DETECTING URBAN MOBILITY BEHAVIORS USING MACHINE LEARNING

by

Dhanush Dinesh, ECE

Priyank Arya, CS

Vinay Shanmukh Aradhya, ECE

Zhe Meng, ECE

Advisor[s]:

Dr. Nektaria Tryfona, Associate Professor, Virginia Tech

May 5, 2022

Submitted in partial Fulfillment of the Requirements for Design Project

Bradley Department Electrical and Computer Engineering

Virginia Tech

Abstract

Although multiple factors can affect urban mobility, it seems that no single factor has changed the world more than the COVID-19 pandemic. Due to the concern of infection, people had to rethink their mobility choices and behavior. Majority of the people were traveling primarily for grocery shopping or for availing medical services during the pandemic. To study these significant changes in the mobility patterns and linked public behavior, we developed a Machine Learning model based on DBSCAN which is a density based clustering algorithm to understand the impact of the pandemic on urban mobility across different sections of the society. We observed that, as expected, fewer people traveled and the average trip length during pandemic was reduced. Usually popular places like tourist spots, airports, etc. were not crowded during the pandemic. At the same time, hospitals and cemeteries received more people than before. For some special federal agencies such as the CIA and Pentagon, no change in traffic flow was observed during COVID-19 as compared to pre-covid flow. We also found that demographic features like race, education level and the unemployment rate also had a significant affect on the general mobility trend. Wealthy neighborhood with mostly white people saw drastic reduction in movement during COVID-19 while most colored neighborhoods saw an increase in mobility. The above findings indicate that the proposed model can effectively analyze urban mobility patterns and help in identifying and understanding the urban mobility changes during the COVID-19 period.

Acknowledgments

This report and the research behind it would not have been possible without the exceptional support of our mentor, Dr. Nektaria Tryfona, Collegiate Associate Professor, Virginia Tech. Her enthusiasm, knowledge and exacting attention to detail have been an inspiration and kept our work on track from our first encounter with the basics of Urban Mobility to the final draft of this paper. This opportunity has taught us more than we could ever give her credit for here. She has shown us, by her example, what a good mentor (and person) should be. We are grateful to the team that happened with whom we have had the pleasure to work during this project. Each of the members of this team has provided extensive personal and professional guidance and taught a great deal about both scientific research and life in general. Our team member's constant support was vital to us in the pursuit of this project. We would like to thank Dr. Vassilios Kovanis, Collegiate Professor and MEng Program Director for Northern Virginia Campus for introducing Project Based Learning for the Spring'22 semester. This learning experience provided a chance for us to work on a major topic related to the general public and understand the effects of Covid-19 on Urban Mobility. Finally, we thank all the teaching staff and technical staff of The Bradley Department of Electrical and Computer Engineering, Virginia Tech for their help and support.

Contents

Introduction

Urban Mobility means with all its grammatical variations "all aspect of transport infrastructure, facilities, vehicles and services available to the general public in an Urban Mobility Region, including but not limited to private transport vehicles and services and all modes and means of transportation within the categories specified under Schedule" [1]

1.1 Background

It is estimated that "by 2050, 89% of the U.S. population and 68% of the world population will be living in urban areas." [2]. With the process of urbanization, urban mobility has been paid more and more attention. The meaning of urban mobility has gone beyond the original definition. It is more about the "accessibility to various urban services." [3] That is why our city planners and policymakers spend vast amounts of money and time on urban mobility.

COVID-19 had a huge impact on urban mobility. Due to the concern of infection, people had to rethink their mobility choices and behavior. Many prior studies exploring the influence of COVID-19 on traveling behavior showed that people decreased their trips significantly. The use of public transportation, such as buses and metro, is also found to have a noticeable drop [4]. Even after travel restrictions were lifted, people were still hesitant to travel due to increased chances of being infected with the virus. Although the pandemic affected people of all ages, the elderly and children were more prone to get easily infected and hence were further refrained from traveling. The pandemic also lead to the rapid evolution of various online businesses and work-from-home was implemented by most Tech Companies. Accordingly, majority of the people were traveling primarily for grocery shopping during the pandemic[5].

1.2 Problem Statement

COVID-19 has caused significant changes in the behavior and the mobility patterns. We developed a model for spatiotemporal assessment to understand the impact of the pandemic on urban mobility across different sections of the society.

1.3 Objective

With the help of a suitable Machine Learning Model, we wanted:

- To verify the general changes summarized by the previous studies
- To analyze the mobility data and predict future mobility trends if there is another disruptive event like COVID-19
- To understand the mobility pattern changes due to COVID-19 across different sections of society and their adaptation to the changes.

1.4 Assumptions and Limitations

While taking only Northern Capital Region (NCR) as the study subject, we can expect some differences in the results across different cities. However, we observed that most COVID-19 countermeasures applied by the NCR local government are also adopted by other cities not only in the U.S. but also all around the world. In this case, we expect our result can be, to some extent, applied to other cities beyond the NCR.

1.5 Organization of the report

The outline of the report is as follows:

- **Chapter 1: Introduction** This chapter deals with introduction of the project with enough details to understand the premise of the project.
- **Chapter 2: Literature Review:** This chapter presents an overview of the recent

related literature in the field of urban mobility and machine learning.

- **Chapter 3: Design and Implementation:** This chapter describes the overall methodology and all the clustering methods we implemented and the reason why we finally chose DBSCAN. This chapter also discusses about data preprocessing and model evaluation.
- **Chapter 4: Analysis** This chapter deals with analysis of the results which includes visualization and story telling.
- **Chapter 5: Conclusion** In this chapter we describe the conclusion drawn from the results and entail the future scope of the project.

Literature Review

Understanding the change in traffic and mobility patterns before, during, and after the COVID period is a trending topic of research today. Applying twelve scenarios, Advani, M, Sharma, N and Dhyani, R (2021) [5] studied New Delhi's mobility changes and the demand with non-motorized transportation. The study shows that compared to the pre-Covid period, unlocking level 3 has a vehicle kilometer traveled (VKT) reduction of 19% in Motorized Two-Wheelers (MTWs), 5% in Cars and 49% in Buses. Private vehicles are more preferred because of the infection concern. The increase in bicycle trips has been estimated to be 5.88 million for the post-lockdown Scenario compared to 1.1 million trips estimated for the pre-Covid Scenario. As a result of the significant modal shift from motorized to non-motorized, a decrease in both vehicular emission and road accidents is observed (Alfredo Aloi et al. 2020; Abdullah, M et al. 2021) [6].

As a powerful tool, machine learning is often used in COVID case tracing and urban mobility modeling separately. Combining clustering and the feature selection techniques, Khmaissia F et al. (2020) determined patterns of Zip code-level increase in the number of new COVID-19 cases in megacities like NYC [7]. And by using the machine learning technique, Kuo, C.-P and Fu, J. (2020) indicated that, compared with the Phase I re-opening, a 1-week and a 2-week lockdown could reduce 4% – 29% and 15% –55% infections, respectively, in the future week, while the 2-week Phase III re-opening could increase 16% – 80% infections [8]. Using ML applications to study and understand the changes in Urban Mobility, Song H.Y and You D (2018) established a method using Clustering techniques (DBSCAN and GMM) to identify and analyze urban mobility models based on real taxi transportation data [9]. In another study focusing on regional mobility patterns, combined with the nonnegative tensor factorization, the clustering technique (fuzzy C-means) is used to provide more meaningful region division and higher interpretability of the extracted data (Qi G et al. 2019) [10]. Moreover, Vidovic K, Mandzuka S and Brcic D (2017) proposed an adaptive neuro-fuzzy inference system

(ANFIS) to build an urban mobility index, enabling a new approach for urban mobility assessment based on a real domain expert's expertise [11].

Compared with the large quantity of research solely about Machine Learning and urban mobility, a few pioneer studies have tried to apply various Machine Learning methods to analyze the impact of urban mobility during this worldwide pandemic period. Using the biclustering algorithm to discover different traffic patterns, Aparicio, J. T, Arsenio, E, and Henriques, R (2021) introduced a dynamic assessment of the effects of COVID-19 on public transport use [12]. Besides, Paiva, S et al. (2022) used the K-means clustering method to analyze how specific age groups' mobility changed during the Covid period [4]. And as one of the results, they found that the mobility in retail and recreation areas is reduced compared to the pre-pandemic period.

To sum up, although previous studies have made outstanding contributions to mobility pattern analysis due to COVID-19 impact, there are still some limitations:

- Some proposed techniques before COVID are not examined with the situation under COVID.
- Some studies are statistics based, which cannot dynamically assess urban mobility changes during the COVID period.
- The general approaches to analyzing the urban mobility changes during the COVID visually, numerically, and dynamically are not well developed yet. In case of these limitations, a spatiotemporal assessment model for understanding the pandemic's impact on urban mobility is introduced in the following section.

Design and Implementation

The below Fig. 3.1 represents the design methodology of our project's. We used metro ridership data from Smartrip, user mobility data from GeoDS, and demographic data from City-Data website. We then preprocessed the data and implemented multiple clustering methods like K-Means, DBSCAN, Agglomerative Clustering, and Affinity Propagation on the cleaned data and evaluated the methods using Silhouette Score to find the clustering method which best suits our data. We then analyzed the clusters to study the changes in the mobility patterns.

