IGL Mid Semester Report
Greek Drivers
Spring 2021

Faculty Mentor: Prof. Richard Sowers
Project Leader: Lloyd Fernandes
IGL Scholars: Anirudh Eswara, Archie Gertsman, Ridha Alkhabaz, Sheil Kumar

March 2021

1 Introduction

In this project we try to understand the driver behaviour based on the trajectories of vehicles. For this purpose we use the pNEUMA dataset-an open large-scale dataset of naturalistic trajectories of half a million vehicles that have been collected by a one-of-a-kind experiment by a swarm of drones in the congested downtown area of Athens, Greece. The scope of the project is restricted to differentiating between a car driver and a taxi driver.

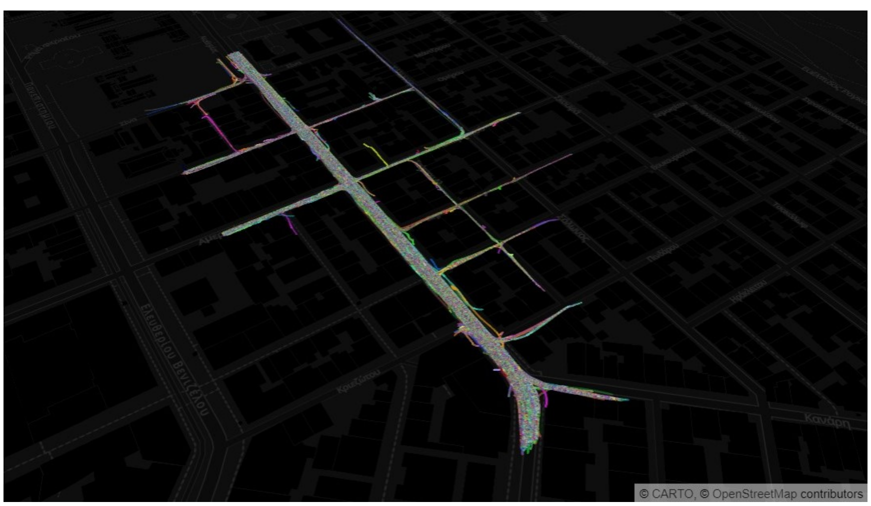

Figure 1: Trajectory of 900 vehicles passing through Block 1 between 8:30 - 9:00am on 24/10/2018

2 pNEUMA Dataset

The data set includes over half a million trajectories over the course of 4 days between 8:00 am to 10:30 am. The data was gathered using a swarm of 10 drones hovering over the central business district of Athens over multiple days to record traffic streams in a congested area of a 1.3 km2 area with more than 100 km-lanes of road network, around 100 busy intersections (signalized or not), many bus stops.[1]

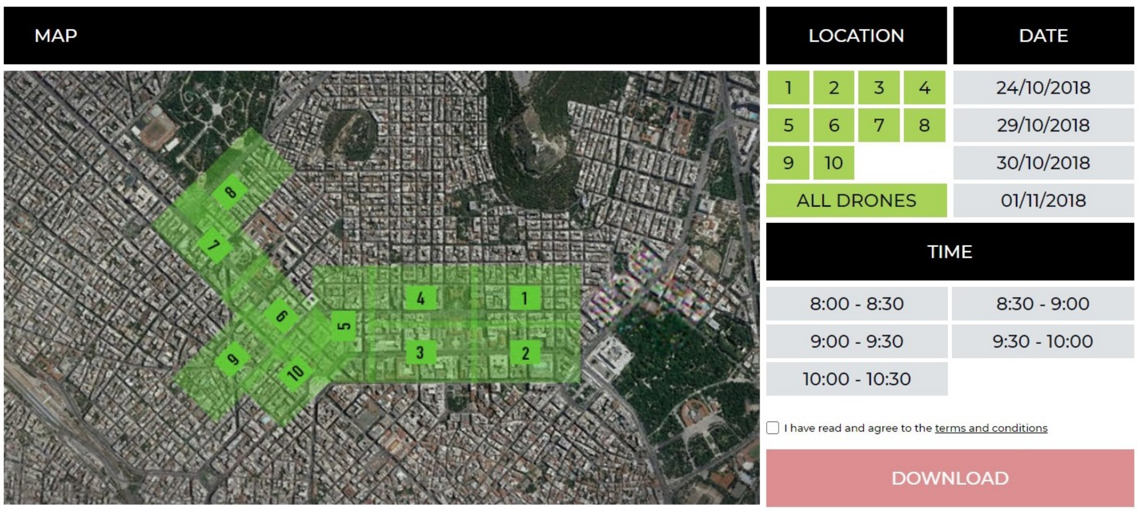

Figure 2: Screen grab of the data available at https://open-traffic.epfl.ch/index.php/downloads/

3 Work done so far

3.1 Understanding the fundamental diagram of Traffic flow

The fundamental diagram of traffic flow is a diagram that gives a relation between road traffic flux (vehicles/hour) and the traffic density (vehicles/km). A macroscopic traffic model involving traffic flux, traffic density and velocity forms the basis of the fundamental diagram. It can be used to predict the capability of a road system, or its behaviour when applying inflow regulation or speed limits.[4]:

