Conceptualising the Multifaceted Nature of Urban Road Congestion

L F J Abberley

PhD 2022

Conceptualising the Multifaceted Nature of Urban Road Congestion

Luke Francis James Abberley

A thesis submitted in partial fulfilment of the requirements of

Manchester Metropolitan University

for the degree of

Doctor of Philosophy

Department of Computing and Mathematics

Manchester Metropolitan University

in collaboration with Transport for Greater Manchester

2022

“The passage of goods carts on narrow city streets so congested them that they became impassable and unsafe for pedestrians” – Caesar.

Declaration

The candidate confirms that the work submitted is his own, except where work which has formed part of jointly authored publications has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.

Jointly authored publications and the contributions of the candidate and the other authors are as follows:

- Gould, N. and Abberley, L. (2017) 'The semantics of road congestion.' *In UTSG*. Dublin.
- Abberley, L., Gould, N., Crockett, K. and Cheng, J. (2017) 'Modelling Road Congestion using Ontologies for Big Data Analytics in Smart Cities.' *In Proceedings of the Third IEEE International Smart Cities Conference*. Wuxi, China: IEEE.
- Abberley, L., Crockett, K. and Cheng, J. (2019) 'Modelling Road Congestion Using a Fuzzy System and Real-World Data for Connected and Autonomous Vehicles' *Proceedings of the 11th IEEE Wireless Days Conference*. Manchester, United Kingdom: IEEE.

The above conference and journal papers are part of the candidate's thesis. Where Luke Abberley is the first author, he wrote the entire article and is the corresponding author. The co-authors, Dr Nicholas Gould, Prof. Keeley Crockett, and Dr Jianquan Cheng contributed by providing their valuable knowledge within their respective disciplines and proofreading the articles.

Abstract

Urban road congestion is not a new phenomenon and remains an outstanding problem that continues to impact people around the world. Road congestion costs the European Union an estimated 1-2% of GDP each year and is responsible for 27% of deadly C02 emissions. In addition, it can cause life-threatening delays in the emergency services response time.

Road congestion has a multifaceted nature and lacks a clear and explicit definition. This makes the problem of tackling it very subjective, time and context dependent. There have been several approaches to both modelling and predicting road congestion. From a physical perspective, road congestion has been modelled using speed, capacity, velocity, and journey time; relatively road congestion has been classified using terms such as non-recurrent and recurrent congestion which tend to be relative to each stakeholder; conceptual models such as the bathtub, traffic flow, and origin to the destination have been used to ascertain the impact of road congestion on a city scale.

This research presented tackles the problem of defining what is meant by congestion within an urban road network through defining a conceptual model that captures the semantics of road traffic congestion and its causes. The model is validated through the construction of a real-world dataset and the development of a visual tool which can be used to identify and alleviate congestion. The final stage of the project uses both the model and the dataset to investigate and implement a series of fuzzy systems to classify three types of congestion (non-recurrent, recurrent, and semi-recurrent). The fuzzy system results are then validated against human methods of classifying congestion.

The main contributions of this thesis to world knowledge can be summarised as follows: The design and development of a novel universal Urban Road Congestion Conceptual (URCC) model. The URCC model is broken down into two main components: Analogical conceptualisation which builds upon the famous ‘bathtub’ model and will integrate with other analogies to create ‘a raindrop hitting a leaf inside the bathtub with ever changing water temperatures’. The second component is an ontological approach to modelling congestion thus providing a better understanding for decision-makers through providing a formal and explicit explanation for concepts within the domain of urban road congestion. Another contribution is the development of a real-world spatiotemporal quasi-real-time big data dataset known as the Manchester Urban Congestion Data (MUCD) dataset which was used to validate the URCC. A visualisation graphical user interface called TIM (Transport Incident Manager) was developed with stakeholders TfGM (Transport for Greater Manchester). TIM has the ability to fill the void left by the clear lack of visualisation tools that are capable of visualising real-world big data datasets, such as the MUCD and models of urban road congestion. The final contribution to knowledge is the design and development of two fuzzy decision-making systems which are not only capable of predicting urban road congestion on a link but the type of congestion occurring on a network of links. Using a fuzzy decision-making system allows for explainable and interpretable decisions, and also provided useful and meaningful qualitative context back to the relevant TfGM stakeholders. The non-optimised multi-classification fuzzy system had slightly worst accuracy than the J48 decision tree algorithm, however, the fuzzy system is easier to interpret and provides meaningful context compared to the J48 algorithm due to only requiring 12 rules compared to the 1184 learned rules in the J48 decision tree. Furthermore, once the fuzzy system has been optimised (future work) it is likely to have similar if not better performance than the J48 decision tree.

Acknowledgements

I would like, firstly, to thank my original Director of Studies, Nicholas Gould for providing the initial inspiration for this project and for giving me the opportunity to embark on a PhD. In addition, I am deeply grateful for his help and encouragement throughout my PhD. His Successor, Keeley Crockett, had the task of taking over the project half-way through and her invaluable knowledge, commitment, attention to detail, and support is one of the few reasons I found the strength to be able to complete this project. Keeley Crockett you are more of a friend than you are a Director of Studies and I hope we work together on many more projects in the future. In addition to thanking both my previous and current Director of Studies, I would like to thank Jianquan Cheng, who spent many hours providing me with insight into modelling spatio-temporal data using spatial and temporal statistics.

Thanks are due to Transport for Greater Manchester, not only for their financial contribution towards this project, but also providing me access to their data and spend time in providing feedback during my project.

In addition to thanking people directly involved in my project, I would also like to thank Soufiene Djahel, who took time to critically review parts of my work and provide invaluable feedback.

I would also like to thank My fiancée Naomi Parry (Soon to be Abberley) and my daughters Lilly Abberley and Thea Abberley for their forbearance and support throughout the years. Furthermore, I would like to Thank Synectics Solutions and especially, Robert Bevington for giving me the opportunity to put my skills that I have learnt during my PhD into practice within the fraud prevention domain as a Senior Data Scientist. Finally, I would like to dedicate this thesis to my grandfather who sadly, passed away a few years ago, but not before instilling within me the drive and confidence to aim big.

Content

List of Tables

List of Figures

List of Equations

Abbreviations

ANPR	Automatic Number Plate Recognition
ATC	Automatic Traffic Counters
DfT	Department for Transport
FPR	False Positive Rate
GIS	Geographic Information System
GPS	Global Positioning Systems
GSM	Global System for Mobile communication
ITS	Intelligent Transport System
MAC	Media Access Control
MUCD	Manchester Urban Congestion Data
NRC	Non-recurrent Congestion
RC	Recurrent Congestion
RFID	Radio-frequency Identification
RSU	Roadside Unit
SRC	Semi-recurrent Congestion
TfGM	Transport for Greater Manchester
TIM	Transport Incident Manager
TPR	True Positive Rate
URCC	Urban Road Congestion Conceptual
V2H	Vehicle-to-Human
V2I	Vehicle-to-Infrastructure
V2R	Vehicle-to-Roadside Unit
V2V	Vehicle-to-Vehicle
VANET	Vehicular Ad-hoc Network
VMS	Variable Messages Signs

Chapter One: Introduction to research

1.1 The importance of a resilient road network

Resilient transport networks are vital for sustainable development and therefore the focus of this research is on road networks. The major threat to resilient road networks is congestion, which has an estimated cost of 1-2% of GDP across the European Union and 3-4% within the United Kingdom (Department for Transport, 2020), including reduced productivity and increased transport costs (Somuyiwa et al., 2015). Congestion also has a major impact on air quality and the quality of life in general. Congestion has two known forms, non-recurrent congestion, which can be the result of a road traffic incident such as traffic accidents and roadworks, and recurrent congestion, which can occur at well-known bottlenecks where traffic demand exceeds capacity (Van Schijndel and Dinwoodie, 2000).

This research will focus on conceptualising and validating the differences between the traditional types of congestion; non-recurrent congestion, recurrent congestion, and semi-recurrent congestion which is a third type of congestion that will be coined within this thesis. Semi-recurrent congestion is the consequence of scheduled events, such as a 'football match', 'music concert', and 'planned roadworks'. These types of events are not cyclical because they do not happen at the same time or on the same day. However, they do tend to be predictable due to schedules, which are created in advance. The coining of semi-recurrent congestion is one of the contributions of this thesis which will help stakeholders to be able to distinguish the difference between a road accident, a 'cup' football match, and unplanned roadworks compared to a concert, 'league' football match, and planned road works. Allowing stakeholders, to respond more adequately depending on the type of congestion. Which traditionally, all the above have been treated as non-recurrent congestion.

There are multiple models of congestion that currently exist, including the 'Bathtub' model (Arnott, 2013), a data-driven agent-based model (Othman et al., 2015) and a dynamic 'bottleneck' model (Silva et al., 2014). However, to help reduce congestion there needs to be more work done to improve the level of resilience, which requires being prepared for a road traffic incident and/or action a recovery plan within an acceptable period, restoring the network to the same level or better. The development of a conceptual framework that can be used for developing an Intelligent Transport System (ITS) which will increase the quality of information being provided to stakeholders, such as transport managers allowing for a faster response to congestion.

1.2 Measurements of congestion

To model urban road congestion, relevant dimensions, such as journey time, density, (vehicle and traffic) speed, and travel time are required. The appropriate dimension/s are determined by academics, researchers, and traffic managers and will depend on what the problem is being solved and what the technical limitations are.

In recent studies various dimensions, such as journey time (Anbaroglu et al., 2014; Anbaroğlu et al., 2015), density (Bauza et al., 2010), (vehicle and traffic) speed (Bauza et al., 2010) and travel time (Bar-Gera, 2007; Li and Chen, 2014) have been used to measure and define urban road congestion on different scales. In this thesis, the adopted dimension for classifying urban road congestion was journey time. This was due to the requirements set out by Transport for Greater Manchester (TfGM). A more in-depth discussion can be found in Chapter Two.

1.3 Scope

This project was 50% funded by TfGM, who provided two data sources that were used in this research. These are the Bluetooth passive sensors and Automatic Traffic Counters (ATC). Due to the location of the data sources and the project sponsor, the domain for this research will be Greater Manchester, UK.

1.4 Problem statement

Transportation systems are a fundamental part of society, providing people with ways to explore the world, commute to work, and visit shops for everyday essentials. There are several types of transport systems, such as land, rail, water, air, space, and intermodal. Intermodal transportation is when one or more mode of transportation is used within the same system, for instance, goods being transported from America to the United Kingdom may travel on a ship and then be loaded onto a lorry to be delivered to the destination.

Land transportation is the linchpin that holds the other modes together (Somuyiwa et al., 2015), however, it generates several challenges, such as bad air quality due to vehicles being stuck in congestion (Transport 2020, 2016), a financial burden that costs the European Union 1-2% GDP (Djahel et al., 2015), and wasting limited fuel resources (Djahel et al., 2015). The crucial issue that these challenges all revolve around is road congestion that many approaches are being taken to address road congestion; however, road congestion remains a multifaceted issue due to existing road networks becoming increasingly more congested because of the growth in the number of people that are using vehicles; and the inability to develop a more sustainable and resilient network (Hartgen and Fields, 2009).

The Highways Agency estimates that 65% of congestion is caused by traffic volumes at or above capacity, 25% as a result of incidents, and 10% by roadworks (Department for Transport, 2014). In 2020, the Department for Transport (DfT) released a further report based on 2019 figures that shows the overall volumes on the road network have increased by an average of 2% compared to the previous year (2018) and in some cases, such as the highway, where the traffic volume has increased by 14.1% compared to 10 years ago (Department for Transport, 2020). Furthermore, transportation is responsible for 28% of all greenhouse gas emissions within the UK, which has only reduced by 3% in 30 years (Waite, 2020).

Road congestion has been a problem for many years with literature going back as far as 1920 by Pigou (Verhoef, 1999) and although there is a vast amount of literature addressing road congestion there still remains a clear absence of a formal and explicit understanding of road congestion. Consequently, this research will attempt to develop a formal and explicit conceptualisation of urban road congestion. To achieve this, the research presented in this thesis will explore the use of analogical and ontological methods to conceptualise urban road congestion and the conceptual model will be validated using a real-world big data dataset and a custom-built fuzzy decision-making system.

1.5 Research questions

This work attempts to address the following research questions:

RQ1: Is it possible to provide a clear conceptualisation of urban road traffic congestion using an ontological model?

RQ2: Can quantitative big data be used to provide qualitative information in conjunction with a road traffic ontology with the support of machine learning?

RQ3: Can quantifiable big data on urban road congestion be visualised to provide quasi-real-time insight?

RQ4: Can a fuzzy rule-based system be designed to predict road congestion through validation of the Urban Road Congestion Conceptual (URCC) model?

1.6 Research aim and objectives

The aim of this research is ‘To develop a conceptual model that captures the semantics of road traffic congestion and its causes and to use the model to better identify and alleviate congestion.’

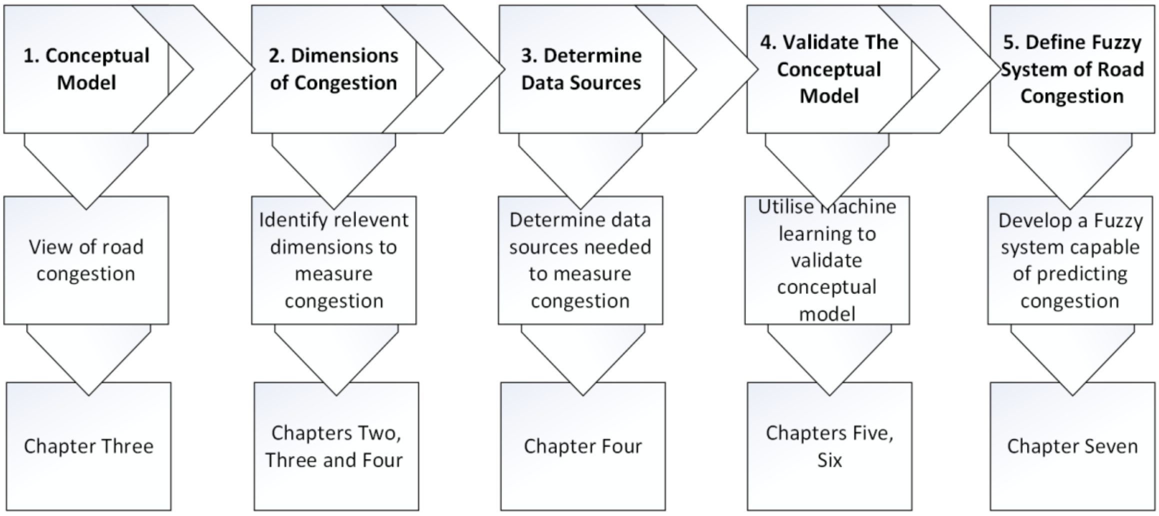
To achieve the research aim, the following objectives have been set:

1) Conduct a comprehensive review of what defines congestion, and how conceptual models have been used with the support of resilience to reduce congestion.
2) Develop an Urban Road Congestion Conceptual model using analogical and ontological approaches that identify the key concepts and the relationships between them.
3) Develop a quasi-real-time dataset using real-world data that has the capability of supporting complexity and volume at a ‘big data’ level.
4) Conduct an unsupervised learning experiment on the dataset developed in objective three, to ascertain whether it is possible to use predictive analytics to predict urban road congestion.
5) Investigate, design, and develop a fuzzy rule-based decision system to validate the Urban Road Congestion Conceptual model.

6) Develop a case study using real-time data provided by Transport for Greater Manchester to conduct a critical evaluation of the conceptual framework and its ability to support resilience in response to a road network event.

1.7 Research methodology

This section details the approach undertaken to achieve the research aim and answer the four research questions. The research is experimental in nature, building on the key findings and gaps in knowledge identified in Chapter Two: A literature review of road congestion. Figure 1 shows the five key stages of the research methodology and the associated chapters.



1. Conceptual Model
2. Dimensions of Congestion
3. Determine Data Sources
4. Validate The Conceptual Model
5. Define Fuzzy System of Road Congestion
View of road congestion
Identify relevent dimensions to measure congestion
Determine data sources needed to measure congestion
Utilise machine learning to validate conceptual model
Develop a Fuzzy system capable of predicting congestion
Chapter Three
Chapters Two, Three and Four
Chapter Four
Chapters Five, Six
Chapter Seven

Figure 1: Research Methodology

Each stage will now be briefly described:

Stage 1: Is the formulation of an urban road conceptual model of congestion leading to the development of an ontology to provide a formal and explicit conceptualisation of congestion and in particular, the impact of road accidents. This rationale, justification, and methodology for the development of this conceptual model is described in Chapter Three.

Stage 2: Identifies the different dimensions capable of defining the distinct types of urban road congestion caused by traffic events by using the ontology proposed in Chapter Three. The specific research methodology for this stage is described in Chapters Three and Four.

Stage 3: Now that the dimensions of congestion have been identified due to the development of the ontology, it is possible to distinguish which big data sources are relevant by evaluating the data sources described in Chapter Two and Four. This stage investigates whether it is possible to calculate journey time using Bluetooth sensors, Global Positioning Systems (GPS), cameras, and traffic volume with Radio-frequency Identification (RFID) and Automatic Traffic Counters (ATC). However, this research will only use Bluetooth sensors and ATC. The complete dataset will be presented in Chapter Four.

Stage 4: Introduces a statistical visualisation toolkit which was developed as part of this research and called Transport Incident Manager (TIM). TIM will utilise the relevant dimensions and their data sources to perform analytics to identify patterns in the traffic volumes and journey times, which can be used to translate quantitative data into qualitative information. Moreover, the conceptual model is also validated using machine learning techniques (Chapters Five and Six).

Stage 5: Design and develop a fuzzy decision-making system, which can predict one of four classifications (non-congestion, recurrent congestion, semi-recurrent congestion, and non-recurrent congestion). This was achieved by creating two separate fuzzy decision-making systems, the first Fuzzy decision-making system is a prototype that uses only two data sources and has a binary classification outcome (congested and non-congested). The second fuzzy decision-making system is an advancement on the previous system, as it introduces four extra data sources and has four classification outcomes (non-congestion, recurrent congestion, non-recurrent congestion, and semi-recurrent congestion). Both fuzzy decision-making systems are discussed in Chapter Seven.

1.8 Contributions

The research presented in this thesis makes several key contributions to the field.

- The first contribution is the development of a novel URCC model which conceptualises the three types of congestion: non-recurrent, semi-recurrent, and recurrent congestion. (Chapter Three)
- The second contribution is the development of the Manchester Urban Congestion Data (MUCD) Dataset which incorporates real-world data from various sources, such as TFGM and the United Kingdom's Governments freely open data. (Chapter Four)
- The third contribution is the development of a Graphical User Interface (GUI) visualisation toolkit called TIM that provides the user with better knowledge of the MUCD dataset. (Chapter Five)
- The fourth contribution is the development of a binary fuzzy decision-making system, to determine if a rule base system could identify congestion at a high level. The two classification outputs are congestion and non-congestion. (Chapter Six and Seven)
- The fifth contribution is the development of a multi-classification fuzzy decision-making system that will be used to predict the type of congestion and then validate the conceptual model. The classification outputs are non-recurrent congestion, semi-recurrent congestion, recurrent congestion, and non-congestion. (Chapter Seven)

The research presented in this thesis has led to the following peer-reviewed publications at the time of submission. A copy of the publications can be found in appendix 3 of the thesis.

Gould, N. and Abberley, L. (2017) 'The semantics of road congestion', In UTSG. Dublin.

L. Abberley, N. Gould, K. Crockett and J. Cheng, 'Modelling road congestion using ontologies for big data analytics in smart cities', 2017 International

Smart Cities Conference (ISC2), 2017, pp. 1-6, Doi: 10.1109/ISC2.2017.8090795

L. Abberley, K. Crockett and J. Cheng, 'Modelling Road Congestion Using a Fuzzy System and Real-World Data for Connected and Autonomous Vehicles', 2019 Wireless Days (WD), 2019, pp. 1-8, Doi: 10.1109/WD.2019.8734238.

1.9 Thesis overview

The research in this thesis is presented over eight chapters.

- Chapter Two provides a background review of existing literature and discusses the current state of research related to the following: Concepts of congestion, data sources used within the domain of road congestion, and existing road congestion models.
- Chapter Three will introduce the URCC model which consists of several analogies and a universal ontology of road congestion. Moreover, Chapter Three will introduce the critical third type of road congestion which has been coined as semi-recurrent.
- Chapter Four provides an insight into the creation of the MUCD dataset and validates the universal ontology of road congestion through a case study using the MUCD dataset.
- Chapter Five introduces the visualisation toolkit developed for this research called TIM and will provide examples of TIMs functionalities, such as real-time visualisation, statistical measurements, and unsupervised learning viewer.
- Chapter Six provides insight into the patterns caused by congestion which are of interest to stakeholders such as TfGM through the application of clustering techniques.
- Chapter Seven demonstrates the use of two rule base decision systems for prediction congestion and validates the conceptualisation model using the multi-classification fuzzy decision-making system. The benefit of using a fuzzy decision-making system compared to a traditional machine learning algorithm, such as a decision tree or a probabilistic model is the explainability of the outcome and the meaningful context to better assist the stakeholders.
- Chapter Eight concludes by summarising the answers to the four research questions and proposes the format of future work.

Chapter Two: A literature review of road congestion

2.1 Introduction

This chapter presents a critical review of previous literature regarding conceptualising and modelling urban road congestion and its causes. The review examines several techniques, such as analogies, ontologies, and machine learning. The review reports on various problems and challenges in the field of road congestion and the impact on urban planning, some of which will be addressed by the research presented in this thesis.

2.2 Overview of congestion

In the past, various dimensions (Measurements) have been used for monitoring traffic flow, network performance, and detecting congestion. These include journey time (Anbaroglu et al., 2014; Anbaroğlu et al., 2015), density (Bauza et al., 2010), (vehicle and traffic) speed (Bauza et al., 2010) and travel time (Bar-Gera, 2007; Li and Chen, 2014), however, this thesis will use 'journey time' to classify urban road congestion at the link level, due to the requirements set out by Transport for Greater Manchester (TfGM). Furthermore, the dimension 'traffic volume' will be used to assist with defining and predicting urban road congestion as this is another data source used by TfGM for conducting manual predictions.

There have been many definitions for defining types of congestion, such as recurrent congestion, which happens when vehicles simultaneously use the road network at peak times, and non-recurrent congestion, which happens when an unpredictable incident occurs. However, existing literature also uses measurements, such as severity (of congestion, weather, and accident) to define congestion. For example, (Bauza et al., 2010) proposes to create a novel cooperative traffic congestion detection system for highways, using fuzzy logic to detect road traffic congestion. The paper does not specify what type of congestion it is trying to detect and appears to classify both non-recurrent and recurrent congestion as a single entity. Furthermore, the paper classifies congestion into four types of severity measurements which are: free, slight, moderate, and severe using two dimensions: Traffic density and Vehicle speed. To validate the performance of the approach, a traffic simulation was conducted using SUMO.

Other papers, such as the one written by Anbaroglu (Anbaroglu et al., 2014) classifies non-recurrent congestion into three types of severity performance measurements, using only one dimension which is journey time. The proposed performance measurements are: high congestion, medium congestion, low congestion, and expected journey time. Other performance measurements for defining congestion at city scale (national) and neighbourhood (regional) levels are fast, smooth, light congestion, medium congestion, and severe congestion and are defined in (Chen et al., 2020).

Over many years, the multifaceted nature of congestion has been expressed in the literature with various definitions and terms being used, such as recurrent congestion that refers to when significant amounts of vehicles simultaneously use the overpopulated road space in an expected period (Arnott, 2013). For example, on weekday mornings and afternoons at peak times, traffic jams otherwise known as ’rush hour’. Rush hour is defined as when substantial amounts of road users are trying to use the same portion of the road network to get to work and drop their children off at schools at the same time (Emmerink et al., 1995) which can cause longer than expected journey times. Non-recurrent congestion is a term that has previously been used to defined unexpected, unplanned, or momentous events, such as traffic accidents, roadworks, extreme weather conditions, and some dedicated events like music concerts and important sports events (OECD, 2006; Djahel et al., 2015).

Throughout the years, many methods have been used, such as diagnosis (Latham, 2011; Uschold et al., 2011) to model different aspects of road congestion, for example, congestion cost (Verhoef, 1999; OECD, 2006), driver behaviour (Kilpeläinen and Summala, 2007; Fernandez and Ito, 2015) and traffic controlling (Pan et al., 2013). One limitation which was consistently observed across most of the literature is a lack of real-world data which meant a lot of the proposed models were created using simulated ‘dummy’ dataset that have equal proportion of data per classification (Emmerink et al., 1995; Sheu and Ritchie, 1998; Romilly, 1999; Arampatzis et al., 2004; López et al., 2017; Djahel, Jones, Hadjadj-aoul, et al., 2018), due to a lack of access to reliable data sources.

In the literature, a distinction is made between ‘direct’, ‘hard' or ‘physical’ data, which is data in the form of numbers or graphs, for instance, the speed or volume, and ‘indirect’, ‘soft’ or ‘relative’ data, which is qualitative information and requires interpolation and lacks the rigor that is implied in statistical data, for instance, a tweet about a ‘major’ road accident. Nevertheless, in recent years, there has been an increase in both hard and soft data sources, such as Bluetooth sensors and social media. This has created more dimensions, which can be used as measurements such as traffic behaviour, waiting time, volume, capacity, journey time etc. These dimensions can then be used to model certain aspects of urban road congestion by performing quantitative and qualitative analysis that can then be used to inform stakeholders, such as road users, policy makers, and traffic managers of potentially congested areas in quasi-real-time.

Table 1 shows the different approaches which have been taken in the literature regarding how congestion has been viewed and dealt with in the past. There are five primary approaches observed in the reviewed literature, which are optimisation, mitigation, traffic control, congestion cost, diagnosis, or a mixture of two or more.

The key findings from the literature are as follows:

35% of literature (shown in Table 1) (Herman and Prigogine, 1979; Emmerink et al., 1995; Sheu and Ritchie, 1998; Thomas, 1998; Sheu, 1999; Fernandez-Caballero et al., 2008; Wang et al., 2009; Riad and Shabana, 2012; Arnott, 2013; Tsekeris and Geroliminis, 2013; Chen et al., 2014; Liang and Wakahara, 2014; Mathew and Xavier, 2014; Othman et al., 2015; Patire et al., 2015; Wu et al., 2015; Steenbruggen et al., 2016) had a single approach to optimising an Intelligent ITS or a Transport Management System (TMS). However, only a single piece of literature (Chen et al., 2014) discussed developing their own unique ITS, with a large amount of focus being on either analysis or reviewing previous ITSs or TMSs for informing road users, policymakers, or transport managers.

Indirect data sources have not been used much throughout the literature (Emmerink et al., 1995; Koetse and Rietveld, 2009; Lécué et al., 2012; Pan et al., 2013; Li and Chen, 2014; Djahel et al., 2015; Chen and Rakha, 2016; Steenbruggen et al., 2016) with events information being used only twice, weather stations being used five times and social media being used three times.

Whilst reviewing the literature from the past 30 years it is noticeable that data sources prior to 2000 (Herman and Prigogine, 1979; Arnott et al., 1993; Emmerink et al., 1995; Pope et al., 1995; Gualtieri and Tartaglia, 1998; Sheu and Ritchie, 1998; Thomas, 1998; Romilly, 1999; Sheu, 1999; Verhoef, 1999; Yasdi, 1999) tended to have limited data available. For instance, traffic volume data was manually collected through physical labour until the introduction of inductive loop counters in the late 90s, which are the most reliable method for detecting traffic flow. moreover, between 2000 and 2010 (Yuan and Cheu, 2003; Arampatzis et al., 2004; Verhoef and Rouwendal, 2004; OECD, 2006; Fernandez-Caballero et al., 2008; Wen, 2008; GUO and HUANG, 2009; Koetse and Rietveld, 2009; Lozano et al., 2009; Wang et al., 2009) alternative data sources became widely available and used, such as Radio Frequency Identification Devices (RFID), probe vehicles and cameras. Finally, post 2010 (de Palma and Lindsey, 2011; Mandal et al., 2011; Lécué et al., 2012; Riad and Shabana, 2012; Arnott, 2013; Bauza and Gozalvez, 2013; Pan et al., 2013; Tsekeris and Geroliminis, 2013; Isa et al., 2014; Li and Chen, 2014; Liang and Wakahara, 2014; Mathew and Xavier, 2014; Agarwal and Kickhöfer, 2015; Djahel et al., 2015; Othman et al., 2015; Patire et al., 2015; Shao et al., 2015; Stefanello et al., 2015; Wang et al., 2015; Wu et al., 2015; Chen and Rakha, 2016; Colak et al., 2016; Grote et al., 2016; Kaddoura and Nagel, 2016; Steenbruggen et al., 2016; Zhang et al., 2016) seen an increase of even more data sources being used, such as Bluetooth, Global Positioning System (GPS) and Global System for Mobile communication (GSM).

With new techniques, continually being developed within both data collection and geospatial analysis and with the introduction of more modern data sources and dimensions as mentioned above, there are now more ways to analyse congestion in urban and rural areas. Finally, with the advancements in data analytic techniques, such as big data, data fusion, data mining, and machine learning, will allow several heterogeneous datasets to be integrated into a single consistent dataset that can be used to model and provide a meaningful representation of the real-world object known as congestion and the events that cause it.

Table 1: Literature Review of Congestion

Approaches	References	Network Structure					Objective Function									Type		Data Sources											Method		
																		Direct ('Hard)								ndirect ('Soft'					
		Whole Network	City Centre	Highways	Intersection	Links	Time Saving	Reduce impact	Detect Anomlies	Analysis	Develop	Review	Inform Road Users	Inform Policy Makers	Improve Transport Management System	Recurrent	Non-Recurrent	Traffic Amount	Bluetooth	loop Counters	GPS	RFID	Probe Vehicles	Cellular Data (GSM)	Cameras	Events Information	Wealter Stations	Social Media	Data Analysis	Spatial & emproal Anylsis	Hybrid
Optimization	(Emmerink et al., 1995)	🟥								🟥			🟥				🟥	🟥									🟥				
	(Sheu and Ritchie, 1998)		🟥						🟥						🟥		🟥	🟥		🟥											🟥
	(Thomas, 1998)					🟥			🟥						🟥					🟥			🟥						🟥		
	(Sheu, 1999)			🟥				🟥												🟥									🟥		
	(Fernandez-Caballero et al., 2008)			🟥				🟥	🟥																🟥				🟥		
	(Arnott, 2013)	🟥								🟥						🟥															
	(Chen et al., 2014)	🟥							🟥	🟥	🟥						🟥											🟥	🟥		
	(Wu et al., 2015)	🟥				🟥							🟥		🟥					🟥				🟥	🟥				🟥		
	(Othman et al., 2015)	🟥						🟥																					🟥		
	(Herman and Prigogine, 1979)	🟥										🟥																			
	(Wang et al., 2009)			🟥										🟥																	🟥
	(Riad and Shabana, 2012)		🟥				🟥	🟥		🟥			🟥								🟥			🟥						🟥	
	(Tsekeris and Geroliminis, 2013)					🟥				🟥																					
	(Mathew and Xavier, 2014)				🟥	🟥						🟥							🟥		🟥	🟥		🟥	🟥						
	(Liang and Wakahara, 2014)					🟥		🟥					🟥		🟥			🟥													
	(Patire et al., 2015)		🟥					🟥							🟥				🟥	🟥	🟥								🟥		
	Steenbruggen, Tranos, & Rietveld, 2016)			🟥					🟥						🟥									🟥			🟥				🟥
Mtigation	(Isa et al., 2014)											🟥					🟥														
	(Shao et al., 2015)		🟥				🟥																								
	(Wang et al., 2015)	🟥						🟥										🟥		🟥											
Traffic Contro	(Wen, 2008)				🟥		🟥	🟥							🟥	🟥						🟥							🟥		
	(Kaddoura and Nagel, 2016)	🟥						🟥								🟥	🟥													🟥	
Congestion Cost	(Romilly, 1999)	🟥						🟥							🟥			🟥													
	(GUO and HUANG, 2009)					🟥	🟥																								
	(Chen and Rakha, 2016)	🟥		🟥			🟥								🟥	🟥				🟥	🟥		🟥	🟥		🟥					🟥
	(Grote et al., 2016)		🟥					🟥						🟥				🟥		🟥											
Diagnosis	(Pope et al., 1995)		🟥					🟥				🟥																			
	(Verhoef, 1999)					🟥				🟥		🟥		🟥						🟥											
	(Yasdi, 1999)					🟥									🟥	🟥				🟥									🟥		
	(Arampatzis et al., 2004)		🟥					🟥						🟥																🟥	
	(Koetse and Rietveld, 2009)			🟥	🟥	🟥								🟥													🟥		🟥		
	(Lécué et al., 2012)		🟥						🟥						🟥											🟥	🟥		🟥		
	(Pan et al., 2013)		🟥		🟥	🟥			🟥												🟥							🟥		🟥	
	(Arnott et al., 1993)				🟥	🟥				🟥		🟥																			
Hybrid	(Stefanello et al., 2015)		🟥				🟥																								
	(Gualtieri and Tartaglia, 1998)		🟥			🟥		🟥						🟥																🟥	
	(Yuan and Cheu, 2003)			🟥					🟥								🟥	🟥		🟥			🟥						🟥		
	(Verhoef and Rouwendal, 2004)	🟥								🟥				🟥																	
	(OECD, 2006)				🟥	🟥		🟥						🟥		🟥	🟥														🟥
	(Lozano et al., 2009)				🟥	🟥			🟥									🟥							🟥				🟥		
	(de Palma and Lindsey, 2011)	🟥										🟥		🟥		🟥							🟥	🟥	🟥						
	(Mandal et al., 2011)					🟥		🟥	🟥						🟥							🟥	🟥	🟥							🟥
	(Bauza and Gozalvez, 2013)	🟥							🟥				🟥	🟥									🟥								🟥
	(Li and Chen, 2014)			🟥			🟥	🟥									🟥			🟥			🟥				🟥				🟥
	(Djahel et al., 2015)				🟥							🟥			🟥	🟥	🟥		🟥	🟥				🟥	🟥			🟥	🟥		
	(Colak et al., 2016)	🟥					🟥																								🟥
	(Zhang et al., 2016)		🟥					🟥							🟥																
	(Agarwal and Kickhöfer, 2015)		🟥					🟥										🟥												🟥	
		14	13	8	8	15	8	19	11	8	1	8	5	10	14	8	9	9	3	13	5	3	7	8	6	2	5	3	14	6	9

2.3 Concepts of congestion

A transportation system is a fundamental part of society with the road network being the linchpin that holds the other transportations modes together (Somuyiwa et al., 2015). Unfortunately, existing road networks have become severely congested due to the increase in new and existing drivers using their vehicles more and the inability to develop a more resilient network (Hartgen and Fields, 2009). The Highways Agency estimates that 65% of congestion on the network is caused by traffic volumes at or above capacity (this is the quantity of vehicles, per hour, per lane that the network can manage before congestion occurs (Hall and Agyemang-Duah, 2000)), 25% of incidents and 10% by roadworks (Department for Transport, 2014).

Congestion continues to remain a long-standing problem in transportation science with literature going back as far as 1920 by Pigou (Verhoef, 1999). Throughout all the vast amounts of literature on the detection of congestion, there is an apparent absence of a clear and consistent definition of what congestion is. This is partly due to the multifaceted nature of congestion and the different perceptions within the various disciplines conducting research, such as Ecology, Economics, Intelligent Systems, Geography, Engineering etc. Moreover, according to the (Department for Transport, 2013), the definition of congestion requires both a physical and relative dimension, because ‘a person living in a rural area might regard an unusually long queue of traffic experienced on their daily commute as severe congestion, while someone living in an urban area might experience much longer hold-ups on a daily basis and regard the same length queue as being almost totally uncongested’ (Department for Transport, 2013). Therefore, congestion is relative and can be dependent on the road users’ personal opinions.

Road traffic congestion has previously been classified as three types: non-recurrent, recurrent (Djahel et al., 2015), and pre-congestion state (Somuyiwa et al., 2015). These definitions are relatively vague because although the terms such as recurrent or non-recurrent are widely accepted, they tend to be viewed from a more personal perspective. Table 2 shows a definition for each type of congestion identified in the literature.

Table 2: Classification of Congestion

Congestion Type	Definition	References
Recurrent congestion	Occurs when significant amounts of vehicles simultaneously use the limited road space on a weekday morning and afternoons peak hours' causing a traffic jam situation.	(Arnott et al., 1993; Cassidy and Bertini, 1999; Verhoef, 1999; Verhoef and Rouwendal, 2004; Arnott, 2013; Fosgerau and Small, 2013; Tsekeris and Geroliminis, 2013; Tadeusiak, 2014)
Non-recurrent congestion	Occurs from unpredictable incidents such as traffic accidents, work zones, extreme weather conditions and some special events like music concerts and important sports events.	(Emmerink et al., 1995; Yang, 1997; OECD, 2006; Chen et al., 2014; Isa et al., 2014; Li and Chen, 2014; Djahel et al., 2015)
Pre-congestion (borderline congestion)	Occurs when free flow conditions breakdown. However, full congestion has not yet occurred. This can happen either side of congestion and can occur either upstream or downstream of congestion which is already occurring.	(Somuyiwa et al., 2015)

2.4 Data sources

A "successful" TMS or an ITS will be largely dependent on how current and newly developing data sources are used. Moreover, it was observed in the literature (Arnott et al., 1993; Emmerink et al., 1995; Pope et al., 1995), that before the 2000s technology was limited and data collected was primarily done manually. During the 2000s, a few newer technologies became more regularly available, and from the 2010s to the present-day the transport industry exploded with more widely available data sources being used for analysing, monitoring, and predicting traffic behaviour and events that have a consequence of congestion. The major data sources being used are inductive loop counters (Thomas, 1998; Sheu, 1999; Verhoef, 1999; Yasdi, 1999; Li and Chen, 2014; Djahel et al., 2015; Chen and Rakha, 2016); Bluetooth (Mathew and Xavier, 2014; Djahel et al., 2015; Patire et al., 2015); GPS (Riad and Shabana, 2012; Pan et al., 2013; Mathew and Xavier, 2014; Patire et al., 2015; Chen and Rakha, 2016); RFID (Wen, 2008; Mandal et al., 2011; Mathew and Xavier, 2014); probe vehicles (Thomas, 1998; Yuan and Cheu, 2003; de Palma and Lindsey, 2011; Mandal et al., 2011; Bauza and Gozalvez, 2013; Li and Chen, 2014; Chen and Rakha, 2016); GSM (de

Palma and Lindsey, 2011; Mandal et al., 2011; Riad and Shabana, 2012; Mathew and Xavier, 2014; Djahel et al., 2015; Wu et al., 2015; Chen and Rakha, 2016; Steenbruggen et al., 2016); cameras (Fernandez-Caballero et al., 2008; Lozano et al., 2009; de Palma and Lindsey, 2011; Mathew and Xavier, 2014; Djahel et al., 2015; Wu et al., 2015); event information (Lécué et al., 2012; Chen and Rakha, 2016); weather (Emmerink et al., 1995; Koetse and Rietveld, 2009; Lécué et al., 2012; Li and Chen, 2014; Steenbruggen et al., 2016) and social media (Pan et al., 2013; Chen et al., 2014; Djahel et al., 2015).

In recent years, with the advancement of infrastructure and technology used on the road networks and communication networks or even a combination of both, such as the introduction of smart motorways (highways) (Department for Transport, 2014; Highways England, 2015) and the 4G network becoming more widely available, has allowed for the creation of modern ITSs, which are a principal component of smart cities and is reliant on having as many data sources as possible available in real-time. The next generation of ITSs has started to incorporate the data sources mentioned in this chapter, allowing for the development of Vehicular Ad-hoc Networks (VANETs) which, is where humans, vehicles, Roadside Units (RSUs), and infrastructure have the potential to become a data source (Golestan et al., 2015).

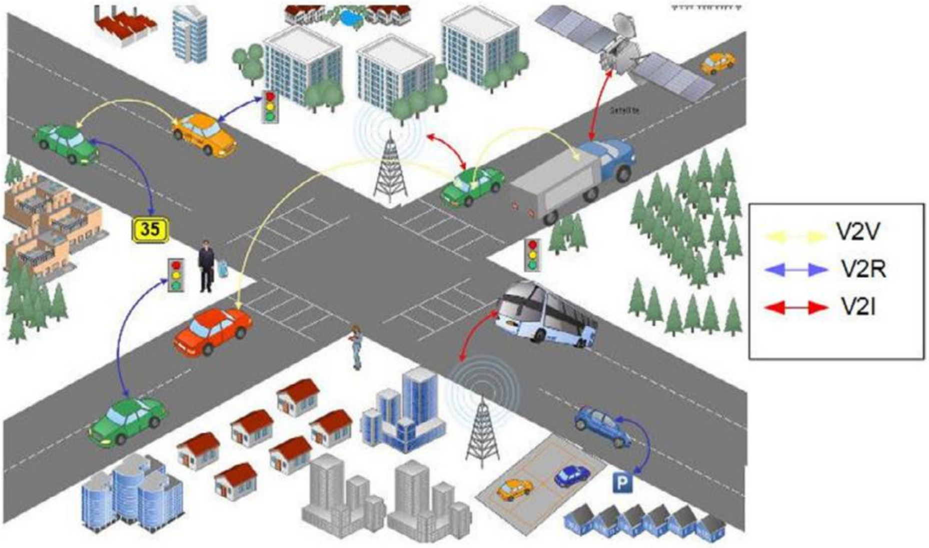


Figure 2: Different types of communication in VANETs [Source: (Golestan et al., 2015)]

Figure 2 illustrates different types of sensors and how they could potentially communicate as a network to transmit data with the aim of improving ITSs. In Figure 2, the connections being used are Vehicle-to-Vehicle (V2V), Vehicle-to-RSU (V2R) and Vehicle-to-Infrastructure (V2I). Furthermore, other possibilities are Vehicle-to-Human (V2H), and Vehicle-to-Sensor (V2S). In recent statistics it has been noted (Golestan et al., 2015) by 2020 an estimated 50 billion “things” will be connected to the internet, which will allow data to be collected from various sources.