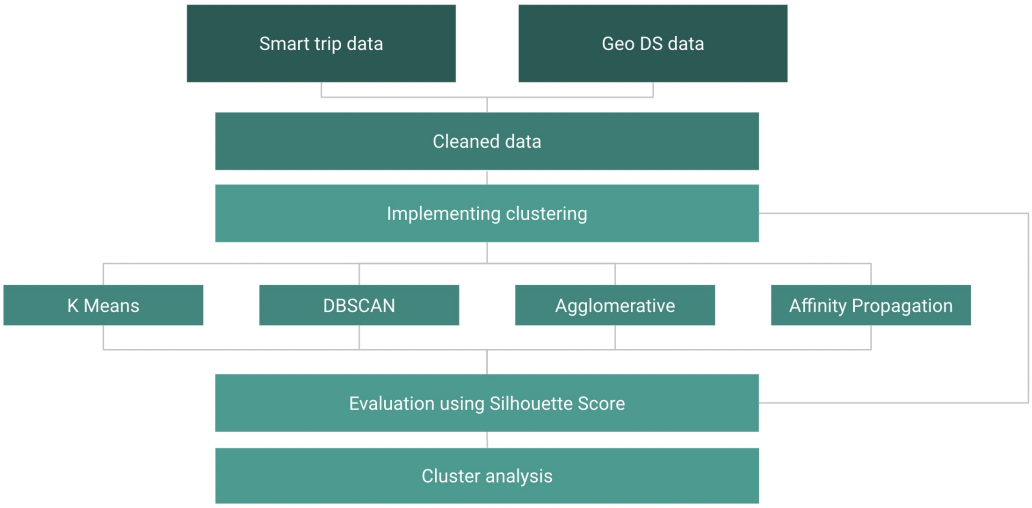

Figure 3.1: Design methodology

3.1 Team Organization

The team was lead by Dhanush Dinesh. The literature survey was lead by Vinay Aradhya. Priyank Arya led the ML modeling part to identify the most suitable algorithm for our project and Zhe Meng took the lead in the visualization. The implementation was done by the whole team led by Dhanush Dinesh. Below Fig. 3.2 shows the organization chart which describes the roles and responsibility.

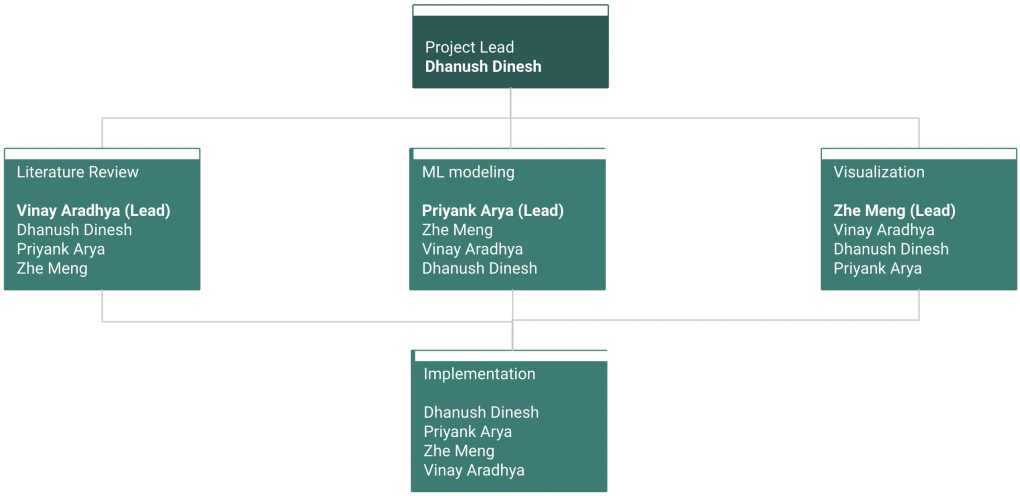

Figure 3.2: Organization Chart

3.2 Data Acquisition

There are two kinds of data used for this project; census data and mobility data. The Census data is available on the official website of City-data.com. Mobility data includes Smartrip data and GeoDS Data[13].

Smartrip is the smart card payment system managed by the Washington Metropolitan Area Transit Authority (WMATA) and Smartrip data here refers to the payments made by a user at the entry and exit checkpoints, thereby recording their trip origin and destination locations.

GeoDS here stands for the Geospatial Data Science lab of University of Wisconsin and the data they provide is openly available on Github and it contains the geographical information of a user's travel trajectories, i.e., their origin and destination latitude and longitude values.

City-data website provides detailed census data with informative profiles such as Living costs, unemployment rate, average income, etc. for every county and city in the United States which can be utilized to get some insight into the users who are making the trip.

3.2.1 Smartrip Data

Our hourly research is based on Smartrip data. The data contains the main table providing Entry and Exit information on dates, times and specific station IDs, and a mapping table containing station IDs, station names, and their geographic information. The smartrip data and mapping data samples are presented in Table 3.1 and Table 3.2. Smartrip data consists of CUSTOMER_CODE, ENTRY_MSTN_ID, EXIT_MSTN_ID, ENTRY_DTM and EXIT_DTM. The CUSTOMER_CODE stands for the unique customer taking the specific trip. Together with ENTRY_MSTN_ID, EXIT_MSTN_ID and the corresponding information in the mapping table, we can get the entrance and exit information of each specific trip by all the customers using the metro service. ENTRY_DTM and EXIT_DTM stands for the date and time of the trip. For example, the first row of Table 3.1 can be interpreted as one person who started their trip from the station Stadium Armory at 6:46 on 5/1/2016 and ended their trip at the station Cheverly at 1:00 on 5/1/2016.

Table 3.1: Smartrip data—main table (sample: raw data)

CUSTOMER CODE	ENTRY MSTN ID	EXIT MSTN ID	ENTRY DTM	EXIT DTM
D46F3D36B96	MSTN_033	MSTN_058	5/1/2016 6:46	5/1/2016 18:00
F4301187B8D	MSTN_036	MSTN_077	5/1/2016 8:42	5/1/2016 18:20
F5805BF1EE7	MSTN_024	MSTN_060	5/1/2016 0:41	5/1/2016 0:57
620F21476261	MSTN_019	MSTN_019	5/1/2016 2:40	5/1/2016 2:47
AC4D98DD5F	MSTN_019	MSTN_055	5/1/2016 0:23	5/1/2016 1:01

Table 3.2: Smartrip data—mapping table (sample: raw data)

STATION _ID	NAME	lat	lon
MSTN_033	Stadium Armory	38.88672	-76.9771
MSTN_036	Civil War Memorial	38.91701	-77.0275
MSTN_058	Cheverly	38.91664	-76.9166
MSTN_077	Branch Ave	38.82645	-76.9115

3.2.2 GeoDS Data

Research around the daily patterns is done using the GeoDS Data[13], which is openly available on Github and is hosted by the Geospatial Data Science Lab of University of Wisconsin, Madison. "By analyzing millions of anonymous mobile phone users' visit trajectories to various places provided by SafeGraph"[13], they estimated "the daily and weekly dynamic origin-to-destination (O-D) population flows". And the data were "computed, aggregated, and inferred at three geographic scales: census tract, county, and state." Here, we chose the level of the census tract. The sample of the initial GeoDS data is presented in Table 3.3. The GeoDS data include geoid_o, geoid_d, lng_o, lat_o, lng_d, lat_d, date, visitor_flows and pop_flows. The geoid_o, lng_o and lat_o stand for the geographic information of the origin, while correspondingly, geoid_d, lng_d and lat_d stand for the geographic information of the destination. Vistor_flows stands for the exact number of moving people reported by the mobile phone and pop_flows stands for the inferred population level of dynamic O-D flows[13]. We use pop_flows rather than vistor_flows as the source data in our project. For instance, the first row of Table 3.3 can be interpreted as a trip of one visitor starting at the location with geo ID 51510200303 and ending at the location with geo ID 24033802001 on Jan 6th , 2020. The inferred number of population flows resulting from this trip is 8.

Table 3.3: GeoDS data (sample: raw data)

geoid_o	geoid_d	lng_o	lat_o	lng_d	lat_d	date	visitor flows	pop flows
515102003	240338020	-77.13	38.82	-76.92	38.84	1/6/20	1	8
515102012	240317028	-77.06	38.84	-77.04	39.01	1/6/20	1	12
110010088	110010072	-76.99	38.90	-76.95	38.89	1/6/20	1	10
110010021	240338036	-77.02	38.96	-76.84	38.97	1/6/20	1	9
240338020	240338013	-77.92	38.84	-76.97	38.76	1/6/20	1	9
240317028	110010075	-77.04	39.01	-76.98	38.86	1/6/20	1	9

3.3 Data Preprocessing

Preprocessing is a data mining technique which is used to transform raw data into an understandable format. This step is important because the raw data (real world data)

is always incomplete and that data cannot be sent through a machine learning model as that would cause errors. Therefore, before implementing clustering, we cleaned our data by performing preprocessing steps of filtering and merging to get the total trip counts for both the Smartrip and GeoDS Data. Fig 3.3 shows the workflow of data preprocessing.

