Project Dissertation

INTERIM REPORT

Prepared by

Rishabh Malik

220512619

June 16, 2023

Annual Average Daily Traffic Estimation in Liverpool

1. Introduction

In recent years, traffic estimation has become a critical aspect of urban planning and transportation management. Accurate and reliable predictions of traffic flow are essential for optimizing road networks, improving transportation infrastructure, and reducing congestion. Researchers have developed various methodologies and models to estimate traffic parameters based on historical data and other relevant factors to address this challenge.

The estimation of Annual Average Daily Traffic (AADT) plays a significant role in understanding traffic patterns and designing effective transportation systems. AADT represents the average number of vehicles passing a specific point on a road segment on a typical day throughout the year. Estimating AADT requires analyzing vast amounts of data and considering multiple variables, such as road characteristics, land use, population density, and other socioeconomic factors.

One notable research paper in this field is "Annual Average Daily Traffic Estimation in England and Wales: An Application of Clustering and Regression Modelling" by Alexandros Sfyridis. This paper proposes a novel approach that combines clustering techniques with regression modeling to estimate AADT in England and Wales. The methodology outlined in this study has proven to be effective in accurately estimating AADT in a large-scale setting [3][4].

To build upon the work of Sfyridis and address the specific traffic estimation needs of Liverpool City, this interim report aims to adapt and apply the concepts and methodologies presented in the paper. By customizing the clustering and regression models proposed by Sfyridis, we intend to provide valuable insights into the traffic patterns specific to Liverpool, ultimately contributing to urban transportation planning and management decisions in the city [5].

The utilization of Geographic Information Systems (GIS) and related tools play a crucial role in understanding and analyzing traffic patterns. GIS technology enables the integration of various spatial datasets, allowing researchers to visualize and analyze traffic-related information in a geospatial context. For this study, the QGIS software will be employed as a powerful open-source GIS tool for data analysis, visualization, and geospatial modeling.

Furthermore, the Python library Geopandas will be utilized as it extends the capabilities of Pandas, a popular data manipulation library, to handle geospatial data. Geopandas provides functionality for reading, writing, and manipulating geospatial data formats and

enables the integration of geospatial data with other data sources for analysis and modeling purposes.

To provide a comprehensive understanding of traffic estimation methodologies and related research, this report will draw upon a range of relevant studies. Some key papers include "Traffic Flow Prediction with Big Data: A Deep Learning Approach" by Li et al., which explores deep learning techniques for traffic flow prediction, and "Spatial-Temporal Traffic Flow Prediction: A Deep Learning Approach" by Ma et al., which focuses on spatial-temporal traffic flow prediction using recurrent neural networks.

Moreover, the study "Traffic Congestion Prediction Model Based on Big Data and Deep Learning" by Zheng et al. presents a model that integrates big data and deep learning for traffic congestion prediction. These papers highlight the advancements and potential of machine learning and deep learning techniques in traffic estimation.

Additionally, "Spatial Clustering of Traffic Flow Patterns Using Kernel Density Estimation" by Wang et al. discusses the use of kernel density estimation for spatial clustering of traffic flow patterns, providing insights into the clustering techniques relevant to our study. The study "A Comparative Study of Traffic Flow Prediction Methods" by Zhang et al. compares various traffic flow prediction methods, including clustering and regression modeling, offering valuable insights into their effectiveness and applicability.

In a short summary from Fig 1.1, we can see each vehicle type and the miles they traveled in 2011-2021.