2.4.1 Bluetooth

Within and around Greater Manchester, UK, 741 permanent passive Bluetooth sensors have been deployed (Atkin, 2016), with the intention of measuring the journey time on key routes to help TfGM to meet their Key Performance Indicators (KPIs) and will be achieved through monitoring vehicles travelling past passive Bluetooth sensors around Greater Manchester.

The benefits of passive Bluetooth sensors are the inexpensive cost of deployment and their ability to recognise and store relevant information from another Bluetooth device within range. Bluetooth devices are now commonly found in cars, smart watches, and mobile phones that tend to be carried by passengers. The main limitation of passive Bluetooth sensors is the signal is limited and can only reach a short distance. Furthermore, due to the sensors collecting all Media Access Control (MAC) addresses from all active Bluetooth devices within range of the sensor location it is inevitably creating a lot of outliers and noise. For instance, if a single vehicle is carrying multiple people (with Bluetooth devices) within range of a sensor, this will create duplicated records in the data. Moreover, when a pedestrian or a cyclist with a Bluetooth device, such as a mobile phone or a smart watch passes a sensor, the MAC address will be logged and stored in the same database as the Bluetooth devices within a vehicle, causing some journey times to appear slower than expected. Finally, if a vehicle is stationary at a set of traffic lights for an extended amount of time, this will once again produce a duplicate record.

2.4.2 Inductive loop counters

Inductive loop counters are also known as automatic data collectors (ADC), automatic traffic recorders (ATR), and automatic traffic counters (ATC). They are considered to be one of the most reliable and trusted methods available for traffic detection (Djahel et al., 2015). In and around Greater Manchester, 286 permanent ATC sensors have been deployed (Atkin, 2016), to count and classify the types of vehicles around Greater Manchester. Due to the reliability and trustworthiness of ATCs, they have become regularly used for the validation of other data sources (Djahel et al., 2015). One of the main limitations of Inductive Loop Traffic Counters is the cost of deployment, which means there is only a limited amount deployed around urban areas which in turn restricts its use within some ITS due to the sparseness.

2.4.3 Cameras

Cameras are used to detect vehicle speeds, capacity on the roads and traffic incidents. An example would be Automatic Number Plate Recognition (ANPR), which is widely used in Police vehicles, area-based schemes and are primarily used to track vehicles speeds and charging travellers for entering a restricted area (Maruyama and Sumalee, 2007), for example, London congestion zone. Furthermore, ANPR cameras have previously been used for identifying traffic incidences, such as foreign objects on the road, vehicles on the hard shoulder, vehicles that are travelling too fast, and vehicle that has stopped or broken down in the middle of the road (Fernandez-Caballero et al., 2008; Lozano et al., 2009).

Moreover, ANPR cameras can be used to monitor traffic flow. This can be achieved with various image processing techniques. Figure 3 and Figure 4 shows two different techniques for monitoring traffic flow.

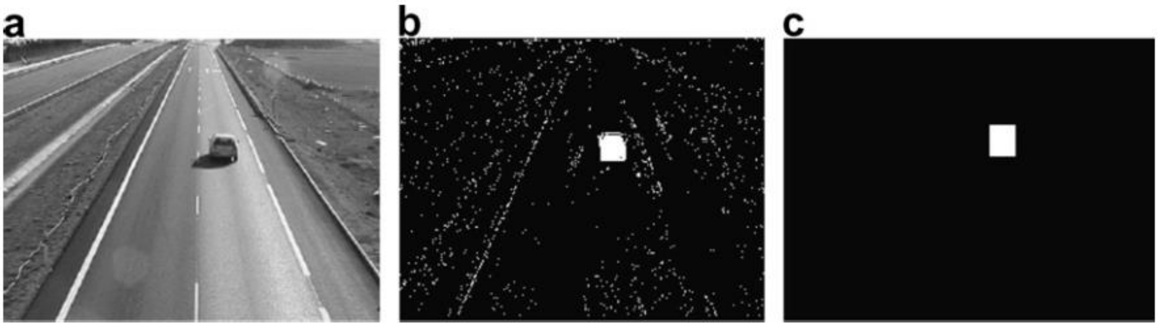


Figure 3: Road traffic monitoring images. (a) The real image is in grey scales. (b) Segmented image. (c) Processed image. [Source: (Fernandez-Caballero et al., 2008)]

Figure 3 is taken from the paper (Fernandez-Caballero et al., 2008) showing the three steps of image processing that is used to analyse traffic on a highway. Figure 3a uses a 256-grey scale image format; this is then processed to provide a black-grey scale, where the vehicle is highlighted as white. Finally, one final process is applied to the image to convert the vehicle shape into a rectangle. This will enable the systems to classify the type of vehicles such as a transit van, a lorry, a small car, or a medium car. In addition, it is possible to calculate the speed of the vehicle depending on the time it takes to progress through two frames captured at separate time intervals.

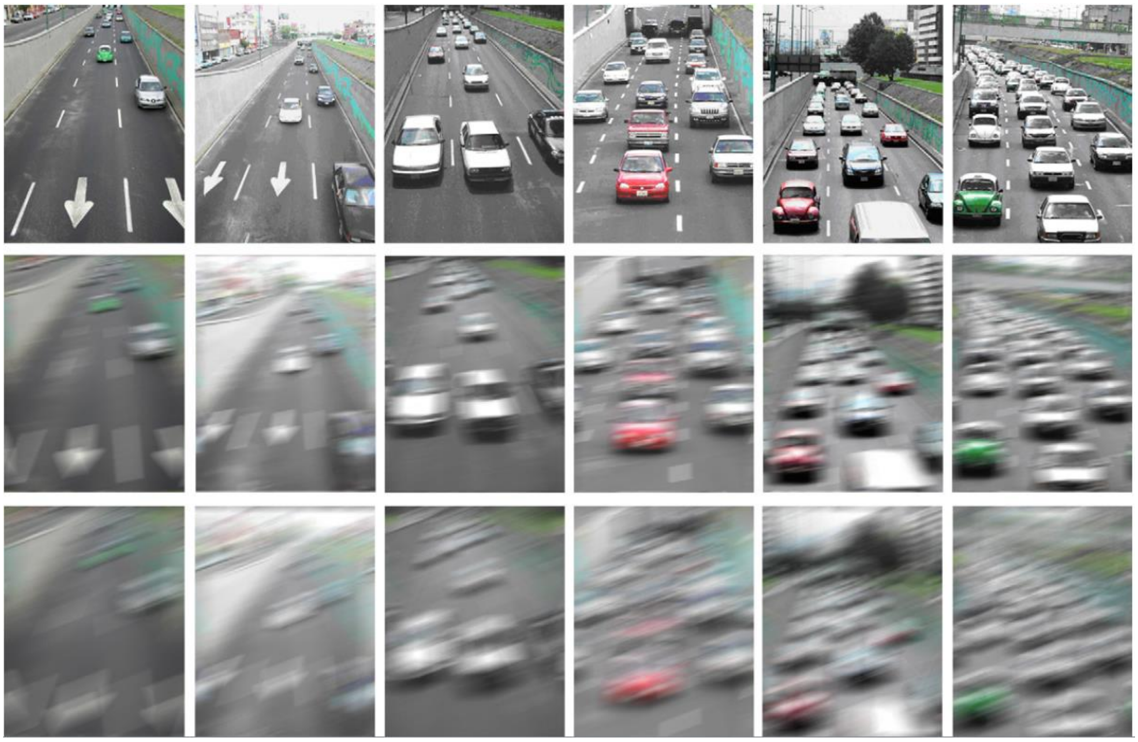


Figure 4: Blurred images. [Source: (Lozano et al., 2009)]

Figure 4 is taken from the paper (Lozano et al., 2009) and shows six different levels of traffic flow that are used to train the system. The six levels of traffic flow are (top row of images) represented starting from left to right are free flow, stable flow (slight delays), stable flow (acceptable delays), approaching unstable flow (tolerable delay, occasional wait), unstable flow and forced flow. The blurred filters are applied to represent the motion of both vehicles and the camera. All these images are then stored in the system as a training set.

Both papers (Fernandez-Caballero et al., 2008; Lozano et al., 2009) managed to achieve their aims, however, it would be almost impossible to implement either of these methods on a large scale because even though cameras are one of the most accurate methods of collecting information and data, through their ability to visibly record congestion and the event that caused it. It is practically impossible to implement an automated process that allows the valuable information to be gathered, transmitted and processed. Furthermore, this would require a multifaceted image processing software suite and image processing would need to be repeatedly done at regular intervals and the cost of transmitting the data to a singular point for transport managers to use this information is a highly expensive process. Considering the physical cost of the infrastructure required and the excessive time required to process and feed the relevant information back to the transport managers would outweigh the value of information being received (Mathew and Xavier, 2014).

2.4.4 Global Positioning System

Global Positioning Systems (GPS) is a global navigational satellite system (GNSS) that is able to compute and provide the location of a GPS capable device, such as a mobile phone and the time the observation was observed from the GPS capable device regardless of weather conditions. The main benefit of using GPS technology is devices are becoming more commonly compatible over recent years with devices, such as satnavs, tablets, mobile phones, smart watches, and vehicles having GPS built in.

Google has taken advantage of GPS data being more widely available within everyday smart devices and uses GPS data to perform traffic analysis (Google, 2016). Whilst the quality of the traffic analysis is generally good for navigating and calculating traffic flow, it is not practical for a typical control system, such as ramp metering, which requires a basic traffic light to be located on the slip roads entering the highway and are designed to stop vehicles from merging onto an overly populated highway unsafely and typically requires density data to work (Patire et al., 2015). However, GPS data could be merged with other data sources such as inductive loop counters, providing the vast potential for developing a hybrid Transport Management System (TMS). The concept of a hybrid TMS has gained traction within recent literature (Riad and Shabana, 2012; Pan et al., 2013; Patire et al., 2015; Chen and Rakha, 2016).

2.4.5 Radio Frequency Identification Devices

Radio Frequency Identification Devices (RFID) are used to automatically identify vehicles and collect data similar to the data collected by ANPR cameras. One of the positives of RFID is its low implementation cost, however, due to the limited traffic information collected by the RFID, it tends to benefit from being used in collaboration with an alternative technology such as GSM.

This theory has been tested in the following literature which was conducted by (Mandal et al., 2011). (Mandal et al., 2011) proposed an ITS capable of monitoring and measuring road traffic congestion using a collaboration between both RFID and GSM technology. Calculations were performed using the data collected from the RFID and GSM data sources, to calculate vehicle speeds over a stretch of road and the average waiting time at an intersection. Although RFID is a relatively old technology, it has not been implemented extensively on all vehicles and due to the maximum range of 10 meters (m), such systems cannot be implemented on highways (Mathew and Xavier, 2014).

2.4.6 Probe vehicles

The concept of probe vehicles, sometimes referred to as floating cars, have been used for collecting real-time traffic data since the late 90s and early 2000s, with a steady increase in the literature presenting probe vehicles as a solution (Mandal et al., 2011; Li and Chen, 2014; Chen and Rakha, 2016).

Probe vehicles are extremely useful because they have numerous traffic sensing technology provided by a single source (the vehicle) (Figure 5), such as GPS, Velocity (speed), Bluetooth, Wipers (Weather Conditions), Lights (Lighting Condition), and RFID. Each source has the potential to gather relevant information that can be fed back to the central processing centre where traffic management experts are located. These traffic management experts will then be able to identify when a road may be closed, what the traffic flow conditions are, whether the driver is experiencing hazardous weather and/or is the visibility reduced due to thick fog.

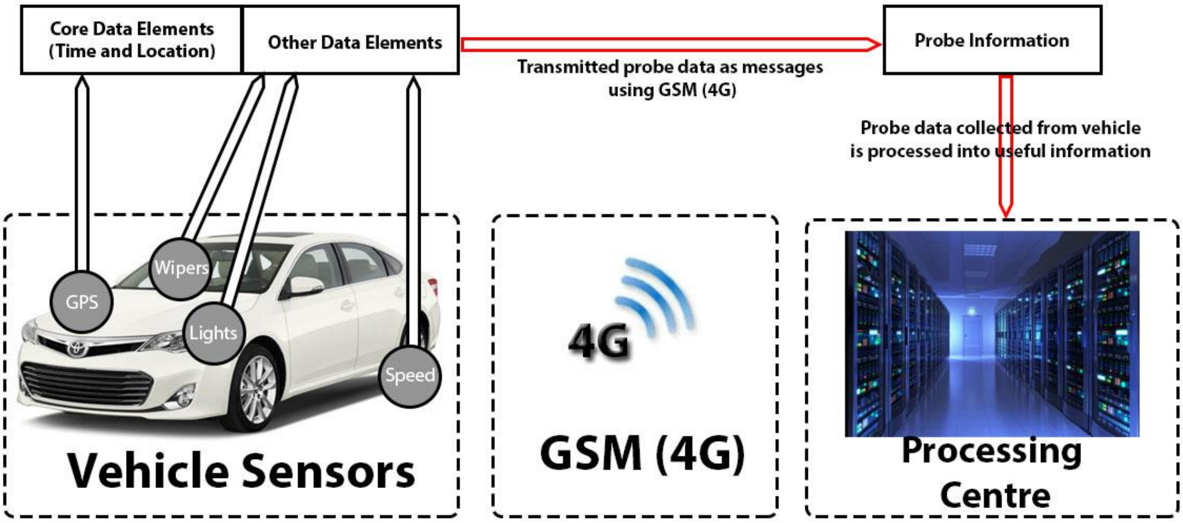

Core Data Elements (Time and Location)
Other Data Elements
Transmitted probe data as messages using GSM (4G)
Probe Information
Probe data collected from vehicle is processed into useful information
Wipers
GPS
Lights
Speed
Vehicle Sensors
4G
GSM (4G)
Processing Centre

Figure 5: Probe Vehicle [Source: Author]

With vehicles becoming ‘smarter’ with the introduction of auto lights, auto wipers, and remote diagnostics with the addition of technologies, such as OnStar (*OnStar*, 2016), more vehicles have the ability to become a probe vehicle and with more manufacturers adding the technology to be able to transmit data back to a central point. There are still limited volumes of vehicles be used as a probe and would require a lot more probe vehicles on the road to produce meaningful information gain. Like other data sources, probe vehicles data is often very noisy and can often be tough to provide an accurate reading (Chen et al., 2014).

2.4.7 Cellular data

Cellular Data such as GSM is a standard developed by the European Telecommunication Standards Institute which allows mobile phones to access a digital cellular network (2G). Over the years, the cellular networks have advanced from 2G to 3G and finally 4G. 4G has the capability to offer potential download speeds of up to 300Mbps and upload speeds of 150Mbps. The continuous improvement of the GSM is relevant because with the increase of newer data sources being implemented in ITSs means more data is being transmitted over mobile communication and having faster speeds will provide better gains and better reliability by having faster data transfer rates for transmitting real-time information. Additionally, the use of cellular data is a breakthrough because the volume of people who have mobile phones and are travelling on road networks has increased rapidly. One of the major limitations of research being conducted into the use of Cellular data is, a large proportion of mobile devices are not set in an active mode. Therefore, researchers are not able to utilise the full potential of cellular data (Mathew and Xavier, 2014).

2.4.8 Event information

When predicting traffic flow or estimating journey time with a route guidance system (RGS) such as a satnav, it requires vital information, such as time of day and day of the week, to allow a comparison against historical data, which will give a more accurate estimation. For example, as part of the initial exploratory phase with TfGM data, Figure 6 shows a comparison of a typical journey time (along with a single link between two passive Bluetooth sensors) against a journey time when there is a football match at the Etihad Stadium. Figure 6 Shows the readings, which are grouped into ten minutes' slots (x-axis) and journey time (y-axis). The black line represents the mean of several typical days, which in this case is four previous Tuesdays prior to the match day. The green line is one standard deviation above and below the mean. The blue line is the day Manchester City are playing at home (Etihad Stadium). When comparing this line to the mean of several days, a spike in journey time is noticeable prior to the five pm kick-off. After five pm this spike settles back down. However, once the match has finished at nine pm and everybody wants to leave the stadium at the same time, another large spike occurs in the journey time.

Figure 6 shows that it may be possible to identify traffic patterns of a football match event and that it may be possible to predict that slight congestion will happen prior to the match starting and hypercongestion after the match finishes.

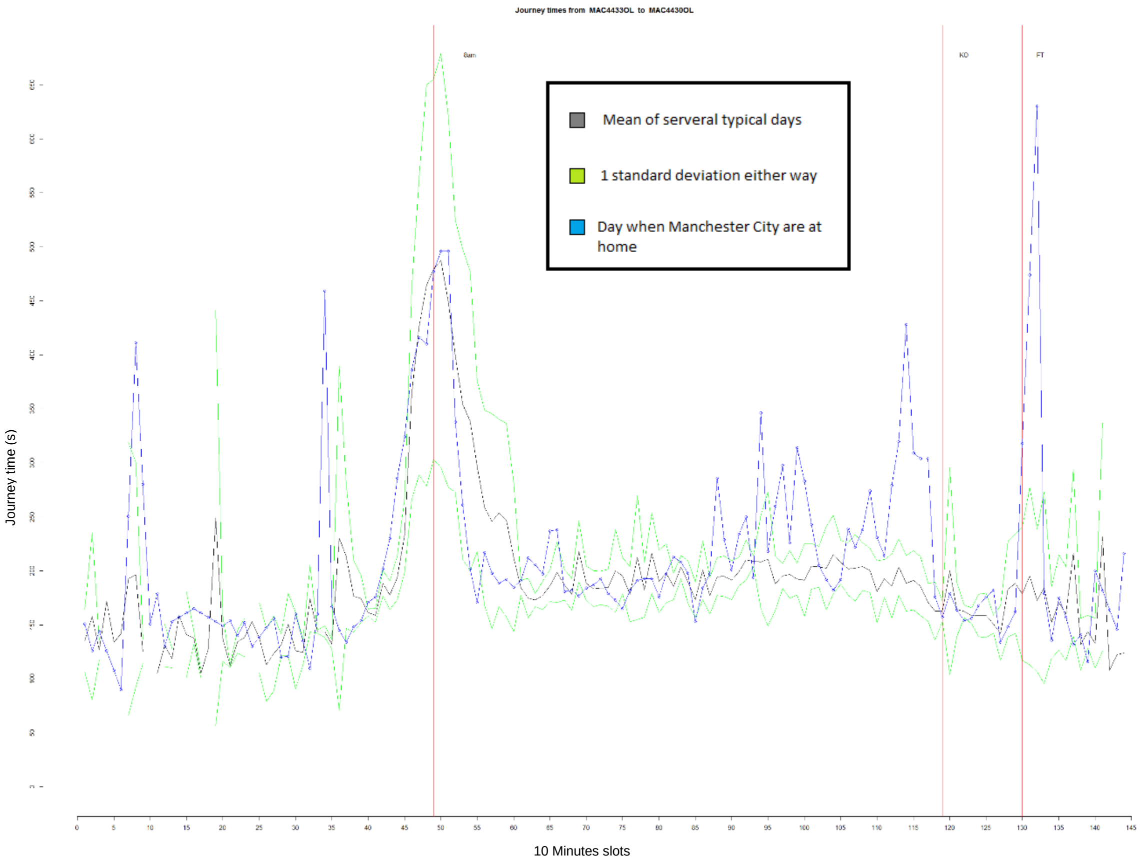


Figure 6: Comparing Journey time on a game day at the Etihad Stadium [Source: Author]

2.4.9 Weather

Weather can have a large impact on the road network and depending on the severity, it can cause wear and tear (damage to the road surface), congestion, and increased journey times which are greatly underestimated by road users and even some researchers. Nevertheless, extreme weather has the power to disrupt free-flowing traffic due to several things, such as damaged infrastructure and reduced visibility which can cause drivers to reduce their speed and may cause the typical traffic flow to change from free flowing to congestion. An example of infrastructure damage by extreme weather would be (News, 2015) during August 2015 in Greater Manchester, UK.

Torrential rain spread throughout the city for days causing widespread flooding and brought the urban road network to a standstill. The primary reason for the impact on the urban road network was a 40ft deep sinkhole that opened on an arterial road known as ‘Mancunian Way’ (Figure 7 and Figure 8).

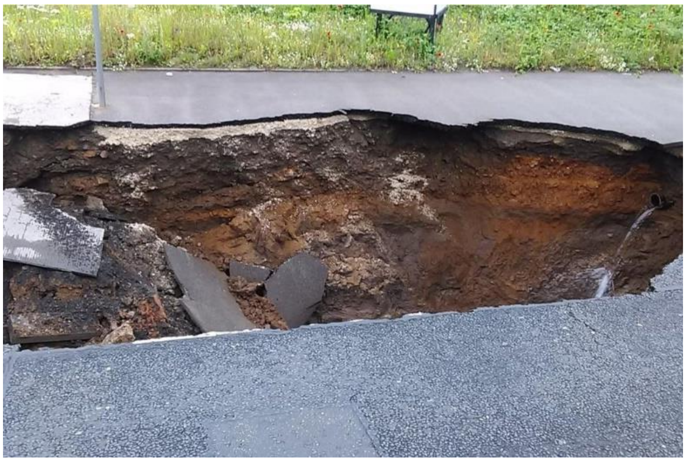


Figure 7: Sinkhole picture 1 [Source: (News, 2015)]

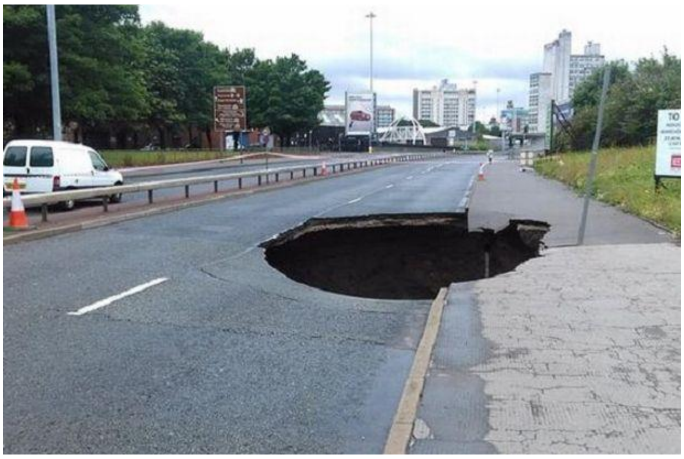


Figure 8: sinkhole picture 2 [Source: (News, 2015)]

The sinkhole caused one of Greater Manchester’s busiest roads to remain closed for ten months and caused congestion around the city due to large volumes of traffic being diverted. (News, 2015, 2016). Heavy rainfall has the ability to affect traffic flow at an alternative location due to the reduction in visibility, causing a driver to reduce their speed and the driver behind them to reduce their speed more than the person in front until traffic builds up causing congestion (Li and Chen, 2014).

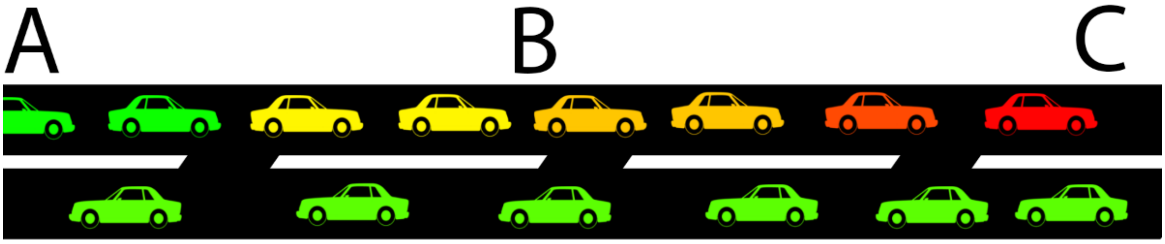


Figure 9: an example of how rain at location A would cause congestion at location C [Source: Author]

Figure 9 shows that when heavy rainfall happens at Sensor A, vehicles speed would be reduced over a short period and by the time the vehicles are monitored at sensor B, the traffic flow would become a bound flow. Due to the cause and effect of traffic flow and the reduction of vehicle speeds, it is inevitable once vehicles reach sensor C there will be heavy traffic due to the reduction of speed prior to sensor C, which is originally caused by the heavy rainfall at sensor A (Li and Chen, 2014). Additionally, if the traffic is monitored in reverse. It is possible free-flowing traffic could still exist because the rain is occurring at sensor A.

Extreme weather such as rainfall is one of the leading causes of non-recurrent congestion (Changnon, 1996; Koetse and Rietveld, 2009; Department for Transport, 2014). Thus, making it extremely difficult to predict without the relevant data source/s to monitor weather changes in real-time and a unique algorithm capable of identifying weather patterns changes.

2.4.10 Social media

Social media applications, such as Facebook and Twitter are becoming a more widely used source of data for analytical and research purposes. Previously, Twitter has been used to predict ‘am’ recurrent congestion (Yao and Qian, 2021) and was chosen to be used because it was claimed that traditional methods, such as autoregressive and spatio-temporal models are ‘extremely limited’. However, (Yao and Qian, 2021) noted one of the limitations of using social media, such as Twitter is spam and advertisement posts which are posted by bots. Although this research has demonstrated it is possible to use social media for predicting traffic, what it demonstrated was extremely limited as it was only capable of being able to predict traffic patterns for the next day.

Furthermore, Twitter has been used to identify non-recurrent congestion by identifying road traffic incidents in China (Luan et al., 2021). In 2014, Twitter was also used to monitor road traffic congestion by observing tweets, and the use of specific words to describe road conditions, such as ‘slow’ or ‘congestion’. Then a model that uses traffic language was developed by (Chen et al., 2014) to help identify large-scale events that have the consequence of congestion. (Chen et al., 2014) observed three technical challenges of using Twitter to monitor road traffic, this is due to the multifaceted nature of Twitter, which requires pedestrians, passengers, and drivers to be treated as sensors. These ‘sensors’ are required to observe the physical world and record observations accurately, which can cause a few technical challenges to arise. These challenges are language ambiguity, geographic location uncertainty, and uncertainty between the interactions of road traffic-related incidents (Chen et al., 2014).

In addition to the challenges mentioned above, a study using Twitter to analyse people's behaviour in a natural disaster found that people tend to add personal feelings and options to their tweets (Hara, 2015). This is a potential problem when identifying congestion from personal tweets.

2.5 Existing models and techniques

Ample research has been focused on a diverse set of aims such as time-saving (Yang, 1997; Tsekeris and Geroliminis, 2013; Colak et al., 2016; Kaddoura and Nagel, 2016); reducing the impact (Arampatzis et al., 2004; Verhoef and Rouwendal, 2004; He et al., 2016); detecting road traffic incidents (Sheu and Ritchie, 1998; Pan et al., 2013; Steenbruggen et al., 2016); analysing and developing new policies for TMSs (Nankervis, 1999; Van Schijndel and Dinwoodie, 2000; Reggiani et al., 2015) and although there are several individual aims mentioned, they all have one primary aim in common, which is to mitigate against the problem of road traffic congestion; using their own unique approaches, such as developing a vehicle-to-vehicle network, comparing similar techniques, data sources and measurements; with the aim of finding the optimal solution by modifying parameters. Additional approaches, such as controlling traffic (Wen, 2008; Kaddoura and Nagel, 2016) by setting restrictions on turning, speeds, and changing signal patterns and mitigation (Isa et al., 2014; Liu et al., 2015; Shao et al., 2015) of traffic towards less congested areas with the objective to lessen the effects of road traffic congestion.

Popular ‘policy’ approaches, which are an alternative to the ‘technological’ approaches, mentioned previously are encouraging behaviour changes. For example, charging to entering congestion zones, increasing parking charges, creating bus-only lanes, and providing cheaper public transport or even a hybrid of some of the mentioned approaches have been explored by transport managers.

2.5.1 Models of congestion

In the earlier years of modelling congestion, a ‘Two-Fluid’ approach was conducted by (Herman and Prigogine, 1979), which looked at the relationship of the evolution of speed, which assesses how road users take different approaches to achieve their desired speed. However, this tends to cause conflict between faster and slower road users. The crucial limitation of this approach was the fragmented and random data being manually collected across several cities within the United States. This was due to the lack of technology able to gather data in 1979.

With the advancement of technology, newer data sources which are more accurate become available. However, even with newer data sources slowly becoming available, the focus on modelling road congestion is still primarily done through modelling demand, using models, such as the bottleneck model, Origin-Destination (OD) model, and the bathtub model which has been reviews and improved up on by other researchers (Arnott and Buli, 2018; Jin, 2020; Bao et al., 2021). These models will now be briefly described.

2.5.2 Bottleneck model

A review was conducted by (Arnott et al., 1993) into numerous demand models known as 'bottleneck models', which were basic, used fixed number of drivers, elastic demand, capacity arbitrary, optimal capacity and the self-financing of capacity. The review argued that peak-period congestion is poorly specified and focused mainly on social costs such as user demand and available capacity; not considering the consumers' behaviour decisions. For instance, where a user trades the convenience of time with the congestion cost such as queuing.

The weakness of these models is the lack of data sources capable of measuring the decision of users and relies only on a single data source, which measures the capacity at a link or intersection vulnerable to a bottleneck occurring.

2.5.3 Bathtub model

20 years later (Arnott, 2013) published a paper with the aim of improving the bottleneck model and producing a new approach to the demand models. He developed a concept of 'A bathtub model of downtown rush-hour traffic congestion' that was built upon a conversation with William Vickrey a few years before his passing. Arnott coined the term 'bathtub model of the road traffic congestion' as a dynamic approach to the demand models, where a disruption at one location can instantaneously spread to all other locations.

The Bathtub model was an improvement on his previous work by simulating a whole city (Manhattan) and with the addition of manually collected real world data to validate the model functionality. The model used two dimensions, which are capacity and flow. These are essential for the model to work. Think of the bathtub as Manhattan. In addition, cars entering Manhattan traffic stream, come from either across the bridges, tunnels or from parking spaces in Manhattan. These cars correspond to the inflow of water into the bathtub. Cars leaving the traffic stream, by either entering parking spaces or exiting Manhattan across the bridges or through the tunnels, corresponds to the outflow of water from the bathtub. The height of the water within the bathtub corresponds to the traffic density.

Figure 10 shows there is a clear connection between traffic velocity and traffic density and how velocity and density relate to the three traffic flow stages, free flow, bound flow, and congestion.

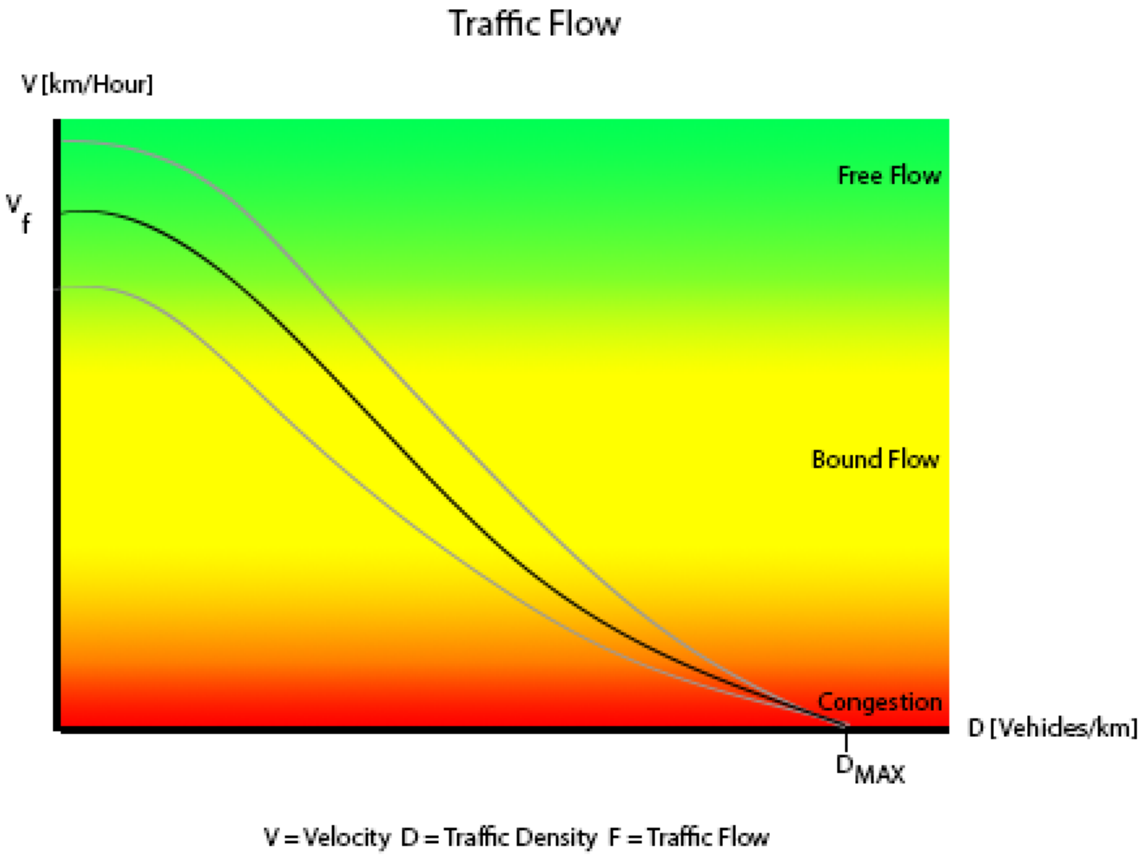


Figure 10: Traffic Flow Diagram [Source: Author]

Equation 1 shows how traffic flow (F), which equals density (D) times velocity (V) is used to try and prevent congestion and to retain a consistent traffic flow.

$$(\mathrm{F} = \mathrm{D}.\mathrm{V})$$

Equation 1: Traffic Flow

The number of vehicles entering the city during the morning rush hour traffic is required to be equal or less than the number of vehicles leaving the city. If these requirements are not met, the free flow state will change to a critical state in which congestion will then occur. The weakness of this model is, once the capacity within Manhattan has been reached, you cannot merely turn off the taps to restrict the flow into the city, without causing a build-up of traffic at the bridges and tunnels into Manhattan, creating numerous bottlenecks around the perimeter.

2.5.4 Origin to Destination model

Numerous researchers such as Guo and Huang, (2009), Wu et al., (2015) and Othman et al., (2015), have investigated OD (Origin to Destination) models. OD models are then used within several TMS, such as RGS which are designed to inform road users of the best route to complete their journey, be it only a link or across a whole network. This is achieved by mapping all links and intersections within a network and calculating a cost for each, with the parameters being the cost of travel, time, speed, flow, and any road events that could increase the time to reach the destination. RGS has been relatively successful over recent years and is constantly improving by implementing new techniques, which the OD models do not incorporate. For instance, road users' behaviours and choices can have a large impact on the optimal route taken by alleviating demand on the network and reducing the time required to reach the chosen destination (Colak et al., 2016).

2.5.5 Data analysis and geospatial techniques

Over the years, many conceptual models of congestion have been developed, using different networks sizes ranging from an intersection (Wen, 2008; Pan et al., 2013; Djahel et al., 2015), link (Thomas, 1998; GUO and HUANG, 2009; Wu et al., 2015), highway (Sheu, 1999; Fernandez-Caballero et al., 2008; Wang et al., 2009), city centre (Sheu and Ritchie, 1998; Riad and Shabana, 2012; Patire et al., 2015) and a whole network (Emmerink et al., 1995; Arnott, 2013; Chen et al., 2014). In addition to the various networks used, various methods, data sources, and dimensions were used in experiments with a combination of different techniques. Although throughout the literature many techniques were observed, it is possible to compartmentalize the techniques used into two separate categories, Data Analysis and Geospatial aspects.

2.5.5.1 Data analysis

Data Analysis (Liang and Wakahara, 2014; Othman et al., 2015; Shekhar et al., 2015) is the process of inspecting, cleansing, transforming, and modelling data with the primary aim of discovering meaningful information that can be used to help support decision-making. Data fusion, data mining, data processing, data interpretation, and machine learning have all been incorporated into data analysis due to the overlapping of these techniques.

Data fusion (Zheng et al., 2014; Radak et al., 2015; Wu et al., 2015) is a process of integrating multiple data sources and dimensions representing the same real-world objects into a consistent and meaningful representation.

Data mining (Kianfar and Edara, 2013; Pan et al., 2013; Li and Chen, 2014) is becoming more regularly used in many disciplines, and it is primarily used within computer science. The primary aim of data mining is to use computational procedures to determine patterns(Pan et al., 2013; Shekhar et al., 2015) within large data sets involving approaches at the intersection of database systems, statistics, artificial intelligence, and machine learning.

Data processing is the carrying out of operation, by either a human or computer to retrieve, transform, or classify information.

Data interpretation is the final stage of data analysis and is a vital stage. Data interpretation is the process of assigning meaning to the processed data, which will allow a conclusion to determine whether the information collected was significant.

Machine learning is interrelated to data mining because data mining is one of the crucial components of machine learning, and both techniques are used in an attempt to find meaningful patterns within the data, however, the main difference is machine learning tries to establish an automated correlation to a classification.

2.5.5.2 Geospatial aspects

‘Is the geographic world a jigsaw puzzle of polygons, or a club-sandwich of data layers?’ was a question asked by (Couclelis, 1992). Figure 11 shows a visual representation of a GIS (Geographic Information System) as a combination of Computer Science and Geography. GIS systems include a database with spatial and temporal characteristics to create computer-based information systems capable of capturing, modelling, storing, retrieving, sharing, manipulating, analysing, and presenting geographically referenced data.

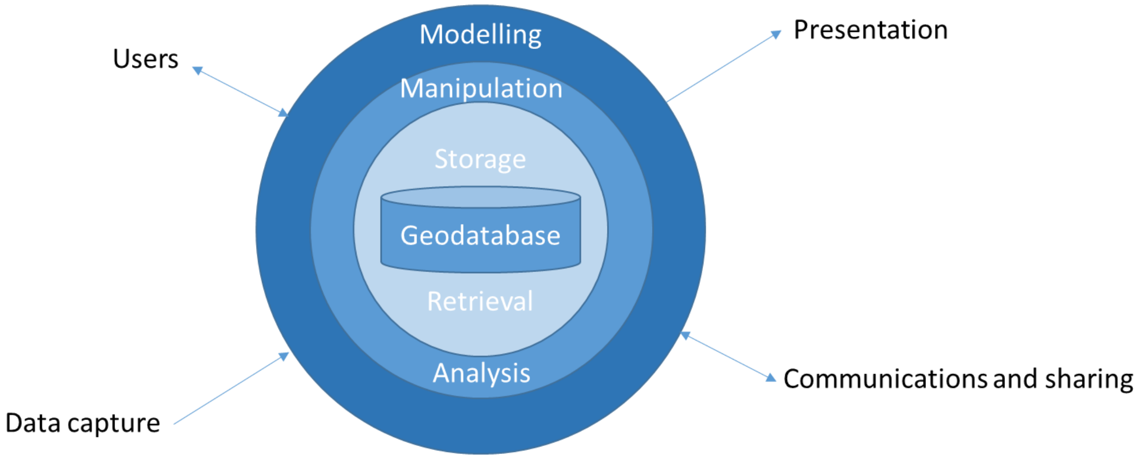


Figure 11: GIS Visual Representation [Source: Author]

A GIS is a simplified view of the real world and has the capability to share geospatial data between different Information systems or even between the various components within a single information system. For applications such as satnav, it is crucial to give the stakeholder the optimal route from point A to point B in an acceptable timeframe. Therefore, making it vital to have a well-maintained database management system (DBMS) that is reliable, accurate, consistent, technology proof and secure.

Two GIS data models are Vector and Raster. Vector data represent space as a series of discrete entity-defined points, polylines, and polygons, which tend to have a static representation regarding X and Y coordinates. Raster data is more appropriate when modelling continuous geographic phenomena such as elevation of land usage. Over the years, GIS has become increasingly more popular and is being more frequently integrated into transportation applications such as RGS and TMS. The use of GIS has become so popular that transportation applications using GIS are routinely referred to as GIS-T (Waters, 1999). An example of what GIS can be utilised for is plotting government data on road transport accidents to identify clusters, the data is published by the Department for Transport and is widely available at (Gov.uk, 2017). These records provide details about the circumstances of all road accidents in Great Britain from 1979, the types (including Make and Model) of vehicles involved and the significant casualties. Figure 12 shows a map of all accidents from 2015 in the Northwest (NW) of England, using the data provided by the Department for Transport.