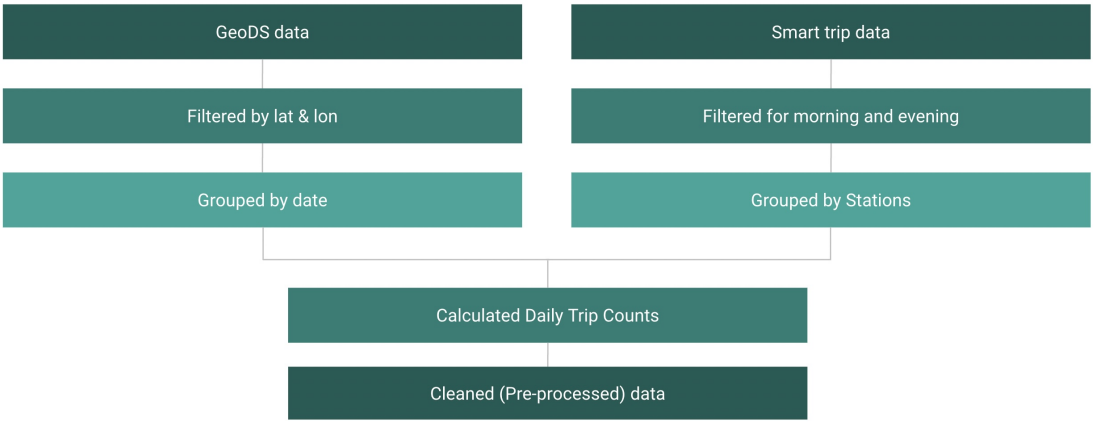

Figure 3.3: Workflow of data preprocessing

Smartrip Data

The filtering step for smartrip data involved grouping the records by date and time and cutting the whole dataset into two pieces for each day. One named date_morning, which included the data from 6 am to 10 am, and another named date_evening, had the data from 3 pm to 7 pm, dropping out data from other periods. Then came the merging step. We converted the data sets from customer basis to station basis. Each row of the raw data was counted in the corresponding entry station and exit station and added separately in ENTRY_COUNT and EXIT_COUNT. We then combined the ENTRY_COUNT and EXIT_COUNT of each station and concatenated the station name and geographic information from the mapping table. For example, if Table 3.1 represented the whole data for May 1st, 2016, Table 3.4 and Table 3.5 would be created using Table 3.2, named as 2016_5_1_morning.csv and 2016_5_1_evening.csv. Joining and combining all such tables will result in Table 3.6 which has a snippet of the cleaned data.

Table 3.4: Smartrip data (sample: data under calculating1)

STATION _ID	NAME	lat	lon	ENTRY _COUNT	EXIT _COUNT
MSTN_033	Stadium Armory	38.88	-76.97	1	0
MSTN_036	Civil War Memorial	38.91	-77.02	1	0

Table 3.5: Smartrip data (sample: data under calculating2)

STATION _ID	NAME	lat	lon	ENTRY _COUNT	EXIT _COUNT
MSTN_058	Cheverly	38.92	-76.92	0	1
MSTN_077	Branch Ave	38.83	-76.91	0	1

Table 3.6: Smartrip data (sample: cleaned data)

STATION _ID	NAME	lat	lon	ENTRY _COUNT	EXIT _COUNT
MSTN_001	Anacostia	38.86	-77.00	1480	2612
MSTN_002	Navy Memorial	38.89	-77.02	5645	726
MSTN_003	Benning Road	38.89	-76.94	379	1164
MSTN_004	Brookland-CUA	38.93	-76.99	1721	2314
MSTN_005	Capitol South	38.89	-77.01	3497	1048

GeoDS Data

The initial GeoDS Data contained records of the whole nation. We first filtered data sets by our target latitude and longitude, only keeping records having their origin or destination inside NCR. The raw data sets were trip-based, and we converted them into location-based ones. So, similar to what we did with the smartrip data, we separated each row into the exit and entry parts. And then, we assigned the number of flows separately according to the corresponding location, under EXIT_COUNT and ENTRY_COUNT. Finally, we added the numbers of EXIT_COUNT and ENTRY_COUNT and assigned them to each location. For example, if Table 3.3 is the data set of Jan 6th , 2020. We would create a table like Table 3.7 below and name it 2020_01_06.csv. Table 3.8 is a real snippet of the cleaned data.

Table 3.7: GeoDS data (sample: data under calculating)

GEOID	lat	lon	EXIT_COUNT	ENTRY_COUNT
51510200303	38.82	-77.13	8	0
51510201203	38.84	-77.06	12	0
11001008802	38.90	-76.99	10	0
11001002101	38.96	-77.02	9	0
24033802001	38.84	-77.92	9	8
24031702800	39.01	-77.04	9	12
11001007203	38.89	-76.95	0	10
24033803607	38.97	-76.84	0	9
24033801305	38.76	-76.97	0	9
11001007504	38.86	-76.98	0	9

Table 3.8: GeoDS data (sample: cleaned data)

GEOID	lat	lon	EXIT_COUNT	ENTRY_COUNT
11001000100	38.90573	-77.0608	8708	19532
11001000201	38.90933	-77.0748	6020	9490
11001000202	38.90619	-77.0696	9155	17904
11001000300	38.91756	-77.0756	10412	7370
11001000400	38.92385	-77.0659	3613	5276
11001000501	38.92656	-77.0517	6168	7482
11001000502	38.92835	-77.0596	5490	4018

Cleaned Data

After preprocessing, the cleaned data consist of GEOID, lat, lon, EXIT_COUNT and ENTRY_COUNT, as shown in Table 3.8, The Data files were created for each date, named after the occurring date. GEOID here is the identifier of NCR at census tract level and 'lat' and 'lon' stands for the latitude and longitude corresponding to that GEOID. EXIT_COUNT stands for the number of people leaving, while ENTRY_COUNT stands for the number of people arriving. For example, assuming the data file is named 2020_01_20.csv, the first row of Table 3.8 indicates that on that day, GEOID 11001000100 had 8708 people arriving in and 19532 people leaving out.

3.4 Clustering

While working with unstructured and unclassified datasets, clustering techniques are often used to find and group similar entities together so that insights can be derived from these groups. We used clustering as we wanted to visualize how the data points are distributed across the space and group similar data points together so that we can analyze and profile the underlying attributes of each group and derive insights in the mobility pattern changes across the socio-demographics.

3.4.1 Different methods of clustering

K-Means, Affinity Propagation, Hierarchical or Agglomerative clustering, and DBSCAN are few of the popular clustering techniques that we implemented and evaluated for our model.

K-Means

K-Means clustering is used for partitioning an N-dimensional population into k clusters, where the value of k is to be specified prior to clustering. It is a widely used technique that makes clusters based on geometric distances between points. The clusters are grouped around centroids, causing them to be globular in nature which is one of the major drawbacks of using K-Means clustering because if the underlying clusters are

not globular then K-Means produces poor results.

Affinity Propagation

Affinity Propagation makes clusters based on the graph distances between the points leading to smaller uneven clusters. The user doesn't have to specify the number of clusters like in K-Means, but it does not produce good results if the underlying clusters are non globular. Also, it is difficult to scale for large datasets as it is computationally expensive.

Agglomerative

Hierarchical or Agglomerative clustering generates a hierarchy of clusters. It is good for non globular clusters and scales well to large datasets but just like K-Means, the user must specify the number of clusters prior to clustering.

DBSCAN

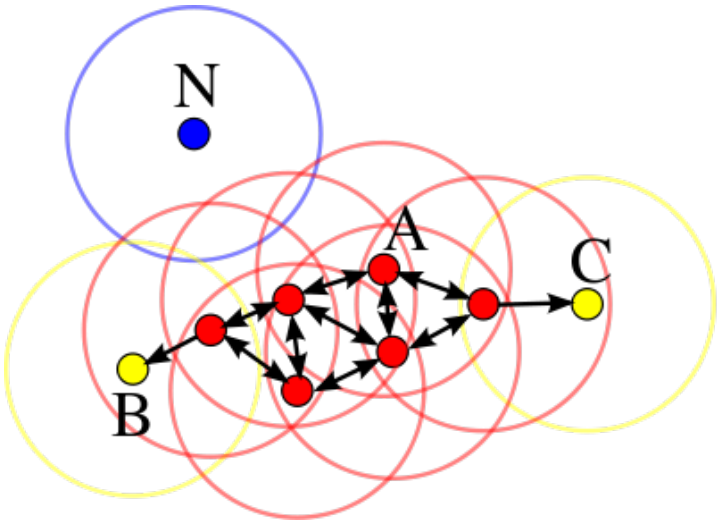

Figure 3.4: Representation of DBSCAN clustering

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a density-based clustering algorithm that works on the assumption that clusters are dense regions in space separated by regions of lower density. Therefore, it is great at separating clusters of high density versus clusters of low density and it performs well with arbitrary shaped clusters. This fits perfectly for our dataset as the data will be dense at some places

like schools, business districts, shopping malls, while being sparse at other places like suburban areas.

DBSCAN does not require every point to be assigned to a cluster thereby reducing the presence of noise in the clusters and based on the parameters of Epsilon and Minimum Points, we can classify each data point as core point, border point or noise point. As shown in the Fig. 3.4, the red points form the core points of the cluster, the yellow points form the border point of the cluster while the blue point is not considered a part of the cluster and hence is marked as a noise point.

3.5 Experiment / Evaluation

Silhouette Coefficient or Silhouette Score is a metric which is used to evaluate the goodness of a clustering technique. We implemented the above four clustering techniques on our datasets and then calculated their Silhouette Score as can be seen in the Table 3.9. We found that a clustering algorithm having a score closer to zero provides a better fit for our data which resulted in the selection DBSCAN for our Machine Learning model implementation.

Table 3.9: Clustering evaluation using silhouette score

SI	Clustering method	Silhouette Score
1.	Affinity Propagation	0.67
2.	Agglomerative	0.56
3.	K-Means	0.53
5.	DBSCAN	0.16

Analysis

4.1 Clustering results

Smartrip Data

In Fig. 4.1 and Fig. 4.2, the data distribution corresponds to the Washington DC metro network for the morning of May 2, 2016. For the data associated with the Entry points, we can see that we have obtained 19 clusters using the DBSCAN technique as shown in the Fig. 4.3 and 6 clusters for the exit points as shown in the Fig. 4.4. The exit clusters are centered around the US Capitol region, Washington National Cathedral and Ronald Reagan International airport. Here the word "Entry" means stations from where people are starting their metro ride while the word "Exit" means places where people are ending their metro ride. For the metro station, the entry station is the origin, and the exit station is the destination. The Fig. 4.2 indicates that the majority of the people's exit is centered around the attraction centers in the DC area. This depicts the fitting representation of the movement of the people with respect to a federal holiday when there are special events organized for the general population around the major attraction centers in DC.

