Indiana University Bloomington

CSCI B 565

Data Mining

IU Bus Route Optimization - The Analysts

Author:
1.Soumik Dey
2.Harish
Annavajjala
3.Ramprasad
Bommaganty

Supervisor:
Dr. Dalkilic

December 18, 2015

Abstract

IU Bus route optimization project focuses on improving the efficiency of the Indiana University Bus transport system running on the Indiana University – Bloomington campus. Owing to the increasing student enrollment at the University over the years, the burden on the existing infrastructure to serve the needs of the commuters has grown proportionally.

The IU Bus route optimization project provides a methodology to adapt to the increasing demand by studying the existing data from daily usage and suggesting schedule modifications and optimum usage of existing resources. The project also focuses on on-time delivery of buses to their respective stops with appreciable reduction in travel times. The data that was used to perform analysis on was real time data collected over a period of time during the academic year 2014-2015.

We transformed the data and performed analysis on the data. We calculated the average differences between actual times and scheduled times(variance) between any two consecutive stops for all the routes and also the average for each days and tried to understand the relation between variance, weather and passenger count.

Contents

1 Introduction

Bloomington has been growing as a college town with the count in student community increasing sharply over the past few years. This combined with the large number of vehicles on road in recent years have made travel during certain time intervals quite taxing and have contributed to the delay in delivering efficient public transportation. IU Bus transport system is one such organization which is facing such an issue with respect to inefficient management of its resources. Hence, there is a greater need to focus on providing an efficient and cost effective mode of commute to the passengers, whilst making optimal use of existing resources.

About IU Bus route optimization project: The details of the architecture of the database system and the data that was used for computation are as follows: Database used: Access, MySQL

We have used the data collected from the time period of August 2014 to March 2015. Weather data was one component of the different sets of data given to us and we made use of it to predict the effects of weather on transportation.

Tables present: Interval data 2014-2015, Route ID, Schedule Data, Stop ID, Weather Data, Work record.
Languages used: Java, R
Information about the data: We have 66 stops in total, 26 buses running on any given particular day owing to resource and maintenance constraints and four routes – A, B, X and E.

2 Problem Description

The current model followed by the IU Bus transport system is as follows:

All the buses run on a static, fixed time schedule which might occasionally vary owing to semester breaks and university events. Each day, a fixed number of buses are run as per the schedule. Students and other eligible passengers can keep track of the buses in real time on a particular route by using a mobile application called DoubleMap or refer the schedule given on the iubus.indiana.edu.

The routes are fixed throughout the semester and are not modifiable. There isn't a two way communication model between the Bus authorities and passengers. Since the existing model is rather influenced by weather and traffic congestions, it does not perform quite well under sustained constraints and is bound to experience a drastic fall in efficiency.

3 Data Description

The following were the data sources which we used during our computation processes:
The tables we utilized were: Interval Data 2014-2015, Route ID, Schedule Data, Stop ID, Weather Data and Work Record.
We also used the Access file: Ridership Spring 2015

Interval Data 2014-2015: This table included eight column attributes as follows:
ID1: an ID assigned to each record collected during real time usage.
from: a numerical representation of the different stops in each of the routes
to: a named representation of the different routes in each of the routes
id: a secondary identifier with respect to each record
time: the number of seconds between two stops(travel time and dwell time can be computed using this attribute)
bus id: each bus has a unique identifier
route id: there are four different routes – A, B, X and E respectively
when: a datetime attribute which specifies the date of the entry log and the exact time instant

Route ID: This table included the following attributes:
ID: unique record identifier
Index: unique numerical representation of each stop
Route ID: specifies route and days of the week
Field3: specifies the semester

Schedule Data: This table included the following four attributes:
ID: an identifier with respect to each record
Route: this attribute identifies the routes and sub classifications such as A1,. . . ,A6
Time: the scheduled time at each stop
Stop: all the bus stops on each of the routes
Stop ID: This table included the following attributes:
ID: unique record identifier
Index: index for each stop
Stop: name of each stop