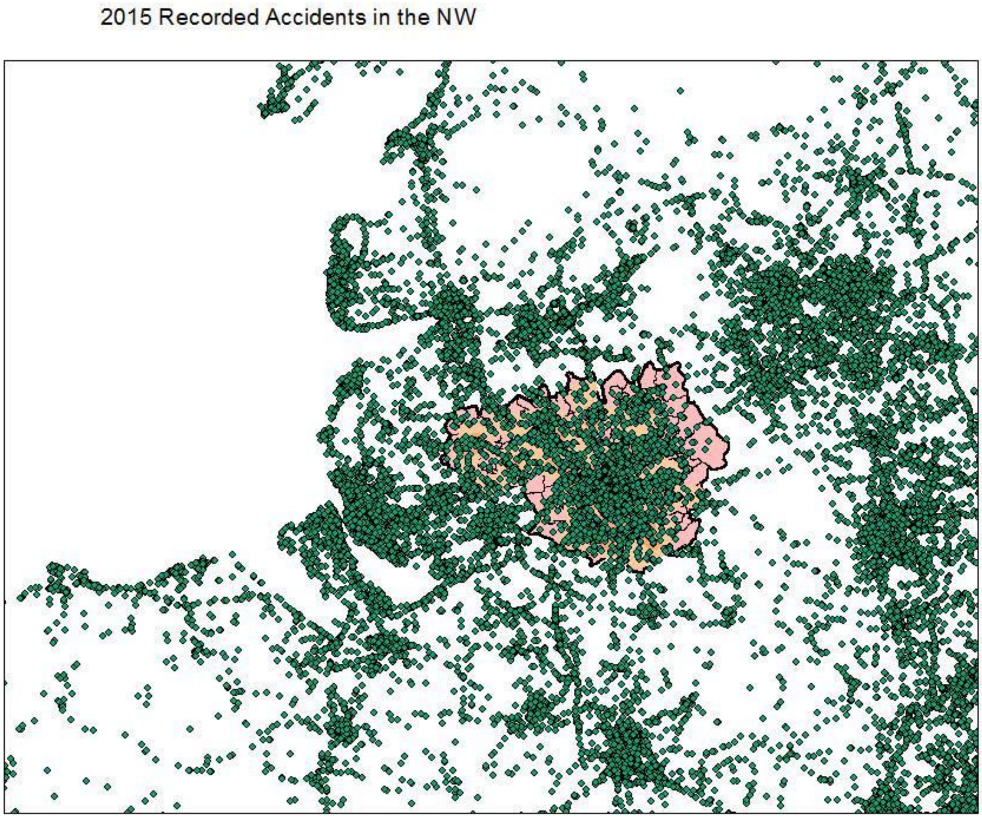


Figure 12: Road Traffic Accident Records for North West England 2015 [Source: Author]

Furthermore, GIS allows for the plotting of sensors such as passive sensors (Bluetooth) and Automatic Traffic Counters in Figure 13, with the powerful spatial analytic tools making it possible to create density maps to show where the majority of sensors are located. See Figure 14.

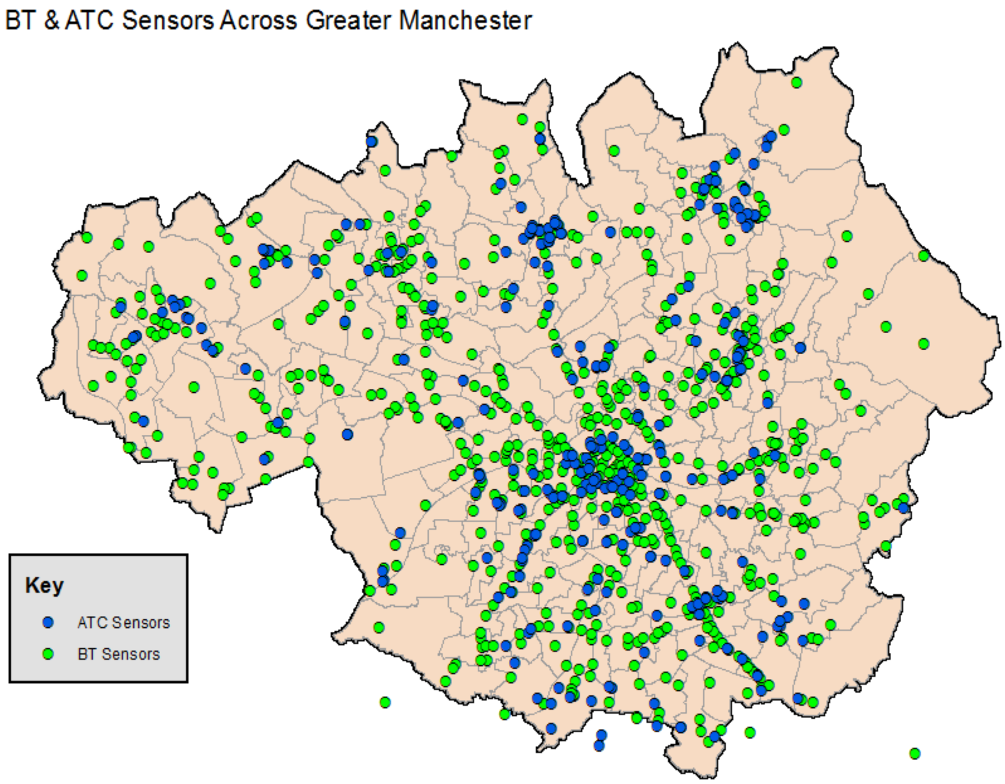


Figure 13: Manchester BT and ATC Sensors Map [Source: Author]

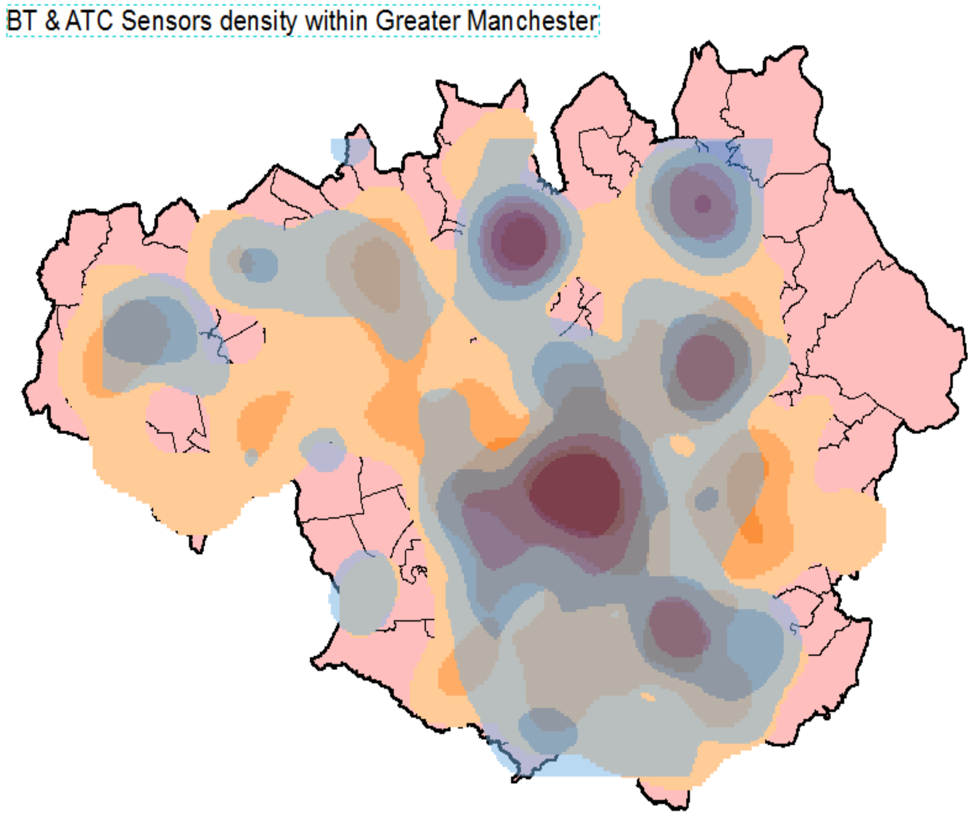


Figure 14: Manchester BT and ATC density Map [Source: Author]

It is vital to use network analysis to calculate the distance between two points on a map to plan routes, calculate driving time and locate facilities. Equation 2: Euclidean distance (d) was used to derive the distance between a and b with the calculation (Figure 15).

$$\mathrm{d}(\mathrm{a},\mathrm{b}) = \sqrt{(\mathrm{a}_\mathrm{x} - \mathrm{b}_\mathrm{x})^2 + \left(\mathrm{a}_\mathrm{y} - \mathrm{b}_\mathrm{y}\right)^2}$$

Equation 2: Euclidean Distance

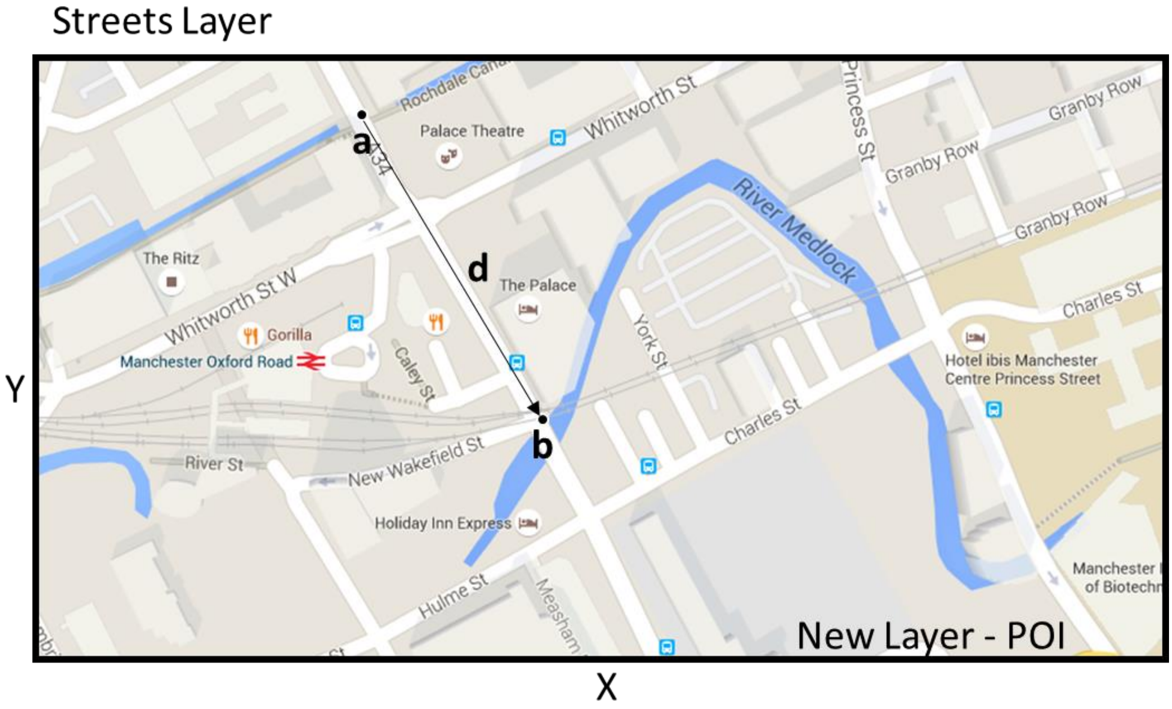


Figure 15: Euclidean Distance Example

Although, the Euclidean distance would normally be fine to calculate the distance between two points, such as ‘a’ and ‘b’. Assume both points were on the same link and did not have any obstacles in the way, such as one-way systems, a road closure due to a traffic incident, or roadworks then Euclidean distance is an easy calculation to understand. However, if the points are on the opposite side of a river with the next nearest bridge 1km away an alternative method of calculating the distance would be needed to calculate an accurate travel time and distance.

2.6 Chapter conclusion

This chapter has identified several gaps within the literature which this research will go on to address. The first gap is the lack of a clear and consistent definition of what is meant by ‘urban road congestion’. The second gap this research will address will consist of trying to re-evaluate whether the ‘generic’ and ‘commonly’ used classifications of congestion known as recurrent and non-recurrent are still relevant and provide a true representation of urban road congestion.

This will be achieved by ascertaining whether a more granular classification is required to provide a better-suited definition of urban road congestion that is clearer and more meaningful. Both the first and second identified gaps will be addressed in chapter three with the development of the Urban Road Congestion Conceptual (URCC) model, which will consist of several analogies and a universal urban road congestion ontology. The third gap identified is the lack of real-world real-time big data datasets in relation to urban road congestion. This will be addressed in chapter four where data from several different sources will be integrated into a single Manchester Urban Congestion Data (MUCD) dataset. The fourth gap identified is a lack of useful visualisation and data analysis tools that can provide high-quality meaningful information. Therefore, a toolkit called Transport Incident Manager (TIM) which will be used for visualising and analysing the MUCD dataset will be introduced in chapter five.

Finally, this research will address the lack of machine learning being used to gain meaningful qualitative information from quantitative data providing useful context regard urban road congestion to a stakeholder. Instead of saying "CONGESTION AHEAD EXPECT DELAYS", it would be more beneficial to say, "CONGESTION AHEAD IN 2 MILES, DUE TO AN MINOR ACCIDENT AT 15:45 CAUSING INCREASED JOURNEY TIMES". Additionally, the review has identified a lack of interpretable prediction models of urban road congestion using real-world, unbalanced, imperfect datasets, such as the MUCD dataset. Both issues will be addressed in chapters six and seven respectively.

Chapter Three: Conceptual model; analogy and ontology

3.1 Introduction

This chapter will attempt to answer the research question (RQ1) – "*Is it possible to provide a clear conceptualisation of urban road traffic congestion using an ontological model?*" by clarifying formally and explicitly what is meant by road congestion which has been previously difficult to clearly define. Furthermore, to answer this question, this chapter will focus on the development of an Urban Road Congestion Conceptual (URCC) model using a mixed-method approach.

The URCC model will consist of two components: the first is an analogical component and the second is an ontological component. Using this mixed-method approach will help to provide a better understanding of the problem as well as provide the foundation for the development of a real-world quasi-real-time spatial-temporal big data dataset and analytics. The main problem with modelling urban road congestion is the lack of a clear and consistent definition of what is meant by 'road congestion' in an increasingly multifaceted urban context and how it relates to the events that cause it.

Due to the complex nature of road congestion, it is not possible to find a single definition that can be used to capture the semantics of the many diverse types of congestion and the events that cause it. This limits both the road user and transport managers ability to make better decisions. For instance, when defining congestion, the UK's Department for Transport (DfT) uses terms, such as physical, which is characterised by considering speeds, volume, and/or journey time on the network and relative, which is defined by the road user's expectation to define congestion (Department for Transport, 2013, 2018). When defining congestion, the U.S Department of Transportation (DoT) uses terms, such as clog, impede, and excessive fullness to describe congestion (U.S Department of Transportation, 2018). Furthermore, when academics define congestion they use terms, such as 'recurrent' (Chen et al., 2014; Djahel et al., 2015; Bifulco et al., 2016), 'non-recurrent' (Anbaroglu et al., 2014; Chen and Ahn, 2015; Chen et al., 2016), 'pre-congestion' (Somuyiwa et al., 2015), 'free-flow' (Knoop et al., 2008; Faro and Giordano, 2016), and 'hypercongestion' (Economics et al., 2003; Fosgerau and Small, 2013; Jin et al., 2015).

Having so many terms to define congestion without any clear and consistent explanation of what they all mean, makes it almost impossible for all types of stakeholders (road users, domain experts, and transportation researchers) to comprehend what is meant by congestion and how it will impact them. Consequently, strengthening the argument concerning the apparent absence of consistency due to the multifaceted nature of congestion and how it is perceived. Furthermore, it has demonstrated how important it is that a universal model needs to be developed that is capable of providing a consistent understanding of what is meant by congestion (including the causes) allowing a variety of stakeholders to gain the knowledge from an explicit and formal description of the many concepts (objects) of the complete domain (Tadeusiak, 2014). It will also enable road users to make better-informed choices before and during their planned journey and allows domain experts to be better equipped to choose an optimal response, such as extending or reducing the traffic light sequences, using Variable Messages Signs (VMS) to warn stakeholders of a road traffic events before they become impacted, and/or diverting traffic to reduce the level of the impact on the whole network based on the knowledge gained from data analytics. All this can be achieved through the development of the URCC model being proposed in this chapter.

This chapter is organised as follows: Section 3.2 will provide a comparison of new concepts of congestion. Section 3.2.1 will introduce a third type of road congestion called ‘semi-recurrent’. Section 3.3 sets out to describe the methodology on how the URCC model is developed. Section 3.4 will provide a brief overview of what the URCC model consists of. Section 3.4.1 introduces and evaluates the four analogies of road congestion which are used to support the development of the associated ontology. Section 3.4.2 describes the methodology for creating the road congestion ontology. Section 3.5 introduces and evaluates the five core ontologies which are fundamental components of the URCC model. Section 3.6 concludes the chapter.

3.2 Comparison of a new concept of congestion alongside the traditional concepts

Road congestion is not a new phenomenon and remains an outstanding problem for road traffic users. With every civilization comes congestion with many unique approaches being taken to try and overcome its consequences. For example, Julius Caesar noticed narrow city streets are becoming unsafe for pedestrians due to the increasing use of good carts and to solve this problem he introduced a ban on good carts during the daylight hours. Nevertheless, this did not solve the problem, it just shifted the time period the problem occurred (Downs, 2005). This example was used to demonstrate how extensive road congestion has been and what seems to be a good idea does not often solve the problem.

Road traffic congestion has a multifaceted nature, and this is evident in the way it has previously been described by road users, domain experts, and researchers to define the perception of congestion (Department for Transport, 2018; U.S Department of Transportation, 2018). All the terms mentioned in section 3.1, appear to be meaningful whilst also remaining vague and does not provide any meaningful knowledge. Additionally, road traffic congestion is typically distinguished between two vague types: non-recurrent and recurrent congestion, the definitions of which are summarised in Table 3.

Table 3 Definition of both types of congestion currently used.

Congestion Type	Definition	References
Non-recurrent	Occurs from unpredictable incidents such as traffic accidents, work zones, extreme weather conditions and some special events like music concerts and important sports events	(Cassidy and Bertini, 1999; Verhoef and Rouwendal, 2004; Djahel et al., 2015)
Recurrent	Occurs when significant amounts of vehicles simultaneously use a limited road space, such as on a weekday morning and afternoons peak hours' traffic jam situations.	(Verhoef, 1999; Hendricks et al., 2001; Arnott, 2013; Fosgerau and Small, 2013)

However, this thesis argues there is a need for a third type of congestion called semi-recurrent. Semi-recurrent congestion will be described in section 3.2.1.

3.2.1 The coining of semi-recurrent congestion

Figure 16 introduces the four characteristics that can be used to distinguish between each type of congestion and convert the current binary classification into a multiclassification. The four characteristics are: predictable, non-predictable, cyclical, and non-cyclical. The term predictable is used when a stakeholder has prior knowledge of an event that will have an impact on the road network. The term cyclical is used when the event happens at the same time of day and day of the week. Non-predictable and non-cyclical are when there is no known knowledge or pattern for an event. For instance, a traffic event that is predictable and non-cyclical can be distinguished separately from events that are either predictable and cyclical or non-predictable and non-cyclical.

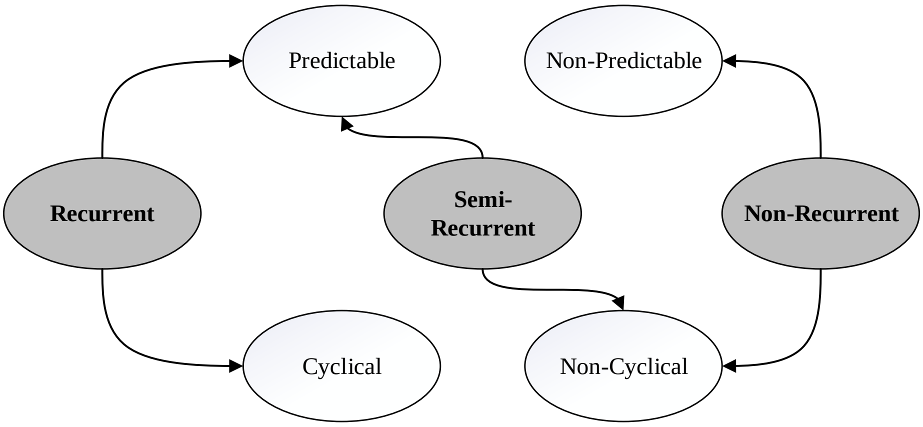


Figure 16: The proposed three types of congestion. [Source: Author]

Recurrent congestion is the consequence of events, such as ‘rush hour’ which are predictable and cyclical occurring Monday to Friday around 8 am in the morning and 5 pm in the evening. Non-recurrent congestion is the consequence of random events, such as ‘road accidents’ and ‘unplanned roadworks’. These types of events are not predictable and not cyclical because they happen at any time of day and day of the week, meaning the impact on traffic cannot be predicted. Semi-recurrent congestion is the consequence of scheduled events, such as a ‘football match’, ‘music concerts’, and ‘planned roadworks’. These types of events are not cyclical because they do not happen at the same time or on the same day. However, they do tend to be predictable due to schedules, which are created in advance.

Table 4 shows a comparison of several events and each event will have its own impact on the road network at different scales. For instance, a concert would impact a neighbourhood around the concert hall. A marathon requires roads to be closed causing an impact at a city scale. A road accident happens at a single point on a link, but the impact diffuses and has a further impact on the surrounding links (known as Point-based diffusion).

Table 4 Comparison of events, the classification of congestion types, and its entities by examples

Event	Congestion Type	Predictable	Temporal	Scales
Concert	Semi-recurrent	yes	Non-cyclical	Neighbourhood
Football Match (Cup)	Non-recurrent	No	Non-cyclical	Neighbourhood
Football Match (League)	Semi-recurrent	Yes	Non-cyclical	Neighbourhood
Marathon	Semi-Recurrent	Yes	Non-cyclical	City
Parade	Semi-Recurrent	Yes	Non-cyclical	City
Road Traffic Incident	Non-recurrent congestion	No	Non-cyclical	Point-based diffusion
Roadworks (Planned)	Semi-recurrent	Yes	Non-cyclical	Point-based diffusion
Roadworks (Unplanned)	Non-Recurrent	No	Non-cyclical	Point-based diffusion
Rush Hour (Weekday's morning and afternoon)	Recurrent	Yes	Cyclical	City-scale
Terrorist Act	Non-recurrent	No	Non-cyclical	Variety

3.3 Methodology for a universal conceptual model

The focus of this section is on developing a universal URCC model that allows different types of stakeholders to understand and benefit from gaining valuable knowledge of the multiple types of congestion, the associated events, and the impact on an urban network.

3.3.1 Overview: Universal conceptual model methodology

The methodology for creating a universal URCC model consists of three key stages, which are described as follows:

1) Perform a comprehensive review of the domain, analogies related to the domain, and any relevant ontologies that could be incorporated within the new ontology (Abberley, 2016; Abberley et al., 2017).
2) Using the review, gain an understanding of previously used concepts and develop new analogies, which are capable of capturing the required knowledge and explaining terminology in a manner a layperson would understand.
3) Using the knowledge gained from the analogies and comprehensive review, develop a road congestion ontology.

In order to validate the universal URCC model in a real-world situation. A quasi-real-time real-world dataset with spatial-temporal characteristics is required. The dataset will need to consist of journey time and traffic volume data, which is generated from real-world sensors around Greater Manchester, UK. Furthermore, additional spatial-temporal data, such as road accident data and event information will be collected and merged with the sensor data. The data being collected and processed needs to be as close to real-time as possible to be able to identify incidents and allow stakeholders to respond in a timely manner. This dataset will be introduced in chapter four.

3.4 Urban road congestion conceptual model

To solve the vagueness surrounding congestion, modelling techniques have previously been used to provide a certain level of clarification of what is meant by road traffic congestion. For instance, the bathtub model of downtown rush-hour traffic which was developed by (Arnott, 2013), only measured recurrent congestion by simulating the volume of vehicles entering or exiting Manhattan at peak times in the morning, which is represented by the water flow. However, this model has numerous weaknesses, which includes only exploring events that cause recurrent congestion, only considering a large ‘unique’ city, and only utilising a single dimension of data which was volume.

Another popular model is the bottleneck model, which signifies a limited or fixed capacity located at a single point on a link where the number of vehicles arriving exceeds this limit, causing congestion, i.e., an entry point to an industrial park (Arnott et al., 1993; Kianfar and Edara, 2013). Again, this model has several weaknesses, for instance, this study analysis a single point on a link, which consequently, does not consider the consequence of traffic building up on the surrounding network. Therefore, the URCC model being introduced will try to address some of these weaknesses and will consist of two main components which will provide two distinctive explanations of what is meant by urban road congestion.

The first component is an analogical approach (section 3.4.1) which will provide a high-level explanation of road congestion and explains how all the concepts interlink with each other, in a manner a layperson would understand. The second component is an ontological approach (section 3.4.2) which will provide a logical solution for creating a formal and explicit definition of urban road congestion, allowing more advanced stakeholders to gain greater knowledge, thanks to its ability to bridge natural language (informal) and programming language (formal).

Thanks to its high degree of expressiveness, the use of ontologies is suitable to ensure greater interoperability among agents and different applications involved in intelligent transportation systems (ITS) (Studer et al., 1998; Fernandez and Ito, 2015). Ontologies also provide a common vocabulary in a given domain and allow defining, with different levels of formality, and the meaning of terms and the relationships between them.

3.4.1 Analogies of congestion

Analogies are vital to producing a conceptual model that is universal and can be understood by anyone, ranging from a layperson with no knowledge of road congestion to a domain expert. Therefore, the use of analogies for the URCC model was chosen. Analogies are used to simplify conceptual modelling (Breitman et al., 2007), allowing familiar conceptual models to be broken down into fragments and reinterpreted providing context to newer conceptual models in alternative domains. Therefore, there is, a need to construct a model, which encompasses a universal understanding of the multifaceted nature of road congestion, overcoming the weaknesses of the previous models by capturing the causes of congestion and the impact congestion has on the network at multiple scales. For example, it can be used to explore the impact of an accident on a local (link) level or explore the impact of a premier league football match on a global (city-scale network) level.

The URCC model introduces a more granular classification of urban road congestion by breaking away from the traditional two types of congestion, which are ‘recurrent’ and ‘non-recurrent’, introducing an extremely important third type of congestion which has been coined by the author as ‘semi-recurrent’ and was introduced in section 3.2.1 and will be discussed further throughout this chapter. The analogy component of the new URCC model proposed in this chapter is made up of four interlinking analogies defined as ‘a raindrop landing on a leaf, which is floating in a bathtub with an ever-changing water temperature’. These analogies can be broken down into three primary concepts (bathtub, leaf, and raindrop) and one secondary (water temperature) concept.

The four concepts are defined as follows: The first is a well-known concept called the bathtub model (Arnott, 2013) where the bathtub represents the whole network and the water within the bathtub represents the number of vehicles using that network. However, due to the limitation of the bathtub model, this research has developed three new concepts to be able to incorporate scalability, non-recurrent congestion, and severity. These three unique analogies are a ‘leaf model’ that represents a set of connected links along a route, a ‘raindrop model’ that represents an event that has the consequence of congestion, and the final concept is ‘water temperature’ that represents the weather condition which can increase the severity of an event and will have an impact on the network. These four analogies are individual components that relate to each other to create a single model.

i. A ‘bathtub’ represents the whole network and the water that represents the volume of vehicles using the network.
ii. A ‘leaf’ represents a set of links between an origin and destination along the route.
iii. A ‘raindrop’ represents a congestion-causing event and its level of severity.
iv. An ever-changing ‘water temperature’ represents the weather condition.

Figure 17 shows a visual representation of the concept ‘Leaf inside a Bathtub’.

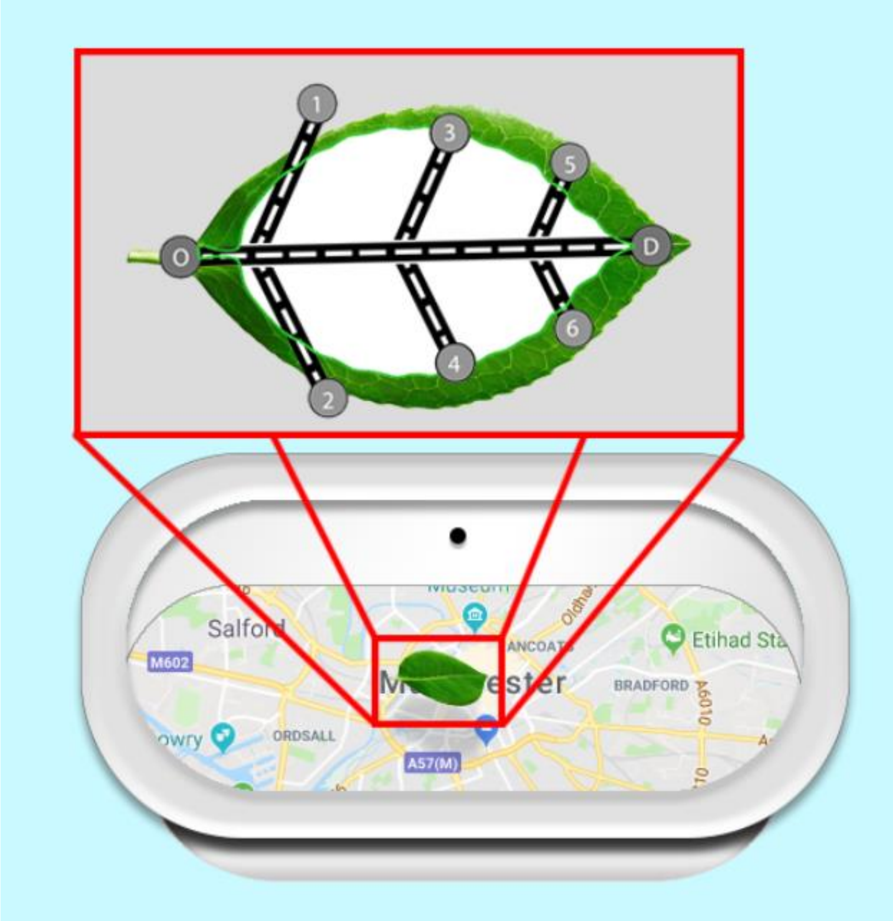


Figure 17: A conceptual model of a Leaf inside a Bathtub

3.4.1.1 Bathtub

The bathtub analogy coined by (Arnott, 2013), is used to provide an understanding of how a major city, such as Manchester, UK. The road network is impacted by large quantities of vehicles entering the city on a daily basis, this phenomenon of vehicles entering the city's urban network from the 6-lane highways that circles the city causes recurrent congestion. These vehicles correspond to the inflow of water into the bathtub, equally, cars leaving Manchester would correspond to the outflow of water from the bathtub. The ever-changing, fluctuating water level corresponds to the density of traffic within the city and as the water increases the volume of traffic becomes higher and speed becomes slower. Once the water level reaches a critical level, the bathtub will take an excessive amount of time to drain. This phenomenon has been referred to as 'hypercongestion' (Verhoef, 1999; Fosgerau and Small, 2013).

3.4.1.2 Leaf

The bathtub analogy provides a theoretically sound method of modelling recurrent congestion at a city scale. However, it lacks the ability to model congestion on a neighbourhood scale, such as a link or a set of links, which are vulnerable to non-recurrent and semi-recurrent congestion caused by road traffic incidents, public events, roadworks, and terrorist attacks. Non-recurrent congestion contributes between 40% and 70% of all congestion

(Kwon et al., 2006) and with the introduction of semi-recurrent congestion, this research has deemed it necessary to develop the following ‘leaf model’ concept.

Figure 18 shows a leaf with an origin, O, destination, D, and six additional nodes {1,...,6}, that represents a set of links within the whole network. The midrib vein that travels through the centre of the leaf corresponds to an arterial road within Manchester (UK), such as the A6 or A57. The lateral veins, which arise from the midrib vein, correspond to the less important roads that tend to lead through housing estates. These lesser important links tend to be used when an incident has occurred, and stakeholders attempt to avoid congestion.

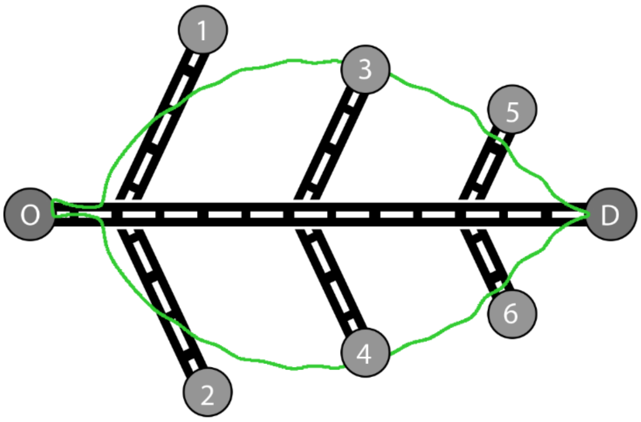


Figure 18: Leaf Concept

3.4.1.3 Raindrop

A raindrop signifies the severity of an event that has a consequence of congestion and has an impact on the road network, whether it is recurrent, non-recurrent, or semi-recurrent. A number of studies have been conducted, identifying an association between road congestion and events, such as football matches (Isa et al., 2014; Gould and Abberley, 2017), concerts (Anbaroglu et al., 2014; Anbaroğlu et al., 2015), and road accidents (Wang et al., 2009; Radak et al., 2015; Abberley et al., 2017). The impact of the events is dependent on the severity of the event. For example, depending on the severity, the impact of an accident could be very minimal, and the road segment could be cleared within minutes, or it could be extremely severe and will require several hours for the road network to return to the expected conditions.

Table 5 shows a scale of severity with regards to a road accident, which ranges from slight to fatality. These road accidents are recorded by local law enforcement (Transport, 2004).

Table 5: Instructions for completion of a road accidents report

Severity	Description
Fatal	Where death occurs in less than 30 days as a result of the accident.
Serious	Injuries sustained include fracture, internal injury, severe cuts, crushing, burns, concussion, severe general shock requiring hospital treatment, detention in hospital as an in-patient, either immediately or later and injuries to casualties who die 30 or more days after the accident from injuries sustained in that accident.
Slight	Injuries sustained include sprains, neck whiplash injury, bruises, slight cuts, and slight shock requiring roadside attention.

Therefore, the size of the raindrop corresponds to the severity of the incident, for instance, a small bump that does not require law enforcement to attend would be represented by a small raindrop with little impact on the traffic flow, speed, or journey time. A fatal accident usually requires several emergency services and would be represented by a large raindrop which, causes a mass disruption to the traffic flow, speed, and journey time of the stakeholders. Furthermore, the concentrated incident at a point would ripple out to the surrounding neighbourhood. Additionally, other types of events are represented by the raindrop and have a similar profile as a road accident. For example, a football match is similar to a road accident with a small raindrop that has little impact being a small team league game, a large raindrop with a moderate impact being a cup game, and a severe impact being a world-ranking match.

3.4.1.4 Water Temperature

The weather has a passive impact on all three of the primary concepts previously discussed above. For instance, if the temperature were to drop to minus degrees Celsius, snow and ice would likely occur impacting on the inflow and outflow of the water within the bathtub causing congestion and hyper-congestion sooner than expected. Additionally, in bad weather, traffic will become slower because of stakeholders requiring to leave extra stopping distance, reducing speeds, and setting of earlier to reduce the chances of being caught up in a road accident. Weather also has an impact on roads similar to the impact it has on leaves. It causes damage to the surface and in some cases, the damage is severe enough that it will cause non-recurrent congestion, such as the giant sinkhole that occurred on one of Manchester’s busiest roads (Gani, 2015).

Finally, incidents are also at the mercy of the weather, road surfaces can become covered in snow or excessive amounts of water as a consequence of extreme rainstorms, the correlation of a possible incident occurring increases and causing the severity to be more serious than if it was good weather. Out of all four concepts, the weather condition has been categorised by this research as a secondary concept. However, it has an influence on all three of the primary concepts.

3.4.1.5 Summary of the four analogies of congestion

To summarise the analogies, they provide a simplistic explanation for things that impact congestion, which can be understood by the stakeholders. For example, the bathtub and leaf analogies refer to the spatial context of a network, such as global scale or neighbourhood scale. The raindrop analogy refers to an event that has different levels of severity and has the consequence of congestion. Finally, the water temperature refers to weather and how extreme weather can have an adverse effect on either the road network or a specific event.

3.4.2 Ontologies of congestion

One of the gaps within the literature is a distinct lack of a clear and consistent definition of what is meant by ‘urban road congestion’. This lack of clear and consistent definition makes it impossible to answer simple questions with some level of clarity, which stakeholders, such as road users or transport managers require the answers to and tend to ask, assisting them with better decision-making. Such questions can be as simple as ‘what is meant by congestion’, ‘what is the cause of congestion’ and ‘where has congestion occurred’. These questions may appear easily answered but if you asked these questions to a layperson, they would provide an implicit and informal response that is almost as vague as what has been identified in the literature written by both academics and transport managers.

When transport managers and academics have previously discussed the aspects of road traffic congestion, they have used vague terms without fully providing a formal and explicit definition of what they mean, these terms are “recurrent”, “non-recurrent”, “pre-congestion”, “free flow”, “bound flow” and “hyper-congestion”. To further support this argument, two of the world’s leading transport departments definitions of road traffic congestion will be evaluated. The two transport departments are as followed, the Department for Transport (DfT) within the United Kingdom (UK) and the United States (US) Department of Transportation (DoT). The DfT (Department for Transport, 2013) identifies the need to provide a clear definition of road traffic congestion, in an attempt to solve this, they provide a distinction between two aspects, which are physical and relative congestion. The latter is defined by the road user’s expectation rather than using a physical definition, which considers characteristics such as speeds, capacity, and traffic flow on the network.

Whereas the report on traffic congestion (U.S Department of Transportation, 2018) by DoT focuses primarily on a relative approach to defining congestion using terms such as ‘clog’, ‘impede’ and ‘excessive fullness’ and adds ‘For anyone who has ever sat in congested traffic, those words should sound familiar’. In addition, in the same report, it is noted that congestion is typically related to an excess of vehicles on a portion of roadway or pedestrians on a sidewalk. Analysing both of these approaches of defining congestion has strengthened the argument of an apparent absence of consistency, which is largely due to the multifaceted nature of congestion and how it is perceived. Because of this, it is vital to be able to develop a way of providing an informal and explicit understanding of the domain, which both a person and non-person such as an Intelligent Transport System (ITS) will be able to understand.

Due to the concerns mentioned above, an ontology is a logical solution, thanks to the ontological ability to bridge natural language (informal) and programming language (formal). In addition, its high degree of expressiveness, makes the use of ontologies suitable to ensure greater interoperability among agents and different applications involved in intelligent transportation systems (Studer et al., 1998; Fernandez and Ito, 2015). Ontologies also provide a common vocabulary in a given domain and allow defining, with different levels of formality, and the meaning of terms and the relationships between them.

3.4.2.1 What is an ontology?

An ontology is defined as a 'formal, explicit specification of a shared conceptualisation' (Kohli et al., 2012; Gould et al., 2014) and is made up of objects, properties, facets, and instances. Ontologies are a logical solution for developing a conceptual model because of their ability to bridge natural language (informal) and programming language (formal). In addition, thanks to its high degree of expressiveness, the use of ontologies is suitable to ensure greater interoperability among agents and different applications involved in ITSs (Studer et al., 1998; Fernandez and Ito, 2015). Ontologies also provide a common vocabulary in a given domain and allow for defining with different levels of formality, the meaning of terms and the relationships between them (Fox, 2015).

3.4.2.2 Ontological methodology

The method of using ontologies for developing a conceptual model has many benefits due to its ability to provide a 'formal, explicit specification of a shared conceptualisation' (Staab and Studer, 2007), meaning it allows integration, decision support, semantic augmentation, and knowledge management. In addition, ontologies provide a visual representation of the relationships between individual objects, making it an ideal choice for developing a multifaceted conceptual model. For the creation of the universal road congestion ontology (which is one of the main components of the URCC model), a highly cited methodology for creating ontologies by (Noy and McGuinness, 2001) will be modified, which will reduce the suggested seven stages down to five stages, these stages are:

Stage 1: Determine the domain and scope of the ontology.
Stage 2: Consider reusing existing ontologies.
Stage 3: Enumerate important terms in the ontology.
Stage 4: Define the objects and the object hierarchy.
Stage 5: Define the Object-Properties.

The two stages that are not being performed are:

Stage 6: Define the facets of the Object-Properties.
Stage 7: Create Individuals.

Stages 6 and 7 are not being utilised because the aim of the ontology is to explore the nature of congestion and help to define what is meant by urban road congestion. In the proposed ontology, there are some high-level facets, such as HighVolume and LowJourneyTime. These facets are as much detail as a stakeholder requires and what is meant by the facets depends on the context. Additionally, because the ontology is not being used to classify road conditions using a reasoner[1] the creation of individuals is not required. Once the ontology has been completed, the final step is to validate the conceptual model using information for individual events that cause road traffic congestion. This will be achieved once the dataset is complete and through production of a case study described in chapter four.

3.5 A universal ontology of road congestion

This section uses both the literature review that was conducted in chapter two and the newly coined concept of 'a raindrop landing on a leaf, which is floating in a bathtub with an ever-changing water temperature' set out in section 3.4.1 to help complete the five stages.

Stage 1: Determine the domain and scope of the ontology.

The scope of the ontology is to provide a well-defined understanding of the conceptual model, which will help to identify the optimal dimensions of congestion, indicating which data sources are required. The author with the support of domain experts from TfGM (TfGM, n.d.) and Transport for the North (TfN) (TfN, n.d.), in conjunction with the comprehensive literature review conducted in chapter two and the knowledge gained from the four analogies, created the following statements to describe the domain.