We can say that the daily flow of the people is mostly from suburbs to the city center

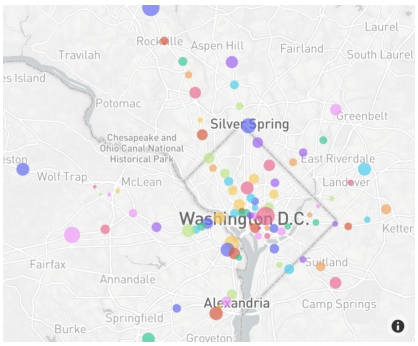

Figure 4.1: Morning Entry, Smartrip data

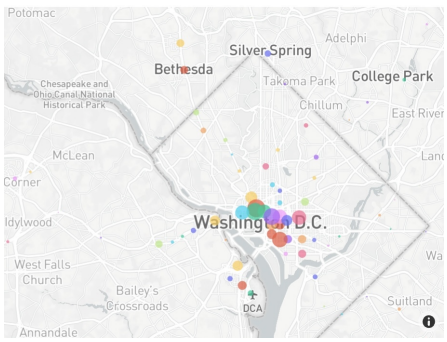

Figure 4.2: Morning Exit, Smartrip data

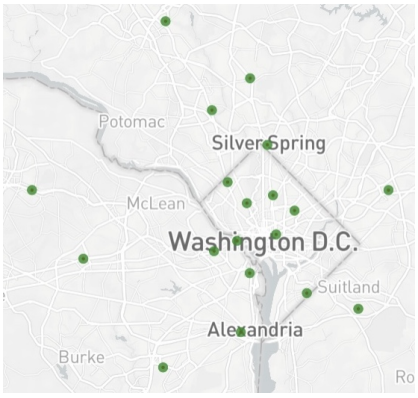

Figure 4.3: Morning Entry, Smartrip cluster data with 19 clusters

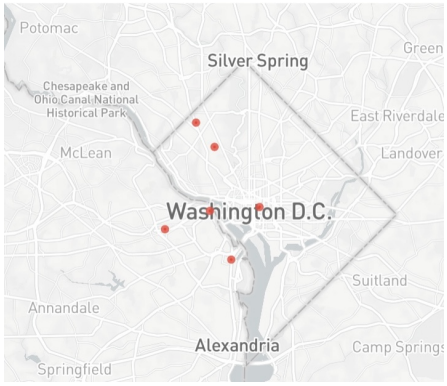

Figure 4.4: Morning Exit, Smartrip cluster data with 6 clusters

in the morning and from city center to the suburbs in the evening as can be observed in the Fig. 4.1, 4.2. The interesting fact that there are more clusters in the morning with more entrances signifies the wide distribution of the points. Due to the exits being more focused in the center of the city we can observe less clusters.

We observed 6 clusters in the central DC area for May 2, 2016 as shown in Fig. 4.7. These are the entry points of people for the evening of Labor day. In Fig. 4.5, we can observe the data points' circles are larger as compared to the exit points of Morning (Fig. 4.2) indicating that there is a larger number of people who are exiting from the central region of DC and are moving to their respective homes. This explains the fact that there are comparatively a large number of people who were present near the major monumental landmarks of DC which includes the White House, United States Capitol, Lincoln memorial and so on.

The exit cluster for the evening of May 2, 2016 in Fig. 4.8 shows the exit points of people who took the metro. This indicates that the people's exits are spread throughout the spread of the metro lines. The major clusters are spread around East Riverdale, Sultland, Alexandria, Fairfax, Gaithersburg which means that people traveled to these places at the end of the day. Major clusters are seen in the central DC area because of the day being a federal holiday and the inflow of people to the capitol could be seen in the evening as well.

On comparing the clustering data of Morning Entry with 19 clusters (Fig. 4.3) and the clustering data of evening entry with 16 clusters (Fig. 4.8), we observe that the

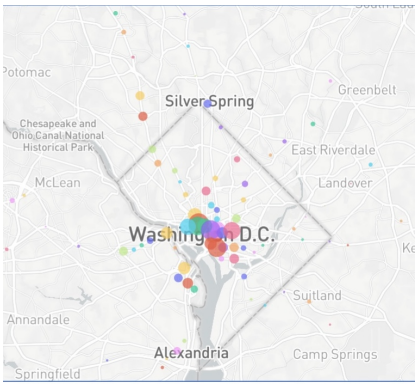

Figure 4.5: Evening Entry, Smartrip data

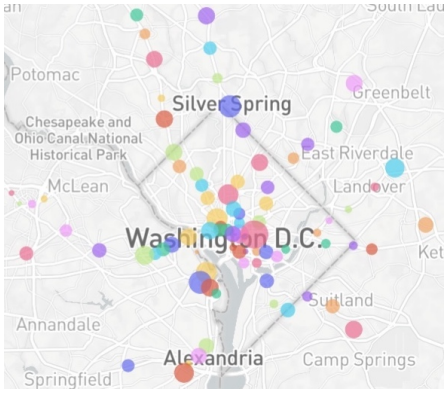

Figure 4.6: Evening Exit, Smartrip data

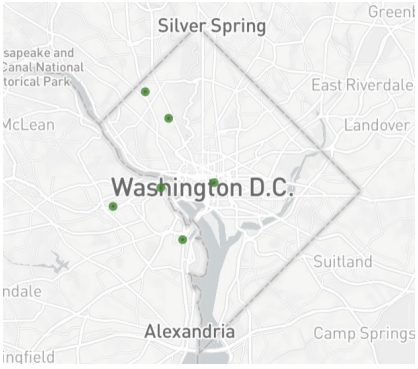

Figure 4.7: Evening Entry, Smartrip cluster data with 6 clusters

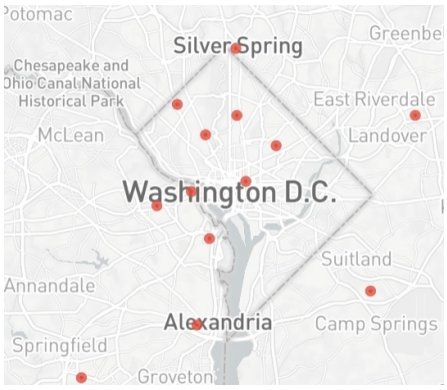

Figure 4.8: Evening Exit, Smartrip cluster data with 16 clusters

clusters are similar but there is a decrease in the number of clusters in the evening and all the 3 clusters which are missing are observed to be near Silver Spring. This suggests that either the people have not traveled back or they have used other modes of transportation.

GeoDS Data

So, based on our definition of "Entry" and "Exit" as defined for the Smartrip data, in the case of GeoDS data, the "Entry" place refers to the destination (as people are arriving to that GEOID) and the "Exit" place refers to the origin (as people are leaving from that GEOID). This is different from the metro station scenario that we discussed earlier. In the plots, we can observe that in Fig. 4.9 the points' circles are much bigger than those in Fig. 4.10 which is of 2021. It can be inferred that the traffic flow was reduced considerably in 2021 due to the pandemic. Here, the points on the plots stand

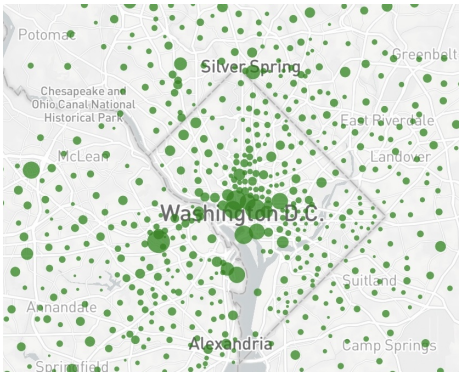

Figure 4.9: Jan 22, 2020 Entry, GeoDS Data

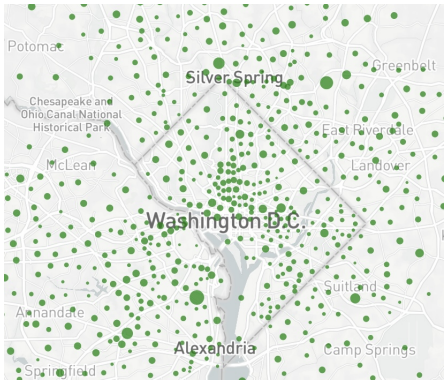

Figure 4.10: Jan 27, 2021 Entry, GeoDS Data

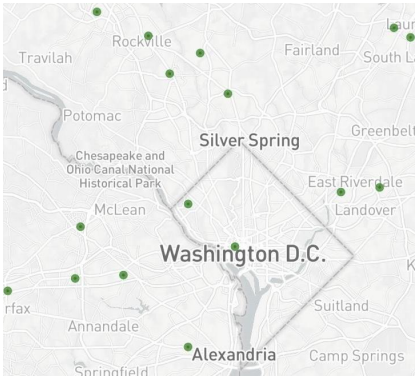

Figure 4.11: Jan 22, 2020, Entry 16 clusters

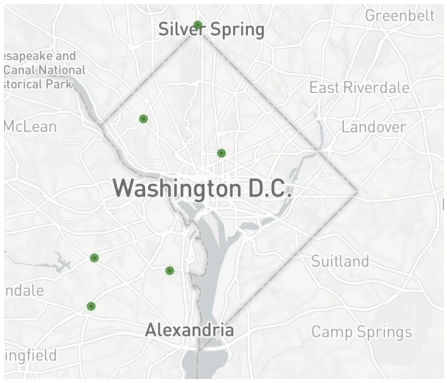

Figure 4.12: Jan 27, 2021, Entry 6 clusters

for different census tracks. In Fig. 4.11 we have 16 clusters formed that are spread throughout the DC region. Jan 22, 2020 was a Wednesday and large data points can be seen in the vicinity of Capitol region, Dulles International Airport (IAD) and Ronald Reagan airport. On Jan 27, 2021 a Wednesday one year after the previous date, we can observe in Fig. 4.10 that less number of people are traveling during the pandemic. Also the clusters are centered in and around the capitol which could mean that mostly people were visiting the DC region.