Weather Data: This table included the following attributes:
ID: each record identifier
EDT: recorded date
MinTemp: minimum temperature readings for a date
Precipitation: numerical representation of precipitation
Events: climate during the log recording, e.g., Rain, Snow

Work record: This table included the following nine attributes:
ID: record identifier
Clock in: time when the employee clocked in
Clock out: time when the employee clocked out
Driver: First name and Last name of the driver
Shift Type: indicates if the shift was a standard one or a substitution
Daynum: day number
Date: date of the log entry
Route: It specifies the day of the week, bus route and order of bus operation
Bus: unique bus identifier

Ridership Spring 2015 Access database file:
Under the qryRouteTotals table we have the following attributes:
Date: Date of the record entry
Sum of Inbound: Number of passengers who got into the bus
Sum of Outbound: Number of passengers who got out of the bus
Total: the summation of inbound passengers and outbound passengers
Route: the specific route e.g., A, B, X or E (includes Special routes)

4 Computational Implementation

We have used Java to model the data and do clustering of the data. We have used Java to model the data and map the Intervaldata to Scheduledata through Workdata. For clustering we have used Java to do that as well. We clustered the intervaldata around the scheduledata to perform the mapping. We also performed joins on the intervaldata mapped table with weather data and Ridership Spring data. Then we used R and the library of arules to do apriori algorithm to generate rules for different routes based on variance, weather anomalies and passenger count.

The following are some of the inferences derived from studying the accumulated data:

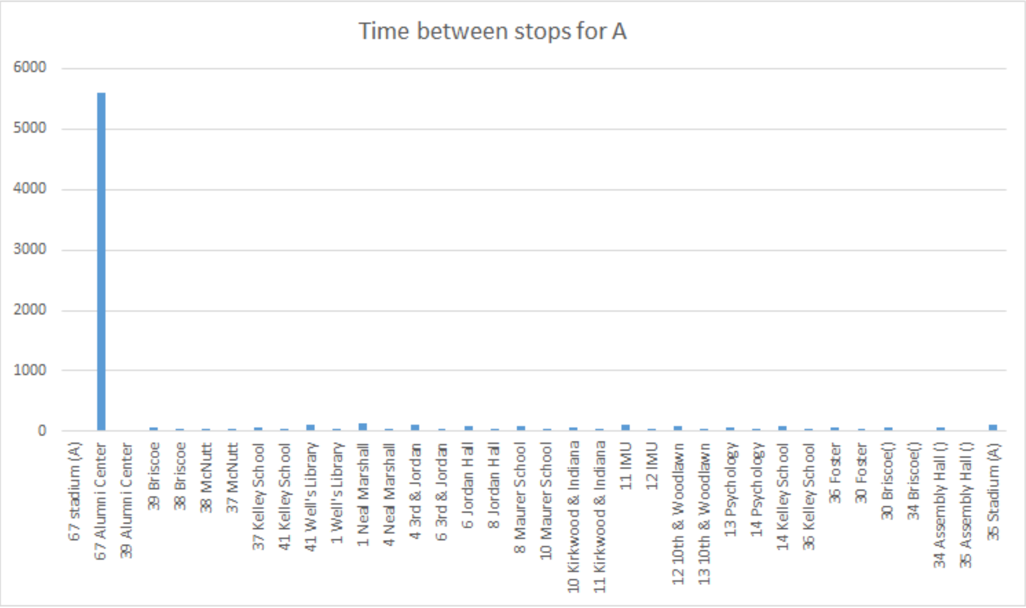
The data was analyzed by comparing the Schedule Data table and Interval Data 2014-2015 table with respect to the expected clock in values and the actual clock in during real time usage. The GPS data accumulated over a period of time is representative of the location parameters stored during service for each bus over a certain route, say A, B, X or E. It was noticed that certain buses were restricted only to certain routes while there were a few exceptions along the way. Perhaps, resource constraints were a prominent factor in such anomalies.

