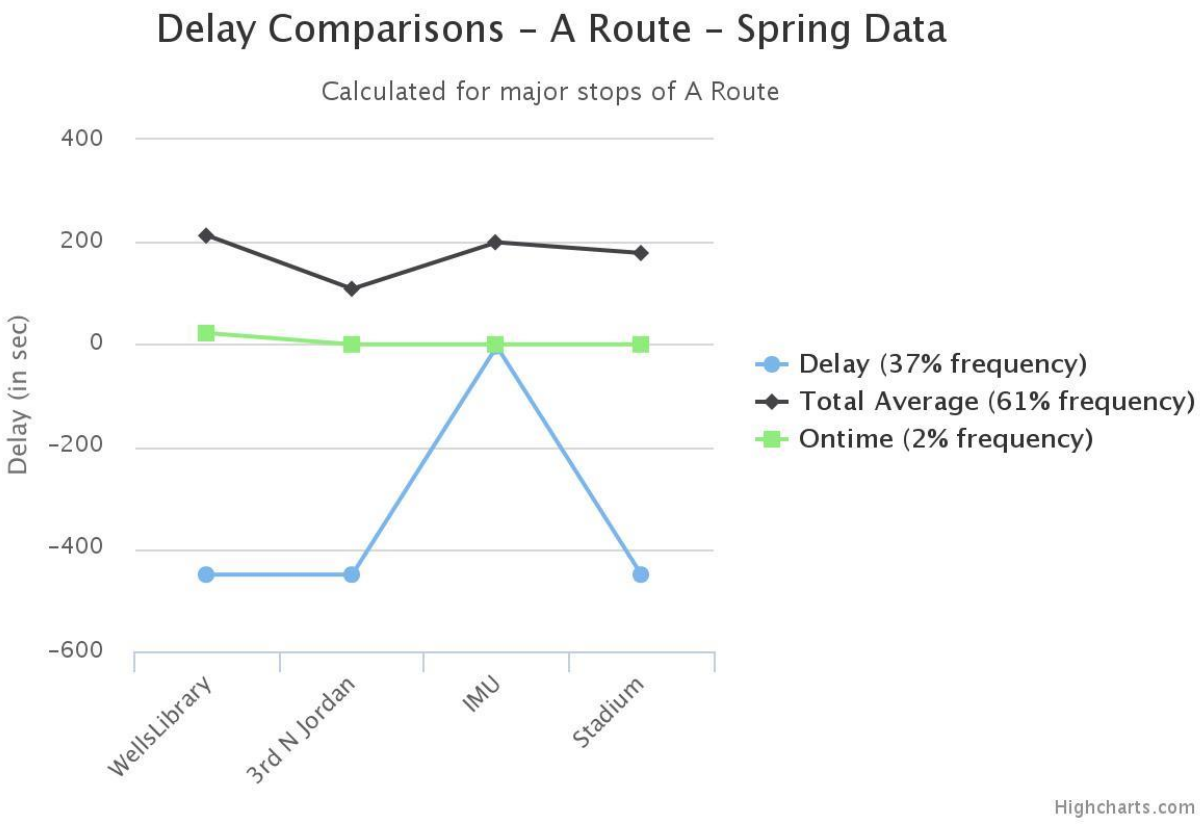
Analyzing the above graphs, the following conclusions are drawn:

- Overall, bus delays for A and B routes are observed to be relatively more on Wednesday and Thursday compared to Monday and Tuesday
- Delay is observed to be more at beginning and end of hours (around 9AM and 5PM)
- Maximum delay did not exceed 15mins for all route buses
- Among all the routes, E route bus delays are the least and is frequently running early than getting delay
- Results of E and X routes are a bit skewed due to possible reasons like insufficient or less accurate data or due to effect of external factors like traffic

2. Delay - Early – On time comparisons of major stops

Spring data is analyzed to calculate the average of total, delay and on time schedule of different routes for entire semester. A comparison is drawn and the graph is plotted as below. The on time plot is generated with delay values between -30 and +30 seconds. A route is chosen for our analysis as it has quality data and is highly commuted route.