The clusters in Fig. 4.11 are evenly spread throughout the Northern Capital Region. Compared to the clusters seen in Fig. 4.12, it can be observed that that visitors and travelers are reduced because of the pandemic. Looking closely at the centers of clusters, Congressional cemetery, Jesuit community, Oak hill cemetery seem to be the hotspots.

In Fig. 4.13 we see that the plot is similar to the entry as shown in the Fig. 4.9 as both represent pre-covid data. We observe that the number of people traveling is reduced as can be seen in Fig. 4.14.

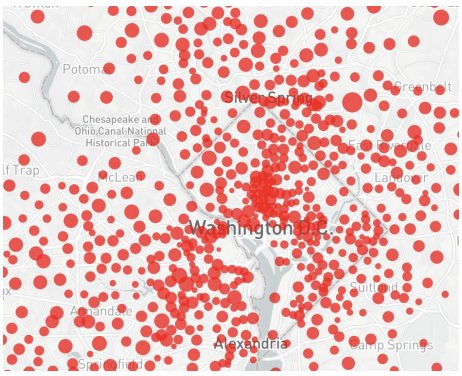

Figure 4.13: Jan 22, 2020 Exit, GeoDS Data

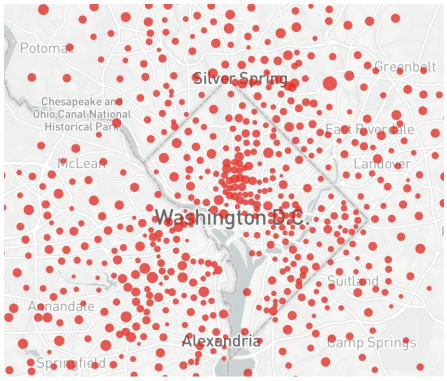

Figure 4.14: Jan 27, 2021 Exit, GeoDS Data

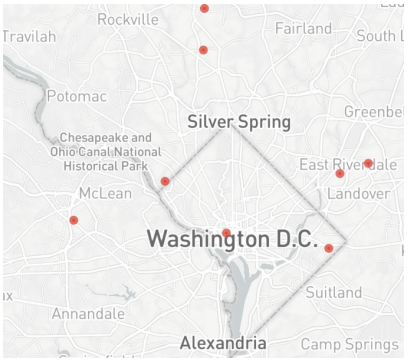

Figure 4.15: Jan 22, 2020 Exit, 10 clusters

Figure 4.16: Jan 27, 2021 Exit, 7 clusters

4.2 Clustering analysis

To determine the impact of COVID-19 across the socio-demographics, we divided all the cluster centers into three main categories as can be seen in Fig. 4.17

- The Blue cluster points refer to the places where traffic flow was reduced during COVID-19 as compared to pre-covid flow
- The Red cluster points refer to the places where no change in traffic flow was observed during COVID-19 as compared to pre-covid flow

- The Orange cluster points refer to the places where traffic flow was increased during COVID-19 as compared to pre-covid flow

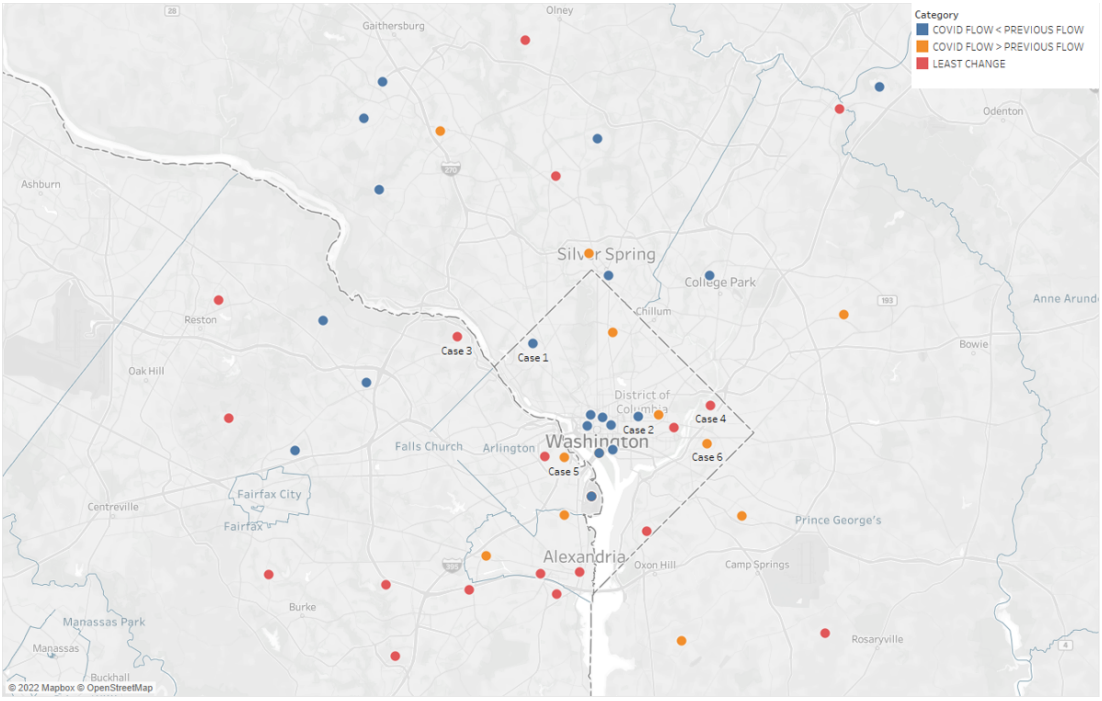

Figure 4.17: All the clustering points with their respective color coding

4.2.1 Places where Traffic Flow decreased during COVID-19

In Fig. 4.17, all the Blue cluster points refer to the places where traffic flow was reduced during COVID-19 as compared to pre-covid flow. One such blue point is in the area of American University, DC as seen in Fig. 4.18. The median household income in this area is around 192,763$. White people constitute the majority in this area. The unemployment rate is at around 2.7%, while the average age of people is around 40 years and percentage of people dropping out of high school is less than 3%. We observed an average reduction of around 13 thousand people traveling in this area.

Looking into Penn Quarter, DC which is another area where we got a blue cluster point, we can see in the Fig. 4.19 that it is predominantly filled with offices, museums and theatres. The median household income is around 161,667$, with well educated people living in the area with no person who has not finished high school. Asian and Black people constitute a minority while White people constitute the majority in this area. It has a low unemployment rate of 7.3% while the nation average being 10.4%.