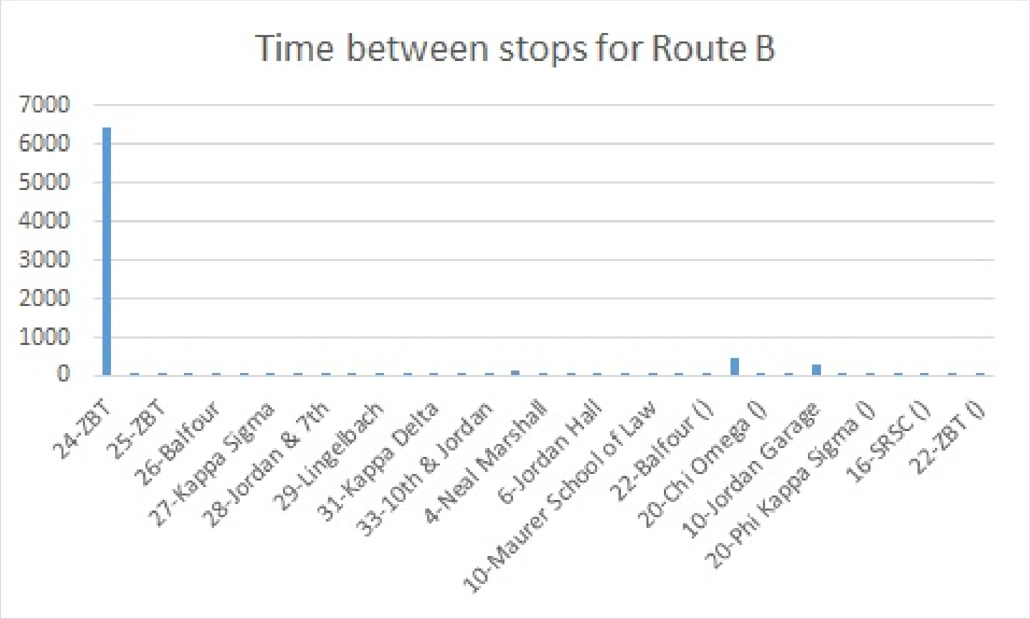
As a team, we first tried to compute the variance of the estimated travel times and the actual travel times. This variance was crucial in understanding the other factors which contributed to a less efficient service. Then we tried to calculate the average travel times and dwell times between any two consecutive stops in a route. Then we calculated variances for all the days and also for all the major stops and we mapped it to weather data and passenger count and tried to find some rules using apriori algorithm.

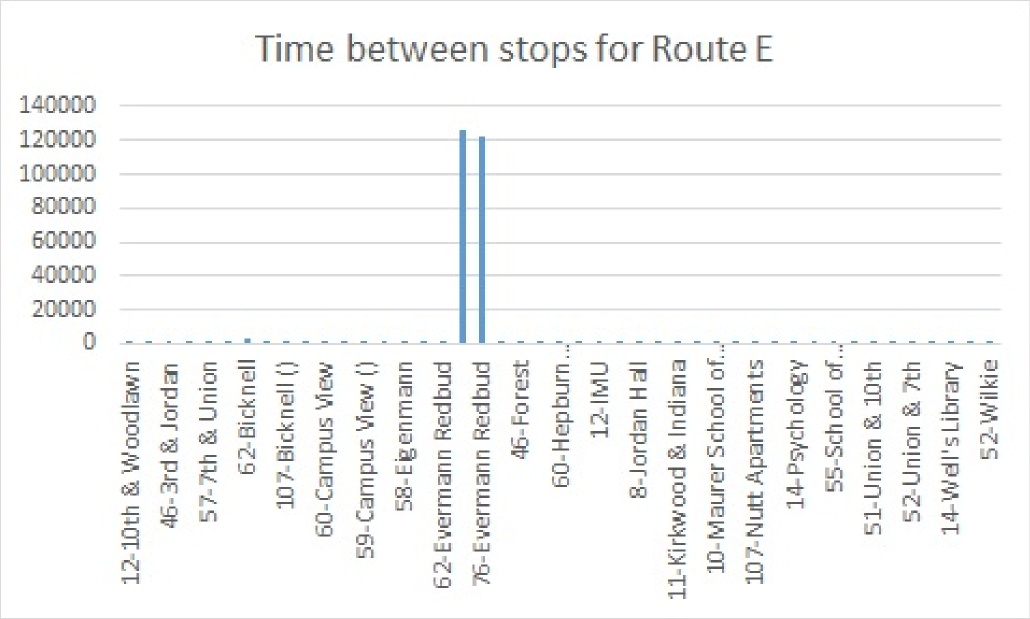
The apriori algorithm gives us the result that for route A we see a increased passenger count for weather anomalies like rain, snow, fog and thunderstorm. For B and E we see an increased variance which might be the result of the bus facing some difficulties to travel due to the weather. We also tried to figure out the average variance between two consecutive stops in a route and have figured out the stops in each route with the highest variance.