Some interesting observations and conclusions are also drawn as explained below in Results and Analysis section.



Results and Analysis

From the above graph, the following analysis is made.

1. Although the buses in A route experience delays 37% of the time, they catch up with schedule most of the times
2. The bus is either running early or getting delayed most of the times and is very rarely on time (which can be understood by 2% frequency)
3. Although the bus is experience delay at other major stops, due to *some reason* it is mostly on time at IMU. This could be affected by external factors like less traffic and less transit of students between 3rd N Jordan and IMU.

   One most interesting explanation for this could be – due to internal walkable distance shortcuts between the buildings near IMU and 3rd N Jordan.
4. On the other hand, the delay is relatively more at 3rd N Jordan. This could be because of high traffic and/or more number of students commuting from Stadium or wells library to 3rd N Jordan. This could also be justified by the high passenger in/out bound as explained later.

3. Factors affecting delays – effect of Passenger In and Out bound on delays:

Average passenger in and out bound is calculated for all routes during M-R for the corresponding dates used for calculating delays as explained above. The data is extracted from qryTotals(A,B,E,X) tables from Ridership Fall database.

The passenger in and out bound for M-R is as shown in below table. The highest tow/three counts for each route is highlighted.

Route	Day Of Week	Passenger In bound	Passenger Out bound
A	Mon	3736	3073
	Tue	4491	3661
	Wed	4859	4096
	Thr	4589	3703
B	Mon	1720	1983
	Tue	**2097**	2328
	Wed	2307	2657
	Thr	**2042**	1185
E	Mon	1006	1108
	Tue	**1282**	**1379**
	Wed	**1255**	1458
	Thr	1262	**1364**
X	Mon	3212	3263
	Tue	1572	1557
	Wed	1585	1600
	Thr	1556	3024

Results and Analysis:

From the above table it can be observed that passenger in and out bound are relatively high on Wednesday and Thursday which are consistent with the high delays during those days.

Hence it can be implied that high passenger count is having positive impact on bus delays.

Visualization

The average delay for all major stops for each route A, B, E and X for Fall and Spring semester is as plotted below. This is generated using the Geocode information generated above. Negative delay implies early arrival of bus at stop than expected schedule.