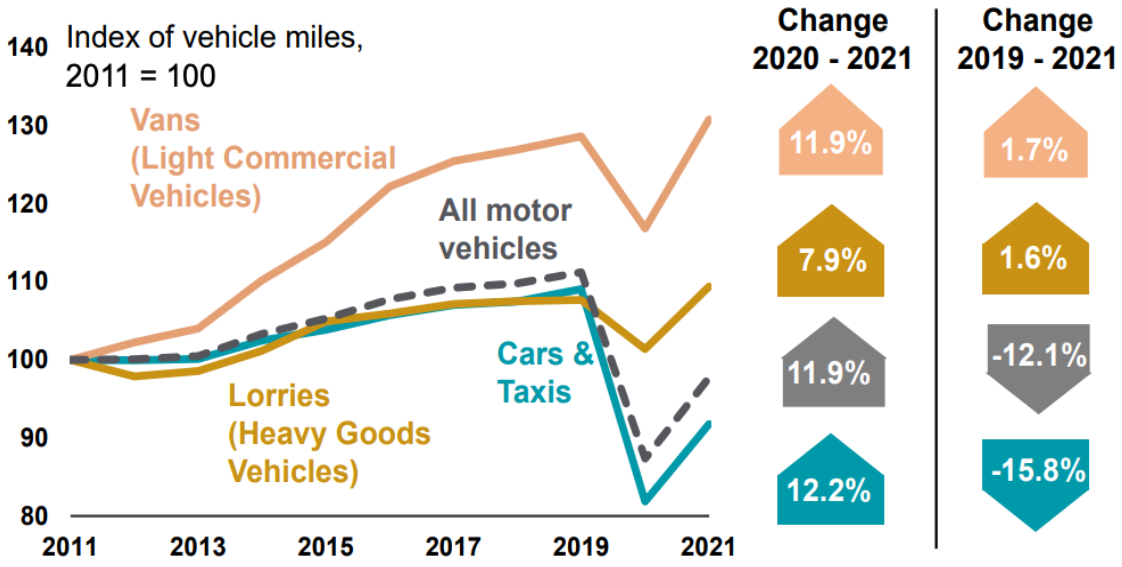

Fig 1.1: Vehicles miles traveled by selected vehicle types in GB 2011-2021

2. Aim and Objectives

2.1 Aim

This project aims to apply GIS technology, and geopandas in Python, and adapt the clustering and regression models presented by Sfyridis to develop a comprehensive and accurate traffic estimation model for the city of Liverpool. This study aims to provide valuable insights into traffic patterns specific to Liverpool by considering methods to classify data with missing variables [1] and exploring models for individual vehicle types, such as LGVs and HGVs, to reveal patterns peculiar to specific traffic flows.

2.2 Objectives

This work will focus primarily on:

- Utilize GIS technology and geopandas in Python to collect, manage, and analyze relevant spatial data related to the road network, traffic counts, and other influencing factors in Liverpool City.
- Adapt and customize the clustering methodology proposed by Sfyridis to classify traffic data into meaningful clusters, considering missing variables and finding the shortest distance from new points to the center of cluster centroids.
- Incorporate regression models to estimate the Average Annual Daily Traffic (AADT) for the road segments in Liverpool City, using the identified clusters and the variables with the smallest degree of overlap across clusters.
- Develop models specifically tailored to LGVs and HGVs, allowing for the identification of traffic patterns unique to these vehicle types and their respective flows within Liverpool City.
- Analyze and interpret the results obtained from the clustering and regression models, identifying the factors influencing cluster formation and traffic flow variations in Liverpool City.
- Evaluate the performance of the developed model through comparison with existing traffic estimation methods, considering accuracy, efficiency, and applicability to transportation planning and management decisions.
- Provide recommendations and insights based on the findings, highlighting the potential implications for transportation planning, infrastructure development, and environmental studies in Liverpool City.

Optional objectives in case time permits:

- Explore the usage of additional explanatory variables, such as mileage estimation for street segments, to improve the accuracy of the traffic estimation model and enable precise calculations of air pollutant emissions [2].

3. Overview of Progress

The progress of this project involved several key steps and activities. Initially, an extensive literature review was conducted to explore the existing research on traffic data analysis and the application of GIS technology. This provided a solid foundation for understanding the relevant concepts and methodologies.

Data collection played a crucial role in this project. Road traffic statistics were gathered to acquire the necessary data for analysis and modeling. The collected data was then subjected to profiling and exploratory data analysis (EDA) techniques to gain insights into its characteristics and identify patterns or trends.

To enhance the understanding and implementation of GIS technology, tutorials were undertaken to acquire the necessary skills in working with spatial data and geopandas in Python. These tutorials served as a valuable resource for leveraging GIS tools and techniques in the subsequent stages of the project.

Data cleaning was a significant focus to ensure the quality and reliability of the collected data. Relevant variables and features required for model training were identified and retained. Efforts were dedicated to handling missing data, removing outliers, and normalizing the data to ensure its suitability for analysis and modeling.

Resolving the resolution of GIS images also required dedicated research. Determining the appropriate resolution for the GIS images used in the study was essential for accurate spatial analysis and modeling.

The impact of the COVID-19 pandemic on traffic patterns will be given special consideration in this study. Separate analysis and discussion will be conducted to understand how the pandemic has influenced traffic flow and behavior in Liverpool City.

As the project progresses, additional data analysis and modeling will be carried out. This includes further exploring different modeling techniques, evaluating model performance, and comparing the results with existing traffic estimation methods.