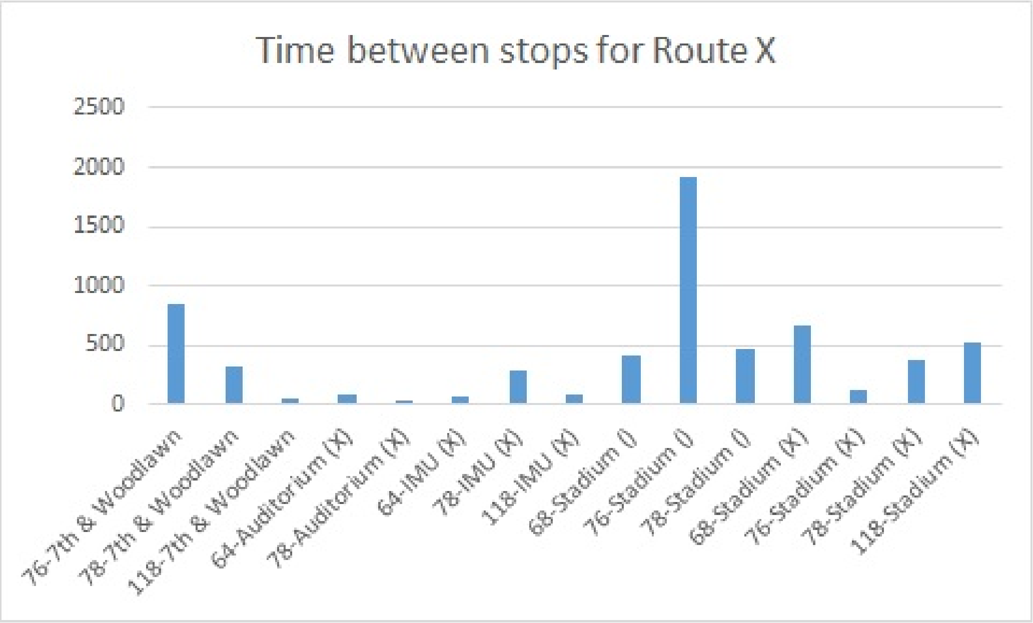
5 Visualizations

The following charts represent the travel times between stops on routes A, B, X and E.