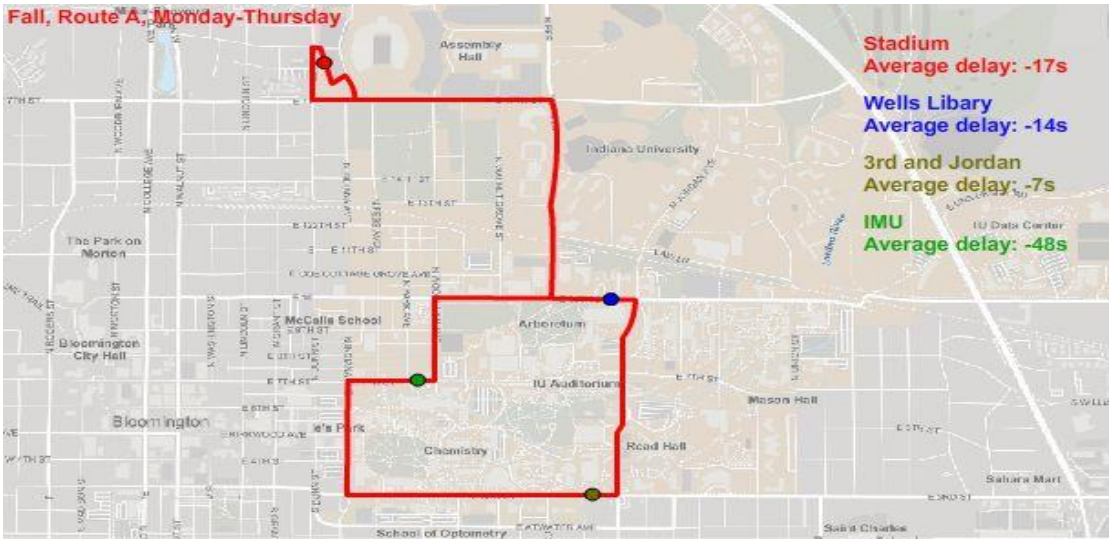


Figure 1: Fall A route, Monday-Thursday

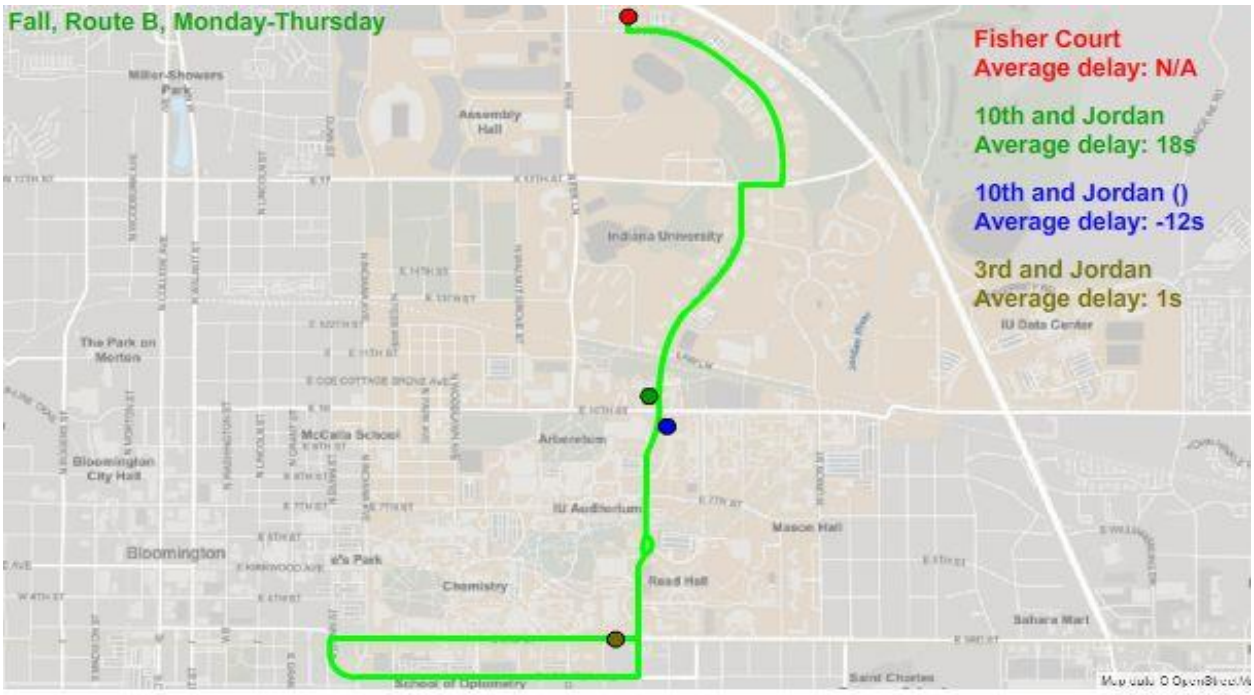


Figure 2: Fall B route, Monday-Thursday

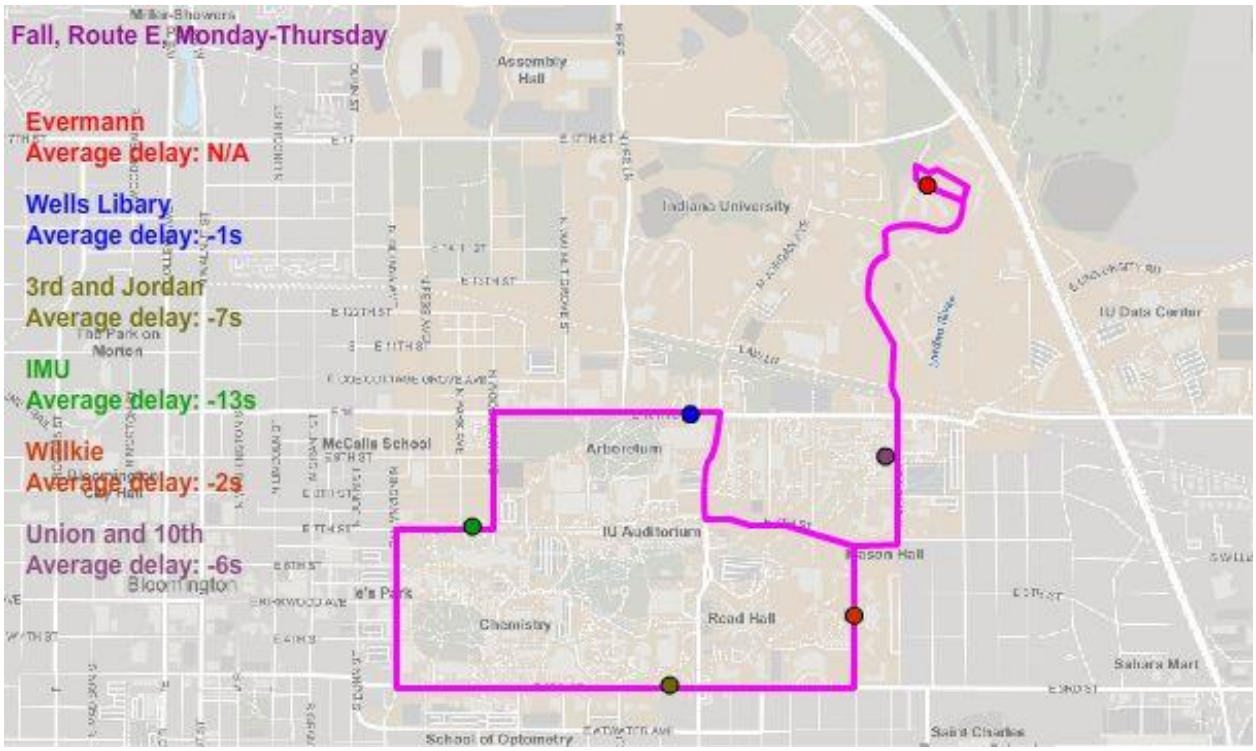


Figure 3: Fall E Route, Monday-Thursday

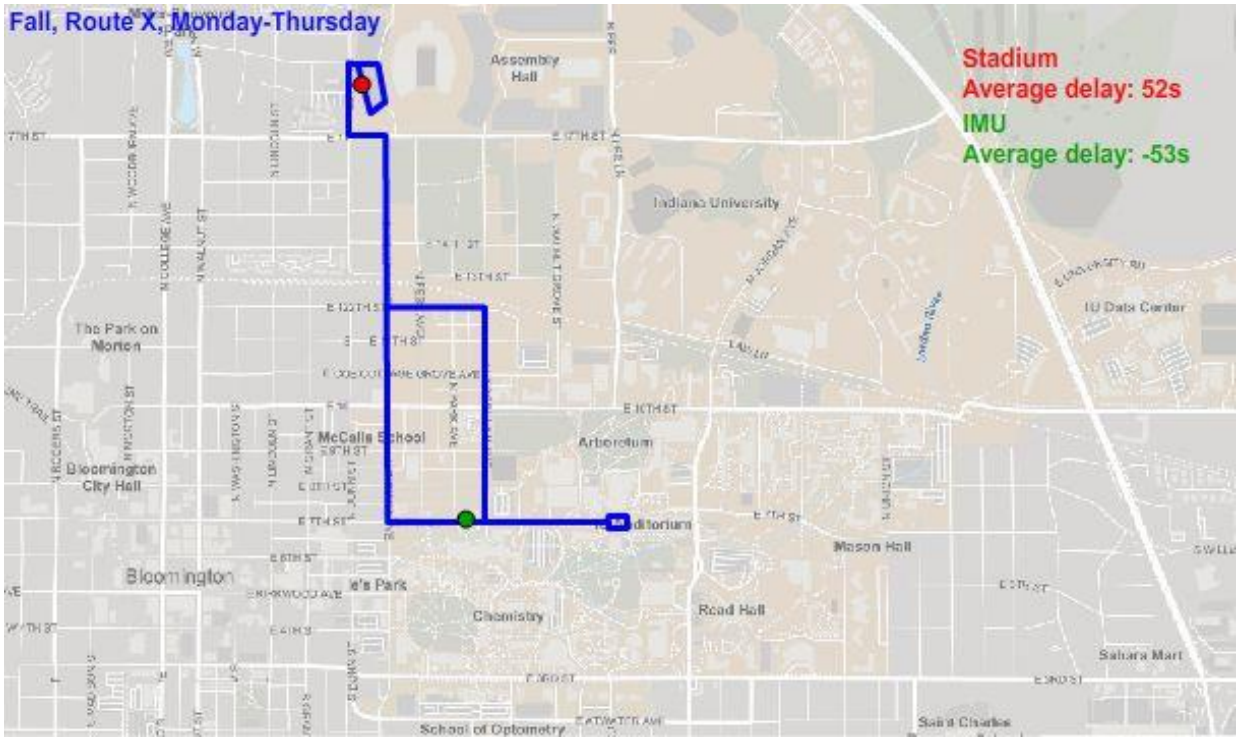


Figure 4: Fall X route, Monday-Thursday

Figure 5: Spring A route, Monday-Thursday

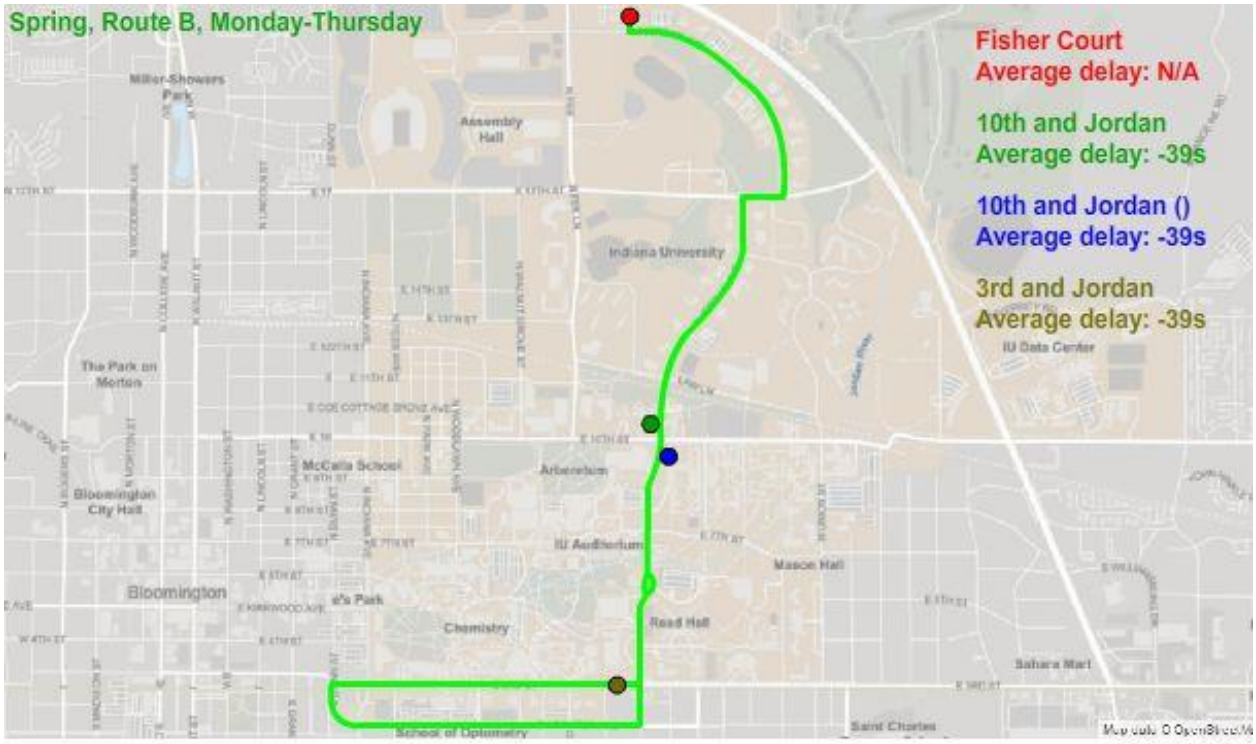


Figure 6: Spring B route, Monday-Thursday

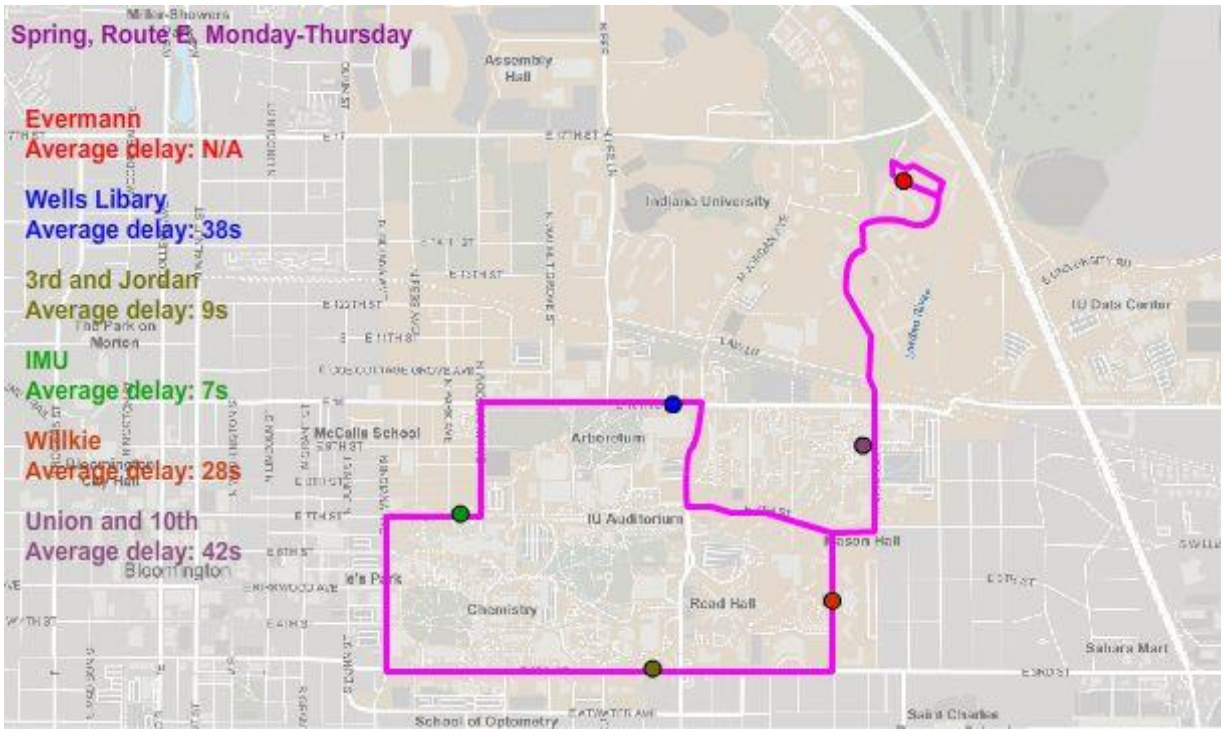


Figure 7: Spring E route, Monday-Thursday

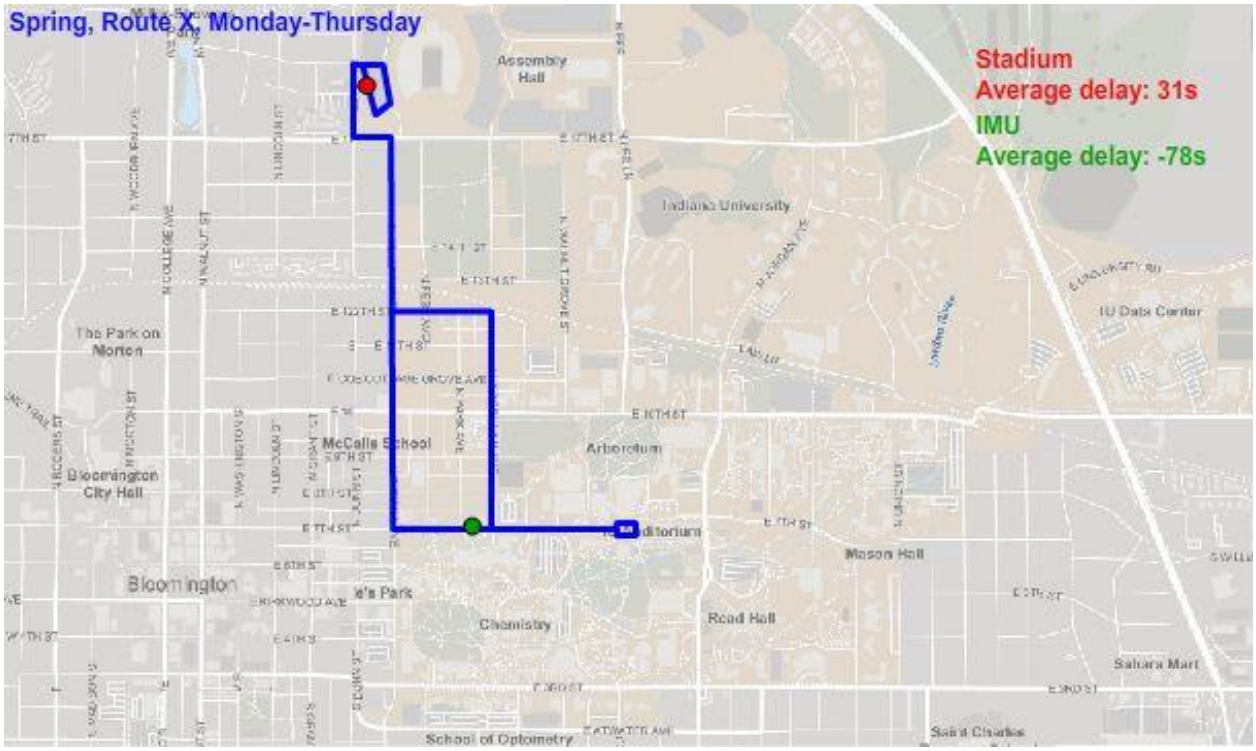


Figure 8: Spring X route, Monday-Thursday

Proposed Changes