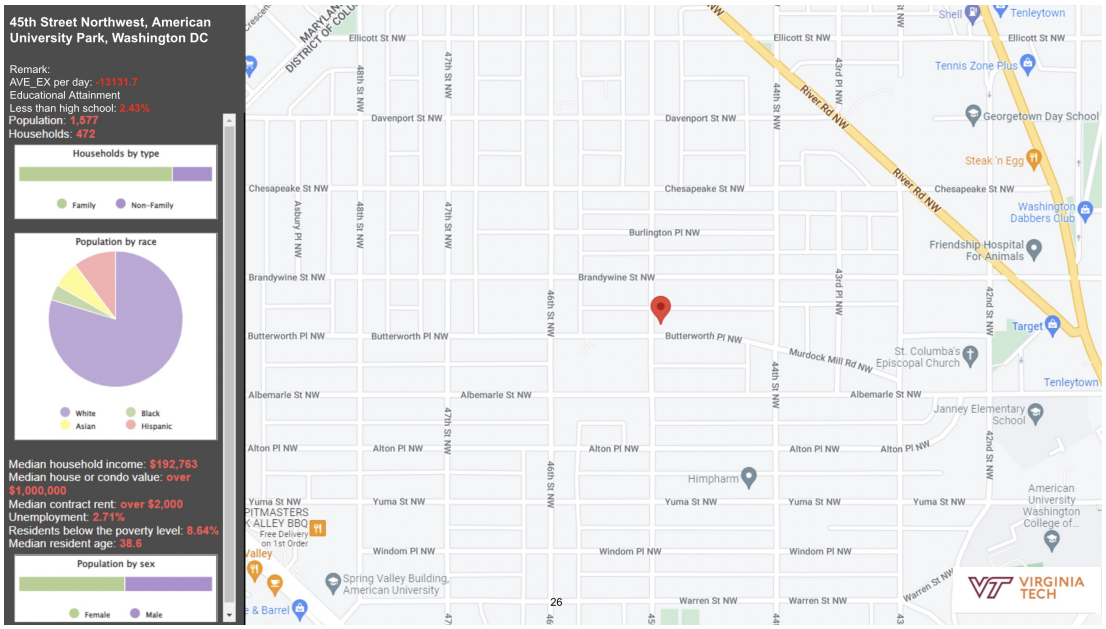

Figure 4.18: Infographic of American University Park, DC

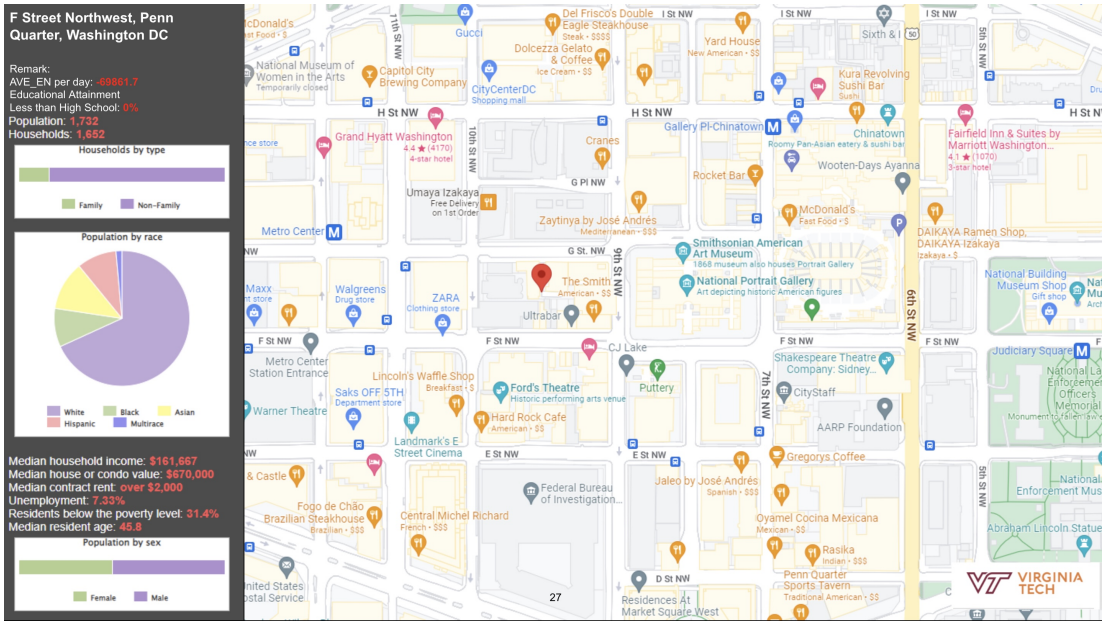

Figure 4.19: Infographic of Penn Quarter, DC

We observed an average reduction of around 69 thousand people traveling in this area, which is a drastic reduction.

Summarizing over all the blue points, we observed that the places where there was a massive reduction of people traveling were universities such as University of Maryland, George Washington University Alumni House, American university park, and The George Washington University, DC. John F. Kennedy Center for the Performing Arts, DC and Thomas Jefferson memorial saw a major decrease in visits which indicates that in 2021 there were a lot less tourists visiting DC. The wealthier parts of the city such as Wesley Heights saw the most decrease, which has a median household income of 250,000 and all the people living are highly educated. We also observe that Union station, DC got a massive decrease in foot fall indicating less use of public transport by the people.

Most of the neighborhoods we analyzed had around 64-100% White residents with an average household income of around 133K$, and an average unemployment rate of around 4.8% with a mean age of 37 years.

4.2.2 Places where Traffic Flow remained same during COVID-19

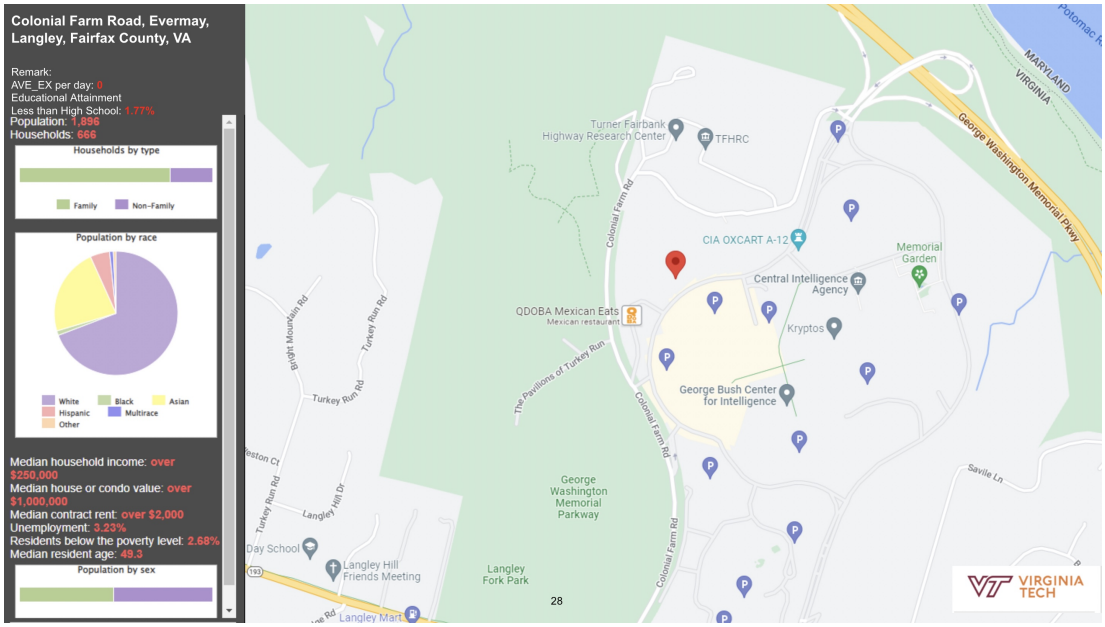

Figure 4.20: Infographic of CIA HQ Fairfax, VA

In Fig. 4.17, all the Red cluster points refer to the places where no change in traffic

flow was observed during COVID-19 as compared to pre-covid flow. Two such red points are in the area of CIA Headquarters Fairfax, VA (as seen in Fig. 4.20) and Pentagon City, VA indicating that intelligence and security agencies were working with the same capacity even during COVID. We can see a well educated population living in these areas, constituting mostly of Whites and Asians. The median income here is over 250,000$ and the unemployment rate is around 3%. Few other red cluster points were found in the areas of parks like National Capital Park, DC shown in Fig. 4.21 and grocery/shopping centres where no change in traffic flow was observed.

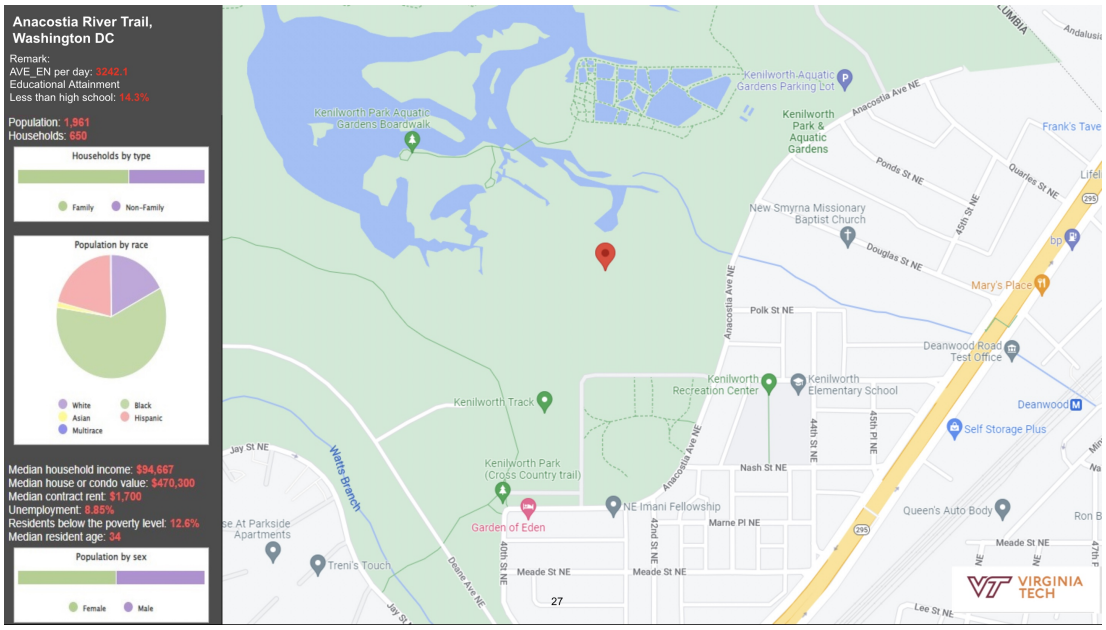

Figure 4.21: Infographic of National Capital Park, DC

4.2.3 Places where Traffic Flow increased during COVID-19

In Fig. 4.17, all the Orange cluster points refer to the places where traffic flow was increased during COVID-19 as compared to pre-covid flow. One such orange point is in the area of Arlington National Cemetery, DC as seen in Fig. 4.18 which saw a huge increase of around a thousand more people visiting this place on an average as compared to the same period prior to Covid-19. Here in Fig. 4.22 we can observe that the male population in this region outnumbers the female population. We see a mixture of White, Black, Hispanic, and Asian people in this region, with the average resident age being around 22 years which is quite young.