- Road accidents have a consequence of congestion on the road network, as described by the raindrop model in section 3.4.1.3.
- A road network is made up of links connected by nodes similar to the leaf model in section 3.4.1.2.
- Several dimensions can be used to measure congestion. These include traffic volume, occupancy, speed, velocity, and journey time. Some of which have been used in the bathtub model.
- A road traffic event, such as an accident is an event that has a duration.
- The event has a consequence of congestion.
- Congestion has three main types: recurrent, non-recurrent and semi-recurrent. These have been explored throughout this chapter and extensively within the four analogies (section 3.4.1).
- An event that has the consequence of congestion happens at a point on the road network, which is made up of several links and nodes.
- Links can have numerous lanes and are segments of a road.
- Having more lanes on a link increases the amount of capacity, which in return will reduce the severity of congestion caused by an event, such as an accident.

[1] A Reasoner is also known as a 'semantic reasoner', 'reasoning engine' or a 'rule engine' and is a piece of software that is able to understand logical consequences from a set of rules or asserted facts. Because a reasoner is not being used.

Stage 2: Consider reusing existing ontologies.

The following three components from existing ontologies will be reused within the universal road congestion ontology to provide a level of consistency across the domain. Geospatial (Lieberman et al., 2015) is reused due to its inclusion of spatial aspects, such as point. Owl-time (Cox and Little, 2017) is reused because it is vital for anything that has a temporal entity. The two main objects used within Owl-Time are instant and interval. Transport disruption (Corsar et al., 2015), which is an extension of the ontology of the event (Raimond and Abdallah, 2007) is reused because it captures a range of events. These types of events cause the dynamic phenomena which the universal ontology is trying to model.

Other ontologies, which have been considered by the author, would be the urban density ontology (Chen et al., 2018), because it introduces objects, such as 'boundary' and 'zone', which will have provided a greater spatial understanding of events that cause road congestion in a specific zone, such as a rural area. Other spatial ontologies (Jung et al., 2013; Jelokhani-Niaraki, 2018), introduce spatial processes, which could be useful in the future.

Stage 3: Describing the important objects within the ontology.

In total, 63 objects were used to create the universal road congestion ontology. A list of the important concepts and their retrospective descriptions can be found in appendix 1.1 and these concepts and their associated descriptions were implemented using Protégé (Stanford University, 2018)[2].

The use of Protégé allows the universal road congestion ontology to be formalised using the Web Ontology Language, which is designed to characterise rich and multifaceted knowledge about things and will allow multiple terms to be used for the same object. This is important because it takes into consideration a mixture of languages, such as American English and British English, many objects could be known by multiple names. For example, Football is also known as Soccer and a motorway is also known as a highway.

Stage 4: Define the objects hierarchy.

Within Protégé, all 63 objects are restructured into a hierarchy based on their relationship with other objects. For instance, Figure 19 demonstrates that both kick-off and full-time are an instance of instant and instant is a type of time. This is important because it demonstrates the 'is-a' relationship between the many objects within the ontology. For example, in Figure 19, 'instant' is a type of time, but time is not an instant.

[2] Protégé is an OWL editor and a knowledge management system that can be used to check the *consistency* of an ontology.

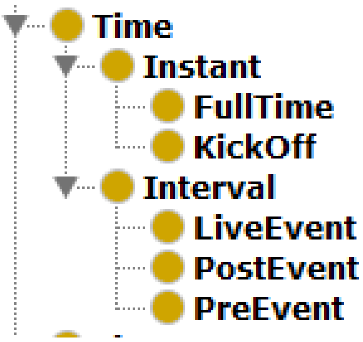


Figure 19: Snapshot of object hierarchy in Protégé

A more in-depth example can be found in appendix 1.2.

Stage 5: Define the Object-Properties

Finally, after all the objects have been described and their hierarchy structure has been defined, the final step is to create the Object-Properties which are used to demonstrate the 'has-a' relationship. For example, the properties for the object Event are 'has-a' beginning and 'has-a' end, which relate to the object instant. Table 20 in appendix 1.3 shows the several domains, their properties, and the range.

3.5.1 Implementation of the universal road congestion ontology

The construction of the universal road congestion ontology is made up of five core ontologies, which are congestion, dimensions of congestion, direction, events, and spatial. The core ontology congestion is visually represented in Figure 20. Recurrent, semi-recurrent, and non-recurrent all have an 'is-a' relationship with congestion. Congestion 'is-a' consequence of an event, which 'has-a' beginning, end, and duration. Additionally, it 'has-a' location, which is a spatial thing.

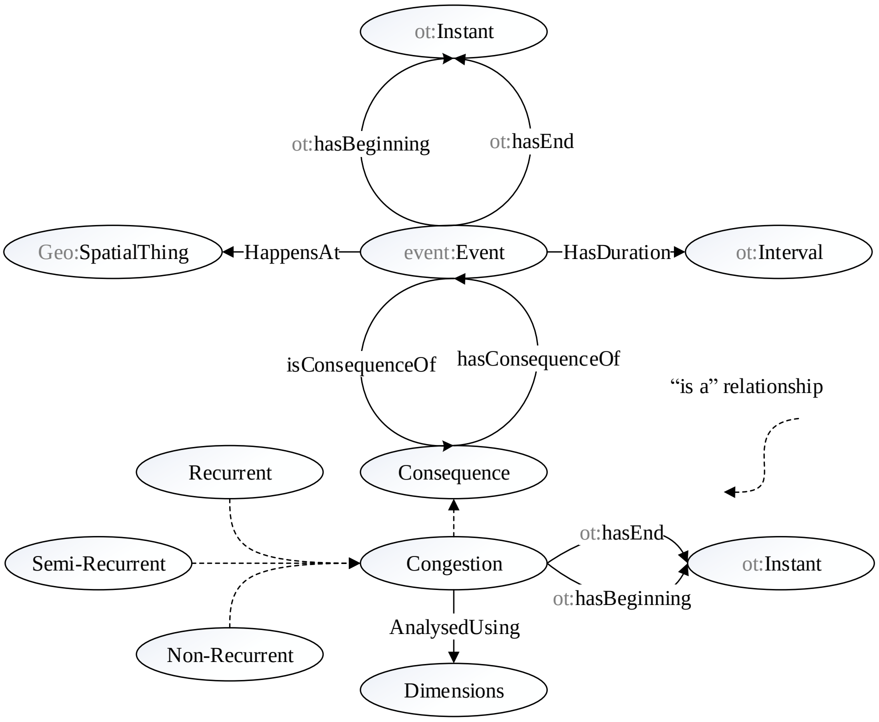


Figure 20: Ontology: Congestion

The other four core ontologies are dimensions of congestion (Section 3.5.1.1), spatial (3.5.1.2), direction (3.5.1.3), and event (3.5.1.4). Finally, how these four ontologies relate to each other will be discussed in section 3.5.1.5.

3.5.1.1 Dimensions of congestion

Figure 21 shows a visual representation of the concept of dimensions and its relevant objects. Congestion can be analysed using several different dimensions, such as speed, velocity, density, capacity, volume count, journey time, and occupancy. Occupancy and journey time are both measured using time. Velocity has a speed in a given direction. Speed can be either speed at a point or an average speed between two points. Additionally, dimensions have a magnitude level that can be used to analyse the road network performance.

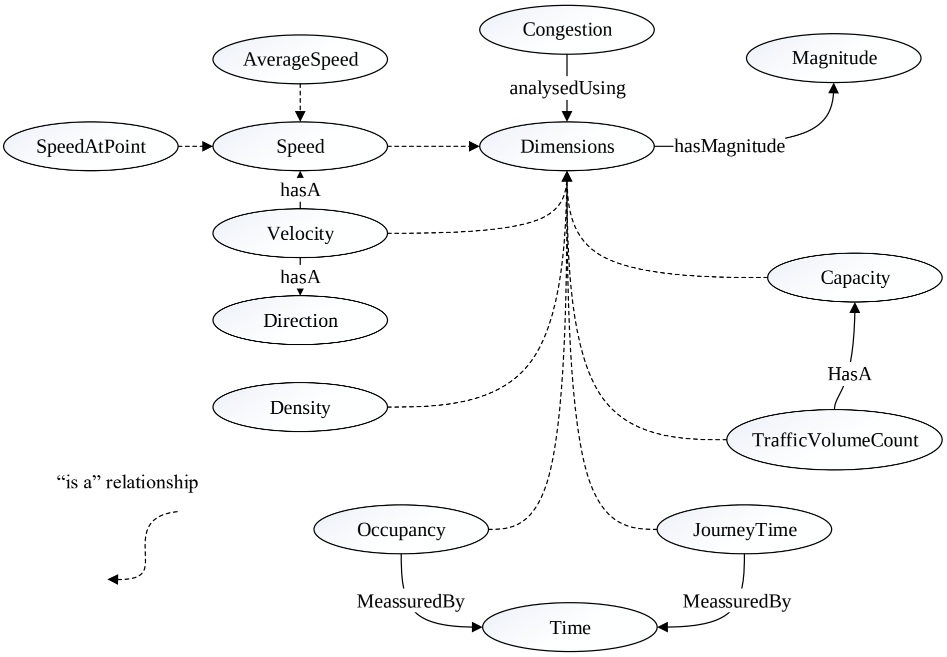


Figure 21: Ontology: Dimensions

3.5.1.2 Spatial

In Figure 22, visual representations of the spatial concepts are presented. The spatial concept is a vital part of the road congestion ontology because it provides a scalable description of the impact of congestion caused by an event. For example, an accident occurs at a point on a link, which is a road that is part of a road network, and it would impact a location.

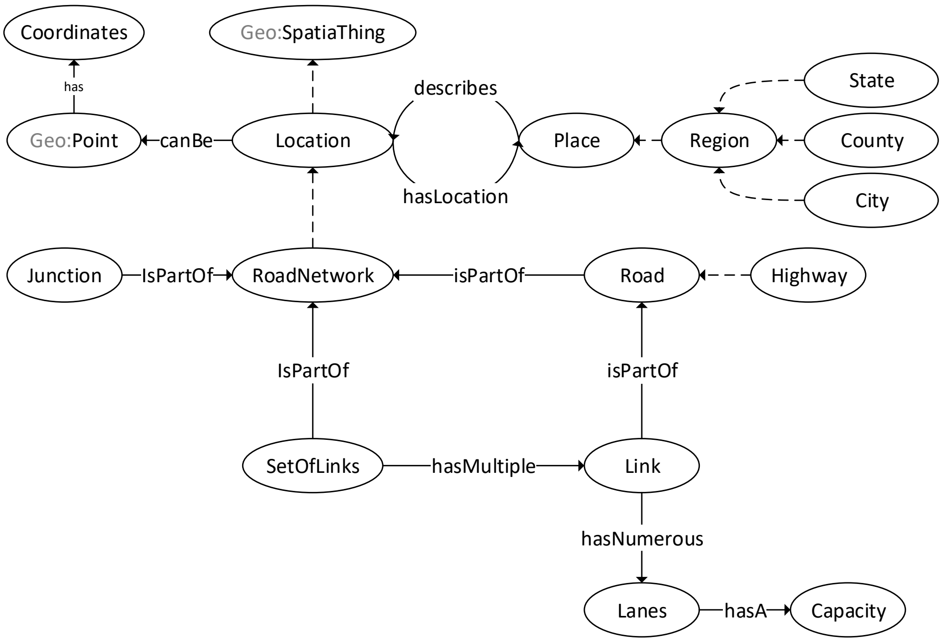


Figure 22: Ontology: Spatial

3.5.1.3 Direction

Figure 23 shows the representation of the concept of direction that has two main types, which are relative and absolute. Absolute is used to provide a precise position of a point or direction. For example, a point has coordinates that are made up of longitude, latitude, and altitude. Other coordinates that are used are degrees, minutes, and seconds. Other absolute directions would be northbound, southbound, clockwise, and anticlockwise. However, these are absolute but at the same time, they are vague. Relative direction is used to provide an extra layer of context that a user would be able to gain valuable knowledge. An example, of a relative direction, would be towards and away. These would be relative to a traveller, event, or attractor.

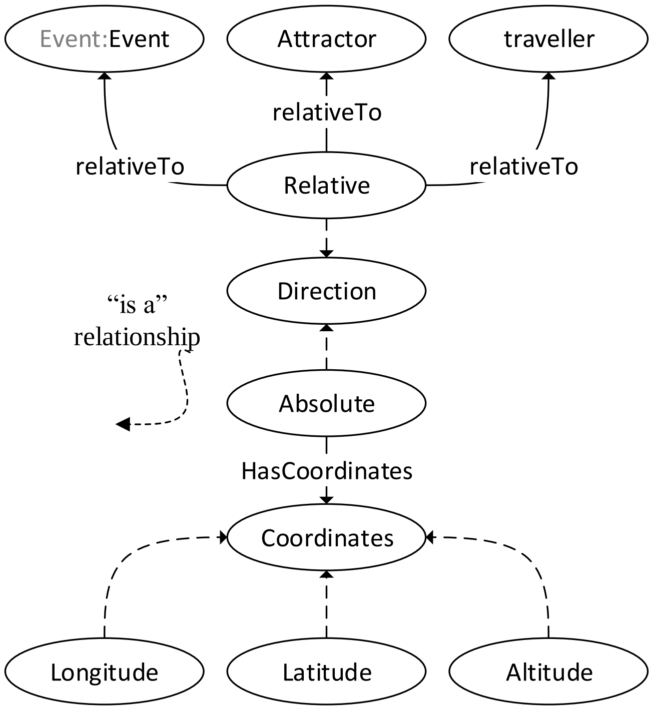


Figure 23: Ontology: Direction

3.5.1.4 Event

Event concepts are visualised in Figure 24. It is important to be able to identify the type of event (a football match) that has occurred and a specific instance of an event (this football match on this day at this time and place) because although, they all have an impact on the road network. Each event or instance of the event has its own unique patterns that can be used to identify what event is or has occurred and been able to predict the impact. For example, football matches, concerts, and planned roadworks are non-cyclical but are predictable. However, accidents, terrorist attacks, and unplanned roadworks caused by sinkholes are non-cyclical and unpredictable. Furthermore, congestion caused by morning AM and PM peak hours is cyclical and predictable.

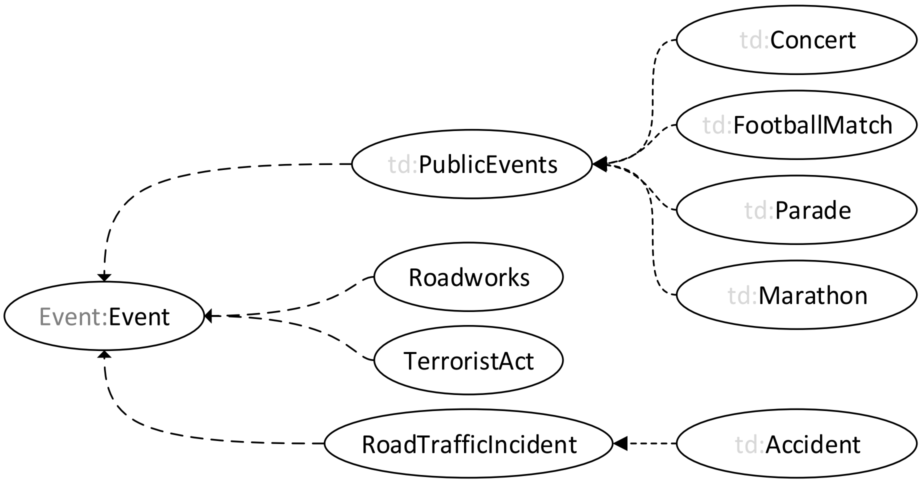


Figure 24: Ontology: Events

3.5.1.5 Combining all four ontologies

Figure 25 demonstrates how the five core ontologies come together to create the universal road congestion ontology which is the second component of the URCC model. For instance, a football event happens in a spatial context, such as a location on or a distance from a road and causes semi-recurrent congestion, which impacts the direction of the traffic towards or away from the event location, such as an attractor or landmark, depending on the state of the event i.e., pre-event, live-event, and post-event. Finally, the congestion caused by the road traffic event can be measured using several different dimensions, such as journey time, volume, and speed.

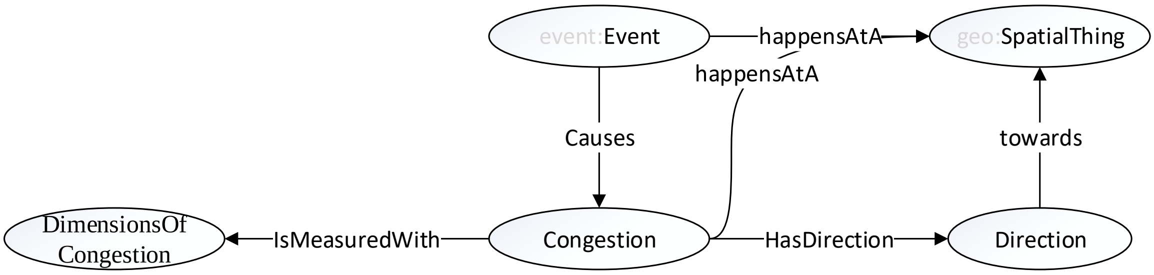


Figure 25: The relationship between the five core ontologies

3.6 Chapter conclusion

This chapter has explained the methodology for creating a URCC model using four analogies and a universal road congestion ontology which is made up of five core ontologies (Dimensions of congestion, events, congestion, direction. and spatial things). Furthermore, this chapter implemented the ontology following a modified methodology set out by (Noy and McGuinness, 2001). Therefore, the next step is to create a dataset capable of validating the universal road congestion ontology using data sources that have the ability to measure the dimensions of urban road congestion, such as speed and volume. To validate the universal road congestion ontology, several seemingly simple questions are proposed based on the gap within the literature.

Questions:

- What is congestion?
- What is recurrent congestion?
- What is semi-recurrent congestion?
- What is non-recurrent congestion?
- What is the cause of congestion?
- Where has congestion occurred?

Although, these questions seem simple, it is vital that when the broad term congestion is used, all stakeholders have the same clearly defined definition because this will assist in modelling urban road congestion and allow for the creation of a better prediction model. Furthermore, as mentioned in the literature, even two leading transport departments (DfT and DoT) define congestion in complete contrast to each other. Although, you could argue each one has a valid definition, it would be impossible to gain knowledge out of a conceptual model without formalising and providing an explicit definition.

The creation of the dataset and the case study to evaluate whether the ontology has the ability to formally and explicitly answer the above questions will be conducted in chapter four.

Chapter Four: Building a dataset from the ontology to validate the urban road congestion conceptual model

4.1 Introduction

This chapter describes the construction of a dataset which will be referred to as the Manchester Urban Congestion Data (MUCD) dataset. The dataset is comprised of several data elements, such as journey time, traffic volume, weather conditions, and event information collected from multiple sources. The MUCD dataset will be used to validate the universal road congestion ontology which is one of the main components of the Urban Road Congestion Conceptual (URCC) model from chapter 3.

The MUCD will attempt to address some of the challenges identified in other studies: using simulated datasets (Yuan and Cheu, 2003; Othman et al., 2015; Lee and Li, 2017; Rui et al., 2018), outdated datasets (Anbaroglu et al., 2014; Anbaroğlu et al., 2015), or datasets collected from expensive data sources (Cheng et al., 2012; Anbaroglu et al., 2014; Anbaroğlu et al., 2015).

Following on from the construction of the MUCD dataset, this chapter will present a case study compromising of several experiments to validate a number of types of congestion using specific events which have a consequence of congestion, such as road accidents, football matches and rush hour traffic. The case study is described in section 4.8.1.

4.2 Types of data sources used to model congestion

As mentioned in chapter 2 to create a “successful” Transport Management System (TMS) or Intelligent Transport Systems (ITS) is largely dependent on the quality of data sources. However, relevant data (i.e., associated with congestion) is not widely available for research and development purposes without several limitations. For example, accuracy of the data which can report incorrect values because of bad weather, cost to deploy new sensors, cost to access the data from currently deployed sensors, sensors get disabled or forgotten about as it sometimes can cost more to maintain them than to replace them). Furthermore, to have a reliable, dynamic, and robust TMS or ITS it is important to use multiple data sources and dimensions in conjunction with each other. The data used should be ethically collected and processed, easy to interpret, and be made widely available within a reasonable time to allow for better collaboration to help reduce the impact of urban road congestion.

Depending on what the TMS or ITS is trying to achieve, having different dimensions is vital, as it is critical for assessing the output from data sources in a meaningful manner that will help to identify traffic incidents that have the consequences of congestion. One of the benefits of using dimensional data is the dynamic aspects that allow TMS or ITS to work with a range of different types of data sources that measure the same dimension instead of being restricted to a single data source. Figure 26 shows the relationship between the dimensions and the data sources that were considered in this research. Where a dimension, such as a journey time can be captured and processed from multiple data sources, such as Bluetooth sensors, Global Positioning Systems (GPS), and road traffic cameras. where the quality of data and cost of deployment can vary. For example, Bluetooth sensors are cheap to deploy yet the data quality is poor compared to a road traffic camera which provides the best quality of data but can cost ten times as much to deploy (Hooke et al., 1996; Sen et al., 2011; Kurkcu and Ozbay, 2017).

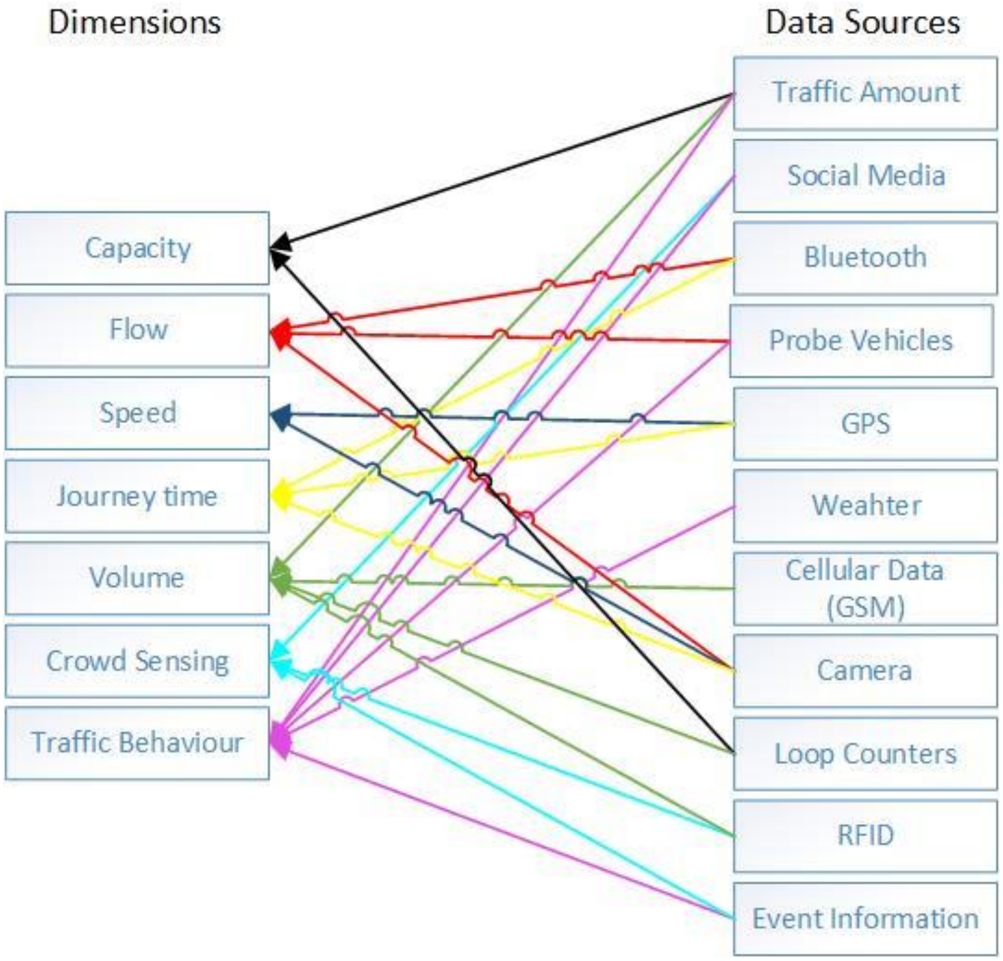


Figure 26: Relationship between dimensions and data sources

4.3 Neighbourhood network topology

A road network topology can range from a local network (two or three links connected to each other), a global network (city scale), and a neighbourhood network which this research is using and is larger than a local network but smaller than a global network. Figure 27 shows the final neighbourhood network topology that is in Manchester (UK) and will be used to test the feasibility and usefulness of the universal road congestion ontology. The neighbourhood network topology used was agreed with Transport for Greater Manchester (TfGM) who assisted in the scoping as domain experts.

The reason it was important to scope out the network before creating the dataset was to ensure the prerequisites provided by TfGM were met and to consider the limitations of the data. For example, one of the prerequisites was the incorporation of the A6 which connects Manchester city centre and Stockport. Moreover, another prerequisite was the need to incorporate event locations, such as the Etihad Stadium. Therefore, sensors around this location were investigated in the hopes of creating a well-distributed neighbourhood network. However, due to some limitations with the Bluetooth sensors, such as the data not being available at all sensors at the same time, data that had been captured was not always complete.

These limitations were due to some Bluetooth sensors being disabled and others being installed at a later period during the data collection phase for this project. Therefore, a total of 25 Bluetooth sensors were used to construct a 64 (32 links in both directions) link neighbourhood network. The blue lines on Figure 27 represent the Traffic Master routed networks in the area and the red links represent the Bluetooth network links being analysed.

For this research, a link is a route between two Bluetooth sensors and the topology consists of a total of 64 links in a two-directional network. Each link is allocated a unique letter combination, such as ‘a’ and depending on the direction a second letter will be allocated for instance upstream (au) and downstream (ad). Amongst the 64 total links being analysed, there is an approximate 68km of the road network with two main attractors, which are the Etihad Stadium (football grounds) and the O2 Apollo (concert hall). Each link has its own heterogeneous characteristics with regards to quantities of lanes, the number of junctions, speed limits, road class, and lengths that varies from 146m to 2,149m. In addition, the volume of traffic and observed journey times differ at spatial and temporal states.

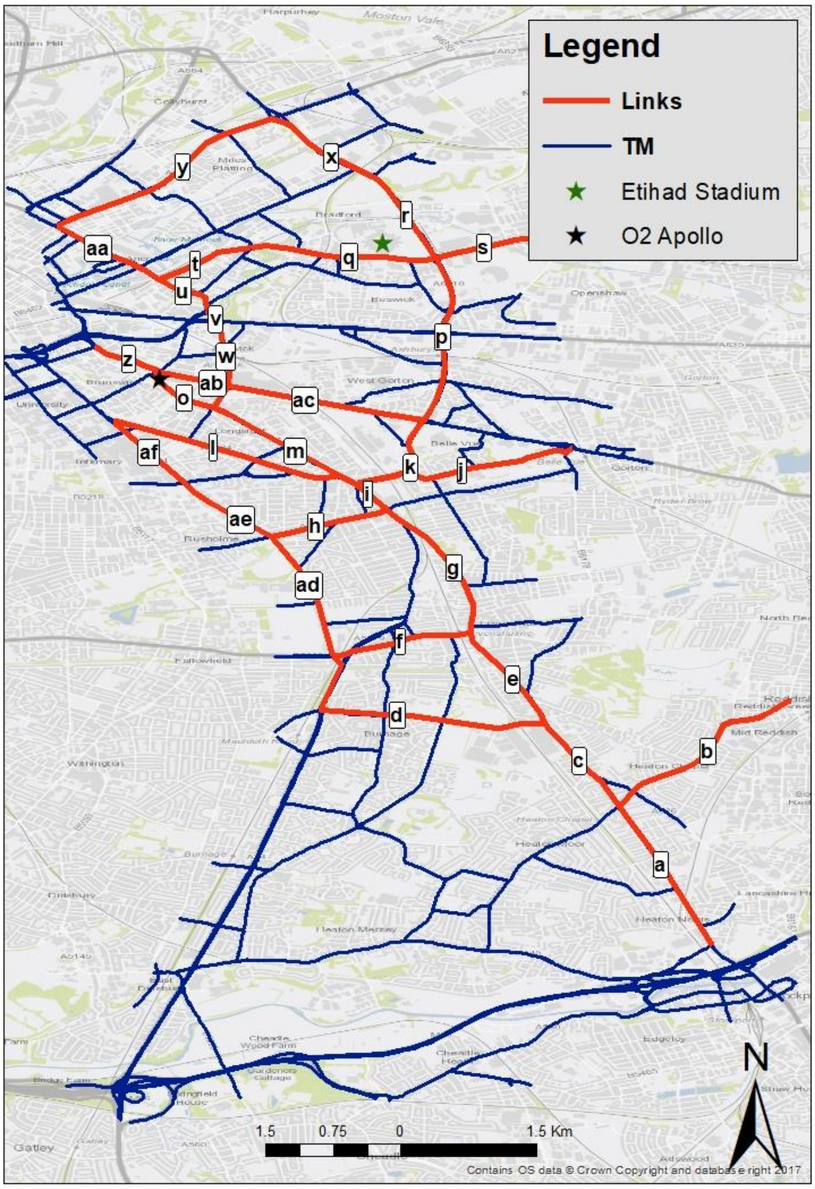


Figure 27: Manchester's neighbourhood network topology (Contains OS data © Crown copyright and database right (2017))

4.4 The creation of Manchester urban congestion data dataset

With the volatile increase of global data in the last 20 years, the term "big data" has become the new 'buzzword' within many disciplines. However, many academics and industry experts confuse 'big data' for 'large data' due to a lack of understanding of what is meant by big data. furthermore, due to the multifaceted nature of big data it has previously been claimed there is no clear definition or understanding for big data (Demchenko et al., 2013) and the more we begin to understand it, the more complicated it becomes, for instance, the Vs of big data, are constantly evolving from 3Vs (Jagadish, 2015), 4Vs (Philip Chen and Zhang, 2014), and 5Vs (Demchenko et al., 2013). The 5 Vs and their characteristics are as followed, Variety (heterogeneity), Veracity (inconsistency and incompleteness), Volume (scale), Velocity (timeliness), and Value (worthiness).

Big data becomes more multifaceted with the addition of geographical data, which accounts for 80% of daily data created in the last few years (Vopham et al., 2018), for instance, the geographical traffic data collected on a heterogeneous urban road network within many smart cities, such as Manchester, UK. Transport for Greater Manchester (TfGM) manages the road network within Manchester, UK and collects data continually from inductive loop counters, Bluetooth sensors, and more recently started exploring the use of Google API data. It was decided not to incorporate the Google API data into the MUCD Dataset due to cost. The API charges a fee for each observation (at 15-minute intervals) for each pre-defined link.

Table 6. Data sources, type of data, provider, range, and location

Data Source	Type of data	Provider	Range	Location
Bluetooth Sensor	Journey time	TfGM	2015-Present	Manchester, UK
Inductive Loop Counter	Volume	TfGM	2015-Present	Manchester, UK
Accident	Slight, Serve, Fatal accidents details	GOV.UK	2010-Present	UK
Etihad Stadium	Football matches, other big events, such as concerts.	Manchester City FC	2017	Manchester, UK
Concert Hall	Comedy shows, concerts	O2 Apollo	2017	Manchester, UK
Weather	Wind speed, humidity, temperature, weather description	Custom Weather	N/A	Worldwide
Bank holiday	School bank holiday details.	GOV.UK	2017-present	UK

The MUCD has 17376 records, each record consisting of 127 attributes and the data ranges from the start of January 2017 to the end of June 2017. The MUCD is primarily an unsupervised dataset, however, for classifying what is meant by congestion, the methodology used by TfGM was implemented to label the dataset. The method used is ‘the Red Amber and Green’ (RAG) method discussed in section 5.4.2.

The MUCD dataset is data collected from five different data providers. TfGM provided access to journey time and volume data, and they are the data owners, GOV.UK provided road traffic accident data and school bank holiday information, Manchester City FC and O2 Apollo provided event information for football matches and concerts respectively, and weather data was collected from Custom weather (www.customweather.com). The data was then stored in two places, the first is a ‘master’ file (CSV) and the second is a SQL Server database which was created to allow the visualisation tool, discussed in chapter 5 to easily access the data.

4.5 Data cleaning and pre-processing

The steps to collecting the raw data are as follows:

1) Gaining the right permissions from TfGM to be allowed to access their data system, known as C2 that contains the relevant data.
2) Extraction of the raw journey time and traffic volume data from the C2 data system. This required manual parameters to be set before each sub-dataset can be populated and downloaded from the C2 system. For example:
    a. Once logged into C2, parameters, such as (Bluetooth) node A and B are required to be identified and set, the time frame of data observe had to be set, the time interval for all observations is required to be set and then the results need to be extracted in a .CSV format file. The selection of nodes was based on the Neighbourhood Network Topology described in section 4.3. Moreover, the selection of the nodes within the C2 system was challenging because not all sensors were active at the same time and there was the additional need to find Bluetooth sensors which overlapped with inductive loop counters.
    b. To cleanse this data all NULL values were replaced with '0' and each .CSV contained a single month for a single link in a single direction for a single data source. Therefore, approximately 888 .CSV files needed to be merged into a single .CSV and once the final dataset containing the journey time and traffic volume data was created, it was loaded into a database.
3) Event information including the dates the events occurred were collected directly from the o2 Apollo, Manchester for music concerts and comedy shows and the Etihad stadium for football fixtures between January 2017 and June 2017.
    a. To cleanse this data, both files (.CSV) provided by the Apollo and the Etihad with regards to events were imported into a database.
4) Accident data was extracted in .CSV format from the Government website which is populated from the stats 19 reports conducted from the police departments.
    a. To cleanse this data, all unnecessary data was removed leaving an estimated start time, end time, severity, and longitude and latitude for plotting the location of the accident. The data was then imported into the database.
5) School term start and end times were extracted from the Government open-source website for school around the area of the Neighbourhood Network Topology.
    a. A list of start and end times were loaded into the database.

Finally, all the data was merged into a single master file(.CSV) and then imported into a SQL Server database table using the date and time value to join sources together, providing an quarter-hourly picture of the Neighbourhood Network Topology (described in section 4.3) performance.

There were restrictions and challenges in relation to using the TfGM system C2. For example, taking into consideration the requirements of TfGM, such as using particular roads and attractors and then trying to find sensors in this area which were all active at the same time. The major challenge was trying to find links that had available data for both the Bluetooth sensors and the inductive loop counters between two set time periods January 2017 and June 2017.

4.6 Data considerations and observations

The inductive loop counters and Bluetooth sensors are the primary data sources used within this research along with several other data sources. This data was collected at source; therefore, it is deemed to be the ground truth data and is fed into the RAG method (discussed in section 5.3) where a human being (from TFGM) classified the data (major congestion, slight congestion, and non-congestion) as domain experts.

Despite, the MUCD dataset having several typical big data issues, such as noise, data sparsity and missing values, the MUCD was still successful in validating the conceptual model for three distinct case studies. Experiment One: Bathtub and leaf modelling recurrent congestion (section 4.8.1.2). Experiment two: Raindrop modelling semi-recurrent congestion (section 4.8.1.3). Experiment three: Raindrop modelling non-recurrent congestion dependent on the severity (section 4.8.1.4). The observations and problems associated with the creation of MUCD can be summarised as:

- There is a lack of consistent distance between the Bluetooth sensors causing each link to have its own heterogeneous characteristics, such as lane quantity, speed limits, road class, number of junctions, and length (which varies from 146m to 2,149m).
- On the urban road network, there is a limited amount of inductive loop counters, which restricts the ability to calculate a volume count for each link.
- Due to the limited number of sensors around Manchester and their position, it was impossible to create a complete network (many of the smaller roads and links are not included). For the purpose of this research, a neighbourhood network has been created and this was discussed in section 4.3
- The data quality of Bluetooth sensors is poor. For example, the capture rates during the night-time or periods where no vehicles pass Bluetooth sensors, the sensors will provide an incorrect average journey time when being observed.
- In bad weather, the sensors which use a mobile network to transmit the data to a central location, can fail and cause the dataset to have missing data.
- The Bluetooth sensor data cannot distinguish the difference characteristics between a bus with 30 people on it or a car with just one person, which causes the level of congestion to be overestimated on several occasions. Therefore, TfGM use a 25% outlier reduction to get a fairer average journey time.

However, despite these challenges, a real word dataset was created. a full description can be found in section 4.7.

4.7 Final MUCD dataset description

After cleansing, the final MUCD dataset will be used throughout the remainder of the thesis. Table 7 lists all attributes that will be used, their sources and data types. The attribute ‘Links’ ‘X’ represents a unique link between two Bluetooth sensors, or the link an inductive loop counter (also known as Automatic Traffic Counter (ATC)) is located on. In Table 7, NB is Northbound, SB is Southbound, NS is Nearside, and OS is Offside and the total number of records is 17376 rows.

Table 7: Attribute description of MUCD Dataset

Attribute			Value	Source	Data Types
Date			Date of record observation	TfGM	Date
Day			Day of record observation	TfGM	String
Time			Time of observation	TfGM	Time
Links	X	Upstream	Average Journey Time between two Bluetooth sensors on each link heading upstream and downstream	TfGM	Numerical
		Downstream			
ATCs	X	NB NS	Traffic volume count for each road link where an ATC is present. Counting individual lanes separately and a sum of both northbound and southbound	TfGM	Numerical
		NB OS			
		SB NS			
		SB OS			
		NB Total			
		SB Total			
Accident		Start date/time	Did an injury accident occur	GOV.UK	Date Time
		End date/time			
		Severity			Categorical
Events	Football	Date	Date of the Football matches, other big events, such as concerts.	Manchester City FC	Date

Attribute			Value	Source	Data Types
		Start time	Starting time of Football matches, other big events, such as concerts.		Time
		End time	Ending time of Football matches, other big events, such as concerts		Time
	Concerts	Date	Date of the musical concert or comedy shows	O2 Apollo	Date
		Start time	Starting time of musical concert or comedy shows		Time
		End time	Ending time of musical concert or comedy shows		
School Term Times	Start date/time		School terms starting times	GOV.UK	Date Time
	End date/time		School terms finishing times		
Weather	Temp(C)		The temperature in degrees Celsius	Custom Weather	Numerical
	Weather status		Recorded weather condition	Custom Weather	Categorical
	Wind(mph)		Wind speed in miles per hour	Custom Weather	Numerical
	Humidity		Humidity percentage	Custom Weather	Numerical
	Barometer		Barometer record at observation	Custom Weather	Numerical
	Visibility(km)		Level of visibility	Custom Weather	Numerical

4.8 Case Studies: validation of the universal ontology of road congestion

Following the creation of a novel universal road congestion ontology in chapter three, real-world data was collected to validate the ontology. The real-world data used in this study, known as the Manchester Urban Congestion Data (MUCD) dataset (and its associated challenges) is different compared to other studies, which tend to use simulated datasets (Yuan and Cheu, 2003; Othman et al., 2015; Lee and Li, 2017; Rui et al., 2018), longstanding datasets(Anbaroglu et al., 2014; Anbaroğlu et al., 2015), or datasets collected from expensive data sources (Cheng et al., 2012; Anbaroglu et al., 2014; Anbaroğlu et al., 2015), which could not solve the practical issues (i.e. noise, data sparsity and missing values) associated with real-world big data analytics for TMS or ITS.

This case study has been developed to address the research question RQ1: Is it possible to provide a clear conceptualisation of urban road traffic congestion using an ontological model?

This will be achieved by validating the universal ontology of road congestion through answering the several questions proposed in chapter three.

Questions:

- What is congestion?
- What is recurrent congestion?
- What is semi-recurrent congestion?
- What is non-recurrent congestion?
- What is the cause of congestion?
- Where has congestion occurred?

This section will look at a series of experiments which are designed to assess the feasibility and usefulness, of the ontology and will investigate the three types of congestion as defined in Chapter three. Section 4.8.1.2 is the first of these experiments and will focus on AM peak rush hour which has the consequence of recurrent congestion. Section 4.8.1.3 is the second experiment that will focus on football matches, which has the consequence of semi-recurrent congestion. Section 4.8.1.4 is the third experiment that will focus on a fatal road accident, which has the consequence of non-recurrent congestion.

For all three case studies, individual links from around Greater Manchester, UK were selected from the neighbourhood network topology defined in Figure 27, and the journey times will be compared to the expected journey times (defined in section 4.8.1). The MUCD dataset as described in section 4.4 will be used to conduct these experiments.

4.8.1 Experimental methodology: Expected journey time detection

To assist in the visualisation of these experiments, a visualisation toolkit which is known as Transport Incident Manager (TIM) was developed as part of this research. A full description of Tim, including the design and justification can be found in chapter five.

To calculate the expected journey time for each link within the data from the MUCD dataset that was developed throughout chapter four and the final dataset presented in section 4.7 was processed to create a 'typical link journey time' for each link. The 'typical link journey time' is defined by aggregating 6 months' of data by link and Time-of-Day-and-Day-of-Week (TOD TOW), which is broken down into 15-minute intervals and then multiplied or divided against the congestion factor to achieve the expected journey time parameters. The congestion factor (c) is derived by using the method outlined in the "Congestion Reduction in Europe: Advancing Transport Efficiency" conducted by (Jones, 2016) and funded by the European Union Horizon 2020 program. The methodology for the congestion factor will be discussed in section 4.8.1.1.

Let $\mathrm{JT_{obs}}\ (t, l)$ be the representative of the observed journey time at link l with TOD DOW, t, $\mathrm{JT_{his}}\ (t, l, w)$ be the representative of the historical data at the same link l with TOD DOW, t, however on a different week w. The typical journey time is represented by $\mathrm{JT_{typ}}\ (t, l)$ and is calculated by aggregating $\mathrm{JT_{his}}(t, l, w_1), \mathrm{JT_{his}}(t, l, w_2), \ldots, \mathrm{JT_{his}}\ (t, l, w_n)$ where $w_1, w_2, \ldots, w_n$ represent each week in the MUCD and then multiply and divide these values by c to create the upper and lower boundaries.