1. In view of analysis of results from Objective 2, the on route time of A route bus can be optimized by modifying its planned schedule to benefit the students commute. This could be done by decreasing the planned travel time between 3rd N Jordan to IMU and balancing that reduced time by increasing the travel time between Wells library/Stadium and 3 rd N Jordan thereby reducing the delay at these stops. This would reduce the early arrival of bus at IMU and benefit students with optimized schedule time.
2. In view of analysis of results from Objective 1, since the delays are observed to be consistent during late weekdays (Wed and Thr), some motivation sessions could be arranged mid-week to help motivate drivers to stick to the schedule.
3. In view of analysis of results from Objective 3, it is rightly justified that delays are positively affected by passenger in and out bound count (for example, explaining the delays between Stadium/Wells library to 3rd N Jordan). In particular, since the delays and passenger counts are observed to be high during 9AM to 11Am and around 5PM, either more buses could be added at the peak times OR more efficiently resolved by addressing the 1st change suggested above.

Conclusion and Future Work

The main objective of identifying delay pattern in buses for different routes has been calculated over Fall and Spring semester and the factors affecting the delays have been measured. High passenger in and out bounds have been found to have positive impact on the bus delays and some postulates has been presented suggesting the reason for bus delays at major certain stops for A route.

The analysis can be drilled down to identify buses (busid's) having frequent delays and in turn identify the drivers causing the delays. Also, In addition to passenger in and out bounds, school academic calendar, weather and traffic data can also be considered into account for identifying the factors affecting delays.

The analysis could further be extended to suggestions on removal of certain stops for each route. This can be achieved by Identifying top 10 dates with highest passenger in and out bound and identify the stop codes where the bus has stopped on those days and make prediction of certain stops (less frequent stops) to be removed.

Technologies and Java Code Files

Technologies used: Java, C++, tools for generating graphs

Java Code Files used:

Project source files.zip

The intermediate files generated for calculating delays and upon which graphs and results are generated are enclosed along with this document as part of submission.