Throughout the study, emphasis will be placed on documenting the progress, findings, and challenges encountered along the way. The results obtained will be critically analyzed, and their implications for transportation planning, infrastructure development, and environmental studies will be discussed.

In conclusion, the progress of this study involved an extensive literature review, data collection, profiling and EDA, tutorials on GIS and geopandas, data cleaning, resolution determination for GIS images, and a separate analysis of the impact of COVID-19. The project will advance by conducting additional data analysis, exploring various modeling approaches, and comparing the results.

4. Project Plan

The project timeline spans from 24th April to 15th August 2023, as illustrated in Figure 6.1. The project initiation phase was conducted and during this phase, the project title and objectives were defined, and an extensive literature review on traffic data and GIS was conducted. This involved more intensive research to gain a comprehensive understanding of relevant methodologies and techniques in the field.

The data understanding and cleaning phase is scheduled from 16th May to 9th June. Road traffic statistics and other relevant data sources were acquired during this phase. The collected data then underwent profiling and exploratory data analysis (EDA) to gain insights into its characteristics and structure. The data cleaning process involved handling missing values, removing outliers, and normalizing the data. Additional research was conducted to determine the appropriate resolution for GIS images used in the analysis.

The project execution phase is scheduled from 10th June to 10th August. During this phase, model training will be performed on selected features using appropriate machine-learning techniques. The implementation of GIS and geopandas will enable spatial analysis and visualization of the data. GIS data will be plotted on maps to visualize traffic patterns and cluster formations. The results obtained from different models and techniques will be compared and evaluated. The impact of COVID-19 on traffic patterns will be analyzed separately. The report writing will also take place simultaneously with the execution phase. This will involve summarizing the overall project findings and results. The report will be structured and organized, including sections such as introduction, methodology, results, and discussion. The analysis and modeling outcomes will be compiled and interpreted. Conclusions will be drawn based on the research objectives and findings. The report will also include recommendations for transportation planning and management in Liverpool City.

It is important to note that the project plan may be subject to adjustments and modifications as the project progresses. Regular monitoring and updates will be carried out to ensure that the project stays on track and meets the desired objectives and deadlines.

5. References

[1] Domingos, P. (2012). A few useful things to know about machine learning. Communications of the ACM, 55(10), 78-87.

[2] Leduc, G. (2008). Travel distance and pollutant emissions: An empirical analysis. Transportation Research Part D: Transport and Environment, 13(4), 249-259.

[3] Junqué de Fortuny, E., Martens, D., & Suykens, J. A. (2013). Interpretability issues in regression modeling by statistical learning. European Journal of Operational Research, 228(1), 178-187.

[4] Labib, S. M., Eluru, N., & Abdel-Aty, M. (2018). Impacts of roadway and traffic characteristics on vehicle emissions: A review. Transportation Research Part D: Transport and Environment, 63, 1-15.

[5] Chen, K., & Wang, W. (2013). An automated weighting clustering algorithm based on genetic algorithm for road traffic flow. Expert Systems with Applications, 40(6), 2300-2308.