Equation 3 and Equation 4 shows how to calculate the typical link journey time for the upper and lower boundaries.

$$\mathrm{E(upper)} = \mu_{(X)} * c$$

Equation 3: Upper boundary

$$\mathrm{E(lower)} = \mu_{(X)}/c$$

Equation 4: Lower boundary

$\mu_{(X)}$ is the mean (typical) journey time for link x, c is the congestion factor, which is a real value multiplied or divided against the typical link journey time. In these case studies, the congestion factor is 1.7 and was calculated using the following methodology discussed in section 4.8.1.1. Equation 5 below shows whether an observation is expected {0} or not expected {1}.

$$\mathrm{f(Expected)} = \begin{cases} 1, & \mathrm{JT_{obs}}\ (t, l) > \ \mathrm{JT_{typ}}\ (t, l)\ * c \\ 0, & \mathrm{JT_{obs}}\ (t, l) \geq \mathrm{JT_{typ}}\ (t, l)\ /c\ \leq\ \mathrm{JT_{typ}}\ (t, l)\ * c \\ 1, & \mathrm{JT_{obs}}\ (t, l) < \mathrm{JT_{typ}}\ (t, l)\ /c \end{cases}$$

Equation 5: Function for detecting congestion

Using the combination of the mean journey time and the congestion factor, it is possible to create a pattern of expected or not expected journey time. Equation 5 has been implemented in Algorithm 1.

Algorithm 1
An indicator of worse than expected journey time.

Variables: i the set of observations, T the set of time periods, L the set of links, x_i: the observation, $\bar{x}_{l,t}$ is the average observed value over the time periods, Expected: An array of outcomes. Congestion factor: 1.7.

```
for $l \in L$ do
    for $i \in I$ do
        $\text{Expected}_i \leftarrow$ false
    if $x_i \geq \bar{x}_{l,t} / \varphi^{-1} * \text{Congestion factor}$ then
        $\text{Expected}_i \leftarrow$ true
    end if
    if $x_i \leq -\bar{x}_{l,t} * \varphi^{-1} * \text{Congestion factor}$ then
        $\text{Expected}_i \leftarrow$ true
    end if
  end for
end for
return Expected
```

4.8.1.1 Methodology for the Congestion Factor

This section describes the methodology for determining the congestion factor. The four steps are as follows:

1) For each row within the MUCD dataset, which is a total of 17376 rows and 64 links. The 95th percentile is calculated for each link journey time. The link journey time for link a at time interval t is denoted as $y_a^{95}(t)$.
2) The congestion factor needs to be calculated for link a at time interval t and is denoted as$c_a^{95}(t) = y_a^{95}(t)/\bar{y}_a(t)$.
3) For each link and time interval, repeat steps 1 and 2. $a \in A$ and $t \in [1, 2, \ldots, T]$, where A denotes the set of links and T denotes the total number to time intervals.
4) Once the 95th percentile has been calculated for all 64-links and all 96-time (15 minutes) intervals, the median of $\sum_{n=1}^{\substack{a=64\\t=96}} c_{a_n}^{95}(t_n)$ is considered the congestion value.

The concept of a congestion factor has been implemented in a couple of studies related to non-recurrent congestion in London (Anbaroglu et al., 2014; Anbaroğlu et al., 2015). These studies have calculated the congestion factor to be 1.2 and 1.4, which are both lower than the 1.7 used within this research. This is because unlike previous studies which only takes into consideration data from within peak times (7am to 7pm), this research has used 15-minute intervals from midnight to midnight (a total of 24 hours) for a total of six months on a network of 64 links. Figure 28 presents a boxplot of the 95th percentile for all 64-links at all 96-time intervals. Then the median of the output (red line) is the congestion factor.

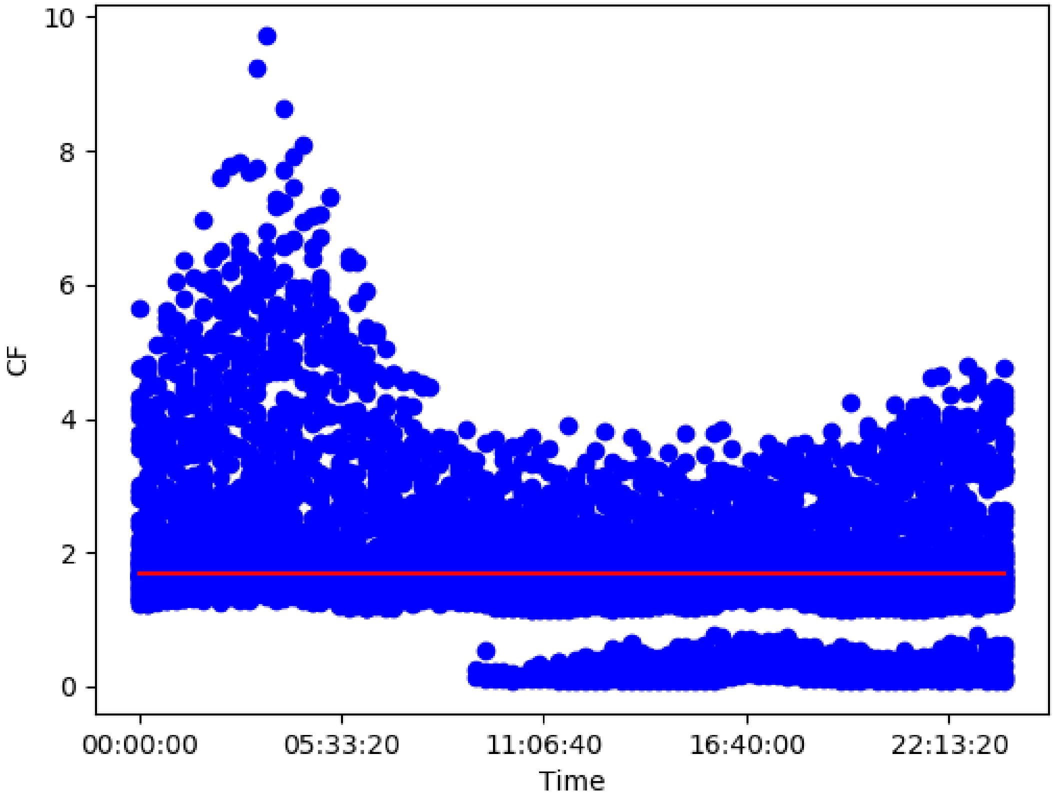


Figure 28: Determining the congestion factor for Manchester case studies topology

4.8.1.2 Experiment One: Bathtub and leaf modelling recurrent congestion

A case study has been chosen to provide a comprehensive understanding and to validate how a universal conceptual model can be used to support stakeholders with regards to recurrent congestion. The bathtub model (in section 3.4.1.1) explains how recurrent congestion is impacted by the inflow and outflow of vehicles causing the network to reach capacity, reducing speeds, and increasing journey times. Therefore, this exploration examines three links toward (inflow) and away (outflow) from the city centre.

This will provide the conceptual model with the ability to capture the semantics of recurrent congestion caused by an event, such as rush hour. The three links are from an arterial route into Manchester similar to the leaf model (in section 3.4.1.2). Figure 29 and Figure 30 were produced using data (from the MUCD dataset) from Tuesdays, and Wednesdays at four different time periods for a total of six months. The x-axis is the time of day, the y-axis is journey time for each link in seconds, the values at the top of the graph and the red line in the centre of each boxplot is the median journey time, and the star is the mean journey time.

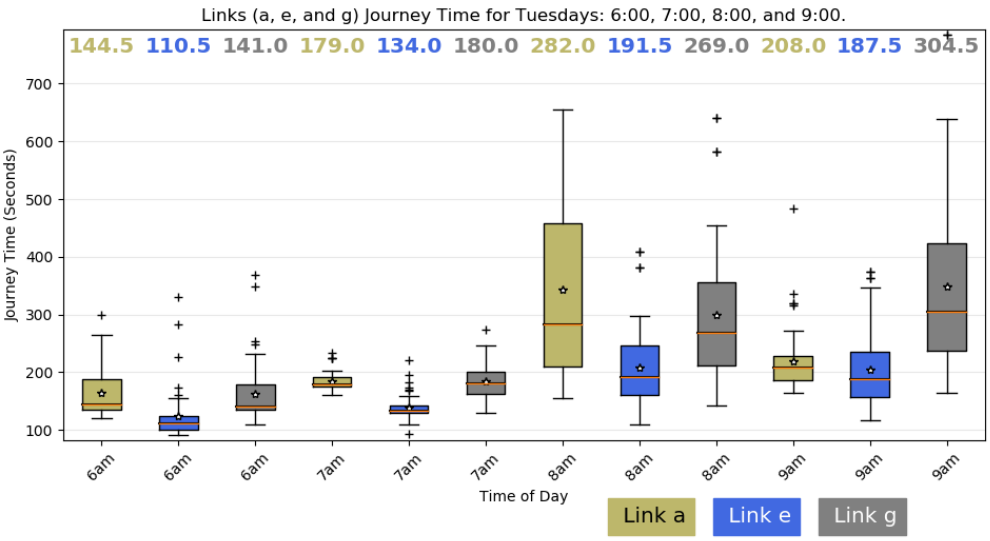


Figure 29: Boxplot of journey time for three links at three different periods (Tuesday)

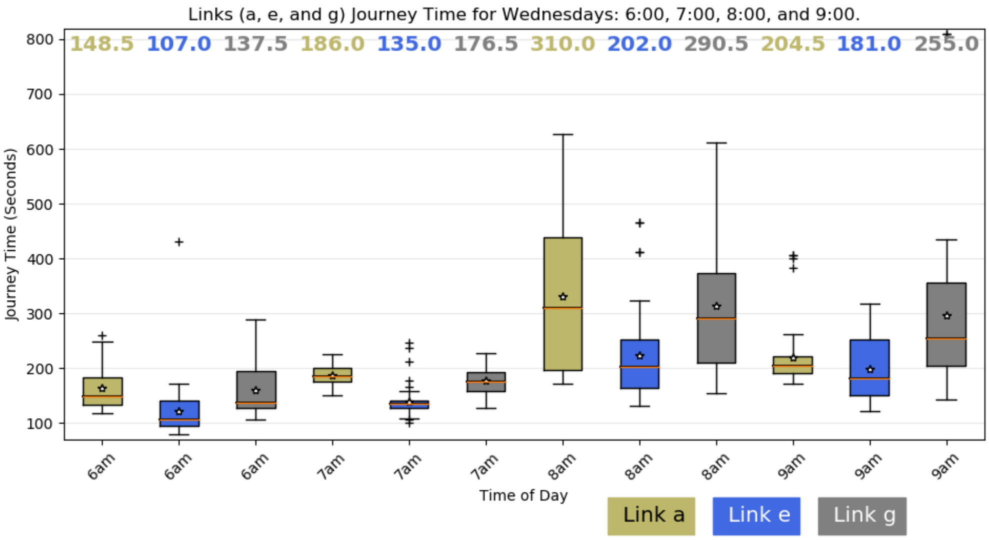


Figure 30: Boxplot of journey time for three links at three different periods (Wednesday)

Both Figure 29 and Figure 30 demonstrate there is a clear and typical behaviour of journey time in the morning rush hour. 6am provides the lowest journey times with tightest clusters and at 7am there is a slightly increased journey time, however, the clusters remain tight. Moreover, at 8am the journey time is at the highest level in the four-hour timeframe and the clusters become extremely sparse. Finally, 9am shows the journey time reduces but the clusters remain sparse.

This information could be used to predict the optimal time for travellers to avoid peak-time congestion and demonstrates what time the inflow of ‘water into the bathtub’ reaches the critical level across specified links. Furthermore, Figure 31 demonstrates that the morning rush hour occurs slightly later on the link nearest to the city centre (link ‘g’) than the link closest to the highway (link ‘a’). Although the observations show the peak occurs at a slightly different time, it is still noticeable that the journey time starts to increase for both around 7:00 am. This is caused by high volumes of vehicles trying to enter the city centre within a short period of time.

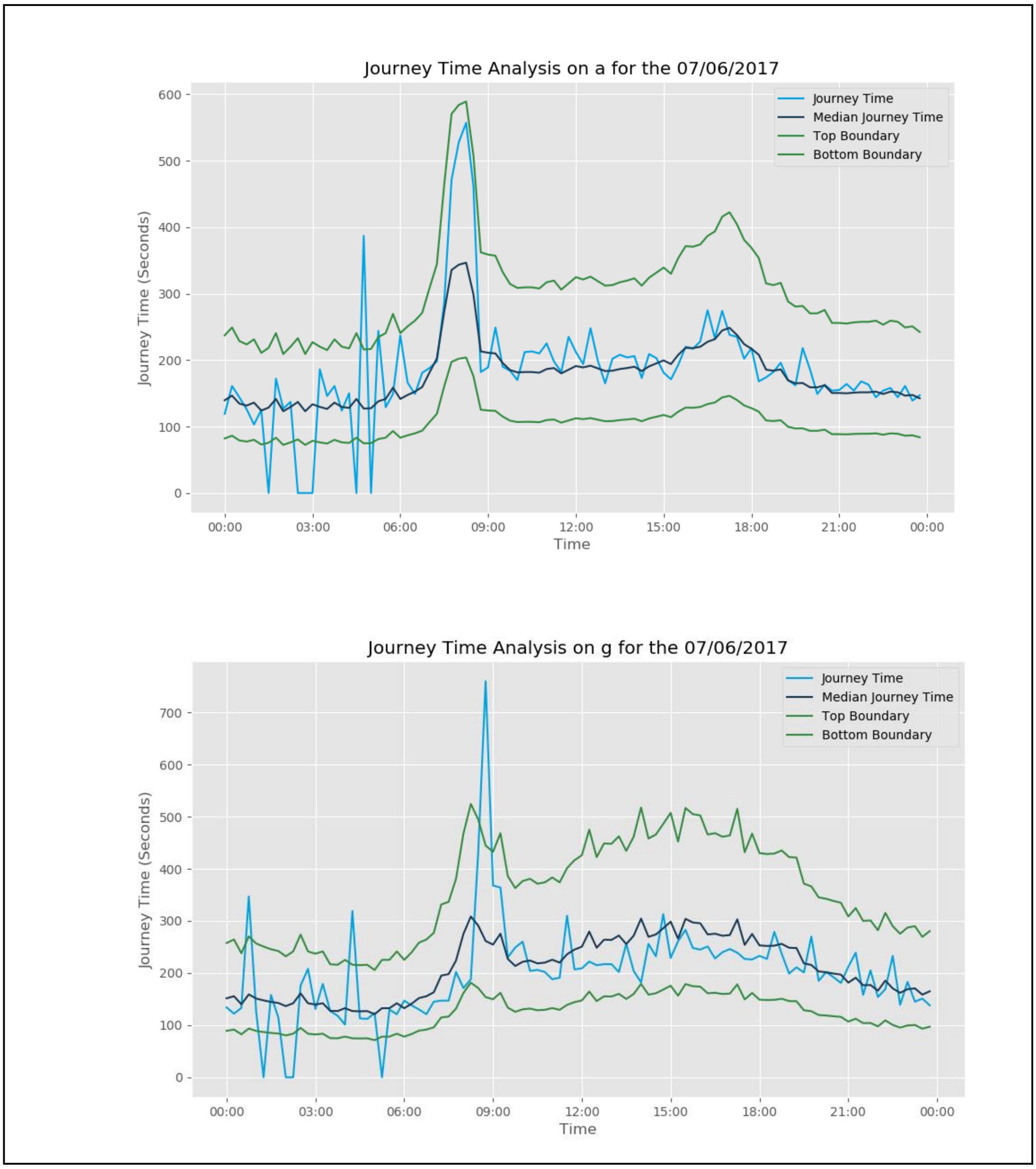


Figure 31: Journey time on 07/06/2017. Link ‘a’ (top) and ‘g’ (bottom). Observed journey time (blue), mean journey time (black), and expected journey time boundaries (green)

Using the semantics captured in Figure 29, Figure 30, and Figure 31 it is possible to validate and display a visual representation of recurrent congestion in an ontology. See Figure 32.

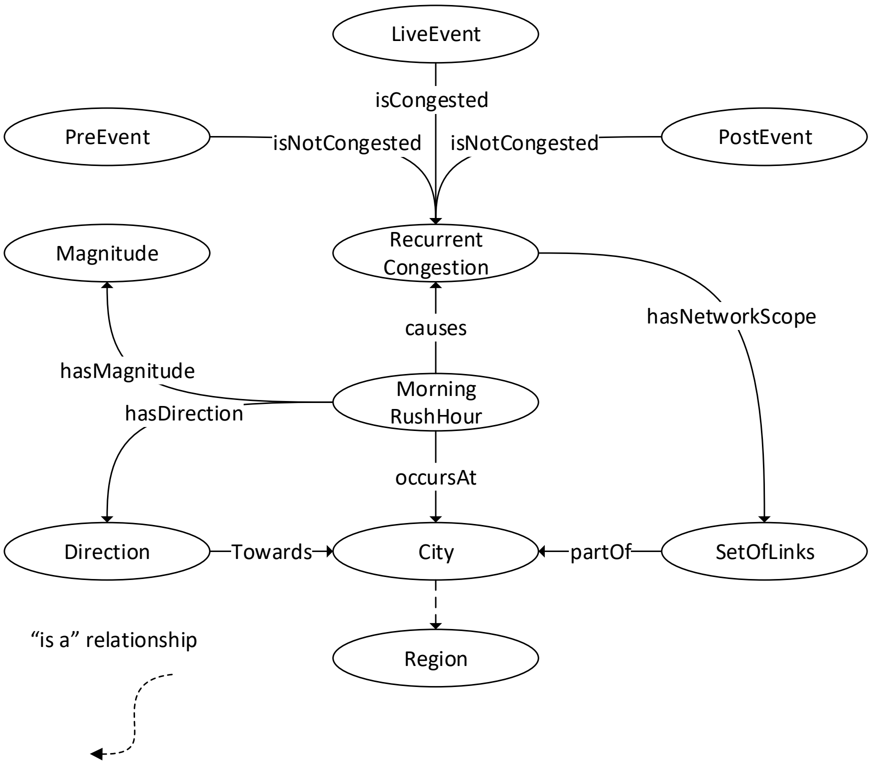


Figure 32: The semantics of a recurrent congestion impact on the road network

When an event, such as rush hour occurs, it causes recurrent congestion, which is predictable and cyclical because it always happens on a weekday between 7am and 9am. Rush hour causes recurrent congestion on a city scale and impacts primarily the traffic going in the direction towards the city centre. During pre and post recurrent congestion, the journey time is at an expected level. However, during the live event, the magnitude triples the expected journey time causing a worse than expected journey time. Moreover, core objects from all five ontologies are present in the construction of Figure 32.

4.8.1.3 Experiment two: Raindrop modelling semi-recurrent congestion

To demonstrate the concept of a raindrop analogy (in section 3.4.1.3) a case study has been chosen. The case study will provide a comprehensive understanding whilst validating the ontology included in the universal conceptual model that can be used to demonstrate the necessity of the newly coined semi-recurrent congestion. The conceptual model captures the semantics of semi-recurrent congestion caused by an event, such as but not limited to, a football match.

For this case study, specific data is analysed from the MUCD dataset that relates to a football match, located at the Etihad Stadium in Manchester, UK which took place on Saturday the 13th of May 2017 with an expected kick-off at 12:30 and full-time at 14:00. A link near the attractor known as the Etihad Stadium has been selected to analyse the journey time and is shown in Figure 33. The expected journey time boundaries were calculated using Algorithm 1.

Algorithm 1 considers the mean journey time on every individual link based on Time-of-Day-and-Day-of-Week (TOD TOW) and then is multiplied by the congestion factor.

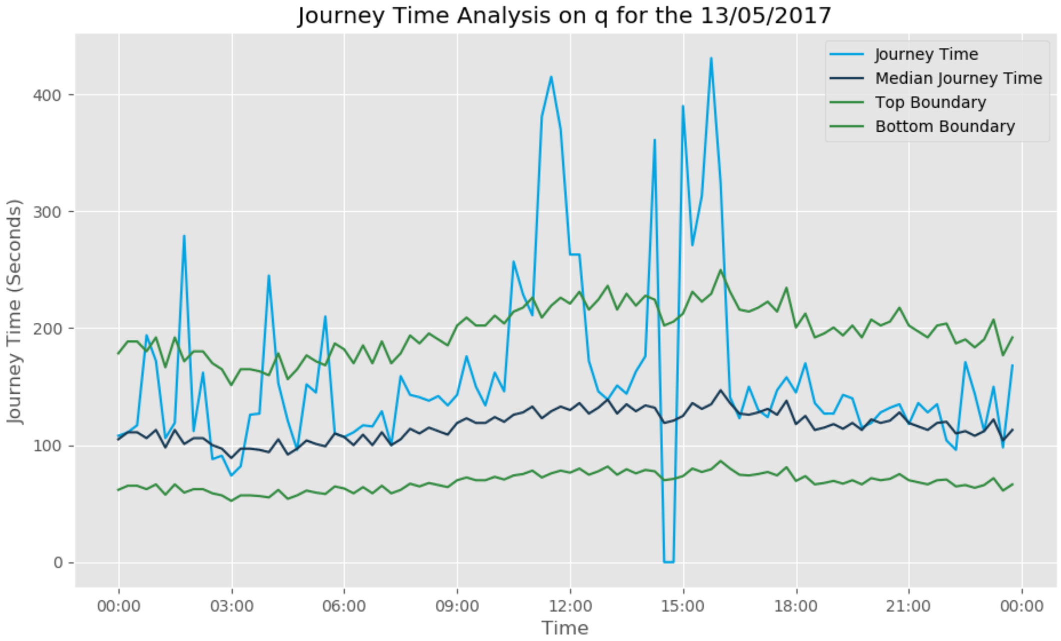


Figure 33: Journey time on 13/05/2017 compared to the expected journey time

One characteristic that separates semi-recurrent congestion from non-recurrent congestion, is semi-recurrent congestion is caused by an event that occurs at an attractor, such as a landmark e.g., a football stadium. Examining Figure 33 shows several other unique characteristics of a football match, such as pre-event, kick-off, live-event, full-time, and post-event. Pre-event is to the left of the 12:30 kick-off and post-event is to the right of the 14:00 full-time where both journey times excessively exceeds the boundaries of an expected journey time. The Live event is between both kick-off and full-time and it is noticeable that the journey time returns to an expected journey time.

Again, these characteristics are different compared to non-recurrent congestion, where pre-event and the post-event journey time is typically an expected journey time, and the live event exceeds the expected journey time depending on the severity. Being able to identify and model these characteristics support the validation of the road congestion ontology by demonstrating the accuracy of the semantics presented in Figure 34 and will help to predict semi-recurrent congestion and the events that cause it.

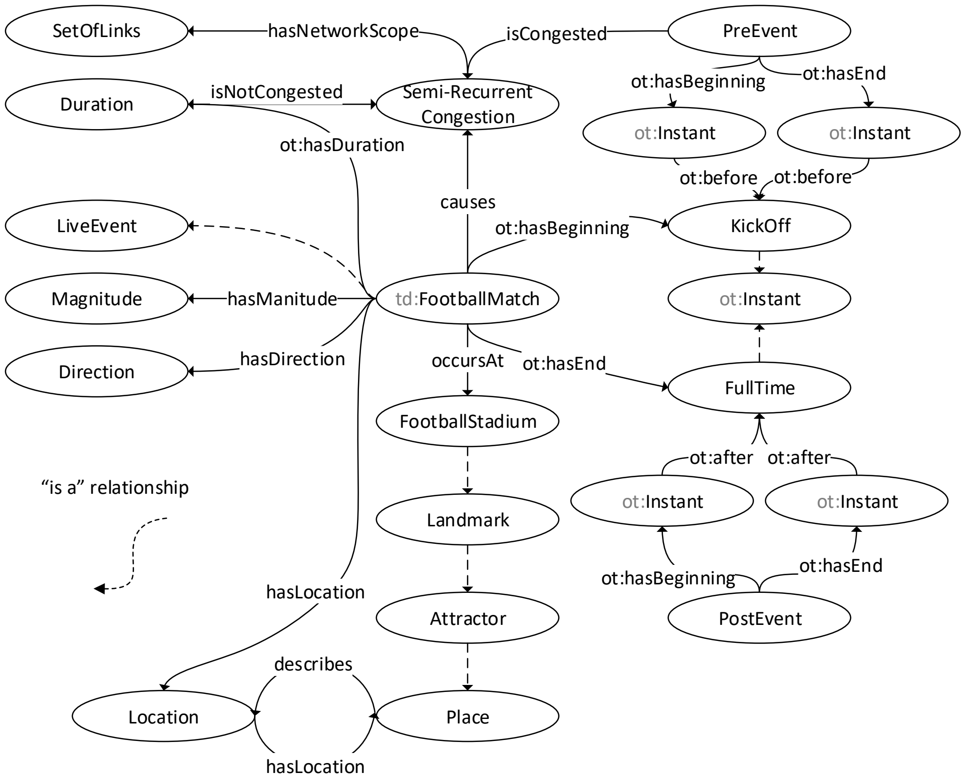


Figure 34: The semantics of a semi-recurrent congestion impact on the road network

Pre-event and post-event have a period of two to three hours of worse than expected journey time. This is due to an increased quantity of stakeholders travelling to and from the same attractor. During the live event, the journey time returns to a state of expected journey time. Post-event journey time is typically worse than pre-event, due to the high volume of stakeholders attempting to leave and gain access to the road network all at the same time. Figure 34 and Figure 35 presents a visual representation of the semantics of a football match and the impact of road congestion using the dimension journey time as a measurement of performance which confirms the characteristics previously discussed in this section to be accurate. Moreover, both Figure 34 and Figure 35 contain core objects from the five core ontologies.

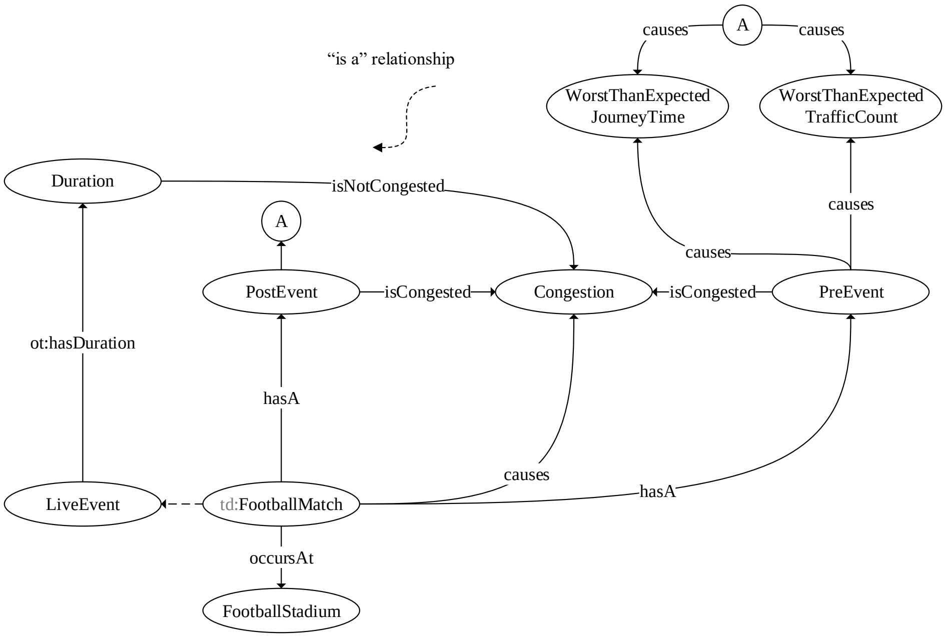


Figure 35: The relationship between a football match, journey time, and traffic volume

4.8.1.4 Experiment three: Raindrop modelling non-recurrent congestion dependent on the severity

The final case study further explores the concept of a raindrop but showed the difference between the large and small raindrop (discussed in section 3.4.1.3), which signifies a 'slight' and 'fatal' road accident. For this case study, the 7th of February 2017 was selected because data analysis indicated two separate road accidents had occurred on the same link. The first accident was classified as 'slight' and the second as 'fatal'. Figure 36 and Figure 37 shows two graphs with the journey time, expected journey time boundaries, and both traffic accidents plotted for the same link in both directions.

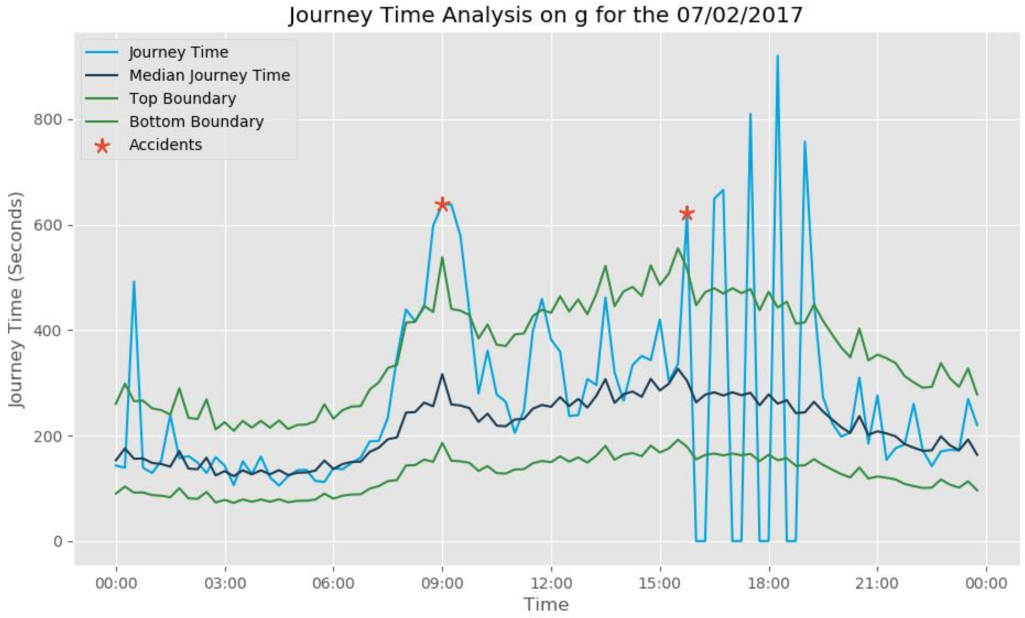


Figure 36: Journey time on 07/02/2017 compared to the expected journey time (towards the City Centre)

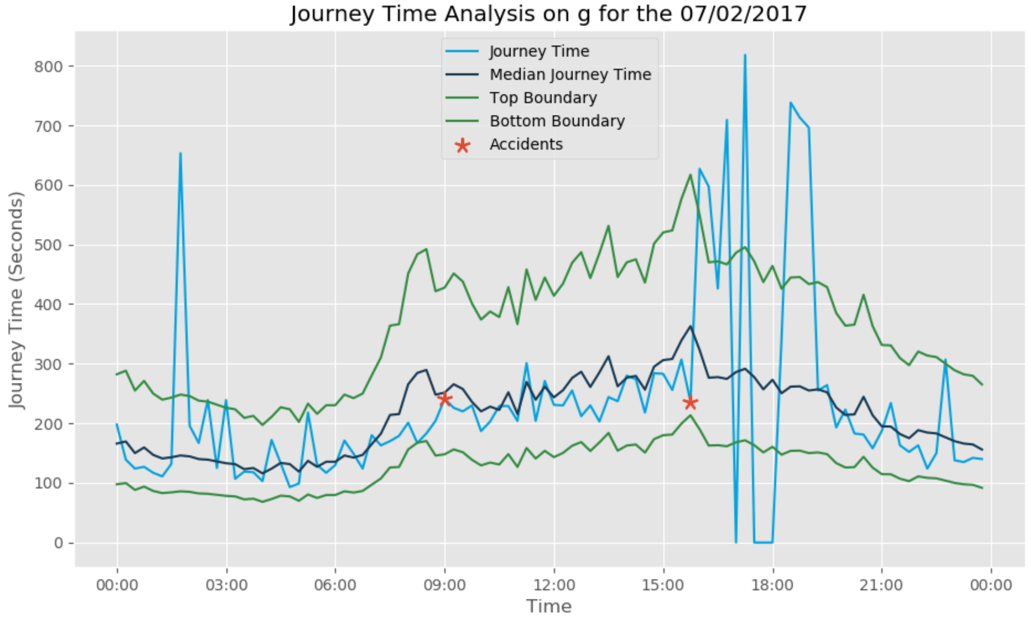


Figure 37: Journey time on 07/02/2017 compared to the expected journey time (away from the City Centre)

Slight road accident analysis

In Figure 36 and Figure 37, the slight road accident (*) is on the left which happened during the AM peak rush hour at around 09:00. The journey time is noticeably worse than expected for traffic heading towards the city centre. However, heading away from the city centre, the journey time is expected and is not impacted by the slight road accident on the opposite side of the road.

Fatal road accident analysis

In Figure 36 and Figure 37, the fatal road accident (*) is on the right which happened around 15:45. The first noticeable difference between the slight and fatal road accident is, the aftermath of the fatal has an impact on both directions of traffic, causing journey time to either be excessively high or not recorded. This behaviour lasts for 4 hours, and diffuses outwards, impacting neighbouring links. Similar to how a raindrop would cause a ripple effect outward.

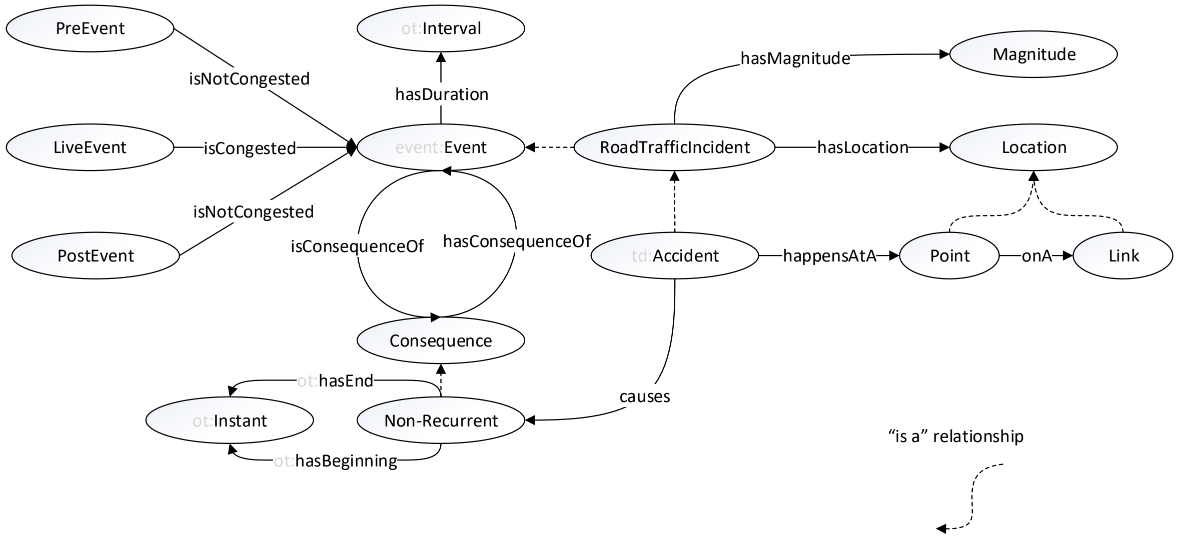


Figure 38: The semantics of a non-recurrent congestion impact on the road network

Figure 38 is the visual diagram of the semantics of non-recurrent congestion caused by a road traffic accident (however, the more serious the accident, the bigger the magnitude will be) and has been validated using the information collected in Figure 36 and Figure 37. A road accident happens at a point on a link, which is a location and where a road traffic incident occurs. A road traffic incident, such as an accident causes non-recurrent congestion because it is non-predictable and non-cyclical. Non-recurrent congestion is the consequence of the live event, which is congested and has a duration that varies on the severity and magnitude of the road traffic incident. Moreover, Figure 38 contains core objects from the five core ontologies.

4.9 Chapter conclusion

This chapter has introduced a real-world big data dataset known as the MUCD, which has many relevant characteristics for identifying events that have the consequence of congestion, such as spatial (location) and temporal characteristics (TOD DOW). When evaluating the five Vs of big data, the MUCD meets all the requirements. Variety (heterogeneity) as the data is extracted from several different sources and provides different types of data. Veracity (inconsistency and incompleteness) as it is real-world data, it is not perfect and contains missing data or incorrect values caused by issues with the sensors. Volume (scale) of the dataset is not the largest data set known to man, however, it does contain 17376 records and 127 attributes. Velocity (timeliness) this data can be collected in quasi-real-time and would need to be analysed in quasi-real-time to gain any Value (worthiness).

This dataset will be used throughout the thesis to validate the universal ontology of road congestion, to experiment with the idea of using supervised learning to gain knowledge and qualitative information from a quantitative dataset. This chapter has demonstrated it is possible to provide a clear conceptualisation of road traffic congestion using both analogical and ontological methods to develop a URCC model, providing vital knowledge to different types of stakeholders.

The universal URCC model uses an analogical approach to provide stakeholders with an unsophisticated explanation of congestion that even a layperson would be able to understand. Additionally, the URCC model provides a more advanced understanding of road traffic congestion by introducing an explicit conceptualisation (ontological) that can be used to capture the semantics of all three types of road congestion. Furthermore, this chapter provides a consistent definition of the many objects that are used to create the overall ontology and explains their relationships with each other, e.g., hierarchy (is-a) and object properties (has-a), allowing for a better understanding of a typical pattern for the many different road traffic events.

The conceptual model was validated using the dataset created within this chapter and a case study which was also introduced in chapter four which demonstrated it was possible to provide a consistent answer to questions that have previously been vague or hard to answer (Abberley et al., 2017; Gould and Abberley, 2017). For instance, “what is congestion?”,” what is the cause?”, and “where has congestion occurred?”. Using the MUCD, the universal ontology, which was validated in section 4.8.1, it is now possible to answer these questions with clarity and consistency, as shown in Table 8.

Table 8. Consistent answers regarding "what is congestion?"

Question	Answers
What is congestion?	Congestion is what impacts the stakeholder's journey and normally consists of excessive journey time and traffic volumes. A more in-depth explanation would require knowing what type of congestion is occurring.
What is recurrent congestion?	Recurrent congestion is the aftermath of an event, such as rush hour (AM and PM peak). During the AM peak, due to large volumes of traffic entering the citing in a small timeframe, causes excessive journey time for all stakeholders in the direction of the city centre.
What is semi-recurrent congestion?	Semi-recurrent congestion happens pre and post events, such as football matches and concerts impacting traffic towards and away from an attractor. The severity of the congestion depends on the type of event.
What is non-recurrent congestion?	Non-recurrent congestion is the effect of a random event, such as a road traffic accident. It happens at a single point on a road network and then diffuses over time and impacts the local neighbourhood by increasing the traffic volumes and the stakeholders' journey time. The level of impact depends on the severity of the initial event.
What is the cause?	The cause of congestion is an event that increases the stakeholder's journey time and causes a concentration of traffic at a single point, neighbourhood, or city scale.
Where has the congestion occurred?	Depending on the type of congestion, the congestion is likely to occur across the whole city centre, at a single point on the network, or an attractor.

Chapter Five: Visualisation of the Manchester urban congestion dataset using the transport incident manager

5.1 Introduction

To gain a greater understanding of how the Manchester Urban Congestion Data (MUCD) dataset (as defined in chapter 4) and the multifaceted nature of the Urban Road Congestion Conceptual (URCC) model (as defined in chapter 3) interact with each other, a Transport Incident Manager better known as TIM was developed. In this research, the main contribution of TIM is the ability to fill the void left by the clear lack of tools that are capable of visualising real-world big data datasets, such as MUCD and models of urban road congestion. This chapter attempts to answer the research question (RQ3) – "*Can quantifiable big data on urban road congestion be visualised to provide quasi-real-time insight?"*

TIM will answer the research question by being a viable visualisation tool which stakeholders could use. TIM was designed to work with spatial and temporal data like the MUCD dataset and provide experts within the domain the ability to visualise their own data. For example, the experts that provided a portion of data for this research and helped to develop TIM are Transport for Greater Manchester (TfGM). Because this data has strong spatial characteristics a spatial measurement will be investigated and incorporated into TIM. The spatial measurement is known as Moran's Index (I) (Lee and Li, 2017). and is used for determining the spatial autocorrelation between links.

5.2 What is TIM?

Due to the multifaceted nature of urban road congestion and the many dimensions that can be used to measure road traffic performance, such as journey time, volume, traffic flow, velocity, density, and spatial correlation. These measurements can all be impacted by indirect consequences, such as weather conditions, road works, social events, and road accidents. All these characteristics have an impact on the type of congestion from non-congestion, recurrent congestion, non-recurrent congestion, and semi-recurrent congestion (as defined in chapter 3). Therefore, TIM was developed to provide a method that allows stakeholders ranging from domain experts to laypersons to visualise the performance of an urban road network or individual road links.

5.2.1 Methodology of the development of TIM

TIM was developed in conjunction with TfGM, who specified the initial requirements. The primary requirement requested by TfGM was the creation of an informative dashboard that is capable of providing transport managers a real-time overview of individual link performance, based on previous trends.

TIM was developed over five stages with multiple different of prototypes being created along the way:

Stage 1: TfGM to provide a list of key requirements.

The main requirements requested by TfGM was the development of an informative dashboard that transport managers can use to analyse critical links. Additionally, it is important that the dashboard is capable of updating in real-time.

Stage 2: Development of an initial prototype.