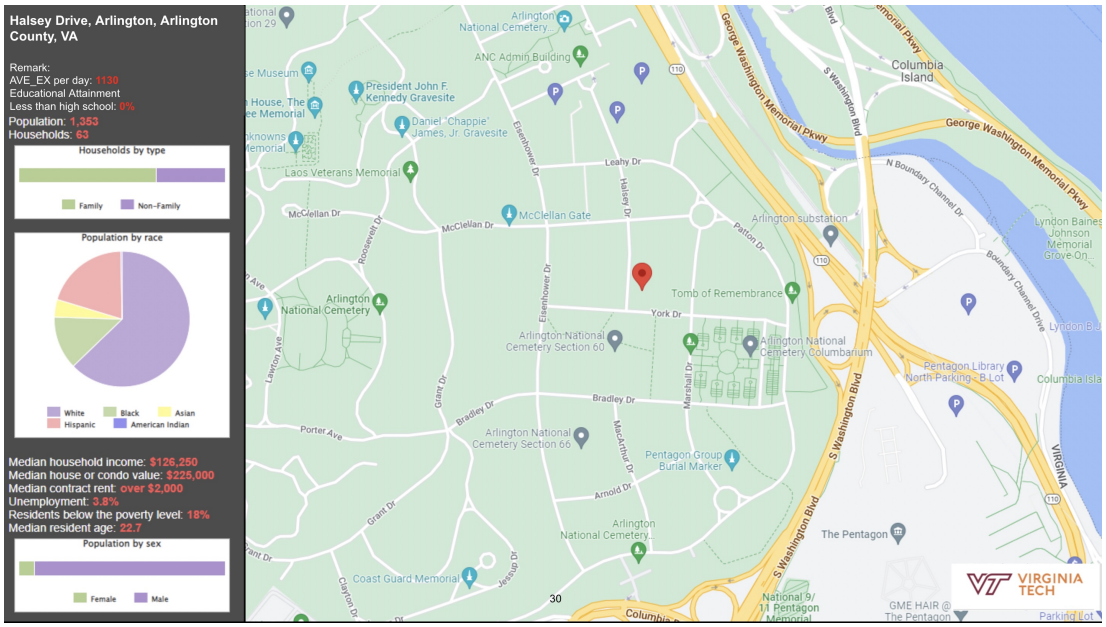

Figure 4.22: Infographic of Arlington national cemetery, VA

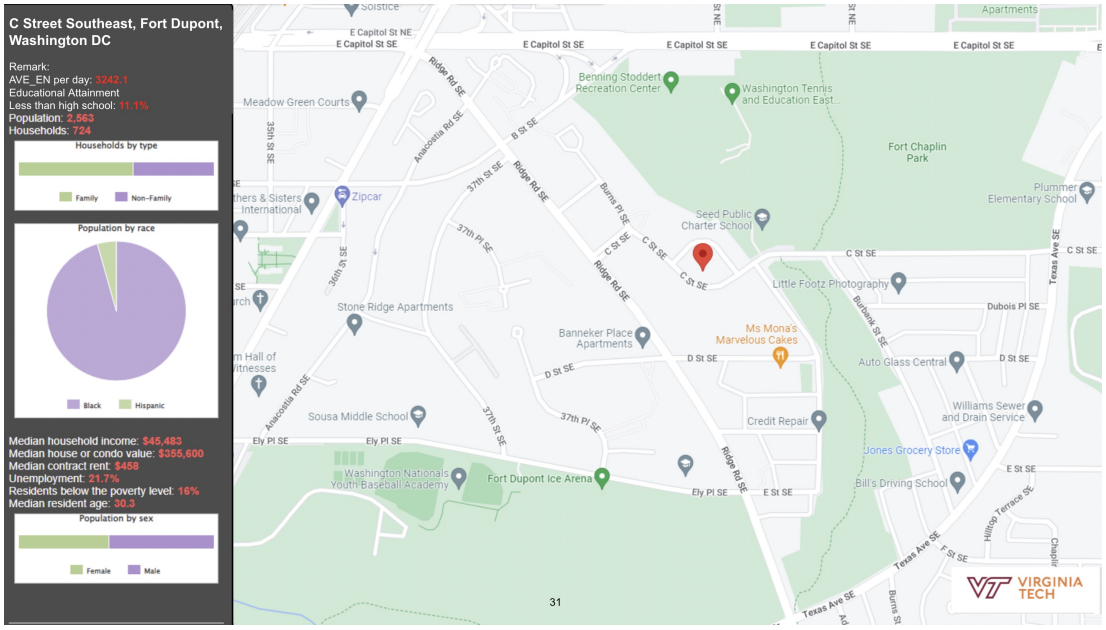

Figure 4.23: Infographic of Fort Dupont, DC

Looking into Fort Dupont, DC (Fig. 4.19) which is another area where we got a orange cluster point. The population is predominantly black around 90%. The number of people not completing high school is around 11%, with average household income of 45,000$ and an unemployment rate of staggering 21%. Summarizing over all the orange points, Trinidad saw the most increase 11,000 more compared to previous year. Brightwood Park, Anacostia DC, Bluemont VA where the percentage of black people was 60-98%, around 8-11% Hispanic people saw an average increase of 4,000 people traveling. 22% of the people living in this area have not completed high school, the average household income in this area is 50,000$ and the unemployment rate is 37% which is very high.

4.3 Trip length analysis

Legend	Low	Medium	High
Colour			

Place	2020					2021				
	20-Jan	21-Jan	22-Jan	23-Jan	24-Jan	25-Jan	26-Jan	27-Jan	28-Jan	29-Jan
UMD College Park, MD	9.3493	9.2074	9.0634	9.0114	9.0169	7.0423	7.9405	7.9744	7.6013	6.8117
American University Park, DC	4.5679	4.0743	3.8090	4.0245	4.1929	4.7817	5.1723	2.2787	4.0373	4.1421
George Washington University Alum, DC	5.2920	4.5318	4.7194	3.5712	4.4684	3.1473	3.7767	3.7533	4.6011	3.7340
John F. Kennedy Center for Arts, DC	4.7874	4.8871	4.7021	5.5309	4.5849	4.7558	3.7486	3.7035	4.0244	3.5959
Wesley Heights, DC	5.6838	5.6573	5.2158	5.0601	4.8919	4.2319	3.5710	2.5418	3.6386	4.0580
Vierra Falls Church, VA	7.6687	9.5461	7.6487	7.9717	7.4819	6.6896	6.5543	3.1125	6.5397	6.9046
Barnaby Woods, DC	4.5640	4.2244	4.7409	4.1276	4.5171	4.8407	4.6873	3.6152	4.3406	3.8209
Yorktown Blvd, Arlington,VA	7.1407	6.6492	6.8706	6.9793	7.8844	6.2051	5.5851	4.5954	5.0417	5.8714
Bluemont, Arlington, VA	5.3993	5.2876	5.1788	5.7531	5.0827	4.5697	4.5322	4.3005	4.6731	5.1839
Forest Glen, Arlington, VA	4.4046	5.7479	5.5590	5.7799	5.8847	4.8832	6.2781	2.5584	4.3146	5.6706
Penn Quarter, DC	4.0525	4.5277	3.9064	4.2716	4.0902	3.2108	2.9362	2.8646	3.6262	4.6585
Ronald Reagan Airport, VA	9.2728	8.3798	8.4396	9.2315	10.6835	7.6649	8.0857	2.8157	8.1697	7.9487
Golden Triangle, DC	6.3204	5.2655	5.8059	5.5256	5.4219	5.2952	6.4810	4.2797	5.5668	4.5376
The George Washington University, DC	5.6636	5.4719	5.2543	5.2504	5.3250	3.7453	4.3626	3.6166	4.5199	4.6380
Union station, DC	5.6813	5.7221	5.7204	4.8123	5.4258	4.1669	4.7408	2.9772	4.9717	4.6729
Southwest Federal Center, DC	4.7406	3.7708	4.2887	3.8428	3.6218	4.6185	6.0832	3.9285	5.2217	5.0161
Downtown, DC	5.2038	5.1004	5.1355	5.1452	5.1084	3.5907	3.8559	4.0065	4.3603	4.3007
Downtown Silver Spring, MD	4.8101	5.2475	4.6363	5.0311	5.0055	5.0393	5.1469	2.5754	5.4053	4.4698
Crystal City, VA	4.9165	5.3148	5.0926	4.7861	4.2143	4.5063	4.9586	2.7405	3.2404	3.5331

Figure 4.24: Heatmap of average trip length on weekdays (2020, 2021)

We divided the data into weekdays and weekends as in general the mobility patterns on weekdays and weekends are completely different as shown in Fig. 4.24 and in Fig. 4.24 respectively. We compared the average trip length of the one week of 2020 (pre-covid) with one week of 2021 (during Covid-19).

In University of Maryland, College Park, Maryland, we observe a massive decrease in the average trip length. On a Friday, the average trip length was around 9 kms and during Covid-19 it's around 6.8 kms which is a decrease of 2.2 kms. We observe a similar decrease in trip length on Monday at Ronald Reagan Airport and Union Station, DC from 9.27 to 7.66 kms and 5.6 to 4.1 kms respectively.

In Fig. 4.25 we have a heatmap of the average trip length on the weekend. In University of Maryland, College Park, Maryland, we observe a massive decrease in the average trip length. On a Saturday, the average trip length was around 10.7 kms and during COVID-19 it was around 6.9 kms which is a decrease of 3.6 kms. We observe a similar decrease in trip length on Sunday at Ronald Reagan Airport with average trip length of 11.9 kms pre-covid and 7.3 kms during COVID-19, which is a reduction of around 4.6 kms on an average. We also see a few places with slight increase like Union Station and Downtown.