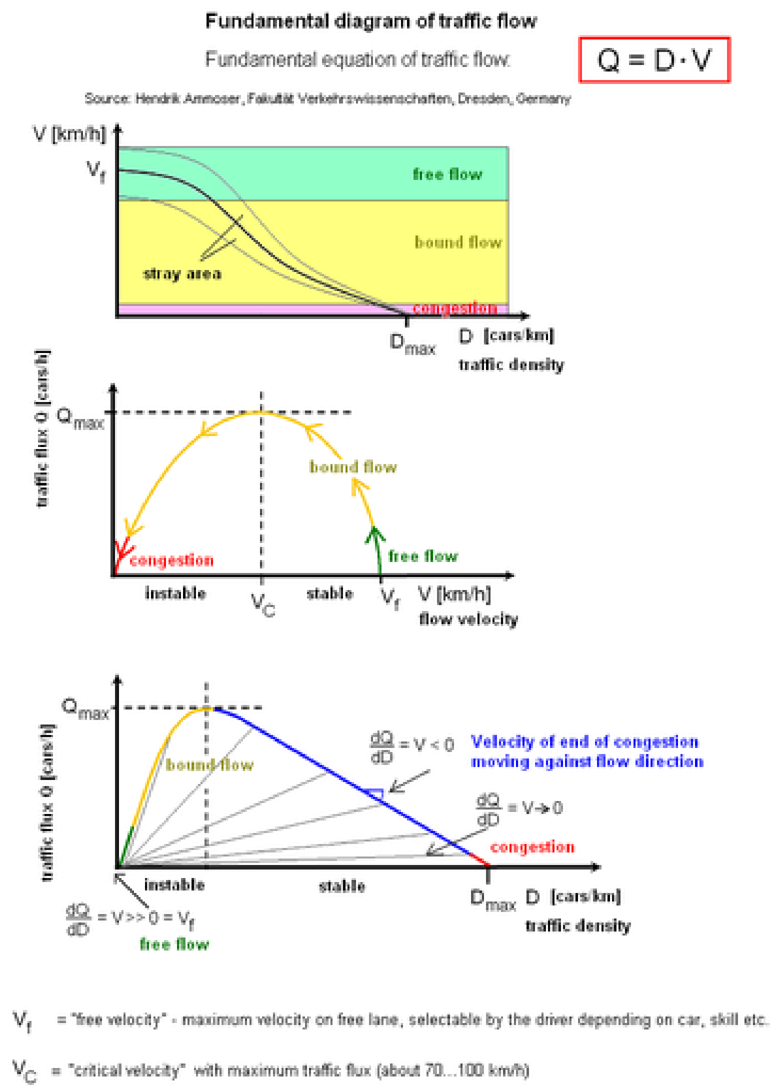

Figure 3: Fundamental diagram of traffic flow [4]

This concept can be summarised using the four statements below [4]:

1. There is a connection between traffic density and vehicle velocity: The more vehicles are on a road, the slower their velocity will be.
2. To prevent congestion and to keep traffic flow stable, the number of vehicles entering the control zone has to be smaller or equal to the number of vehicles leaving the zone in the same time.
3. At a critical traffic density and a corresponding critical velocity the state of flow will change from stable to unstable.
4. If one of the vehicles brakes in unstable flow regime the flow will collapse.

3.2 Data Preprocessing

The csv file by pNEUMA, consists of a string of variables collected at every 0.04 sec. These rows of series were converted to a Multi-index dataframe, with trajectory ID and time as its indices. This dataframe was then augmented with additional features which could be useful with predicting driver behaviour. The features are as follows:

1. Nearest street and intersection: the nearest street and intersection data is collected by querying OpenStreetMaps[2]. We used osmnx, a dedicated python library, which fetches map data information from OpenStreetMaps. The map data is codified as a multi-directed graph with edges representing the road and nodes representing intersections.(we would use the term edges/nodes and road/intersections interchangeably).

2. Bearings: It is the direction in which the vehicle moves with respect to the true north. This feature was calculated using dedicated function in osmnx.

3. Cross-Track distance: This is the transverse distance of the vehicle with respect to the edge. This feature was calculated using a dedicated function in osmnx.

4. Edge progress: This estimate describes the percent distance travelled by a vehicle on a road.It was calculated by getting the distance of the vehicle from one of the nodes. This node was selected based on its unique ID (fetched from the multi-directed graph). The smaller of the two node ID's was decided to be the first node. As most of the roads in the region of concern are straight, the inaccuracies induced by curved roads are not present.

5. Vehicle direction: In order to get an estimate of the lane changes, the direction of the vehicle up or down an edge was calculated using the trend in the edge progress. If this estimate reduced with successive time stamps, the vehicle is moving towards the first node, else it is moving towards the second node.

