# Abstract

todo

# Introduction

## Traffic in The District of Columbia

Add source and info here for how bad traffic is

In order to combat the heavy volume of traffic into the District of Columbia, Virginia has added tolling lanes to Interstate Route 66. These lanes flow eastbound in the morning and westbound in the evening, according with the heaviest traffic. These lanes function as both High Occupancy Vehicle (HOV) lanes, which can be used for free, and tolling lanes for those travelers who wish to pay.

## Problem Statement

The goal of adding tolling and HOV lanes can be interpreted as to adjust traffic patterns of travelers. The goal of this study is to measure if consumers have changed their traffic patterns. This study is focusing on measuring if traffic path patterns have changed based on the origin and destination of the trips. This means that the entire traffic of roadways will not be the main focus of the study but rather the proportion a pathway is used relative to the entire traffic for specific starting and ending locations.

## Inrix Data

Inrix operates as a third-party data compiler to mediate the data exchange between GPS service companies and data analysts. These GPS service companies provide consumers with GPS traveling information, which include but are not limited to: route planning, travel times, and police locations. These companies store the data that the consumer devices send to them. These data include the consumers position and speed. Companies likely store this data to make travel time predictions and sell to third parties such as Inrix. Inrix compiles traffic data from multiple different service companies including add companies that inrix pulls from. This compilation of data sets is then cleaned and processed. This process involves encoding user identifications and company identification for privacy.

The Virginia Department of Transportation (VDOT) has purchased multiple Inrix data sets. The data sets used for this study were the Inrix waypoint and the Inrix trips data sets. The time period for the data sets purchased were two three-month periods, March through May, for 2017 and 2018 resulting in a total of six months’ worth of data. The data sets provided were for all twenty four hours seven days a week. The Waypoint data set includes locations where the consumer was identified throughout their trip. This data includes: unique trip identifier, unique device identifier, XD segment, time stamp, and others not used in this study. waypoint data. The Trip data set is organized by individual trips made by a consumer. Each unique trip contains a unique trip identifier, trip start and end coordinates, start and end time of trip and other attributes not used in this study.

The Inrix data was provided to VDOT via “gz” zipped files. The size of the trips data set was size of trips data set (zipped or unzipped). This data set was provided in 5 files for each year, for total of 10 files. The size of the waypoint data was size of waypoint data. This data set was provided in number of files.

# Literature Review

Todo

# Methodology

## Date Filtering

More on this from Simona

As the problem the study attempts to solve involves the I-66 HOV and toll lanes the dates of the study were carefully selected to best isolate the possible changes.

Need to have the only one dependent variable, and that is that there are now tolling and HOV lanes. This means we can only look at days where the tolling is active. Only dates where tolling was active were dates that were considered for the study.

## Origin and Destination Mapping

The purchased data sets were for Northern Virginia (NOVA) traffic. Both data sets also included any through traffic that may have occurred, meaning any trips during the time period that may have passed through NOVA. To best fit the study at hand, the trip search area was condensed. The study was condensed into two separate grids, one for the Northern Virginia (NOVA) area, and the other for the DC area. Describe the two girds, size, number of points, area covered These two regions were identified in order to best represent the commuter traffic into and out of the DC area. See grids

This does not have to go here: talk about how there are overlapping sections, but this is ok, it is handled in the analysis

Each point on each grid (NOVA and DC) represent both an origin and a destination thus capturing both traffic inbound and outbound of DC. This allows, for example a NOVA point, to be an origin in the morning, and a destination in the afternoon. The two grids, NOVA and DC, interconnect completely with one another. Meaning each point in the NOVA grid connects to every point in the DC and visa versa. The points to not connect to other points from the same grid. There is no origin destination pair that has both points being from the same grid.

These grids, and resulting origin and destination pairs, further known as OD Pairs, function to capture traffic patterns in their general area. Each point functions as a bucket that will represent all the traffic starting or ending in its area. This area is defined as a circle radiating from the point. Any starting or ending traffic that happened within this circle was designated to have happened at the point exactly.

## Time Mapping

The Inrix trip data set contains the start time for each trip. As the study aims to identify possible changes in traffic patterns, these trip times needed to be isolated. Time intervals of thirty minutes were looked at. Time slots were created every half hour for 24 hours resulting in 48 separate time slots. A time slot would capture all the traffic that started during its period. An example time slot being 8:30 AM to 8:59 AM. This example time slot would represent all trips that happened during this time period.

Time slots were then assigned a day of the week. These week assignments were only assigned Monday through Friday as the study did not include weekends, see *Date Filtering*. Each day of the week had 48 time slots assigned to it. These combined, days of the week and times, resulted in 240 separate time intervals for a trip to have occurred in.

Visual to show timing breakdown, something like a google calendar

The study recognizes that it may have been more beneficial to create smaller time intervals to be looked at, such as fifteen, or five-minute intervals to gain a deeper understanding of the trip time distribution. The study choose this time interval because why did we do this

## Path Mapping

Simona

The critical part of the study was to analyze changes in traffic paths. To measure this, trips must be mapped to a pathway. This was done using the Inrix Waypoint data set. A trip was mapped to only one pathway.

## Data Structure

The data structure for capturing and storing the raw data from Inrix was built on a hierarchical structure. Each data structure built off the previous. A custom data structure was made to best facilitate the ease of the data manipulation post processing. This structure allows for information to be found or searched for that was not planned on from the beginning. This allows for flexibility in the analysis post processing. The data structure can also be used for future processing of Inrix data sets, as it is non-biased towards the project.