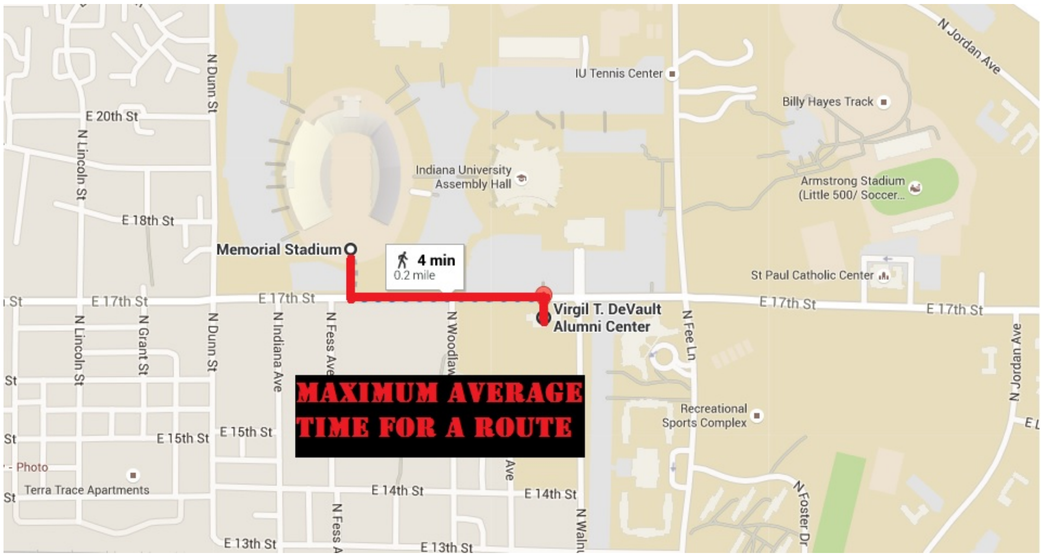


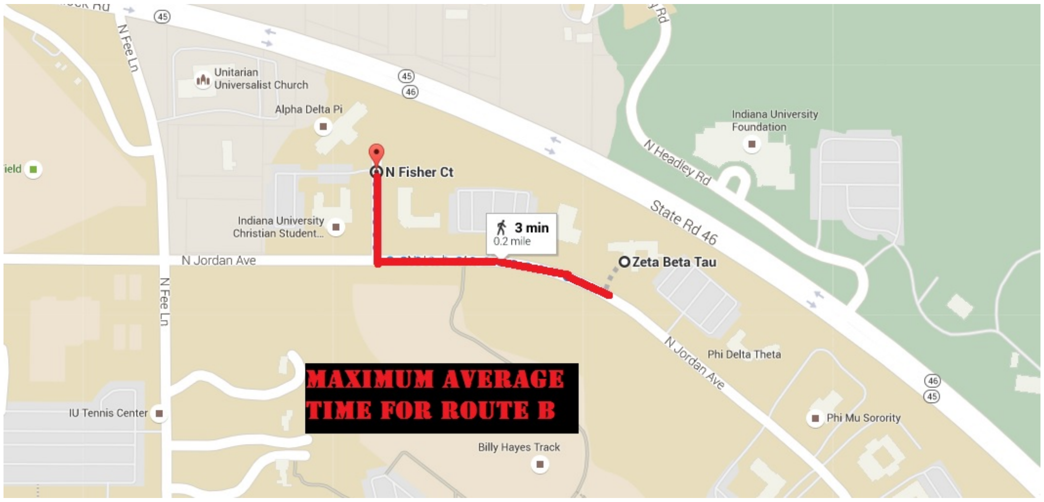

Time between stops for Route B
7000
6000
5000
4000
3000
2000
1000
0
24-ZBT
25-ZBT
26-Balfour
27-Kappa Sigma
28-Jordan & 7th
29-Lingelbach
31-Kappa Delta
33-10th & Jordan
4-Neal Marshall
6-Jordan Hall
10-Maurer School of Law
22-Balfour ()
20-Chi Omega ()
10-Jordan Garage
20-Phi Kappa Sigma ()
16-SRSC ()
22-ZBT ()