The initial prototype created was an informative dashboard that provided a 24-hour view of individual links (selectable by the user) performance. The real-time performance is compared against the typical performance by comparing to historical data in real-time for the same 15-minute period on the same day of week.

Stage 3: Demonstrate prototype to TfGM.

When the initial prototype was developed, it was then presented to TfGM, who then requested the creation of a second dashboard, which required the RAG method for classification to be incorporated. The reason for this requirement was to help the transport managers to better identify urban road congestion in a timely manner when a link is performing inadequately.

Stage 4: Addressing feedback from TfGM.

This stage was conducted over many months and was iterative due to the nature of adding extra functionality into several refined versions of TIM. these refined versions of TIM were presented to TfGM for feedback and led to numerous additional functionalities being incorporated. These additional functionalities included the ability to analyse link performance over longer periods. For example, 24 hours, three days, one week and one month. Other functionalities, such as spatial analysis was included, to help identify high journey time and high traffic volume clusters which are an indicator of congestion on the network. Additionally, TfGM requested the ability to be able to monitor the overall network performance, therefore, this was added. Lastly, the author incorporated an unsupervised machine learning algorithm to help classify the characteristics of traffic volume and journey time into five classifications: Very low, Low, Median, High, and Very High.

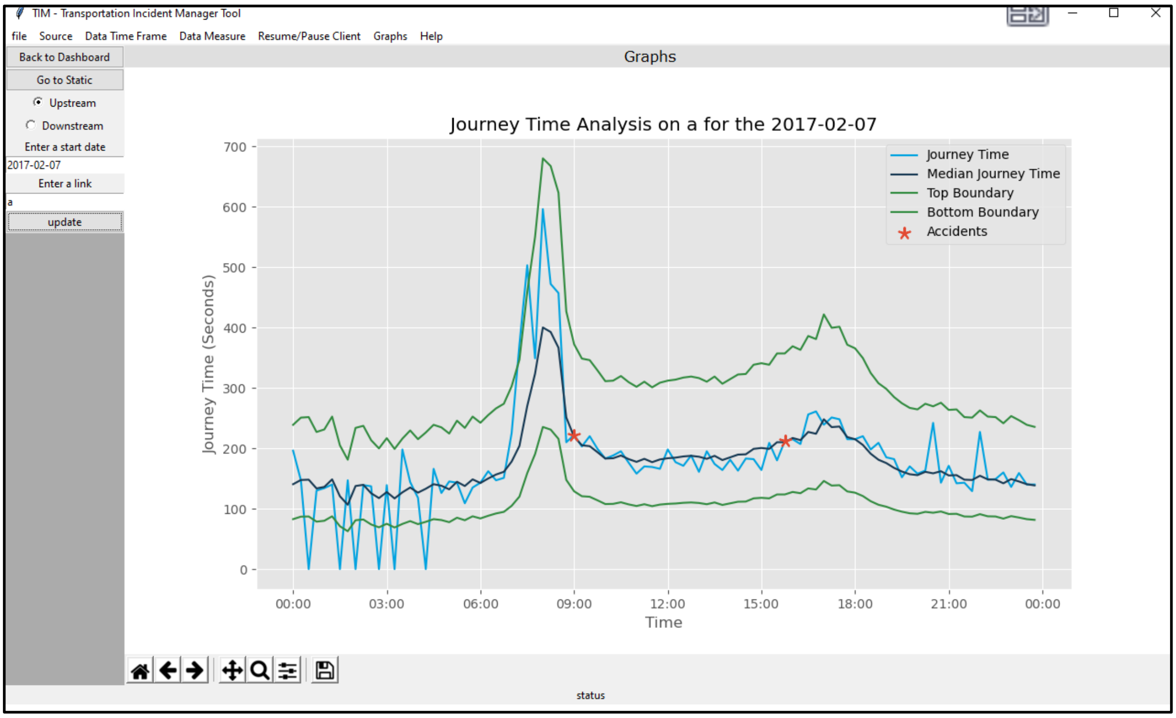
Stage 5: validating the final prototype.

Once all the extra functionalities had been added, the author, created the ability to visualise multiple performance metrics on the same dashboard at different time scales. The final prototype was given to TfGM to ascertain the benefit of having a real-time dashboard that is capable of saving transport managers time by removing the need to collect relevant data and conduct ad hoc analysis on individual links. After evaluating TIM, TfGM deemed TIM to be a success because it was capable of providing instant feedback on the network and individual link performance, allowing them to make decisions quicker than manual operation. The only minor criticism was it was limited to only 64 links (this will be address in future work).

5.3 What data is used in TIM?

The data used within TIM is a subset of the MUCD dataset as defined in chapter 4. The main attributes used within TIM are journey time and traffic volume which were supplied by TfGM and non-recurrent event information, such as a road traffic accident provided by Greater Manchester Police (through the GOV.UK portal). However, TIM was developed with the ability to integrate other sources of journey time, such as the API offered by Google.

The data is presented in 15-minute intervals for a period of either 24 hours, three days, one week, or one month.



● Upstream
○ Downstream

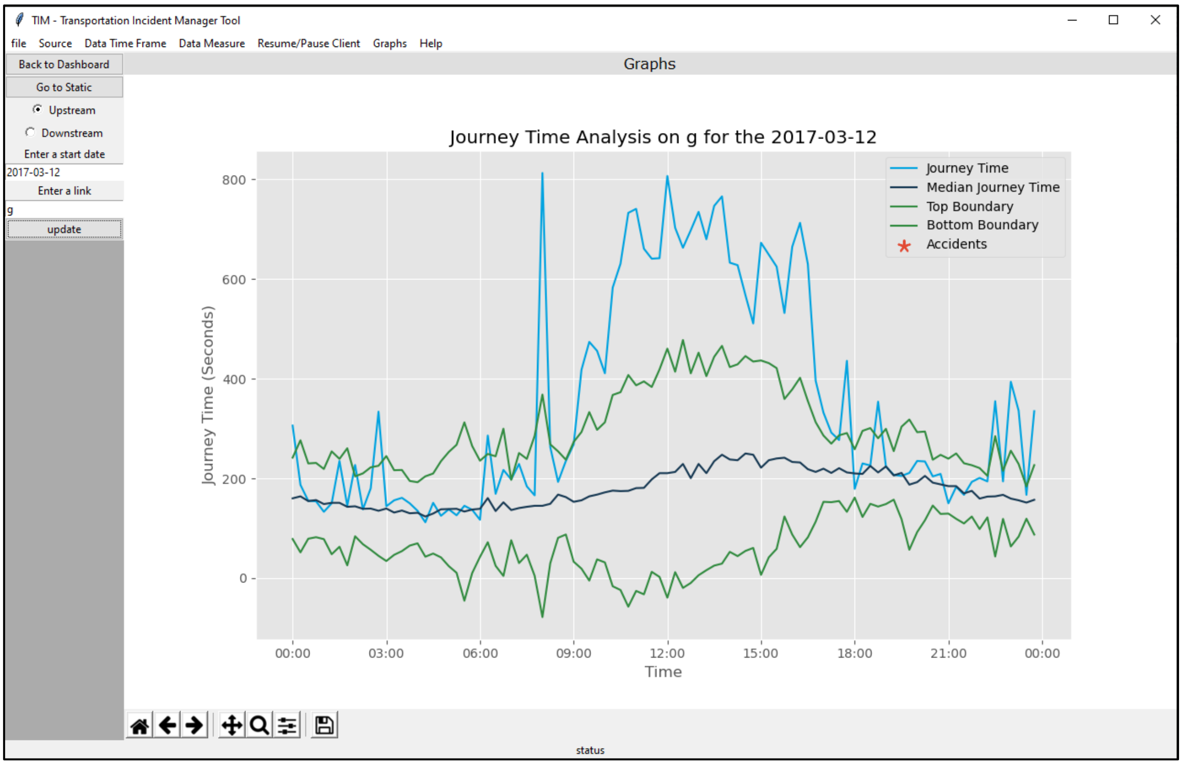
Enter a start date
2017-02-07

Enter a link
a

Figure 39: Transport Incident Manager (TIM)

Figure 39 shows an example of journey time analysis for link ‘AU’ (‘A’ Upstream). Link ‘a’ goes upstream towards link {b} and {c} as described in section 4.3. The analysis presented in Figure 39 is a 24-hour period on Tuesday the 7th of February 2017 from midnight to midnight. The main viewgraph (Figure 39) has five attributes:

- The observed journey time is plotted in 15-minute intervals.
- The typical journey time is used to plot the median journey time every 15-minutes for link ‘AU’ every Tuesday for a six-month period.
- The top and bottom boundaries are used to plot what a typical journey time is for each link at each 15-minute intervals.
- Finally, the main view graph in Figure 39 also plots road accidents that have occurred anywhere on the network and are not specific to an individual link.



● Upstream

○ Downstream

Enter a start date: 2017-03-12

Enter a link: g

Journey Time Analysis on g for the 2017-03-12

Figure 40: Output when using Equation 6 and Equation 7

To determine the top and bottom boundaries, empirical experimentation was conducted to explore a few options, such as Equation 6 and Equation 7 where μ is mean and σ is the standard deviation which is multiplied by φ (phi) to the power of -1 and is added and subtracted and finally multiplied by 1. Figure 40 shows the output when using Equation 6 and Equation 7. The second option is Equation 8 and Equation 9 where M is the median journey time multiplied and divided by 1.7.

$$\mathrm{Top} = \mu + (\sigma * (\varphi^{-1})) * 1$$

Equation 6: Upper (Top) Boundary

$$\mathrm{Bottom} = \mu - (\sigma * (\varphi^{-1})) * 1$$

Equation 7: Lower (Bottom) Boundary

$$\mathrm{Top} = \mathrm{M} * 1.7$$

Equation 8: Upper (Top) Boundary 2

$$\mathrm{Bottom} = \mathrm{M} \, / \, 1.7$$

Equation 9: Lower (Bottom) Boundary 2

The results for both options were provided to the domain experts from TfGM, UK to analyse and compare the values against their currently used RAG (Red, Amber, and Green) method (Abberley et al., 2019). Red, R, (Equation 10) where JT which is the observed journey time in 15-minute intervals, is greater than the M which is the typical (medium) journey time multiplied by 1.5. Amber, A, (Equation 11) where M multiplied by 1.25 less than JT and JT is less than or equal to M multiplied by 1.5. Green, G, (Equation 12) where JT is less than or equal to M multiplied by 1.25. A decision was then made to use Equation 8 and Equation 9, because they are easier equations to implement and have similar behaviour to their current method of classifying congestion using RAG, where (R)ed is major congestion, (A)mber is slight congestion, and (G), is non-congestion.

$$\mathrm{R} = \mathrm{JT} > \mathrm{M} * 1.5$$

Equation 10: RAG (Red) Major Congestion

$$\mathrm{A} = \mathrm{M} * 1.25 < \mathrm{JT} \leq \mathrm{M} * 1.5$$

Equation 11: RAG (Amber) Slight Congestion

$$\mathrm{G} = \mathrm{JT} \leq \mathrm{M} * 1.25$$

Equation 12: RAG (Green) No Congestion

5.4 Functionality

TIM was developed to provide a better way to visualise the multiple concepts of urban road congestion, such as recurrent, semi-recurrent, and non-recurrent and the events that cause urban road congestion, such as ‘rush hour’, football matches, and road accidents.

The functionalities of TIM and the sections where each function is discussed within this thesis can be defined as follows:

- Visual functionality
    - Real-time (graphs) visual views (Section 5.4.1)
        - Main view (Section 5.4.1.1)
        - Classification (Section 5.4.1.2)
        - Moran’s I (Section 5.4.1.3)
        - Network performance (Section 5.4.1.4)
    - Static classification (RAG) view (Section 5.4.2)
    - Unsupervised learning view (Section 5.4.3)
- Temporal measurement (Section 5.4.4)
    - 15-minute intervals within 24 hours
    - 15-minute intervals within 3 days
    - 15-minute intervals within a 1 week
    - 15-minute intervals within a 1 month
- Statistical measurement (Section 5.4.5)
    - Medium journey time for every 15-minute interval for individual links over a 6-month period.
    - Mean journey time for every 15-minute interval for individual links over a 6-month period.
- Pause and resume real-time visual views.

Because of the complexity of the real-world big data (MUCD) dataset and the different concepts of congestion, three views were explored. The first is a real-time visual view (Section 5.4.1) which has the ability to update in real-time as additional data is added to the database, the second is a static view (Section 5.4.2) which is used to visualise the RAG method which is currently used by TfGM, UK. The final view is an unsupervised (k-means++) visualiser (Section 5.4.3).

5.4.1 Real-time visual view

The real-time visual view, first illustrated in Figure 39, was designed with the addition of a ticker which allows the data to refresh every 3000 milliseconds and populates the visual views. The objective of the real-time visual views is to provide a way for the researchers and domain experts to visualise real-world big data and conduct urban road network performance. There are four sub-types of real-time visual views:

1) Main view (Section 5.4.1.1)

2) Classification (Section 5.4.1.2)

3) Moran's Index (Section 5.4.1.3)

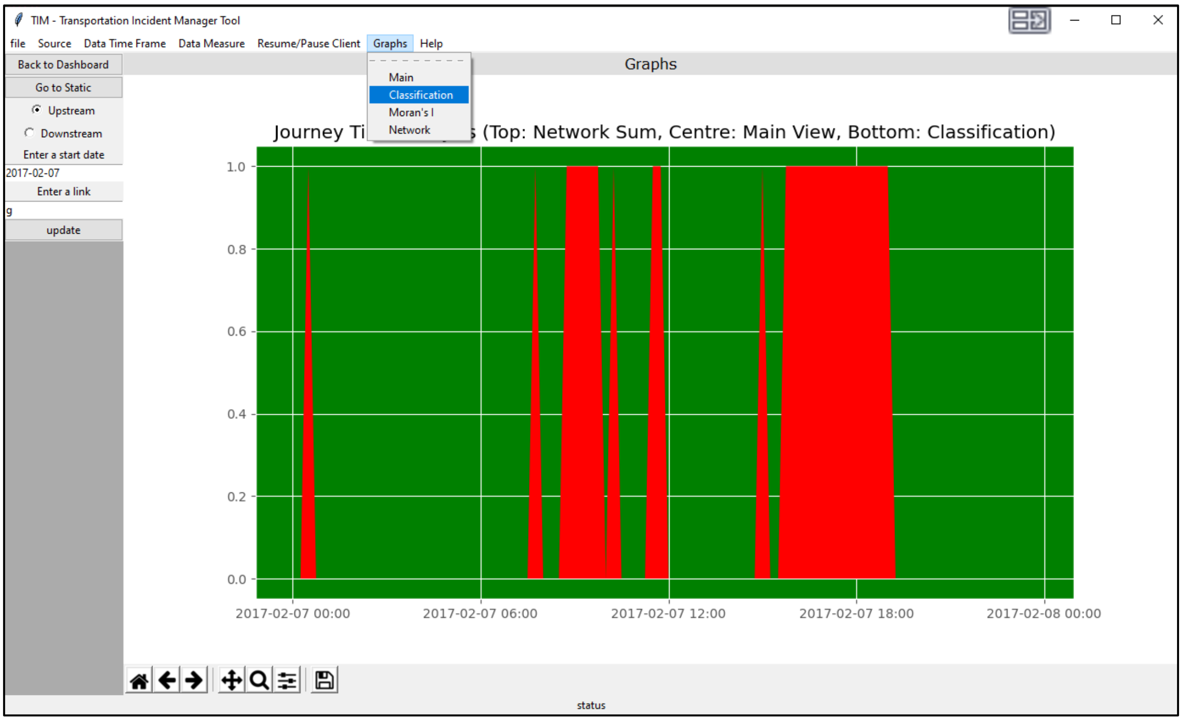
4) Network Classification (Section 5.4.1.4).

5.4.1.1 Main view

The first real-time visual view created was the main view that was partially discussed in section 5.2. Figure 39 Shows an example of journey time analysis for link 'AU'. The analysis presented in Figure 39 is one observation every 15-minutes for a 24-hour period on Tuesday the 7th of February 2017 from midnight to midnight. In addition to the functionality discussed in section 5.2, the main view also plots road accidents that have happened anywhere on the network and are not specific to an individual link.

5.4.1.2 Classification

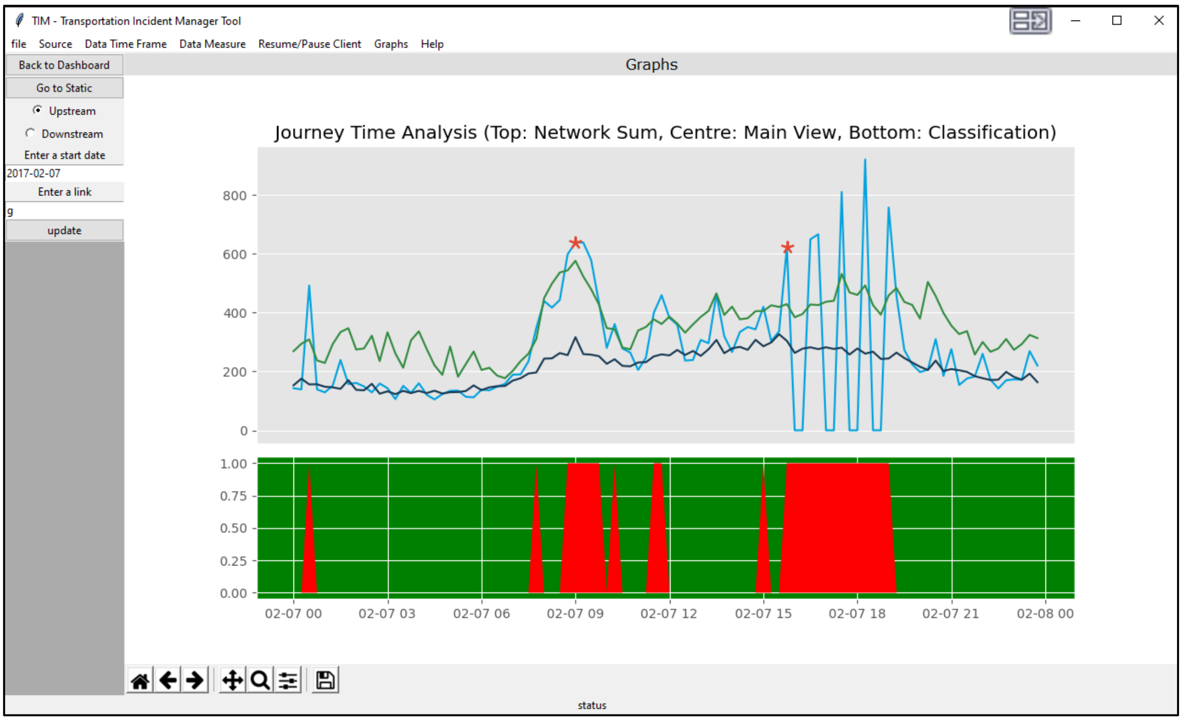
The second real-time visual view (Figure 41) shows the classification of congestion which is either 'congestion' (red) or 'non-congestion' (green) over the period of 24-hours with observations at every 15-minute intervals. The classification is identified when the journey time exceeds the top boundary Equation 8, or the bottom Equation 9 boundary discussed in section 5.3.



Enter a start date
2017-02-07
Enter a link
g

Figure 41: Classification of Congestion

Figure 42 shows how the four visual views can be visualised together as well as by themselves. Figure 42 shows the classification and main view graphs stacked on top of each other for better visualisation.



Enter a start date
2017-02-07
Enter a link
g

Figure 42: Main (top) and Classification (bottom)

5.4.1.3 Moran’s Index

Moran’s Index (I) is used for determining spatial autocorrelation at every observation. Moreover, Moran’s I is just one of many spatial statistical measurements and has previously been used to model spatial autocorrelation for many things, such as the impact hurricane sandy had on HIV, pollution hotspots, and road accidents on Belgium motorways (Black and Thomas, 1998; Zhang et al., 2008; Acharya et al., 2018; Wilt et al., 2018; Chen, 2020). Spatial autocorrelation is extremely multifaceted because it models the spatial correlation in a multi-dimensional space and can be multi-directional at the same time. A good example of multi-dimensional and multi-directional spatial correlation is a road network due to the many different road links (multi-dimensional) and the ability for traffic to travel upstream and downstream on these links (multi-directional).

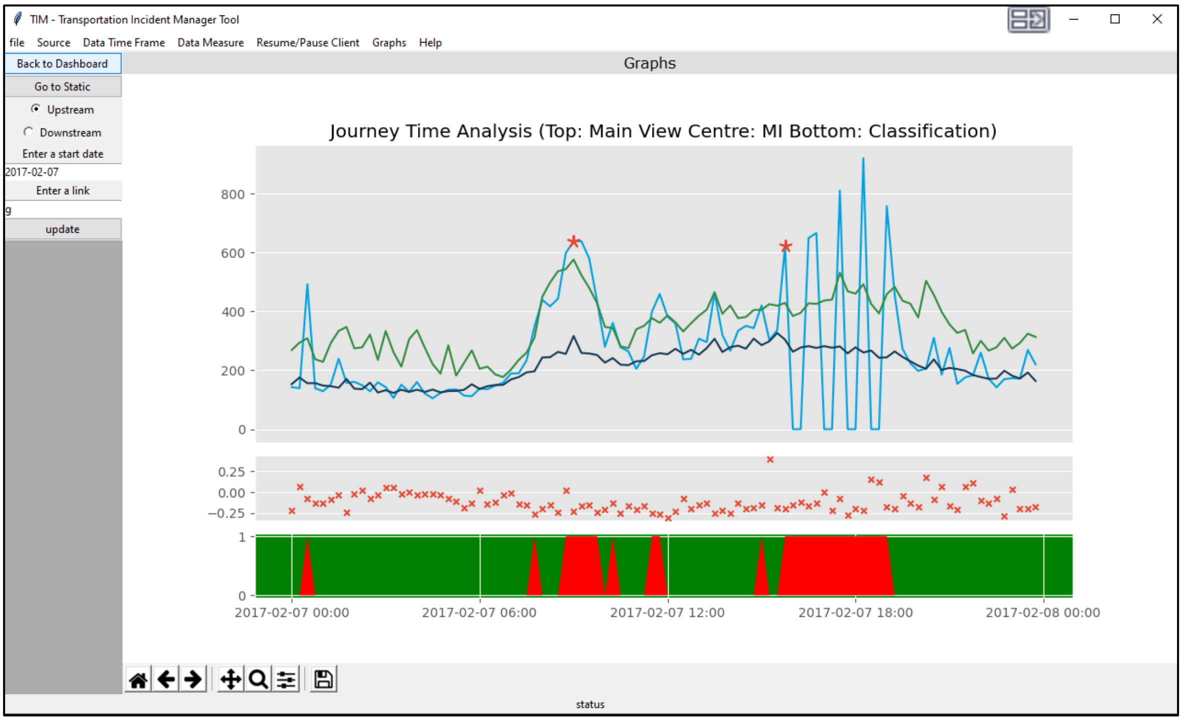
There are two statistical versions of Moran’s I, which are global and local Moran’s I. This research will focus on using global Moran’s I because it is designed to identify clusters on a whole network or region instead of individual links, which local Moran’s I was designed to do and is used to identify the exact area the cluster is located. (Yang et al., 2018; Xiong et al., 2021). Therefore, because this research is trying to model the concept of urban road congestion and its causes and not identify the location of the cause, this research, and the development of TIM, only focused on the inclusion of Global Moran’s I.

Global Moran’s I (Equation 13) is defined as:

$$\mathrm{I} = \frac{\mathrm{N}}{\mathrm{W}} \frac{\sum \mathrm{i} \sum \mathrm{j}\, \mathrm{w_{ij}} (\mathrm{x_i} - \bar{\mathrm{x}})(\mathrm{x_j} - \bar{\mathrm{x}})}{\sum \mathrm{i}\, (\mathrm{x_i} - \bar{\mathrm{x}})^2}$$

Equation 13: Global Moran's I

Where N is the number of spatial units indexed by i and j; x is the variable of interest; $\bar{\mathrm{x}}$ is the mean of x; $\mathrm{w_{ij}}$ is a matrix of spatial weights with zeros on the diagonal, for instance, $\mathrm{w_{ij}} = 0$; and W is the sum of all $\mathrm{w_{ij}}$. Moran’s I ranges between -1 and +1. When I am equal to +1, it is an indication that all the observations on the whole network or within the region are clustered in space. However, if the I is equal to -1 then it implies all observations are randomly scattered (Tepanosyan et al., 2019). Figure 43 shows the TIM visual view of Moran’s I alongside the main view and the classification view.



● Upstream
○ Downstream

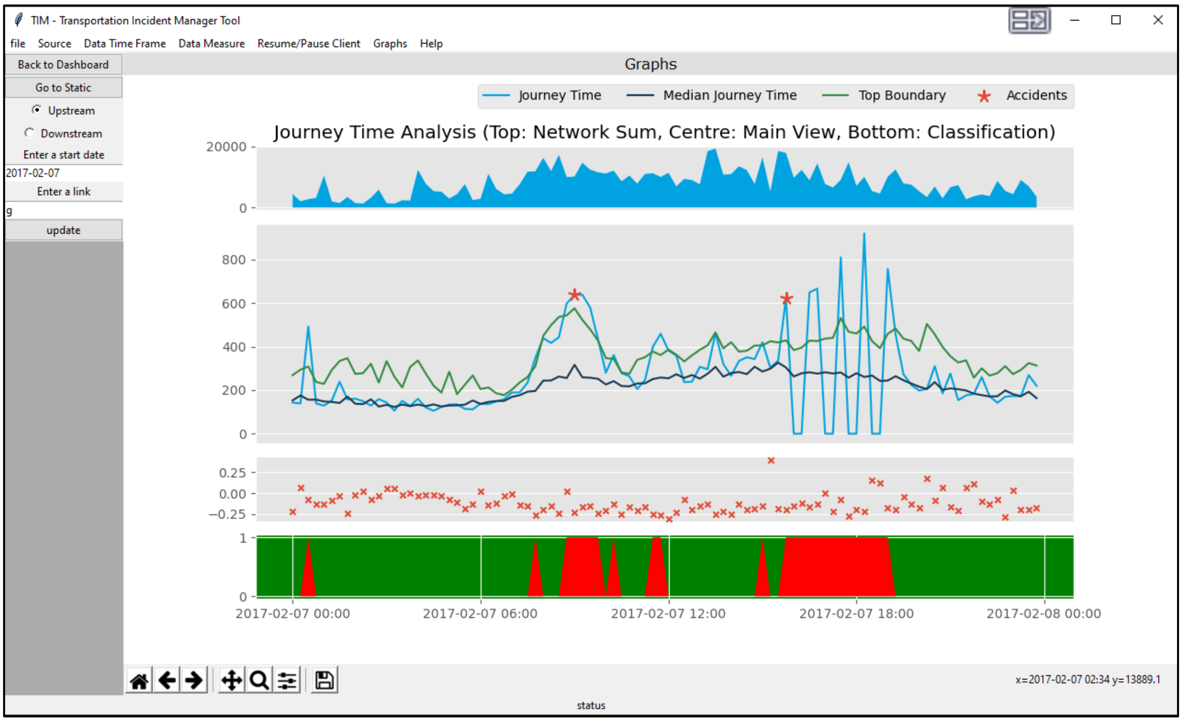
Enter a start date
2017-02-07

Enter a link
g

Figure 43: Main (top), Moran's I (centre), and Classification (bottom)

5.4.1.4 Network Performance

The network performance view provides a total journey time recorded on the whole neighbourhood network at every 15-minute interval. The total journey is the sum of all 64 links on the network (32 upstream and 32 downstream) as described in section 4.3. The benefit of domain experts using the network performance is it provides a summary of the overall performance and will allow domain experts to easily identify anomalies on the network.



● Upstream
○ Downstream

Enter a start date
2017-02-07

Enter a link
g

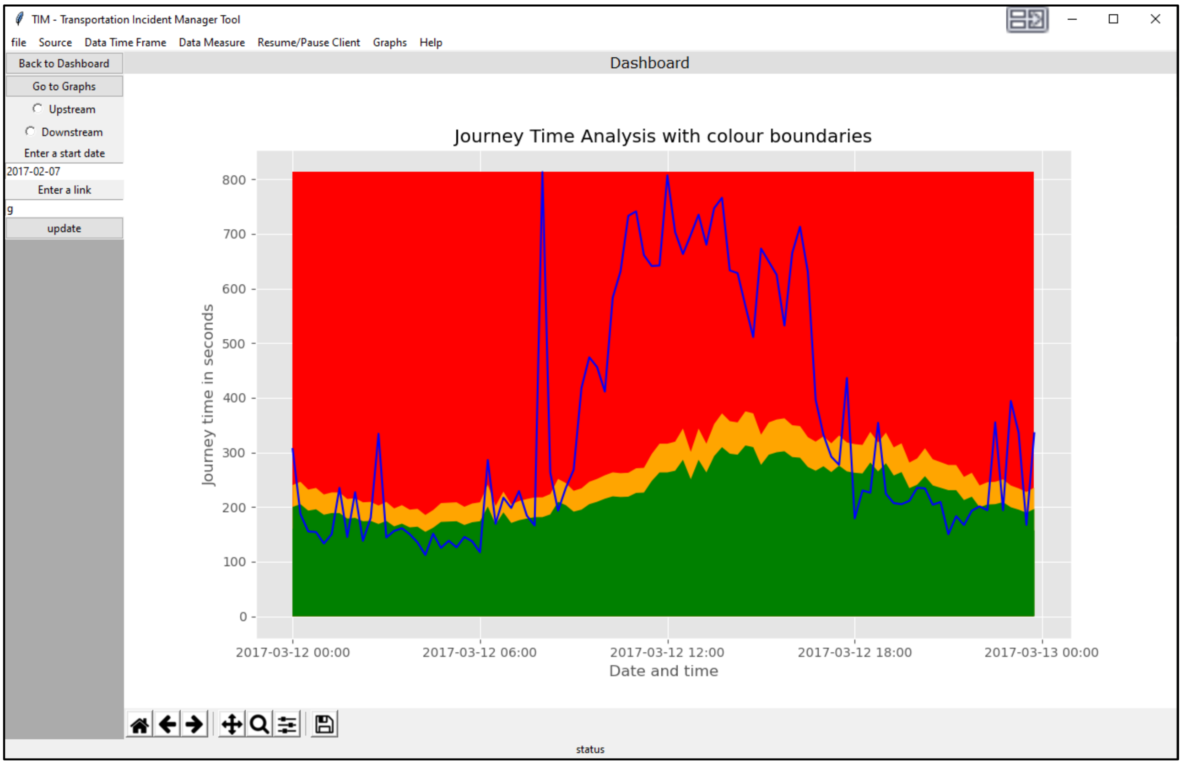
x=2017-02-07 02:34 y=13889.1

Figure 44: Network Performance (top), Main (top centre), Moran's I (bottom centre), and Classification (bottom)

Figure 44 shows the whole network (total journey time) performance (top) for link ‘g’ upstream on the 7th of February 2017. Below that is the main view (top centre), next is Moran’s I (bottom centre), and finally the classification (bottom).

5.4.2 Static classification (RAG) view

The static classification view is illustrated in Figure 45 and Figure 46, which shows the RAG classification which is implemented by TfGM, where the red (Equation 10) represents major congestion, amber (Equation 11) slight congestion, and green (Equation 12) non-congestion. Figure 45 shows the classification for the journey time upstream and Figure 46 shows the classification for the journey time downstream on the 7th of February 2017 on link g. Each observation is at every 15-minute interval.

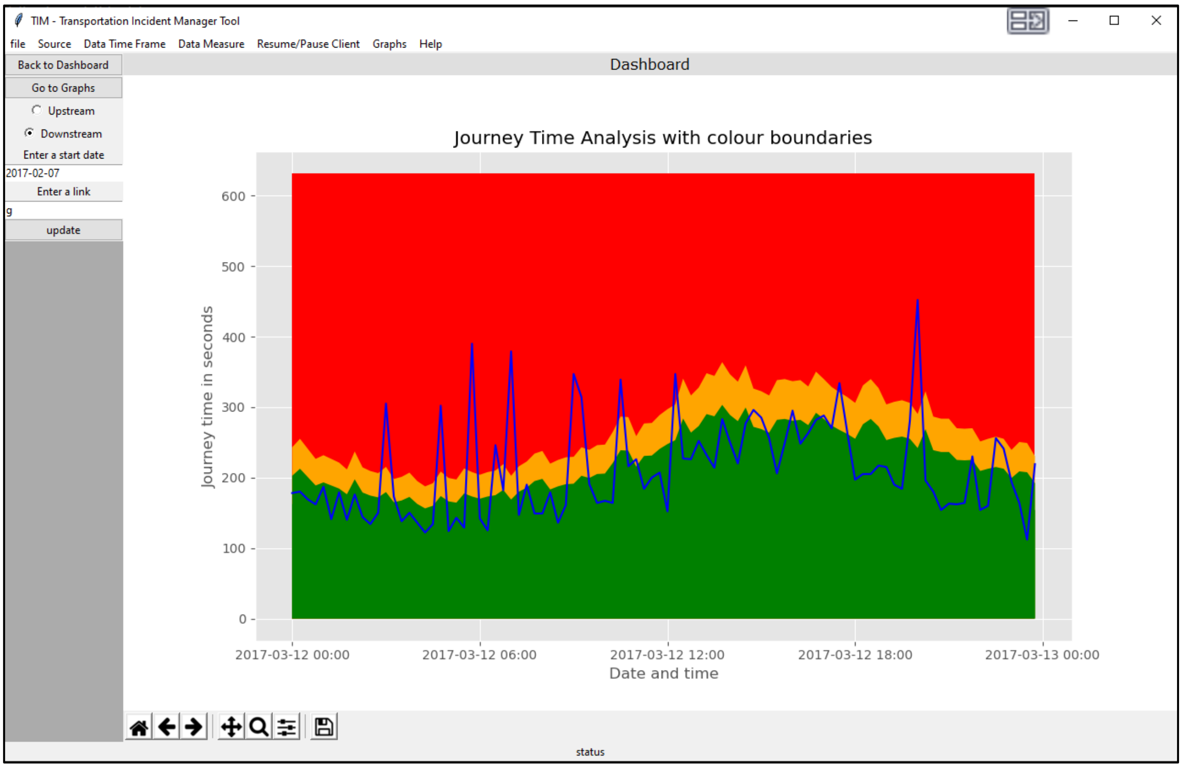


○ Upstream
○ Downstream

Enter a start date
2017-02-07

Enter a link
g

Figure 45: Static classification view (upstream)



○ Upstream
● Downstream

Enter a start date
2017-02-07

Enter a link
g

Figure 46: Static classification view (downstream)

5.4.3 Unsupervised learning visualiser view

The unsupervised learning visualiser view was developed to assist with investigating the use of unsupervised learning to identify the characteristics of urban road congestion in chapter six. Therefore, the supervised learning algorithm introduced in this chapter will contribute towards answering the research (RQ2) - "*Can quantitative Big Data be used to provide qualitative information in conjunction with a road traffic ontology with the support of Machine Learning?*".

5.4.3.1 Why was k-means++ used?

TIM is used to visualise the unlabelled real-world big data (MUCD) dataset to help achieve the focus of the following paper (Abberley et al., 2017), which was to gain knowledge and understanding from an unlabelled subset of the MUCD. Therefore, to analyse the unlabelled data an unsupervised learning approach, such as clustering was taken. Clustering was chosen because it is one of the most common types of machine learning algorithms used when dealing with unlabelled data (Philip Chen and Zhang, 2014).

Some of the most popular clustering algorithms are k-means, k-medians, Expectation Maximisation (EM), and Hierarchical Clustering (Aggarwal, 2013). However, for this research, k-means++ algorithm was chosen because according to (Arthur and Vassilvitskii, 2007) it has previously achieved functional values of 20% compared to k-means and performed 70% faster when conducting experiments on four different datasets, the first two were synthetic and are known as 'Norm-10' and 'Norm-25' dataset and the remain two are known as 'Cloud' and 'Intrusion' dataset which the latter is the largest dataset with 494019 data points in 35 dimensions.

5.4.3.2 K-means++ algorithm

The k-means++ algorithm steps below have been adapted from the method set out in (Arthur and Vassilvitskii, 2007). Where k is the number of centres, which is used to defined how many clusters (c) will be created. C is a set of clusters. X is a set of data points and x is a single data point. *i* and j are both sets of observations.

1a. Take one centre c_1, chosen uniformly at random from X .

1b. Take a new centre c_i, choosing $x \in X$ with probability $\frac{D(x)^2}{x \in X\ D(x)^2}$ where $D(x)$ denotes the shortest distance from a data point.

1c. Repeat Step 1b. until we have taken k centres altogether.

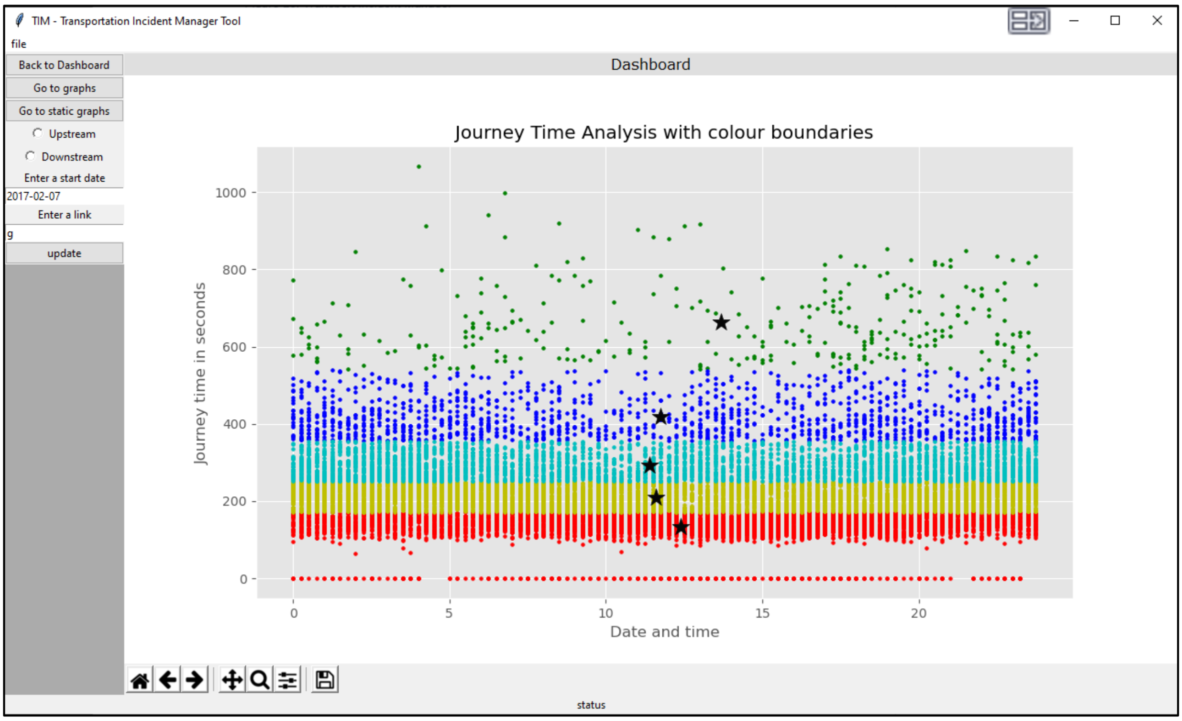
2. For each $i \in \{1,\ldots,k\}$, set the cluster C_i to be the set of points in X that are closer to c_i then they are to c_j for all $j \neq i$.where k centres $C = \{c_1, c_2, \ldots, c_k\}$.

3. For each $i \in \{1, \ldots, k\}$, set c_i to be the centre of mass of all points in $C_i : c_i = \frac{1}{|C_i|} \sum x \in C_i{}^x$.

4. Repeat steps 2 and 3 until C no longer changes.

5.4.3.3 Example of K-means++ clustering algorithm output

Figure 47 provides an example of the K-means++ algorithm output on link ‘g’ for every Tuesday, for a 6-month period from January 2017 and June 2017. The output provided five classifications (which are represented by the stars (*)): Very low (red), Low (yellow), Medium (turquoise), High (purple), and Very high (green) journey times. These five classifications were calculated using the method discussed in section 5.4.3.2. The distance between the points were measured using Euclidean distance (Equation 2).



Enter a start date
2017-02-07
Enter a link
g

Figure 47: K-means++ clustering

Observing Figure 47 would provide TfGM the ability to classify observed journey times, such as a ‘high journey time’ in a statistical manner instead of using their current ‘gut feel’ approach which is relevant to the individual at point of analysis.

5.4.4 Temporal selection

When analysing the data and the network performance, researchers, and domain experts such as TfGM, may want to look at the data for a longer period than 24-hours to help identify any anomalies. Therefore, TIM was developed to display data for four different time periods: 24-hours, three days, one week, and one month. Figure 48 shows a three-day period from the 12th of February 2017 till the 15th of February 2017 on link ‘g’ at 15 minutes intervals.