6. Gantt Chart

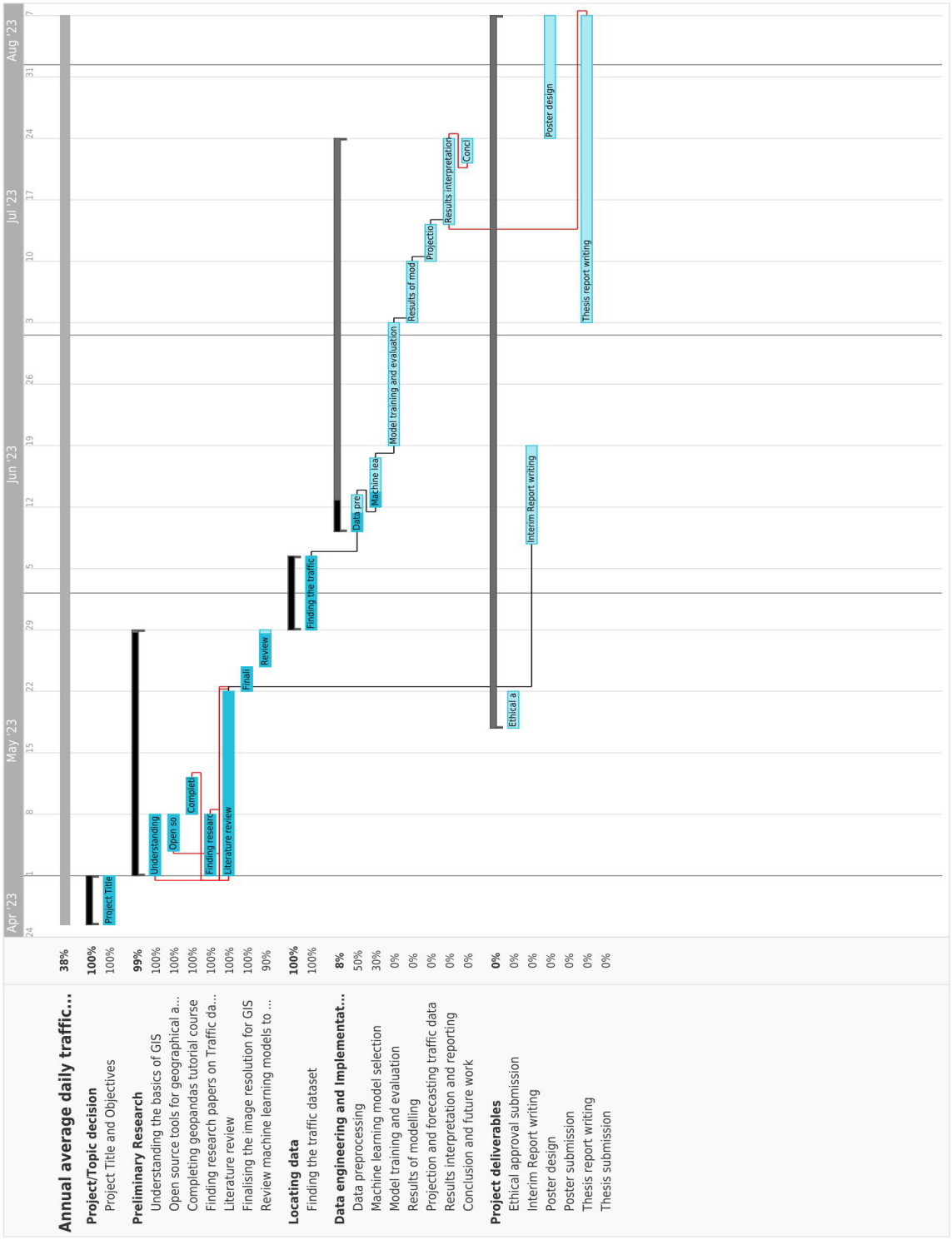

Fig 6.1: Gantt chart showing the timeline of the project plan

7. Data Management Plan

0. Proposal name

Project:

Author: Rishabh Malik	**Version: 1**	**Date:** 13th June 2023

1. Description of the data

1.1 Type of study

Traffic estimation is crucial for urban planning and transportation management. This report aims to adapt the clustering and regression models proposed by Sfyridis to estimate Annual Average Daily Traffic (AADT) in Liverpool. The utilization of GIS and QGIS, along with the geopandas library, will facilitate data analysis and geospatial modeling.

1.2 Types of data

The dataset used in this study consists of road traffic statistics collected by the Department for Transport (DfT) in Great Britain. The data is obtained from sensors installed on highways and link roads, capturing information about different types of vehicles that travel through these roadways. The DfT employs various methods to estimate road link-level traffic, considering factors that contribute to more accurate estimates.

1.3 Format and scale of the data

The traffic data consists of 32 variables, captured in a CSV (Comma-Separated Values) file. The dataset is updated annually and includes the following variables:

count_point_id, direction_of_travel, year, count_date, hour, region_id, region_name, local_authority_id, local_authority_name, road_name, road_type, start_junction_road_name, end_junction_road_name, easting, northing, latitude, longitude, link_length_km, link_length_miles, pedal_cycles, two_wheeled_motor_vehicles, cars_and_taxis, buses_and_coaches, lgvs, hgvs_2_rigid_axle, hgvs_3_rigid_axle, hgvs_4_or_more_rigid_axle, hgvs_3_or_4_articulated_axle, hgvs_5_articulated_axle, hgvs_6_articulated_axle, all_hgvs, all_motor_vehicles.

These variables provide information on various aspects of road traffic, including the unique ID of the count point, the direction of travel, the year of the data, the date and hour of the count, region details, local authority information, road names, road types, geographic coordinates, link lengths, and vehicle counts categorized by vehicle type.