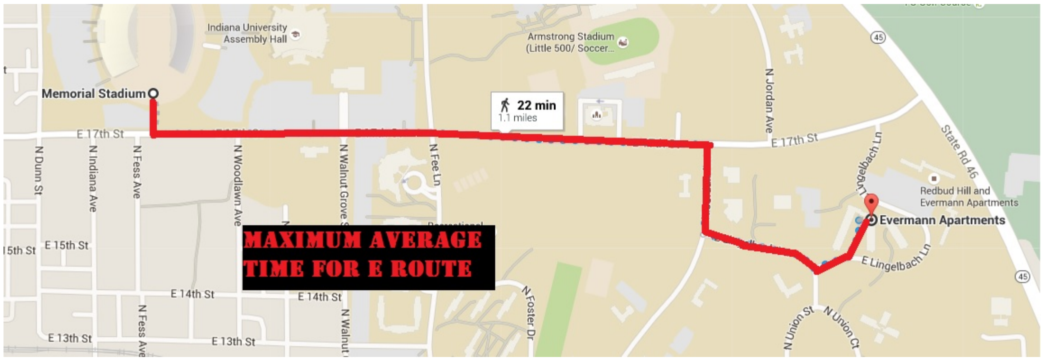

Time between stops for Route E
140000
120000
100000
80000
60000
40000
20000
0
12-10th & Woodlawn
46-3rd & Jordan
57-7th & Union
62-Bicknell
107-Bicknell ()
60-Campus View
59-Campus View ()
58-Eigenmann
62-Evermann Redbud
76-Evermann Redbud
46-Forest
60-Hepburn…
12-IMU
8-Jordan Hall
11-Kirkwood & Indiana
10-Maurer School of…
107-Nutt Apartments
14-Psychology
55-School of…
51-Union & 10th
52-Union & 7th
14-Well's Library
52-Wilkie