6. Traffic around the vehicle: This was calculated by converting edge into smaller segments and calculating the number of vehicles in the segment at a given time stamp. The location of the vehicle can then be used to look up the traffic on that segment at a given time. The segments are categorised using the edge progress estimate.

7. Average vehicle speed: This estimate shows the average vehicle speed for all the vehicles in a given segment of an edge at every time-stamp. This average speed is then used to center all the vehicle speeds in that segment at every time-stamp.

3.3 Data Analysis

So far we have used t-distributed Stochastic Neighbor Embedding to check whether the data can be segregated or not using the features available. t-SNE [3] is a tool to visualize high-dimensional data. It projects multidimensional data onto a 2/3-dimensional plane while preserving underlying clustering structure of the dataset. Using t-SNE plots can help us pNeuma Dataset | Added Features

id	time	lat	lon	speed	lon_acc	lat_acc	type	traveled_d	avg_speed	nearest_node	nearest_edge_start_node	nearest_edge_start_node_lon
1	0.00	37.979838	23.733420	11.6261	0.0	-0.0	Car	1.81	10.850992	250698924	250698924	23.73382
	0.04	37.979839	23.733419	11.6261	-0.0	-0.0	Car	1.81	10.850992	250698924	250698924	23.73382
	0.08	37.979840	23.733418	11.6261	0.0	-0.0	Car	1.81	10.850992	250698924	250698924	23.73382
	0.12	37.979841	23.733418	11.6261	-0.0	-0.0	Car	1.81	10.850992	250698924	250698924	23.73382
	0.16	37.979842	23.733417	11.6261	-0.0	0.0	Car	1.81	10.850992	250698924	250698924	23.73382

nearest_edge_start_node_lon	nearest_edge_start_node_lat	nearest_edge_end_node	nearest_edge_end_node_lon	nearest_edge_end_node_lat	edge_progress	vehicle_dir
23.73382	37.979335	250698925	23.732912	37.980373	0.470782	1
23.73382	37.979335	250698925	23.732912	37.980373	0.471786	1
23.73382	37.979335	250698925	23.732912	37.980373	0.472789	1
23.73382	37.979335	250698925	23.732912	37.980373	0.473461	1
23.73382	37.979335	250698925	23.732912	37.980373	0.474465	1

Figure 4: Data frame post data preprocessing (red box indicates the data processing done on the original csv file while green box highlights the added features

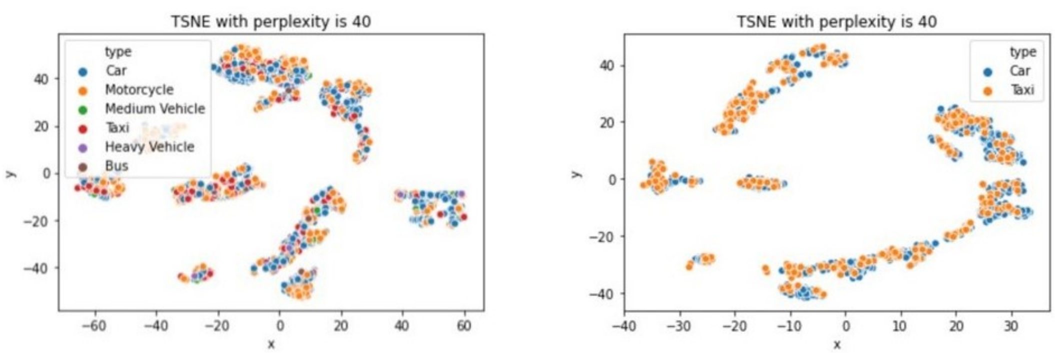

Figure 5: Left is a t-SNE plot of over 3000 vehicles at 8:30 am in Athens in 24/11/2018. Right is a t-SNE plot of over 900 cars and taxis at 8:30 am in Athens in 24/11/2018.

visualize clusterability of the data. 5 is the t-SNE plot for the data-frame generated from pre-processing.

The ellipsoidal shapes in the underlying clustering structure shows that the data can be clustered using Unsupervised Learning Algorithms.

4 What next?

We would be on look out for other features and use all of these features to distinguish between taxi and car drivers. We would survey a range of Supervised and Un-supervised learning algorithms like logistic regression, Gaussian Mixture models and k-means clustering over t-SNE etc for classification. We would also be checking for robustness of the dataset.

References

[1] Emmanouil Barmpounakis and Nikolas Geroliminis. On the new era of urban traffic monitoring with massive drone data: The pneuma large-scale field experiment. *Transportation Research Part C-Emerging Technologies*, 111:50–71, 2020.

[2] OpenStreetMap contributors. Planet dump retrieved from https://planet.osm.org . https://www.openstreetmap.org , 2017.

[3] L.J.P. van der Maaten and G.E. Hinton. Visualizing high-dimensional data using t-sne. 2008.

[4] Wikipedia contributors. The fundamental diagram of traffic flow— Wikipedia, the free encyclopedia, 2021. [Online; accessed 22-March-2021].