UML or visual diagram of data structure

The smallest object created was a path node. This node represented a unique path that could have been used. This node was used to keep track of the paths consumers used, as well as its frequency.

Built above the path node was a time node. This node represents a single time interval explained above, an example being Wednesday 8:30 AM to 8:59 AM. A time node contains multiple path nodes. The list of paths is used to measure what, and how much, separate paths are used during this specific time interval. A time node also keeps track of how frequently it is used. It is possible for a time node to have a higher count than the sum of the count of its paths.

Above the time node in the hierarchy is an origin destination (OD) node. This OD node represents one combination of the NOVA DC grids. An OD node exists for each combination of OD Pairs, resulting in count of OD nodes. Along with the origin and destination points, represented as latitude and longitude coordinates, an OD node contains a complete list of time nodes. A complete list is defined as all possible time nodes, all 48 time periods, Monday through Friday. Each individual time node contains all of its respective attributes. An OD node also keeps track of its: total, AM and PM frequency. Each OD node will represent all the traffic for its specific OD pair. Maybe more info or description

A custom hash table was built to best facilitate the adding of trips to the proper OD node. This was implemented for program efficiency. A loading factor of 0.5 was used. Should this be included

## Program

Overview

The general program flow is outlined as follows:

1. Trip to path matching using waypoint data
2. Trip data set preparation
3. OD Path analysis by small input
4. OD Path analysis aggregation
5. OD Path Data Validation should this be included

Include visual for program flow

Trip to Path

Simona about how the waypoint data was used to map trips to paths

The Inrix waypoint data set matches the customer location to XD segments on a universal road map. This allows for ease of mapping consumer location to road names. List XD maps used.

First, paths of interest were identified. The paths selected were those that were likely to have the most traffic volume and having to do with I-66. Paths that were considered as alternate routes to I-66 and paths involving I-66 directly were considered. This includes paths that may use I-66 entirely, those that use I-66 for some of the trip, and those that solely use I-66. A total of number of paths paths were identified as subject to study. These paths were chosen to give a more complete picture of what has happened (if anything) to the traffic since I-66 has been updated.

Visual of paths chosen

The information output of this process was the unique trip identifier matched with a human readable path title. An example of a path title being “I-66-W”.

During post processing these paths were validated as having enough traffic. If not enough path data was able to be collected, more paths were to be added. The study recognizes that to gain a total picture of traffic flow, all paths must be considered. A concise list of paths were chosen to give the best possible picture of the change to I-66 traffic. As the project is not concerned with the traffic patterns of all of NOVA, not all roadways were considered. The path choices are further validated below in figure containing path counts.

Each path was broken up into four relatively even segments. A trip was said to be on a path if, it was located, via the waypoint data, in all four segments at any given time. For the study this is considered a rigorous definition of a path.

System used – trip data preparation

The entire rest of the programmatic analysis was done on the University of Virginia’s High Performance Computer “Rivanna”. This method of analysis was chosen to best facilitate the projects progression forward. The programming language used was Python 3.6. The code was written and tested using an at home grade laptop, then sent to Rivanna for the data analysis. Github was used to facilitate this exchange seamlessly. Rivanna was controlled via an at home grade laptop through a Secure Shell environment.

Rivanna allows for multicore processing to be done, allowing multiple files to be analyzed separately simultaneously. To best take advantage of Rivanna’s computing power, the trip data set files provided by Inrix were joined into one complete file then split into roughly two hundred (200) separate files (203 for 2017 and 208 for 2018) with fifty thousand (50,000) data entries each.

Each smaller trip data file was analyzed independently using the technique outlined below.

Trip Analysis – od path analysis via small input

The trip data set analysis was done in layers, if the trip data entry passed the current layer, it was considered for the next layer of analysis. These analysis layers were simply identifying whether the trip was of interest to the study using the mapping techniques discussed in the “Mapping” sections above.

The first layer of analysis was to identify if the trip had occurred during the time period of interest of the study. See figure blank for reference to the time period of the study. This was done using Pythons built in data object “datetime”. As GPS is measured using satellite (get reference for this), time stamping is done using UTC times to keep consistency across the globe. As the Inrix trip data originates via GPS satellite data, the time stamps provided by Inrix were in UTC format. Adjustments were made to the Inrix trip data entries to convert them to local time (EST). Daylight savings time was considered as the change occurred during the period of interest. The data entry was considered for further analysis if the start time of the trip was during the period of study.

The second layer of analysis was to match the trip to an OD pair. This was done using the custom hash table discussed in the “data structure” section. Using the trip start and trip end coordinates provided in the trip data entries. This was done using by comparing the position of the start and end coordinates of the trip, with all of the possible OD pairs. A trip was said to belong to an OD pair if both the start and end coordinates were within the tolerance of the OD pair coordinates. A trip data entry was only mapped to one OD pair.

The third layer of analysis was simply mapping the time of the trip departure to the correct time interval. A trip was only mapped to only one time interval.

The final layer of analysis was matching the unique trip identifier to a path title. Using the output of the trip to path analysis, this was a simple one to one correlation between trip identifiers. If the trip had been mapped to a path during the trip to path analysis, then this path title was assigned to the trip during this final layer.

Aggregation

As the trip data set was analyzed in pieces, it was necessary to cumulate all the results. This was done by creating a master set of empty data objects, then continuously populating these master data objects with the results of the smaller data objects. These smaller data objects were the ones found during the smaller analysis’s.

# Findings

# Conclusion

# Sources