The vehicle types covered in the dataset include pedal cycles, two-wheeled motor vehicles, cars and taxis, buses and coaches, large goods vehicles (LGVs), heavy goods vehicles (HGVs) with 2 rigid axles, HGVs with 3 rigid axles, HGVs with 4 or more rigid axles, HGVs with 3 or 4 articulated axles, HGVs with 5 articulated axles, HGVs with 6 articulated axles, and the total count of all HGVs and all motor vehicles.

Each row in the CSV file represents a specific observation of traffic data, providing valuable insights into the patterns and characteristics of road traffic at different locations and time points.

2. Data collection / generation

2.1 Methodologies for data collection / generation

The dataset is collected and maintained by the Department for Transport (DfT), covering various locations throughout Great Britain. The DfT oversees the deployment of sensors that capture road traffic statistics. The dataset contains a comprehensive range of variables related to road traffic and vehicles. The availability of this dataset is highly valuable for research and analysis in the field of transportation, and we acknowledge the Department for Transport's efforts in providing this important resource.

2.2 Data quality and standards

A robust quality assurance strategy is used to guarantee the dataset's integrity. This strategy comprises thorough verifications and tests to validate the dataset. To find and correct any potential discrepancies or abnormalities in the data, several strategies are used. Data cleansing, outlier detection, consistency confirmation, and cross-validation against other trustworthy sources are a few examples of the checks that may be made.
The reliability and correctness of the dataset are upheld by carrying out a thorough quality assurance process, ensuring that the data may be utilised with confidence for analysis and study.

3. Data management, documentation and curation

3.1 Managing, storing and curating data.

To make data maintenance easier, a backup copy of the downloaded dataset is kept both locally on a different project disc and remotely on Google Drive. The risk of data loss is reduced thanks to this strategy, which ensures redundancy. A further layer of security is added by storing the backup on Google Drive, which gives cloud storage and the flexibility to retrieve the data from any location with an internet connection. By maintaining a local copy on a distinct project disc, you may retrieve the dataset quickly and easily without relying entirely on the cloud. Data maintenance is made easier by having numerous copies in several places, maintaining data availability and reducing the effects of any potential hardware problems or data corruption problems.

4. Data security and confidentiality of potentially disclosive information

4.1 Main risks to data security

Since training data is entirely public data, security issues with third party storage solutions (raised by ncl.ac.uk https://www.ncl.ac.uk/library/academics-and-researchers/research/rdm/working/) are not a concern. Indeed, even any trained networks we create push towards a philanthropic cause meaning public distribution is encouraged.

5. Data sharing and access

5.1 Suitability for sharing

The dataset that will be utilized in this project is freely accessible, has passed thorough peer review, and has been cited in several academic publications. The goal of this research is to remove bias from datasets seen in the real world. Sharing experiments and code will therefore be highly encouraged in order to promote transparency and reproducibility.

This project intends to improve collaboration and enable the replication of results by encouraging the exchange of code and experiments, resulting in a more thorough and trustworthy examination of the dataset.

5.2 Discovery by potential users of the research data

All code libraries created throughout the project will be posted to a GitHub repository once all project goals have been met. Additionally, my LinkedIn profile will include a link to this repository.

Others will have access to the developed code for reference, collaboration, and future research by publishing the code libraries on GitHub and posting the link on LinkedIn. This encourages openness, encourages knowledge exchange, and permits the results of the project to be replicated.

5.3 Data preservation strategy and standards

The research data will be kept and archived indefinitely in accordance with established preservation guidelines. The precise data retention duration will be chosen and followed, guaranteeing the data's ongoing accessibility and availability. The research data will be securely archived and maintained in accordance with recognized preservation standards in order to be used for future reference, analysis, and possibly future research.

5.4 Restrictions or delays to sharing, with planned actions to limit such restrictions

There are no significant limitations or delays in the dissemination of the data, as it is accessible to the public. Researchers have unrestricted access to these datasets and can utilize them for their studies without the requirement of additional licenses or permissions. The data collected from the sensors maintained by the Department of Transport in the UK is readily available to anyone for research purposes.

6. Responsibilities

- Are there any resources (e.g. storage/ training) that you will require to fulfil the plan?

  Google Colab Pro+ ≈ £45 / month

  Google Drive Storage (100GB) @ ≈ £1.59 / month

7. Relevant institutional, departmental or study policies on data sharing and data security

Policy	URL or Reference

Data Management Policy & Procedures	https://www.ncl.ac.uk/media/wwwnclacuk/research/files/ResearchDataManagementPolicy.pdf
Data Sharing Policy	https://www.ncl.ac.uk/library/academics-and-researchers/research/rdm/working/