Chicago Traffic Data Analysis

Samiya Qasmi

*Fundamentals of Engineering II*

*Purdue University Fort Wayne*

Fort Wayne, IN, USA

[qasms01@pfw.edu](mailto:qasms01@pfw.edu)

*Abstract*—This project is to analyze the traffic trends in 29 routes within Chicago after 2018. After obtaining the data and cleaning it, further analyses were made using graphical and logical interpretations with the help of MATLAB. Every route was analyzed using similar techniques to find the best time to head out while obtaining minimum traffic and maximum speed.

# Introduction

Traffic patterns are ever changing. They are dependent on multiple factors like time, day, week and even month. Apart from that every route has different traffic patterns and it can get disorganized to maintain and keep track of all the times and routes.

This project aims to undertake the task of organizing data collected to portray relevant information. Using MATLAB plots, different types of graphs were plotted to analyze the various aspects of the traffic patterns.

The data was incorporated from ‘https://data.cityofchicago.org/’[1]. It is titled “Chicago Traffic Tracker - Historical Congestion Estimates by Region - 2018-Current” and includes 17 fields

An interactive program was also curated which intakes the user input in the form of a route name and gives back relevant information about that route. Using this program, a user can figure out what is the best time to go on the route. The best time is defined as the time of day with the least traffic and maximum speed. It also includes the week of the day and the month of the year which has the least traffic and the maximum speed.

# Procedure

## The dataset

The dataset has 17 columns and 6.53M rows. Out of the 17 columns, 7 columns were vital to analyzing the trends. It was downloaded as CSV file and then imported to MATLAB. It was further converted into a table with the name ‘ChicagoTrafficTrack’. The table’s columns were kept the same as the main file, and every field was defined. The dataset is stored in the script – dataset\_chicago.m

* SPEED – This column represents the average speed of the traffic at the given timestamp.
* BUS\_COUNT – This column represents the number of buses received to estimate traffic at the given timestamp.
* NUM\_READS – The number of GPS probes received to estimate the traffic at the given timestamp.
* HOUR – This column represents the hour of the time when the data was recorded.
* DAY\_OF\_WEEK – This column assigns a numerical value to the day of the week when the data was recorded. 1 refers to Sunday and so on.
* MONTH - This column assigns a numerical value to the month of the year when the data was recorded. 1 refers to January and so on.
* DESCRIPTION – This column represents the route in the city of which the data is recorded.

## Cleaning the data

In the cleaning\_dataset.m script, unreadable data was removed from the table to provide ease of analysis. Unreadable data was defined as the rows wherein the data fields were equal to Not a Number wherein the data fields were defined to be numbers. These rows were counted and deleted so that the table would be updated.

## Creating functions

Two functions were created for ease of use in this program – num2day and num2month. The dataset had day of the week and month of the year defined as a number value with 1 as Sunday and so on for the days and 1 as January and so on for the months. These functions used if-else statements and switch cases to convert the number into the corresponding day or month.

## Creating the graphs

The script graph\_chicago was created to plot various graphs and analyze the data to find trends and patterns in traffic across Chicago from the year 2018 onwards. 3D histograms and scatter plots were made for better visualization of data. The graphs are talked more in detail in part III of the paper

## Creating the main script

The main script consists of a menu that presents the user with all the 29 unique routes whose information they can extract at the command window. This script takes user input of the route, checks it against the data and finds all the rows where the number of GPS probes received was the least(indicating least traffic) and then finds when the maximum speed was achieved in these cases. With this information, it returns the best time for the user to travel their choice of route.

The dataset provides us two options to get an estimate of the real time traffic- bus count(number of buses used to estimate the traffic) and GPS probes(number of GPS probes received to estimate the traffic). GPS probes were chosen to find the maximum traffic as they proved to be more accurate over the course.

Furthermore, the user can also choose to view the various graphs provided to draw their own conclusions and analyses about the route of their choice.

# Analysis of plots

Various histograms and scatter plots were made to analyze the data. A summary of many graphs is presented in this section of the paper.

Chart, bar chart

Description automatically generated

Fig 1. Histogram to define the route frequency

Fig 1 describes the route frequency of each route in the table. This describes how many instances of each route were recorded.

Chart

Description automatically generated

Fig 2. Histogram to analyze the relationship between time of the day and speed.

In Fig 2 it is seen that from 2018 till present, the most traffic, on average, is between 12 to 5 am with average speeds between 0 to 25 mph.

Chart

Description automatically generated

Fig 3. Histogram describing the relationship between the speed and day of the week

In Fig 3 it is portrayed that on an average from 2018 till present, the most traffic, was on Mondays with an approximate speed of 0 to 20 mph. This aligns with the fact the Monday is the first business day of the week, wherein everybody is rushing to work.

Chart

Description automatically generated

Fig 4. Histogram describing the relationship between the speed and month

In Fig 4 it is portrayed that on an average from 2018 till present, the most traffic was in July with an approximate speed of 20 to 30 mph. This aligns with the fact July falls in the summer break and people tend to travel withing that time.

Chart

Description automatically generated

Fig 2. Histogram to analyze the relationship between time of the day and GPS probes received

Fig 5 describes that on average, from 2018 to present, 12 am to 5 am was the busiest time. Comparing Fig 1 and Fig 5, we can say that our conclusions drawn from Fig 1 align with the conclusions of Fig 5.

Several other graphs were made to draw similar conclusions. These graphs can be found in the program file attached.

# Conclusions

Traffic patterns seem random but can be categorized and analyzed using various tools to find trends and patterns in their chaotic system. This project found relationships between, speed of the traffic, and the traffic itself with the time of the day, day of the week, and month for the year at an average from 2018 to present and for individual routes as well.

#### References

1. https://data.cityofchicago.org/Transportation/Chicago-Traffic-Tracker-Historical-Congestion-Esti/kf7e-cur8
2. https://neuraljojo.medium.com/a-few-matlab-table-tricks-i-had-to-learn-the-hard-way-ed5cfa876793
3. https://towardsdatascience.com/la-traffic-data-analysis-part-1-53f24e21067d
4. https://www.mathworks.com/help/matlab