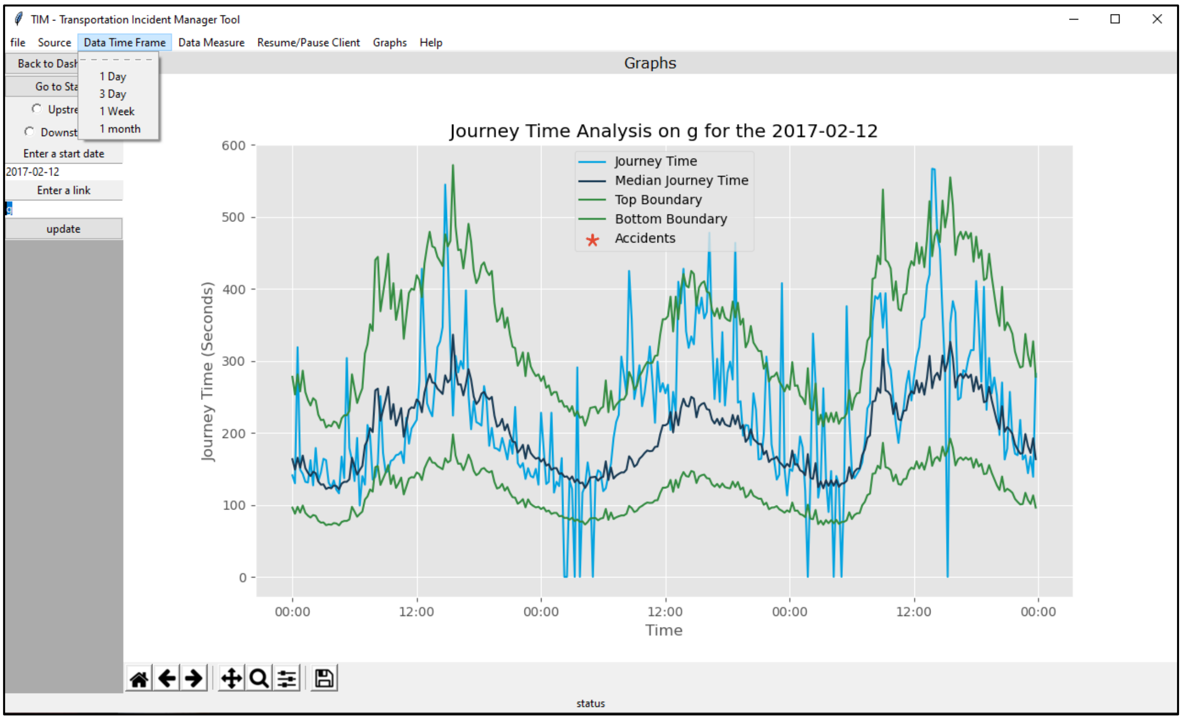


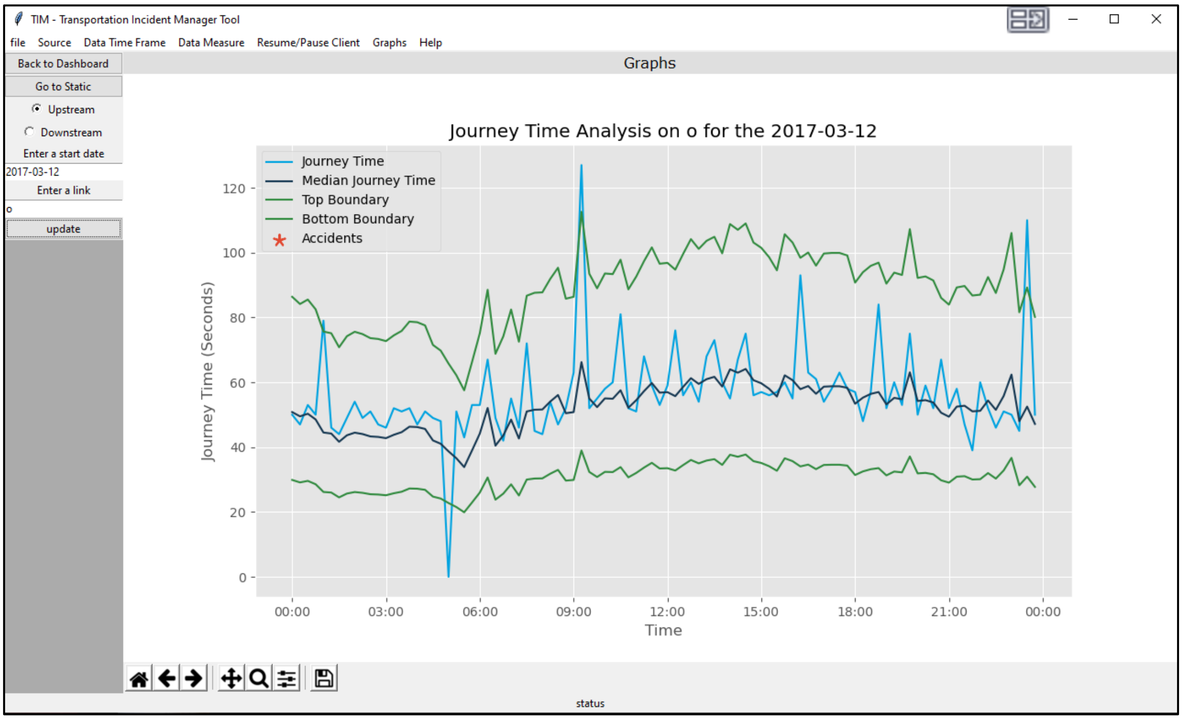
Figure 48: Temporal selection of three days

Observing Figure 48 would provide TfGM the ability to see if any of the days appear to be an outlier, which could be identified because it doesn't follow the typical behaviour of the previous days.

5.4.5 Statistical measurement

Figures 49 and 50 show the different visualisation of the mean average journey time and median average journey time which is represented by the dark blue line. Additionally depending on if you choose to use the mean or median measurement, the calculation for the upper and lower boundaries would be different and give a slightly different classification of congestion.

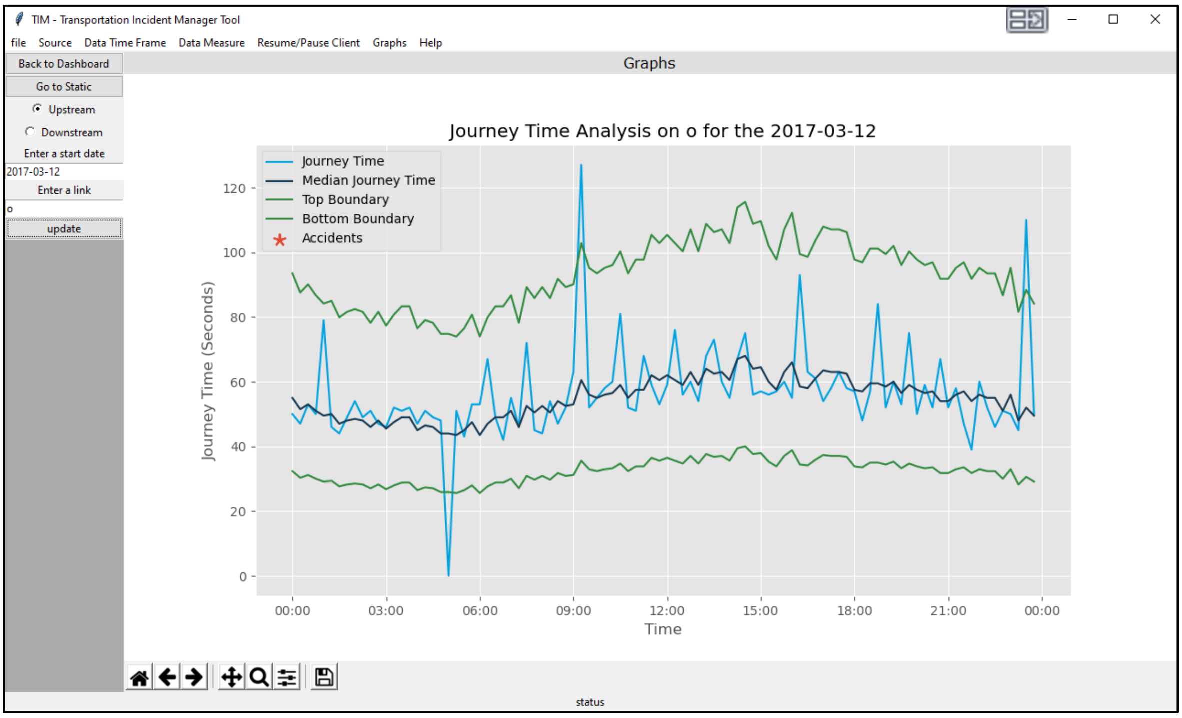
5.4.5.1 Mean



● Upstream
○ Downstream
Enter a start date
2017-03-12
Enter a link
o

Figure 49: Statistical measurement (mean)

5.4.5.2 Median



● Upstream
○ Downstream
Enter a start date
2017-03-12
Enter a link
o

Figure 50: Statistical measurement (median)

5.5 Why is TIM important?

There is an obvious lack of visualisation tools that are capable of visualising real-world big data datasets, such as MUCD dataset. Furthermore, there is a lack of tools that are capable of visualising models of urban road congestion. Making TIM a vital component of this research because it is able to provide a platform for visualising the MUCD dataset in several different ways. Therefore, TIM is an essential tool kit for exploring urban road congestion, analysing data to identify patterns and characteristics, thus, supporting the conceptual model; analogy, and ontology discussed in chapter three. For instance, it has been theorised that urban road congestion has three types of congestion: recurrent, non-recurrent, and semi-recurrent (which was coined by the author).

In the next section (5.6), an event that represents each type of congestion will be evaluated using TIM. These events are: rush hour (recurrent), a road accident (non-recurrent), and a football match (semi-recurrent).

5.6 Evaluation of TIM and the data set provided by TfGM.

As previously mentioned in section 5.5, this section will use the real-world big data visualiser tool (TIM) and the MUCD to evaluate three different types of congestion which are the consequences of three different events. The three scenarios being presented are: rush hour (Section 5.6.1) which is classified as recurrent congestion because it is predictable and cyclical, a road accident (Section 5.6.2) which is classified as non-recurrent congestion because it is non-predictable and non-cyclical, and a football match (league) (Section 5.6.3) which is classified as semi-recurrent because it is predictable but non-cyclical.

5.6.1 Rush hour

The case study being used for the road traffic event that has the consequence of recurrent congestion is rush hour am and pm which is predictable and cyclical, and this example occurs on the A6 within Greater Manchester, UK, on the 8th of March 2017 and the 9th of March 2017. Figure 51, Figure 52, Figure 53, and Figure 54 all show a 24-hour period with 15-minute intervals of journey time for link {z} on a Wednesday and a Thursday.

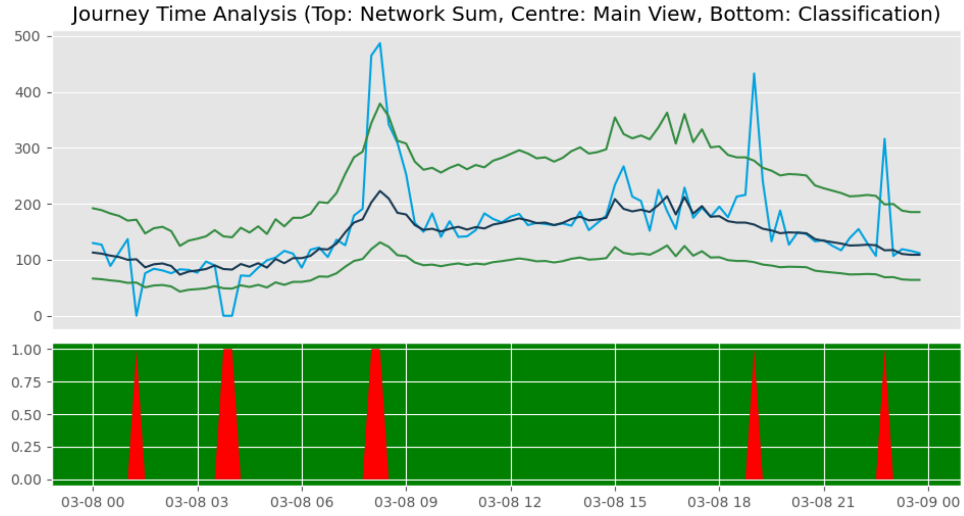


Figure 51: Link z upstream 8th of March 2017

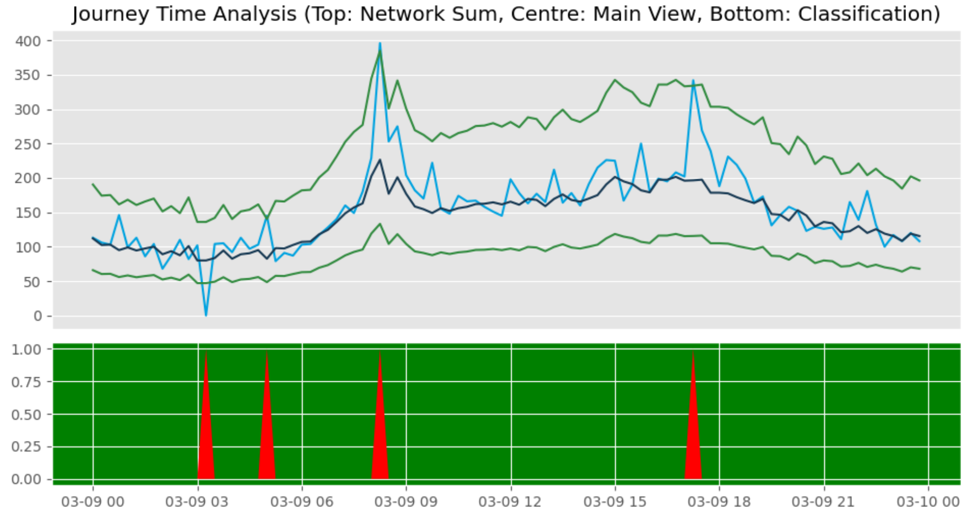


Figure 52: Link z upstream 9th of March 2017

Looking at both Figure 51 and Figure 52 you can see an obvious increase in journey time where the average observed journey time on both days range from 100 to 150 seconds. However, the observed journey time between 8 am and 9 am sharply increase to around 400 and 500 seconds. Then between 9 am and 10 am the journey time starts to revert to a typical journey time between 100 to 150 seconds.

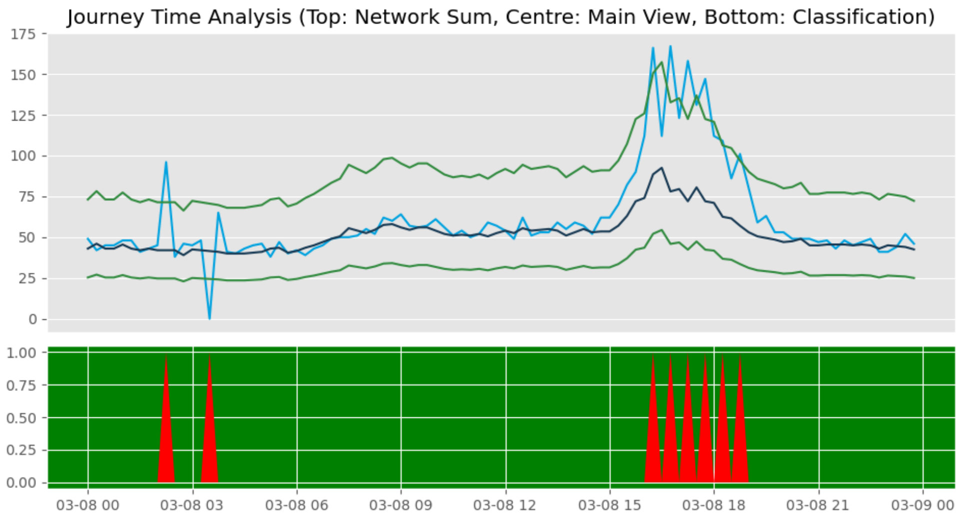


Figure 53: Link z downstream 8th of March 2017

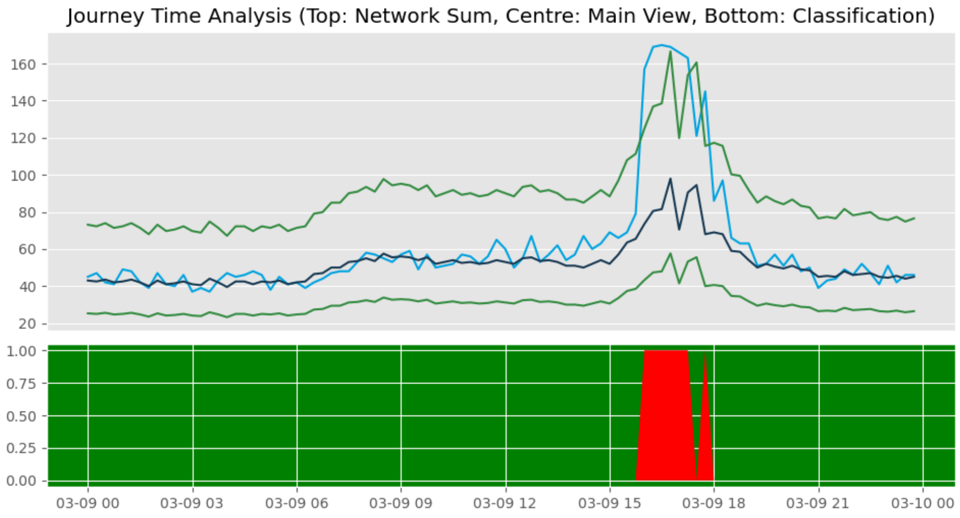


Figure 54: Link z downstream 9th of March 2017

Figure 53 and Figure 54 shows a similar behaviour as Figure 51 and Figure 52, however, due to this being in the opposite direction and heading out of Manchester city centre. The pattern of journey time sharply increasing occurs in the afternoon around 4 pm and returns to a typical journey time around 7 pm.

5.6.2 Road accident

The case study being used for the road traffic event, which has the consequence of non-recurrent congestion is a fatal road accident that happened on the A6 within Greater Manchester, UK, on the 7th of February 2017 at 15:40 and is non-predictable and non-cyclical. The A6 road consists of the following links {a, c, e, g, i, m, o, z} (Figure 27). However, because the fatal accident happens on link {g}, this analysis will focus purely on this link.

Figure 55 and Figure 56 shows a 24-hour period on the 7th of February 2017 where two road accidents (the red star (*)) occurred on the urban road network the first one occurred around 8:55 am which was classified as slight and is not the focus of this section. The second road incident was classified as fatal and caused a more significant impact on the network performance. As you can see from both Figure 55 and Figure 56 the journey time for upstream and downstream exceed the upper boundary significantly or is recorded as a zero-journey time. This means traffic is not passing both Bluetooth sensors within the 15-minute interval because the road link is closed, and vehicles need to divert around the area of the fatal road accident. At 19:30 the road network returned to an expected classification of non-congestion and remain that way for the rest of the 24-hour period.

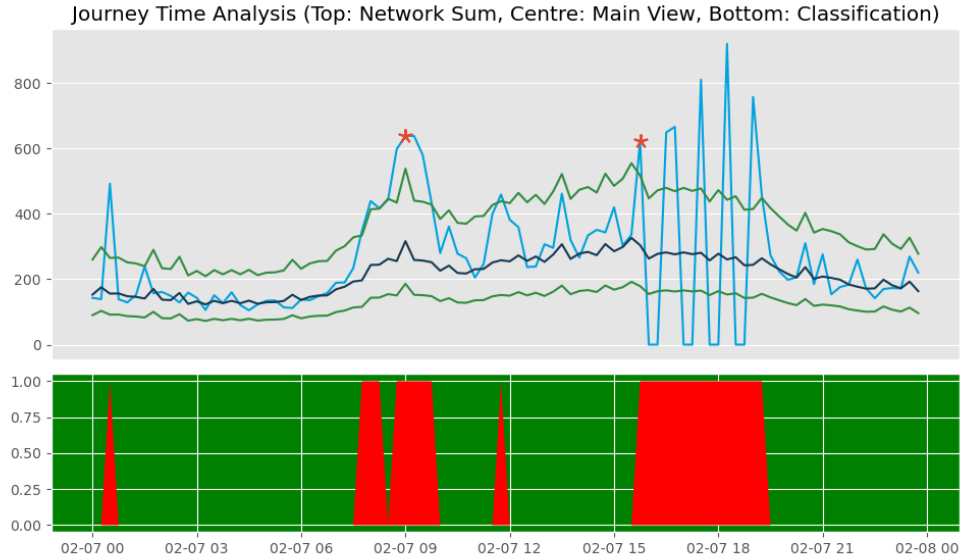


Figure 55: fatal road accident on link {g} (upstream)

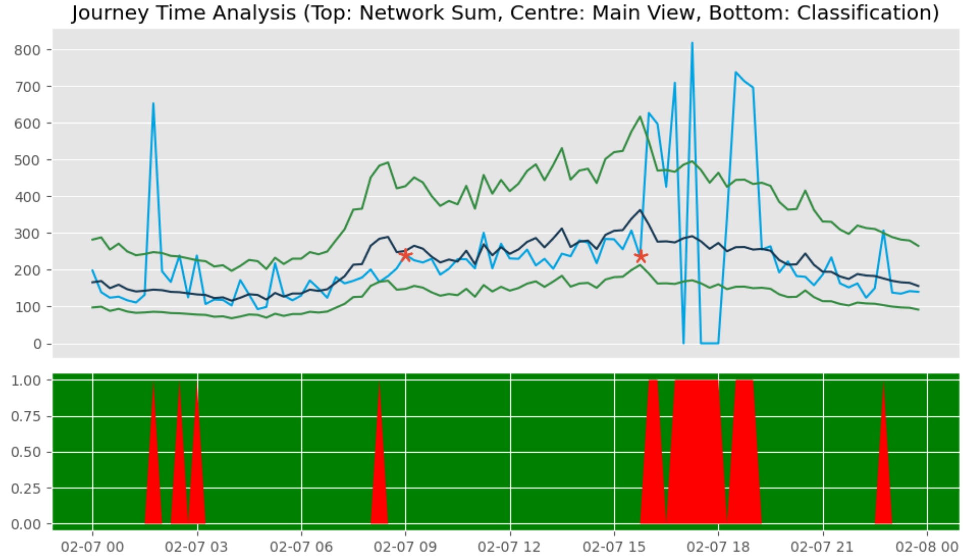
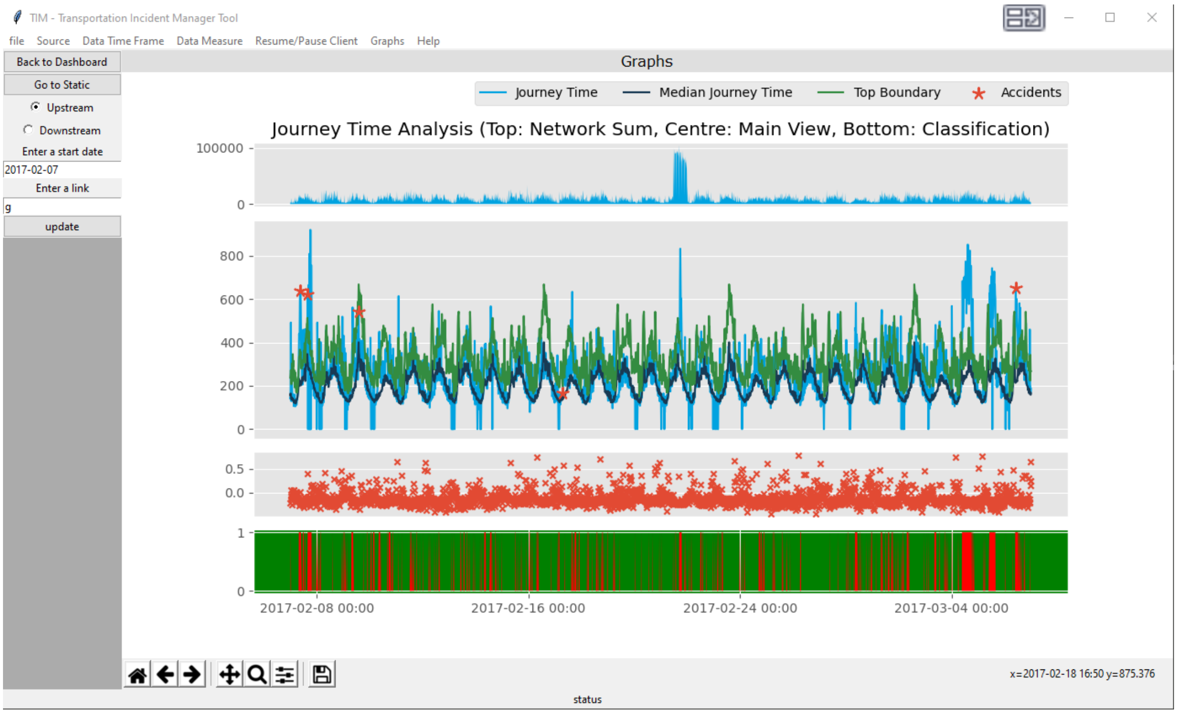


Figure 56: fatal road accident on link {g} (downstream)

5.6.3 Football match (league)

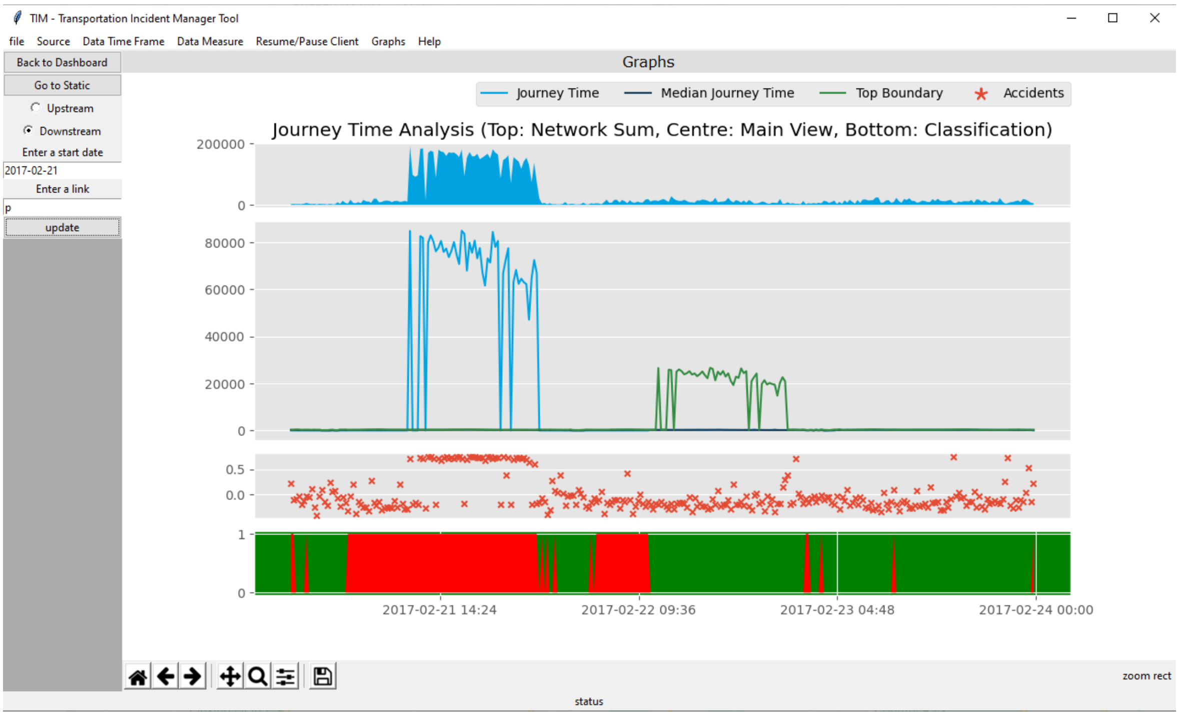
The case study being used for the road traffic event that has the consequence of semi-recurrent congestion is a football match (league) which is predictable but non-cyclical. The league match occurred at the Etihad Stadium in Manchester, UK on the 21st of February 2017 and where Manchester City FC were one of the last 16 teams in the champion league and beat Monaco 5-2. Figure 57 presents an overview of the network performance between the 7th of February and the 7th of March 2017. When viewing Figure 57, there is an obvious anomaly of a sharply increased journey time over the whole network which is visible in the network performance view (top row).



Enter a start date: 2017-02-07

Enter a link: g

Figure 57: A one-month analysis of the road network



○ Upstream
● Downstream

Enter a start date
2017-02-21

Enter a link
p

Figure 58: A three-day analysis of the road network

In Figure 58, link {p} (near to the Etihad stadium) would typically take around 60 seconds to travel, is now taking between 60000 to 80000 seconds and has become extremely congested from around midday until after midnight. Additionally, to the increase in journey time, there is an obvious spatial autocorrelation with a Moran's I value of 0.75 for the majority of the day.

Figure 59 shows a simplistic way to visualise the score (which is a normalise scaling between -1 and 1) as a chess board where each square would represent a road link. -1 represents when a high value, such as high journey time repels other high values. 1 represents when a low or high value is clustered nears similar values. 0 represents when the low and high values are randomly distributed across the network. Therefore, having a constant Moran's I value of ~0.75 means there is a large cluster of links next to each other with very high and high journey times. This implies something major is impacting most of the neighbourhood network topology.

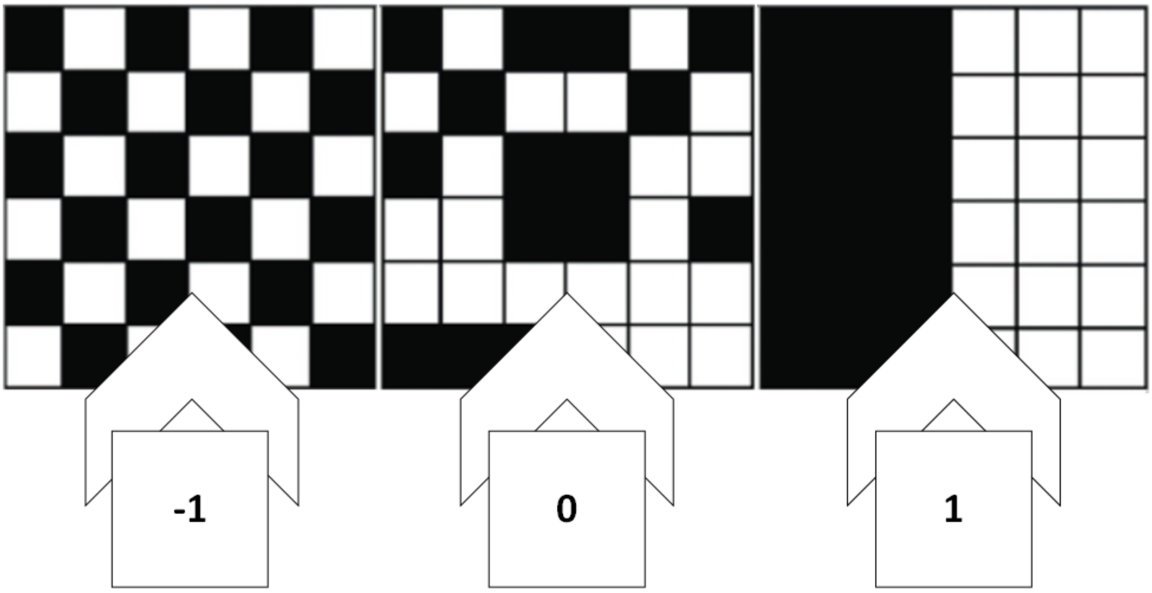


Figure 59: Moran’s (I)ndex simplified

5.7 Chapter conclusion

This chapter has introduced the design and implementation of TIM - a visual Transport Incident Manager tool. Furthermore, this chapter discussed the several different functionalities of TIM, such as plotting data in real-time to allow domain experts to have a real-time view of individual links and an overall network performance. Other functionalities are the ability to classify the data using the RAG method developed by TfGM, conduct unsupervised learning, the ability to look at the data in different spatial and temporal states, and different statistical measurements, such as mean and median.

The key contribution to this chapter is TIMs ability to fill the void left by the clear lack of visualisation tools that are capable of visualising real-world big data datasets, such as MUCD and models of urban road congestion, such as the URCC. Therefore, this chapter answers the research question (RQ3) – “*Can quantifiable big data on urban road congestion be visualised to provide quasi-real-time insight?*”

This chapter has demonstrated that it is possible to take quasi-real-time data such as journey time and implement several statistical functions to gain insight into the behaviour and characteristics of congestion causing events, such as rush hour, a road accident, a football match. The feedback from the stakeholders at TfGM with regards to the functionality of TIM were positive, they were happy that their current RAG method was included because it was one of their requirements for assessing performance of individual links. Furthermore, TfGM has suggested in the future work, they would like TIM to look at their whole network instead of the subsample chosen for this research, and believe it would be a vital tool for daily use.

Chapter Six: An investigation of unsupervised learning to predict urban road congestion.

6.1 Introduction

The aim of this chapter is to investigate the use of unsupervised learning to ascertain whether it is possible to use predictive analytics to identify the characteristics of urban road congestion and gain (qualitative) context from quantitative data within the Manchester Urban Congestion Data (MUCD) dataset.

The reason it is important to extract qualitative context from the MUCD is because current Intelligent Transport Systems (ITS) lack the capability to provide stakeholders, such as road users with meaningful context to allow them to make more informed and better decisions. For example, road users when driving on a highway would notice Variable Message Signs (VMS) declaring, "CONGESTION AHEAD EXPECT DELAYS". However, this message lacks any meaningful context and creates more questions for the road users. For instance, what type of congestion? Where has the congestion occurred? What is the cause? When did it start? When will it end? Are there any alternative routes? How will it influence the overall journey? A more meaningful message would be "CONGESTION AHEAD IN 2 MILES, DUE TO AN MINOR ACCIDENT AT 15:45 CAUSING INCREASED JOURNEY TIMES". This would allow the road users to make better decisions, such as coming off the highway early and diverting. This behaviour would then reduce the consequence of the minor accident, allowing the non-recurrent congestion to be cleared sooner.

The experiments described in this chapter attempt to answer the following research question (RQ2) - "*Can quantitative Big Data be used to provide qualitative information in conjunction with a road traffic ontology with the support of Machine Learning?*"

This chapter will contribute to answering the question by using the Urban Road Congestion Conceptual (URCC) model and the relevant data from within the MUCD dataset to conduct a series of empirical experiments using k-means++. The empirical experiments will focus on using a single road within the Greater Manchester region which consists of several links. Once the empirical experiments have been conducted, the outputs will be visually interpreted to ascertain whether qualitative context can be gained from qualitative data. This methodology can be reproduced, assuming a similar set of links with similar data is used.

The results of this paper have been published in

- L. Abberley, N. Gould, K. Crockett and J. Cheng, "Modelling road congestion using ontologies for big data analytics in smart cities," 2017 International Smart Cities Conference (ISC2), 2017, pp. 1-6, Doi: 10.1109/ISC2.2017.8090795

6.2 Experimental methodology

A subset of the MUCD dataset described in Chapter Four was used for the empirical experiments in this chapter. Two methods for classifying the data have previously been discussed in section 4.8.1 using Algorithm 1 and section 5.4.2 using the RAG method which has been visualised using the MUCD dataset within TIM. The labels are: Non-congested ((G)reen) and Congested ((R)ed and (A)mber).

The aim of the experiments described in this section is to understand the characteristic of urban road congestion, such as ‘high journey time’. The series of experiments will be treating the data as non-labelled. A non-labelled dataset is best suited for working in conjunction with unsupervised learning algorithms such as clustering (Zhang et al., 2016). Clustering is a type of machine learning algorithm and is one of the most commonly used techniques when a user has a non-labelled data problem and requires a solution (Philip Chen and Zhang, 2014). Clustering models the relationship between variables using approaches, such as centroid-based and hierarchical. All clustering methods use the inherent structures in the data to best organize the data into groups of maximum commonalities. Some of the most popular clustering algorithms are k-means, k-medians, Expectation Maximisation (EM) and Hierarchical Clustering (Aggarwal, 2013)

Traditionally, congestion has been human monitored by measuring several different dimensions, such as speed, traffic volume, and occupancy on the road network. However, these dimensions are not without limitations; for example, speed as opposed to journey time is a measure at a single point on a link and cannot be used as a constant or to evaluate the whole link due to the possibility of a traffic incident further down the road. Traffic volume and occupancy require frequently deployed ‘expensive’ equipment, for instance, inductive loop counters.

The series of experiments in this chapter have used data from inexpensive technology (e.g., Bluetooth sensors vs traffic cameras) that can be used to calculate journey times rather than speed and identify changes in journey time and traffic volume depending on the day and time providing information that is more useful and meaningful.

The following hypothesis will be evaluated.

Hypothesis one

$H_{A}0$: Clustering an unsupervised dataset creates clusters that make it possible to predict journey time.

$H_{A}1$: Clustering an unsupervised dataset creates clusters that cannot be used to predict journey time.

Hypothesis two

$H_{B}0$: Clustering an unsupervised dataset creates clusters that make it possible to identify differences between a weekday and a weekend.

$H_{B}1$: Clustering an unsupervised dataset creates clusters that cannot be used to identify differences between a weekday and a weekend.

6.2.1 Experimental methodology

The first step is to create a subset of the MUCD dataset, consisting of journey time, traffic volume, and road accident data in 15-minute intervals for a 3 month (13 weeks) period (January until March 2017), and uses links on the A6 ({a},{c},{e},{g},{i},{m},{o})(see Figure 60) and will be discussed in section 6.2.2. Once the subset of data was created the next step was to model this data by performing clustering using the k-means++ algorithm, which was discussed in section 5.4.3.

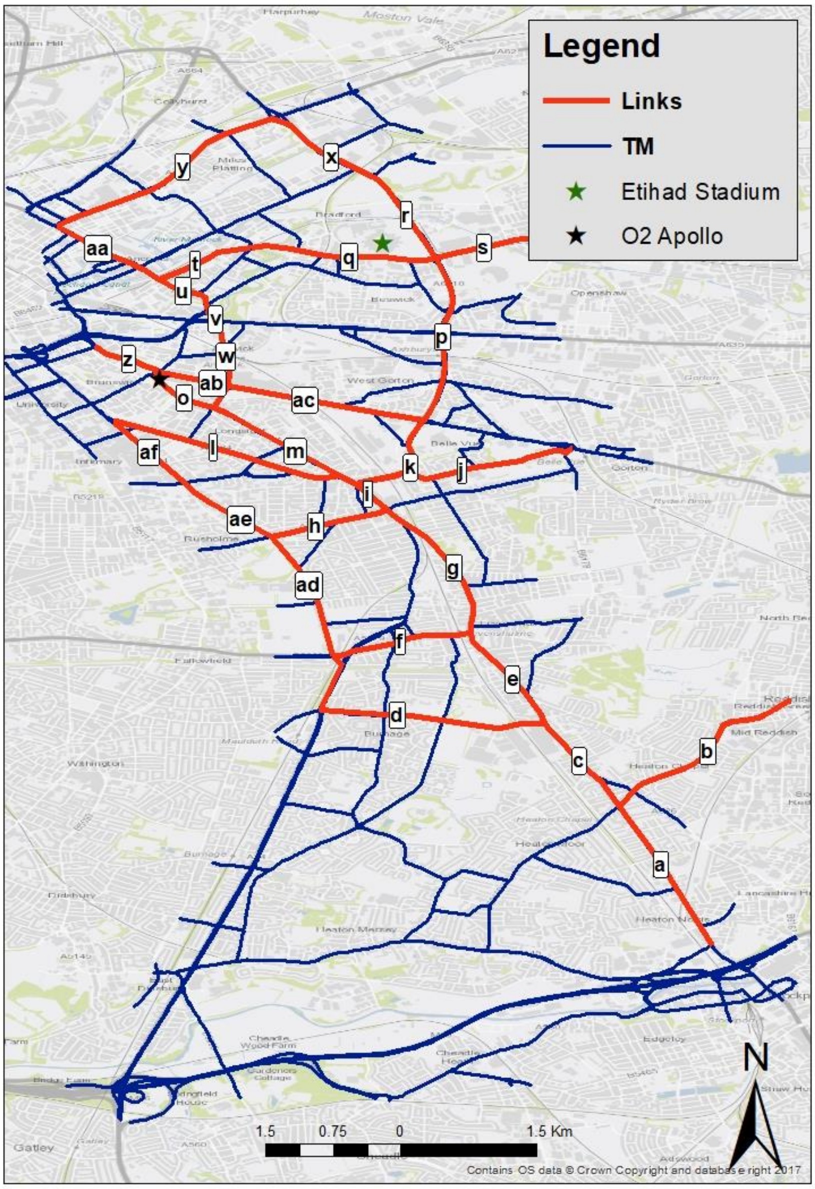


Figure 60: Manchester's neighbourhood network topology (Contains OS data © Crown copyright and database right (2017))

6.2.2 Dataset

This dataset used in these series of experiments is a subset of the MUCD dataset and consists of data for the following links on the A6 ({a}, {c}, {e}, {g}, {i}, {m}, {o}) which can be seen in Figure 60. Furthermore, the data subset uses three primary data sources (Bluetooth sensors, inductive loop counters, and accident data). Table 9 shows the name of the three data sources, where the data was extracted from, and the relevant location for the data, the time frame the data was available from at the point of the experiments, and finally the dimensions in relation to the data sources.

Table 9: Data Sources

Data	From	Location	Timeframe	dimension
Bluetooth	TfGM	Manchester, UK	2016-Current	Journey Time
Inductive Loop Counter	TfGM	Manchester, UK	2015-Current	Traffic volume
Accident Data	STATS19	UK	2005-current	Casualty accidents only

Table 10 shows a more in-depth description of the MUCD dataset, which describes the data, the data type, and the valid values.