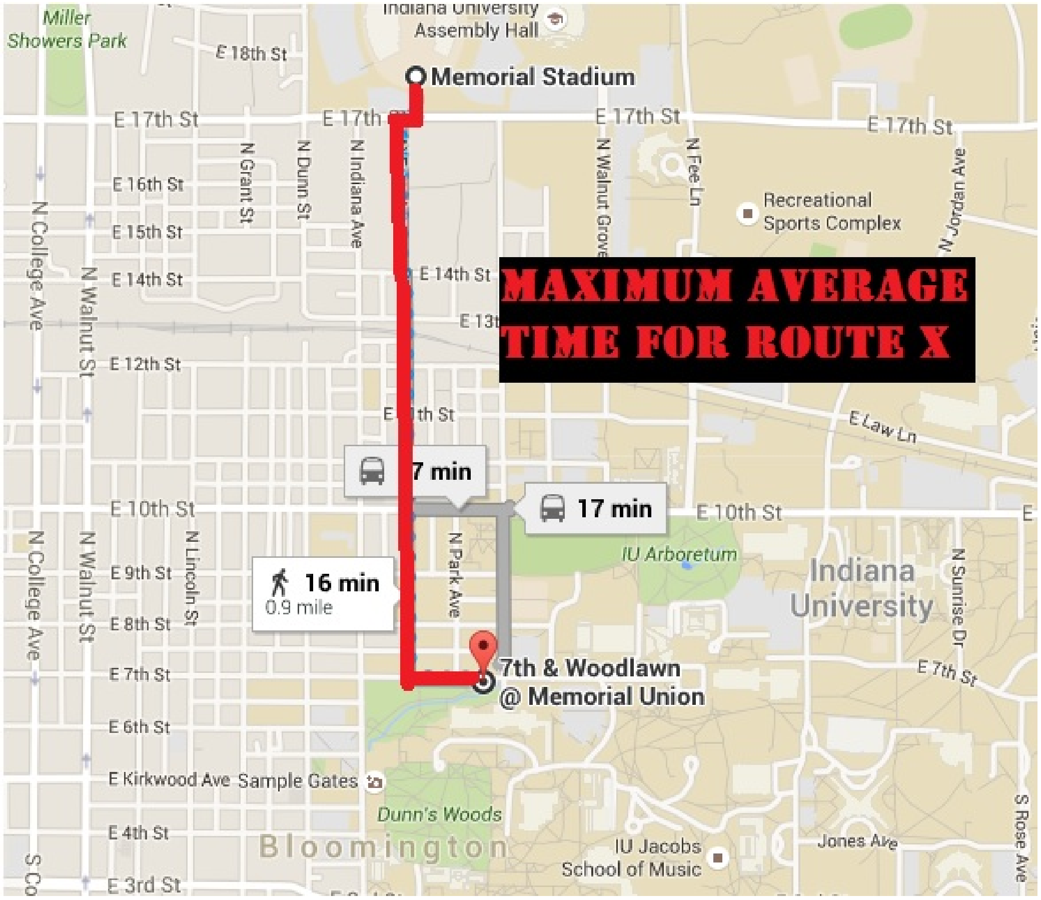


The following figures represent the maximum travel time between two consecutive stops on routes A,B,X and E.




Unitarian Universalist Church
Alpha Delta Pi
N Fisher Ct
Indiana University Christian Student...
3 min
0.2 mile
Zeta Beta Tau
N Jordan Ave
N Fee Ln
State Rd 46
N Headley Rd
Indiana University Foundation
Phi Delta Theta
Phi Mu Sorority
IU Tennis Center
Billy Hayes Track
MAXIMUM AVERAGE TIME FOR ROUTE B


Indiana University Assembly Hall
Armstrong Stadium (Little 500/ Soccer...
Memorial Stadium
22 min
1.1 miles
E 17th St
N Jordan Ave
N Dunn St
N Indiana Ave
N Fess Ave
N Woodlawn Ave
N Walnut Grove St
N Fee Ln
E 15th St
E 14th St
E 13th St
N Foster Dr
Lingelbach Ln
Redbud Hill and Evermann Apartments
Evermann Apartments
E Lingelbach Ln
N Union St
N Union Ct
State Rd 46
MAXIMUM AVERAGE TIME FOR E ROUTE



6 Proposed Changes

Though the DoubleMap application is quite useful in determining the location of the bus at any point of time, it does not involve two way communication which is actually useful in dynamic scheduling of buses.
The following is a proposition which can be implemented over a course of time:

An application, preferably mobile, can be used to offer need based services. If the data is accumulated from a mobile app, say a person wants to take 'A' route bus between 8:00 am to 8:15 am, and more such requests arise for that route and on that particular time frame, then assign more buses to satisfy those requests. Consider the data as streaming data and with a sliding window consisting of the number of buses that can run on that particular day. Fill up the bus schedules till they are all occupied with trips. Implement the property of Heavy- Hitters that is usually used for streaming data.

Also what we observe that for A route, the weather heavily affects the passenger count. And in the other routes we observe that weather affects the variance, i.e. that the buses take longer time to reach than the scheduled time. So what we can do is reallocate a few buses from the other routes during bad weather in such a manner that it compensates for the high variance for the other routes and can be dedicated to A route and the frequency of A route be increased.

7 Concluding remarks

Seeing the data, and observing the rules it is very evident that the A route is the busiest and has the most number of passengers as it covers the entire campus. The apriori algorithm shows that bad weather increases the passenger count for A which again confirms the fact that weather does play an important role.

For the other routes of B and E we get the effect of weather on variance as the weather increases the variance. Weather might hamper the driving of the bus and hence the bus might move slower than usual.

We saw that for A,E and X routes we have a positive average variance

For A it is 416 seconds, for E it is 125 seconds and for X it is 307 seconds. For B we see a negative variance of -284.8084 that is the bus arrives earlier than usual.

8 References

1. CRAN Project ”arules” https://cran.r-project.org/web/packages/arules/index.html
2. Introduction to Data Mining Tan, Steinbach, Kumar ISBN-978-93-325-1865-0
3. IU bus website http://www.iubus.indiana.edu/
4. Doublemaps website http://iub.doublemap.com/map/