Legend	Low	Medium	High
Colour			

Places	2020			2021	
	25-Jan	26-Jan		30-Jan	31-Jan
UMD College Park, MD	10.7014	8.7199		6.9106	6.3824
American University Park, MD	4.0215	4.1961		5.2350	3.1277
George Washington University Alumni House, DC	4.8159	4.8433		3.8591	2.9592
John F. Kennedy Center for the Performing Arts, DC	4.2703	5.3706		3.9543	3.0753
Wesley Heights, DC	4.7208	5.0774		3.8429	4.8782
Vierra Falls Church, VA	8.6443	6.7478		7.3753	3.5599
Barnaby Woods, DC	5.0931	4.5376		3.9258	4.2122
Yorktown Blvd, Arlington, VA	5.9929	6.6765		5.6036	3.8336
Bluemont, Arlington, VA	5.3315	5.3521		5.1122	4.1459
Forest Glen, Arlington, VA	5.3519	5.6531		3.4637	2.8245
Penn Quarter, DC	4.2059	4.5315		4.5121	3.1020
Ronald Reagan Washington National Airport, VA	10.1702	11.9601		7.0129	7.3078
Golden Triangle, DC	5.0114	5.1773		5.1445	4.7720
The George Washington University, DC	4.8345	5.0242		5.1110	3.1954
Union station, DC	5.5768	5.1884		5.3062	5.4127
Southwest Federal Center, DC	4.0574	3.4441		3.3424	4.2739
Downtown, DC	6.0364	5.7011		5.1388	5.3057
Downtown Silver Spring, MD	4.5381	4.9469		4.7790	3.0430
Crystal City, VA	4.8742	4.4092		4.5123	3.5657

Figure 4.25: Heatmap of average trip length on weekend (2020,2021)

In Fig. 4.26 the green bars represent the entries, the red bars represent exits for each of the stations and the brown bars represent the common region where there is both entry and exit. We observe the number of people taking the metro increases from 6 am and peaks at around 10am. We also see that at 12 noon there are very few people using the

metro. Similarly, we see an increase in people taking the metro back from downtown to suburbs increases from 3pm and peaks at around 7pm and reduces gradually after 7pm. In the Fig. 4.27, we compare total number of entries per day pre-covid to the total number of entries per day during COVID-19. We can observe a significant decrease in the total number of people traveling on weekdays and weekends during COVID-19 as compared to pre-covid.

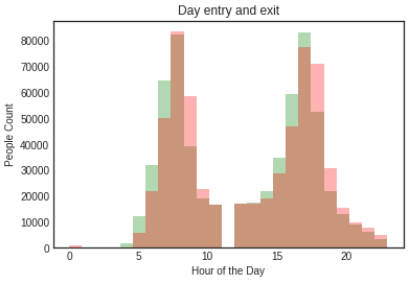

Figure 4.26: Hourly entry and exit count using bar graph

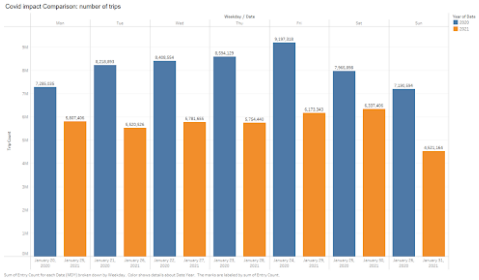

Figure 4.27: Total number of entries per day pre-Covid vs during COVID using bar graph

Conclusions

5.1 Challenges

- One of the major challenges we faced was with the data set size. The dataset we used was pretty huge, each file being around 22 GB in size. We had to collect and analyze many months of pre-Covid (2020) and during COVID-19 (2021) data.
- The initial data we acquired was not in the standard format and hence data cleaning and standardizing was required. The data standardizing took more time than needed as we had to combine different datasets into one for our requirement.
- The availability of the most recent data (2022) was a challenge, making it difficult to analyze the current behavior to which our society is now adapted.
- Selecting the best clustering algorithm was another challenge as we had to experiment with various hyper parameters and clustering methods to determine the best technique for our model.

5.2 Summary

We initiated the project by acquiring and preprocessing the data and then determining the most suitable clustering algorithm for the datasets. We implemented KMeans, DBSCAN, Agglomerative and Affinity propagation techniques. We calculated Silhouette scores for each clustering technique and got 0.53, 0.16, 0.56 and 0.67 respectively. The more the score is closer to zero, the better it fit for our model. Based on the Silhouette score we concluded that DBSCAN was the best clustering technique for our machine learning model.

By analysing the clusters, we observed the following findings. There was significant

reduction of traffic around universities, offices, downtown, tourist spots and wealthy neighbourhoods. An increase in traffic around cemeteries and hospitals was observed while the flow around supermarkets remained the same. The average trip length during-Covid had reduced drastically compared to pre-Covid. Some of the unique findings of this project were that demographic features like race, education level and the unemployment rate also had an affect on the general trend. Wealthy neighborhood with mostly white people saw drastic reduction in movement during COVID-19 while most colored neighborhoods had low income, high unemployment, and low education and saw an increase in mobility indicating daily wage workers who were working or providing services during the pandemic. CIA, Pentagon saw the least amount of change although all the other government offices went online. The heatmaps corresponding to the average trip lengths also provided us insights into the change in mobility patters observed during the pandemic. Overall, this was a learning method to understand how a disruptive event can cause changes in the normal behavior of public. Though few of the results were seemingly expected, the visual representation of the clusters and heat maps provided the suitable justification that supported the claim.

5.3 Future work

This project lays the foundation to the future mobility work to improve transportation in the National Capital Region. In future, we would like to collaborate with domain experts (e.g., urban planners, WMATA) and public stakeholders (e.g., Counties) to acquire more recent data and study the current travel behavior of the citizens. By analysing more recent data, we could predict future mobility trends for an extended timeline and provide recommendations on how to readjust public transportation intervals to match the new mobility patterns. Using the results of our project we can raise awareness about the socio-economic factors which affect modern urban mobility trends.

References

[1] Law Insider. *Urban Mobility definition*. 2020. URL: https://www.lawinsider.com/dictionary/urban-mobility.

[2] University of Michigan. *Center for sustainable systems*. 2021. URL: https://css.umich.edu/factsheets/us-cities-factsheet.

[3] Ke Fang. *World bank blog, "Smart Mobility": is it the time to re-think urban mobility?* 2015. URL: https://blogs.worldbank.org/transport/smart-mobility-it-time-re-think-urban-mobility.

[4] Sara Paiva et al. "Analysis of Mobility Changes Caused by COVID-19 in a Context of Moderate Restrictions Using Data Collected by Mobile Devices". In: *IEEE Access* 10 (2022), pp. 8906–8915. DOI: 10.1109/ACCESS.2022.3141083.

[5] Mukti Advani, Niraj Sharma, and Rajni Dhyani. "Mobility change in Delhi due to COVID and its' immediate and long term impact on demand with intervened non motorized transport friendly infrastructural policies". In: *Transport Policy* 111 (2021), pp. 28–37. ISSN: 0967-070X. DOI: https://doi.org/10.1016/j.tranpol.2021.07.008. URL: https://www.sciencedirect.com/science/article/pii/S0967070X21002092.

[6] Alfredo Aloi et al. "Effects of the COVID-19 Lockdown on Urban Mobility: Empirical Evidence from the City of Santander (Spain)". In: *Sustainability* 12.9 (2020). ISSN: 2071-1050. URL: https://www.mdpi.com/2071-1050/12/9/3870.

[7] Fadoua Khmaissia et al. *An Unsupervised Machine Learning Approach to Assess the ZIP Code Level Impact of COVID-19 in NYC*. 2020. DOI: 10.48550/ARXIV.2006.08361. URL: https://arxiv.org/abs/2006.08361.

[8] Cheng-Pin Kuo and Joshua S. Fu. "Evaluating the impact of mobility on COVID-19 pandemic with machine learning hybrid predictions". In: *Science of The Total Environment* 758 (2021), p. 144151. ISSN: 0048-9697. DOI: https://

doi.org/10.1016/j.scitotenv.2020.144151. URL: https://www.sciencedirect.com/science/article/pii/S0048969720376828.

[9] Ha Yoon Song and Dabin You. "Modeling urban mobility with machine learning analysis of public taxi transportation data". In: *International Journal of Pervasive Computing and Communications* 14.1 (Jan. 2018), pp. 73–87. ISSN: 1742-7371. DOI: 10.1108/IJPCC-D-18-00009. URL: https://doi.org/10.1108/IJPCC-D-18-00009.

[10] Geqi Qi et al. "Analysis and Prediction of Regional Mobility Patterns of Bus Travellers Using Smart Card Data and Points of Interest Data". In: *IEEE Transactions on Intelligent Transportation Systems* 20.4 (2019), pp. 1197–1214. DOI: 10.1109/TITS.2018.2840122.

[11] Krešimir Vidović, Sadko Mandžuka, and Davor Brčić. "Estimation of urban mobility using public mobile network". In: *2017 International Symposium ELMAR*. 2017, pp. 21–24. DOI: 10.23919/ELMAR.2017.8124426.

[12] Joao T. Aparicio, Elisabete Arsenio, and Rui Henriques. "Understanding the Impacts of the COVID-19 Pandemic on Public Transportation Travel Patterns in the City of Lisbon". In: *Sustainability* 13.15 (2021). ISSN: 2071-1050. DOI: 10.3390/su13158342. URL: https://www.mdpi.com/2071-1050/13/15/8342.

[13] Yuhao Kang et al. "Multiscale dynamic human mobility flow dataset in the U.S. during the COVID-19 epidemic". In: *Scientific Data* 7.1 (Nov. 2020), p. 390. ISSN: 2052-4463. DOI: 10.1038/s41597-020-00734-5. URL: https://doi.org/10.1038/s41597-020-00734-5.