Table 10: Subset of the MUCD dataset data dictionary

Field Name			Description	Data Type	Valid Value
Date			Date of the observation.	Date	DD/MM/YYYY
Day			Day of observation.	Character	May only contain letters, digits, and periods with limited variable length.
Time			Time of observation. Each recorded observation is in 15-minute intervals.	time	Lowest value: 00:00:00. Highest value: 24:00:00.
Link	Upstream		Each link represents a section of a road. The journey time between two Bluetooth sensors at the start and end of each Link is recorded in both directions called upstream and downstream.	Integer	0.00…999,999.99
	Downstream				
ATC	1	Volume	Total volume count at Sensor	Integer	0.00…999,999.99
	…	Volume			
	5	Volume			
Accident			Did an injury accident occur	Integer	0,1

6.2.3 Methodology

The following steps were taken to conduct the empirical experiments:

- A subset of data was extracted from the MUCD dataset.
- Define a suitable distance measurement for k-means++. The distance chosen for measuring the distance between the data point was Equation 14: Euclidean Distance.
- Conduct experiments to determine the optimal number of clusters using the silhouette method. Five clusters were selected as they had the highest silhouette score of 0.609.
- Analyse the output to ascertain whether or not the clusters represent the expected characteristics. For instance, does the expected clusters appear at the expected time of day and provide linguistical value.

$$d(a,b) = \sqrt{(a_x - b_x)^2 + \left(a_y - b_y\right)^2}$$

Equation 14: Euclidean Distance

Equation 14 shows how the distance between two points in the k-means++ experiments are measured. Where a and b are both points on a multiple dimensions plane relating to time of day, day of week, and journey time. Where x and y are the axis in each of the series of experiments.

6.2.4 Results

The purpose of these experiments was to use unsupervised learning to identify characteristics of urban road congestion and to gain qualitative information from quantitative data, such as the journey time and traffic volume discussed in section 6.2.2. Figure 61 displays 13 weeks of data in a scatter graph with the following weekdays: Tuesday, Wednesday, and Thursday across the x-axis, the observed journey time along the y-axis, and each data point being grouped into four time periods are 6:0, 7:00, 8:00, and 9:00. Each period group represents a 15-minute interval. For example, 6:00 until 6:15.

From Figure 61, it is apparent that the group 6:00 and 7:00 are a lot more consistent with regards to journey time than 8:00 and 9:00 which appear to have a lot more variation ranging from 0 to 2600 seconds. 9:00 is positioned sparsely between 7:00 and 8:00 demonstrating a visible temporal pattern in the journey time data. For instance, apart from a single outlier the 6 am and the 7 am times are the most clustered and have the quickest journey times throughout the morning, this is because the network is less occupied and allows for the traffic behaviour to be considered free flow compared to 8am and 9am where more vehicles are using the network at the same time causing a larger variation of journey times. Therefore, the visible temporal pattern demonstrates, it is easy to predict the expected journey time for pre-recurrent ‘am rush hour’ congestion.

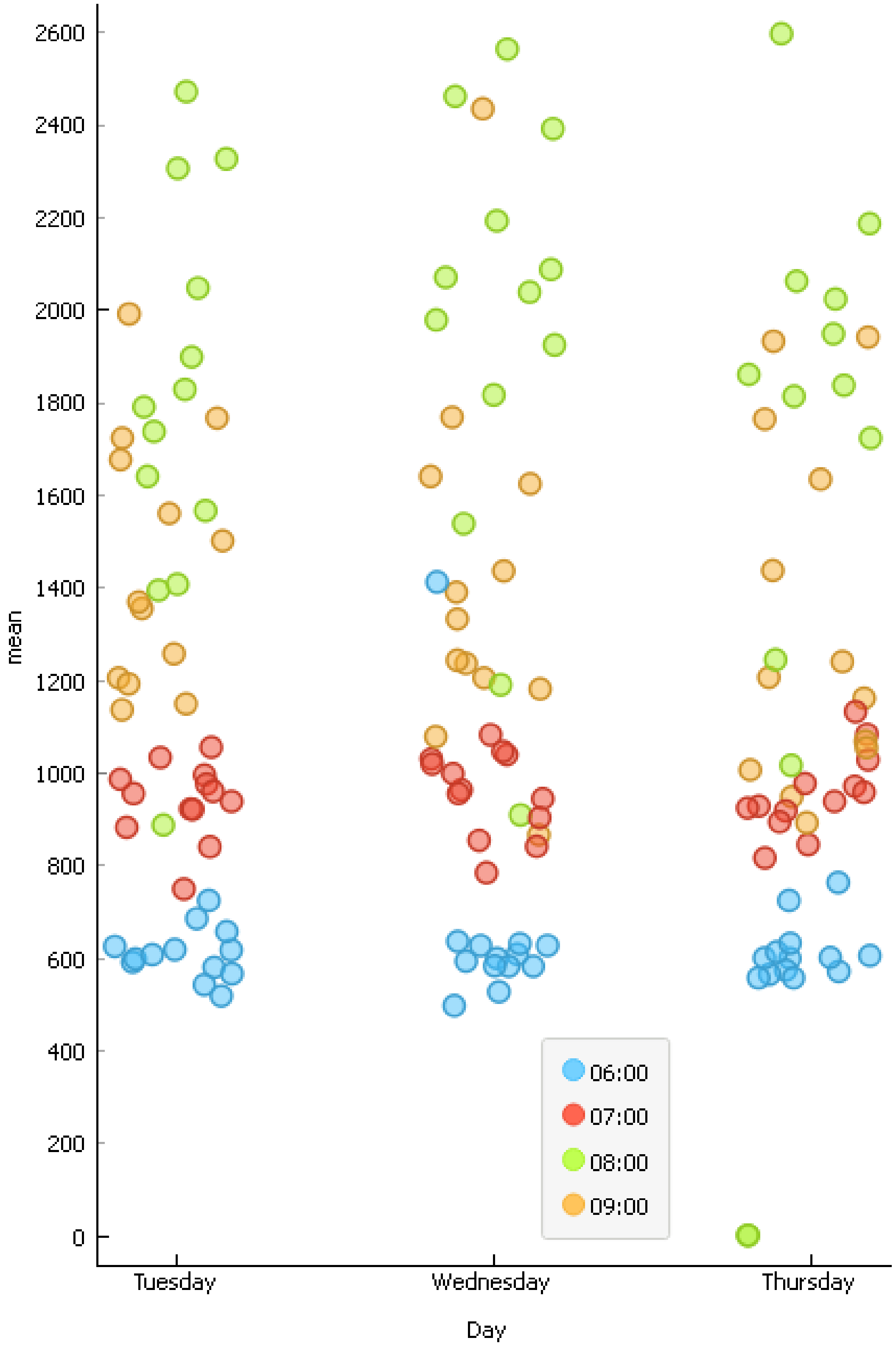


Figure 61: Scatter graph of journey times

Following the exploration of journey time in Figure 61, the next phase was to try to prove whether hypothesis one is true or not. This was achieved by clustering the journey time data for a six-month period over a 24-hour period and Figure 62 is the outcome.

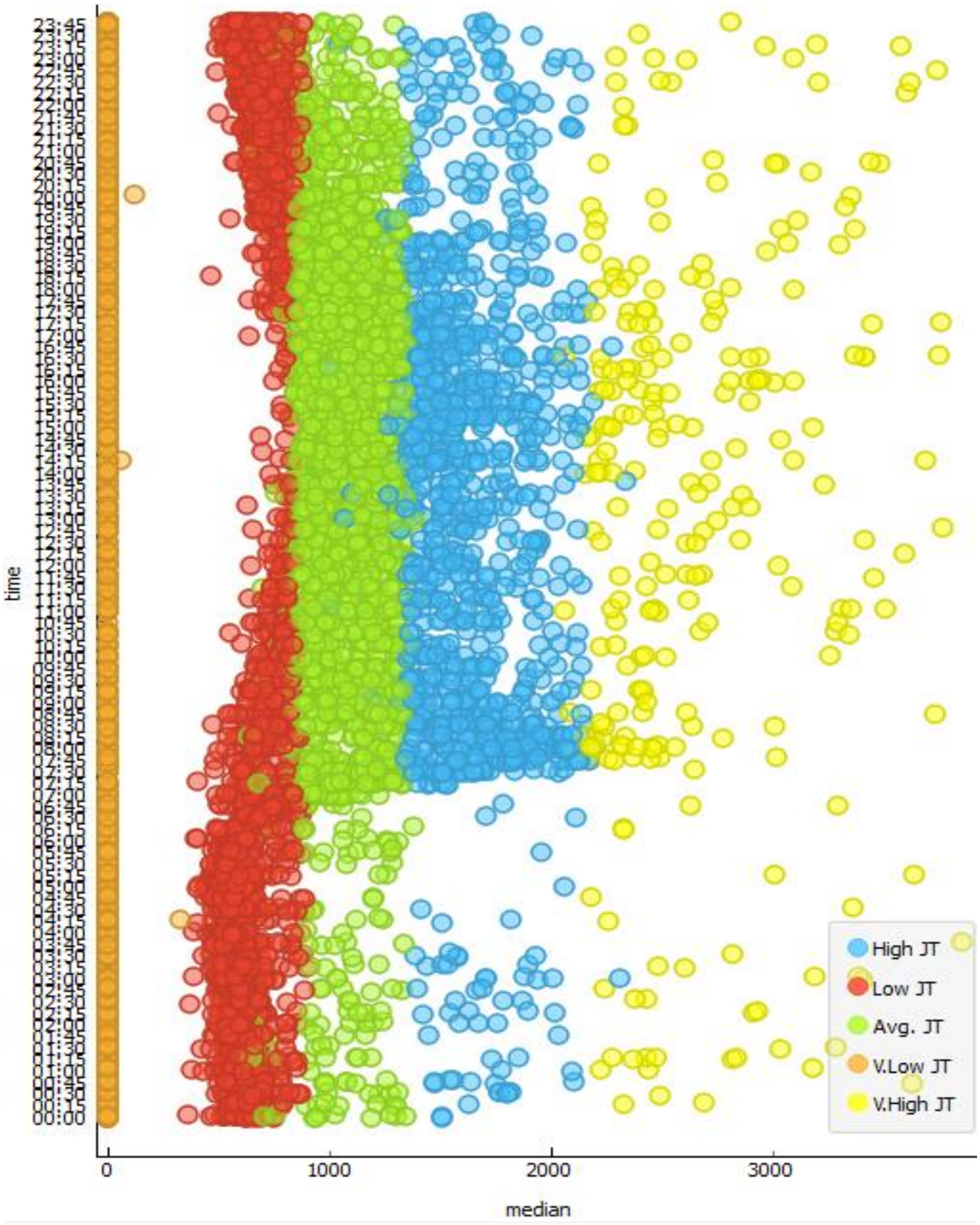


Figure 62: Clustered journey time into five categories 1) V. High JT 2) High JT 3) Avg. JT 4) Low JT 5) V. Low JT

Figure 62 was produced by using the k-means++ algorithm to choose the initial centroid, in addition, Euclidean distance (Equation 2) was used for calculating the distance between points. Five clusters were selected as they had the highest silhouette score of 0.609 which indicates that the data is well-distributed and 'far away' from its nearest cluster. Additionally, using five clusters provides a good level of resolution and resolution is vital to be able to prove hypothesis one true because without the ability to classify journey time into meaningful classes it would be impossible to predict the level of journey time.

In Figure 63 clustering has been used for the purpose of visualising the relationships between the five classifications of journey time, Very High, High, Average, Low and Very Low. The five classes were based on the silhouette score. The size of each observation relates to the volume of traffic on the network at the same time as the journey time observation. Figure 63 has many interesting patterns, such as the journey time between 00:00 and 06:30 remained dense in the Very Low or Low journey time. Then as expected between Monday to Friday around 7am the journey times become more average and around 8am high journey times become more prominent, which is expected behaviour considering the definition of recurrent congestion: “Occurs when significant amounts of vehicles simultaneously use a limited road space, such as on a weekday morning and afternoons peak hours’ traffic jam situations.”.

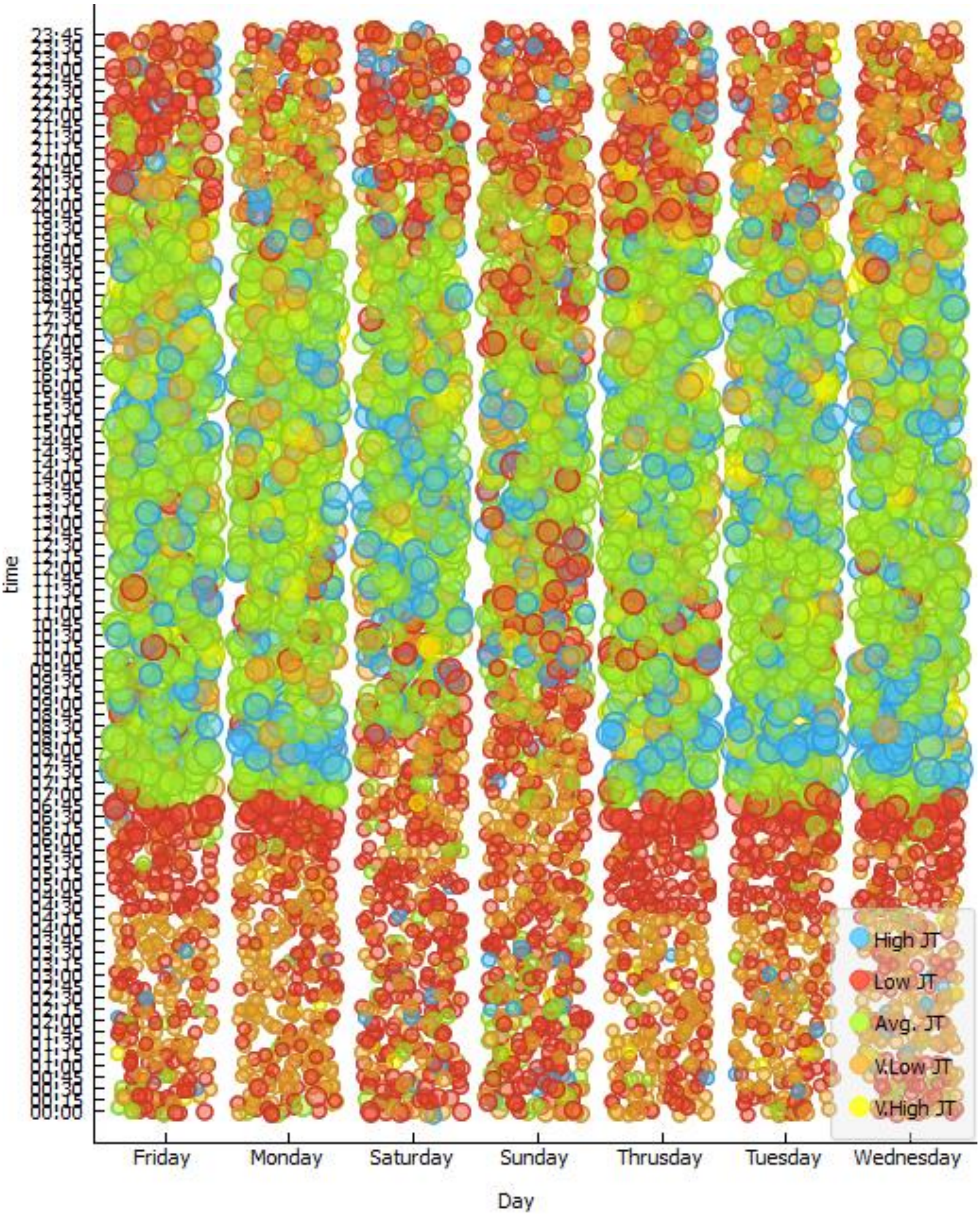


Figure 63: Clustered daily journey times into five categories 1) V. High JT 2) High JT 3) Avg. JT 4) Low JT 5) V. Low JT

To test hypothesis two, a different approach was used with regards to the visualisation of the data. In Figure 63, the x-axis is used for all 7 days of the week and the y-axis is used for time of day in 15-minute intervals. The size of each point is used to refer to the traffic volume.

Figure 63 shows it is possible to use clustering to identify differences between weekdays and weekends. For example, on Saturday and Sunday, there are longer periods of lower journey times and fewer vehicles using the road in the morning. In addition, on Monday, Tuesday, Wednesday, and Thursday there is a noticeable High journey time at around 8:00 each morning, which is expected because people are going to work and dropping children off at school. Finally, it is worth noting the volume levels typically become high at 7 am during the week and does not reduce until around 8 pm proving hypothesis two true (H_B0) "*Clustering an unsupervised dataset creates clusters that make it possible to identify differences between a weekday and a weekend*".

Proving both hypotheses true (H_A0) "*Clustering an unsupervised dataset creates clusters that make it possible to predict journey time*" and (H_B0) "*Clustering an unsupervised dataset creates clusters that make it possible to identify differences between a weekday and a weekend*" is vital when it comes to defining the differences between the consequence of a recurrent event such as morning rush hour and non-recurrent event such as a road traffic accident. It is also important to TfGM to be able to identify the differences between the spike in journey time and a reduction in traffic volume caused by both of these congestion types.

6.3 Case Study

A case study was chosen to attempt to answer the following research question (RQ2) - "*Can quantitative Big Data be used to provide qualitative information in conjunction with a road traffic ontology with the support of Machine Learning?*"

The case study will look at a fatal road accident on the A6 on the 7th of February 2017, using the data sources mentioned in section 6.2.2. Figure 64 was populated using python and shows the mean journey time, the time of the day in 15-minute intervals, and the (road) accident (the green line). The first (top) graph is the day of the accident, and the second (bottom) graph is the mean of 13 weeks (January until March 2017). Looking at Figure 64, there is a noticeable difference at the time of the fatal accident between the average journey time, which is around 1200 seconds (low journey time), and the day of the fatal accident that fluctuates between either 0 seconds (very low journey time) and around 3500 seconds (very high journey time).

For a road user, these values mean very little but after using the clusters created in the experimental analysis, we can say the journey time has changed from a low journey time to either no journey time (road closed) or a very high journey time state due to diversion, which lasts for around three hours overall before returning to the expected journey time. Furthermore, examining the 8 am period, both the average journey time and the single day are both average journey times according to the classification from the clustering which matches up with the typical behaviour of recurrent congestion. Both these examples of congestion (non-recurrent and recurrent) and the measurements match up to what was identified in the road accident ontology within the URCC.

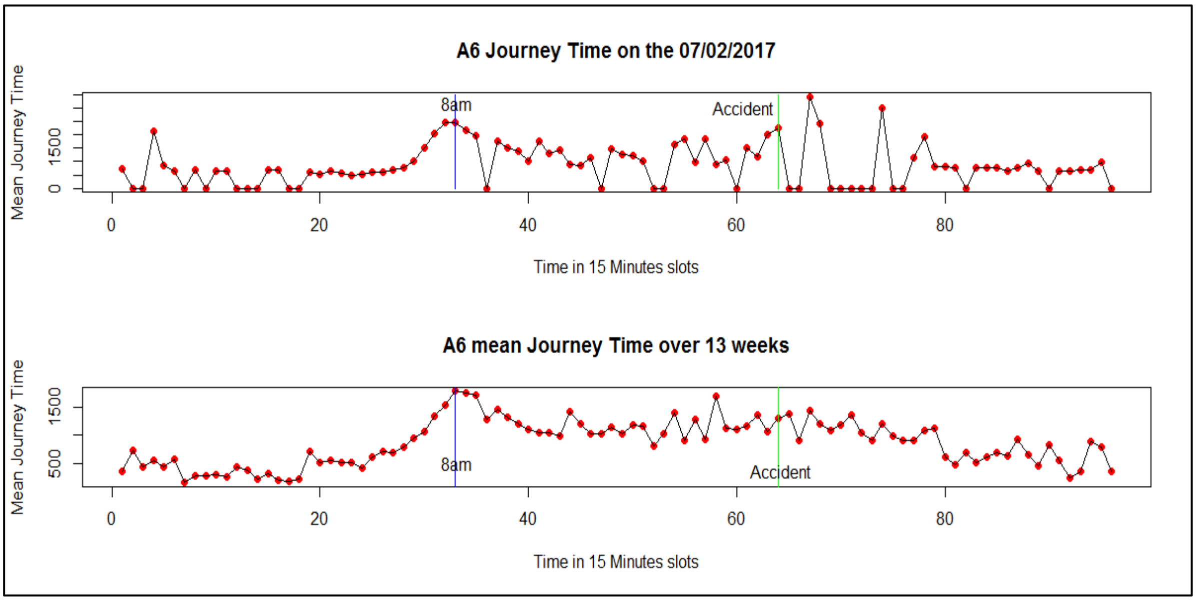


Figure 64: Journey time on the a) 7th February 2017 b) over a 13-week period

6.4 Chapter conclusion

This chapter has introduced a series of empirical experiments which uses a subset of data from the MUCD dataset in conjunction with the URCC and its main component (the universal ontology of road congestion) to prove both the hypothesis and answer the research question (RQ2) - “*Can quantitative Big Data be used to provide qualitative information in conjunction with a road traffic ontology with the support of Machine Learning?”*

This chapter has demonstrated that by interpolating the outcome of the series of empirical experiments it is possible to prove both hypotheses.

H_A0: Clustering an unsupervised dataset creates clusters that make it possible to predict journey time.

H_B0: Clustering an unsupervised dataset creates clusters that make it possible to identify differences between a weekday and a weekend.

And in turn, demonstrated that it is possible to take *quantitative* data and extract *qualitative* information, which can be provided to the stakeholders, such as road users or transport managers. The stakeholders could then use the meaningful information to make better decisions. Therefore, contributing to answering RQ2.

However, despite the promising results, further work is required to establish whether it is possible to identify similar patterns within the larger MUCD dataset and be able to predict the different types of road congestion using a rule-based system such as a fuzzy decision system. The feasibility of using a fuzzy decision-making system in this context is explored in chapter seven.

Chapter Seven: Validating the conceptual model using a fuzzy decision-making system.

7.1 Introduction

This chapter describes the conceptualisation, design, and implementation of two fuzzy-based decision-making systems, which have been designed to validate the Urban Road Congestion Conceptual (URCC) model (described in chapter three). Both fuzzy systems are novel contributions of the work presented in this thesis. First, a binary fuzzy decision-making system is proposed which focuses on classifying a binary output between congestion (recurrent congestion, non-recurrent congestion, and semi-recurrent congestion) and non-congestion using only journey time and traffic volume as the inputs. This work has been published in

- Abberley, L., Crockett, K. and Cheng, J., 2019, April. Modelling Road Congestion Using a Fuzzy System and Real-World Data for Connected and Autonomous Vehicles. In 2019 Wireless Days (WD) (pp. 1-8). IEEE.

Figure 65 shows an overview of the methodology for determining if it is possible to use the binary fuzzy decision-making system for predicting urban road congestion. The steps were as follows:

- Extract a subset of the real-world spatial-temporal dataset, known as the Manchester Urban Congestion Data (MUCD) dataset. This subset of data is then processed and store within a database. (Chapter four).
- The MUCD subset was then labelled using the Red, Amber, and Green (RAG) method proposed by Transport for Greater Manchester (TfGM) (Chapter five).
- Membership functions within the Fuzzy decision-making system were designed using clusters obtained from experiments in chapter five.
- The MUCD subset was then partitioned into two sets: training and test. The training data was used to create both a decision tree and naïve bayes models. The test data was then used to predict if congestion has occurred or not against all three types of machine learning: fuzzy decision-system, decision tree, naïve bayes (Section 7.3.6).
- Predictive results are then compared using several statistical measurements, such as True Positive Rate (TPR), False Positive Rate (FPR), Precision, F-measure, and Efficiency (Section 7.3.7).

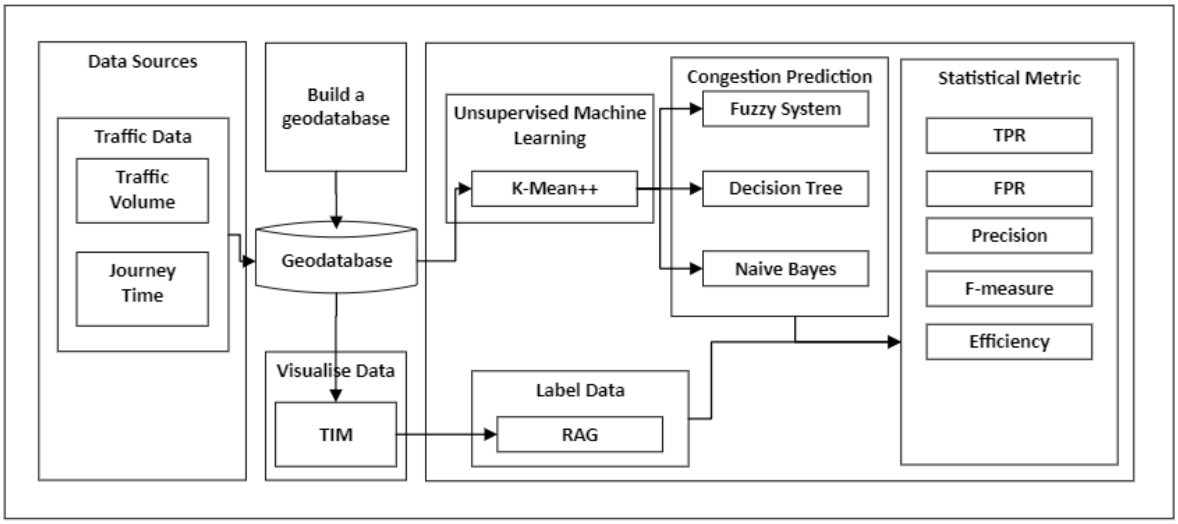


Figure 65: Methodology for the binary fuzzy decision-making system

The second fuzzy decision-making system will focus on classifying a multi-classification output between recurrent congestion, non-recurrent congestion, semi-recurrent congestion, and non-congestion using journey time and volume, time-of-day, day-of-week, and distance from attractor as the inputs.

Figure 66 shows an overview of the methodology for determining if it is possible to use the multi-classification fuzzy decision-making system for predicting urban road congestion. The steps are as followed:

- Extract a subset of the real-world spatial-temporal data from MUCD dataset. This subset of data is then processed and store within a database. (Chapter four).
- The MUCD subset was then labelled using the definitions defined in the conceptual model (Chapter three) and the expert defined method, such as RAG which was proposed by Transport for Greater Manchester (TfGM) (Chapter five).
- The multi-classification Fuzzy decision-making system was built using a percentile model to standardise the journey time and traffic volume data. This standardised data was then used to create the required fuzzy membership function (Section 7.4.1.4).
- The MUCD subset was then partitioned into two sets: training and test. The training data was used to create both a decision tree and naïve bayes models. The test data was then used to predict what type of congestion has occurred against all three types of machine learning: fuzzy decision-system, Decision tree, Naïve Bayes (Section 7.4.1.8).
- Results are then compared using several statistical measurements, such as Recall, Precision, F-measure, and weighted average (Section 7.4.1.9).

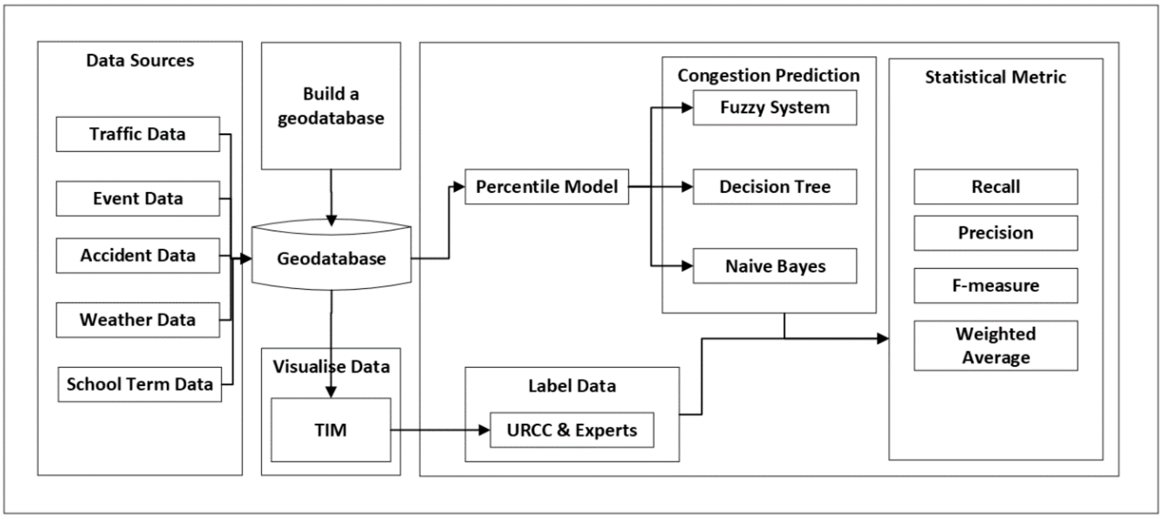


Figure 66: Methodology for the multi-classification fuzzy decision-making system

The experiments described in this chapter attempt to answer the following research question (RQ4) - "*Can a fuzzy rule-based system be designed to predict road congestion through validation of the Urban Road Congestion Conceptual (URCC) model?"*

7.2 Fuzzy systems in transportation

For centuries, we as people have naturally been migrating from rural to urban areas. This natural occurrence of urbanization has contributed to one of the biggest challenges' societies faces each day, which is road congestion. Road congestion in urban areas is estimated to cost the UK economy a total of £307 billion by 2030 (Djahel, Jones, Hadjadj-Aoul, et al., 2018). Furthermore, road congestion contributes enormously to damaging the environment, due to air pollution which has an impact on people's well-being (Gould and Abberley, 2017; Rui et al., 2018).

In an attempt to reduce the impact of road congestion, many large corporations, such as Google, Tesla, and Uber are developing *'smart vehicles'*, such as connected and autonomous vehicles (CAVs) that will be implemented as part of an Intelligent Transport System (ITS) of the future. Smart vehicles are expected to reduce congestion levels and the number of fatal accidents on the roads, with an estimated 37,000 lives a year predicted as being saved in the United States (U.S.) alone (Mudge et al., 2018). This is due to a smart vehicles ability to communicate faster than a human and make better decisions based on information collected by sensors embedded within the vehicles and infrastructure (Djahel et al., 2015). However, due to the limited access to these smart vehicles and their associated infrastructure, this study will use alternative data sources, which comprise of data similar to what is collected by CAVs and Roadside Units (RSUs) that will be used within an ITS of the future, such as a VANETs. Furthermore, these types of ITS will provide data from vehicle to vehicle (V2V) and vehicle to infrastructure (V2I) providing a constant stream of big data (Isa et al., 2014; Djahel et al., 2015; Golestan et al., 2015), which can be used to provide different information, such as volume, journey time, speed, and weather conditions, which are also known as dimensions.

Little work has been conducted using fuzzy systems to model urban road congestion (Pongpaibool et al., 2007; Li et al., 2018; Sun et al., 2018; Amini et al., 2021; Singh et al., 2021; Toan and Wong, 2021). However, this limited work has indicated that fuzzy models of road congestion are better for a stakeholder, such as a domain expert to understand that the conventional quantitative models previously implemented, such as the probability model (Li, 2015) and the spatial-temporal model (Anbaroğlu et al., 2015). Fuzzy sets are the ideal choice for modelling road congestion because of their ability to handle the ambiguity, multifaceted nature, and uncertainty within traffic data. They have the ability to capture such characteristics through the use of linguistic variables and hedges which are easier for a domain expert to understand (Zadeh, 1968).

7.2.1 What is a fuzzy decision-making system?

A Fuzzy decision-making system is a typical control system based on fuzzy logic (Chen et al., 1993; Xuan, 2022). The term "fuzzy" refers to the system's ability to deal with terms that are not binary or predefined and often referred to as linguistic variables. For instance, a humans' understanding of the phrase, near or far, could imply very near, near, not near, far, and very far depending on the context and the environment. Hence, fuzzy terms are subjective and mean different things to different people. The main advantage of a Fuzzy decision-making system is that the model itself is made up of a number of fuzzy rules, which can model a problem, such as urban congestion and the model can be expressed in terms a human operator can understand.

7.2.1.1 Methodology for a fuzzy decision-making system

The focus of this section is on the development of a fuzzy decision-making system using the Mamdani fuzzy inference system method developed in 1975 (Mamdani and Assilian, 1975). Mamdani fuzzy inference systems are typically applied to control-based problems. For example, manufacturing (Pourjavad and Mayorga, 2019), Supply chain management (Pourjavad and Shahin, 2018), groundwater prediction (Saberi et al., 2012) and smart city control problems (Iqbal et al., 2018).

Mamdani is one of two main fuzzy inference systems used in control base sceanarios. The second system is known as Sugeno (Takagi and Sugeno, 1985; Sugeno and Kang, 1988) and was developed in 1985. Mamdani and Sugeno both vary somewhat in how the outputs are determined. One of the main differences is the way the fuzzy rules are determined. Mamdani uses a set of linguistical control rules obtained from expert knowledge, whereas Sugeno uses a systematic approach for generating the rules from a given input-output dataset. Other differences are, for Mamdani there is an output membership function. However, Sugeno has no output membership function. Mamdani maintains a high level of interpretability due to the logistical nature of the control rules. However, due to the systematic approach Surgeno uses to creating the rule set, there is a loss in the interpretability of the output.

Both fuzzy inference systems have their own merits, however, due to TfGMs requesting the ability to be able to interpret the output and gain better understanding, it was decided the Mamdani approach would be used. Figure 67 shows the method for creating a Fuzzy decision-making system which consists of six stages and is described as follows:

1) The first step is to determine a set of fuzzy rules.
2) The second step requires fuzzification of the non-fuzzy input (crisp). The inputs are fuzzified according to the determined membership functions.
3) The third step is to combine the fuzzified inputs in according to the fuzzy rules defined in step one and establish a rule strength.
4) The fourth step is to calculate the consequents of each rule by applying a fuzzy operator (and/or/not) to the antecedents (If-Then). For instance, the top rule has two parts in the antecedent, so an AND operator was used to identify the minimum value as the result.
5) The fifth step is to aggregate the consequences to get a single output distribution.
6) The sixth and final step is to defuzzify (using the centroid of area method). The single defuzzified output will be a crisp value. Although, if a crisp classification is not needed, then this step can be skipped.

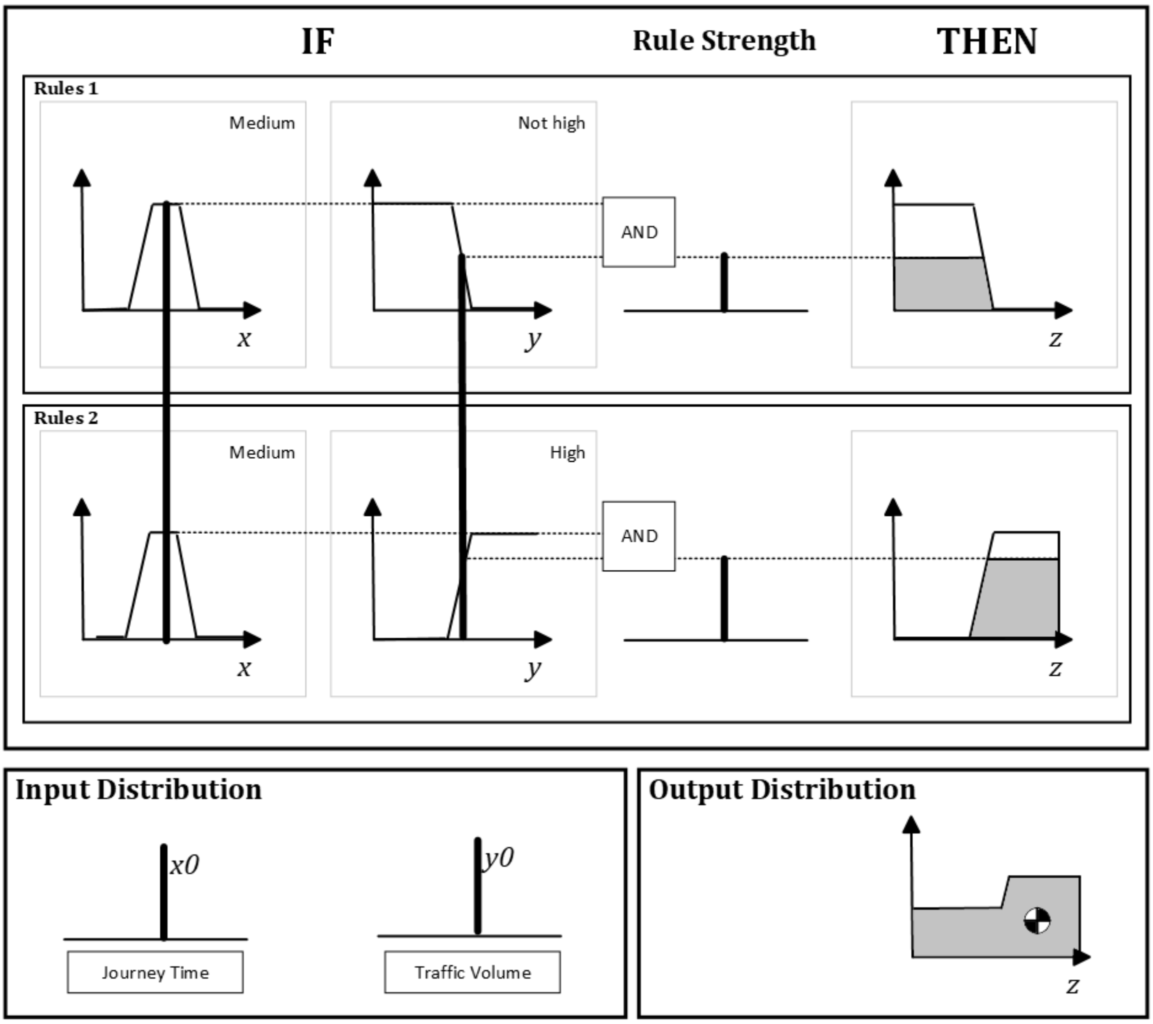


Figure 67. Mamdani inference system [Source: Author]

In the work presented in this thesis, the binary (section 7.3) and multiclassification (Section 7.4) fuzzy decision-making systems have been developed using the six stages outlined In Figure 67. Both systems will be compared against two other algorithms. The first is a decision tree and the second is a probabilistic algorithm. The performance of all models will be validated using several statistical measurements, such as recall, precision f-score, etc. mentioned in Figure 65 and Figure 66.

In addition to the methodology mentioned above, during the crucial stages, such as determining a set of fuzzy rules, determining the membership functions (fuzzification), and determining the final classifications or crisp value (defuzzification) TfGM will be periodically testing both systems, providing expert knowledge and feedback to aid the calibration of the fuzzy control rules to manually optimise the overall performance. Furthermore, to assist with calibrating the fuzzy control rules, empirical experiments were conducted, such as using every combination of rules sets and using subsamples of data to ascertain whether the output was expected or not.

7.2.2 Transport application

The approach to using a Fuzzy decision-making system within the discipline of transportation to classify urban road traffic congestion is relatively new with very few papers primary focus being on congestion. For instance, a study (Bauza et al., 2010) into cooperative a vehicle to vehicle (V2V) road traffic detection congestion on freeways. This study uses a level of service metric created by a third party who collected aerial surveys to define the levels of congestion: slight, moderate, and severe. The author then created a new metric that uses four membership functions: Very Slow, Slow, Medium, and Fast, two inputs: Speed and Density, and 16 rules to define an output for one of three levels of congestion. However, this study does not consider non-congestion as an output and has reported only using the model in a simulation with simulated data, furthermore, the focus of the study is on highways and does not reflect an urban road network, which has very different characteristics.

Another study (Li et al., 2018), investigates road traffic anomalies that contribute to congestion at a single junction using a one-way traffic video sequence. This study uses two data inputs: Traffic flow and traffic density. Traffic flow has three membership functions called low, medium, and high. These functions are calculated using linear increasing, decreasing, and trapezoidal-shaped membership functions with the fuzzy boundaries calculated using $\mu \pm \sigma$ and $\mu \pm 2\sigma$. Where μ defines the mean and σ defines the standard deviation. Traffic density also has three membership functions, which are sparse, normal, and dense. These memberships are calculated using a statistical analysis of the pixels. This study uses a total of nine rules, which were obtained through experience and experiments. The output classifications were either: Normal traffic, slight congestion, and heavy congestion. The main limitation of this experiment was only evaluating on three different scenes and in total only had 142 observations. Another limitation was in the results, where the authors only report the accuracy, false detection rate, and the ‘average’ of three scenes without calculating an average. The results claimed to achieve 100 per cent accuracy for normal traffic but claims a 0.11 per cent false detection rate, which contradicts the 100 per cent claim. Furthermore, the slight congestion classification had an accuracy of 93.4 per cent and heavy congestion had an accuracy of only 72.2 per cent.

Reviewing recent literature (Amini et al., 2021; Singh et al., 2021; Toan and Wong, 2021) have demonstrated some of the limitations of trying to model urban road congestion, which includes the unavailability to obtain good quality data. Therefore, they have either simulated their own data, not used any data, or chosen to investigate highways using toll data to count the number of vehicles entering a zone. Another limitation of these studies was they developed their respective fuzzy systems in a simulated environment which would not be transferable in a real-world environment.

7.3 Binary fuzzy decision-making system

The development of a binary Fuzzy decision-making system contributes to knowledge by creating a novel way to predict urban road congestion. Additionally, the use of an unbalanced real-world big dataset is rare, as majority of literature use synthetic balanced datasets. Furthermore, the use of domain experts’ knowledge to construct a fuzzy model of road congestion requires no training data for the model to learn from unlike traditional machine learning models which require training and test data, some even require validation data. The Fuzzy decision-making system is achieved through the construction of a set of fuzzy membership functions and fuzzy rules that can be used to identify road congestion. An experiment is then conducted using the real-world data to determine whether the fuzzy model can be used to analyse traffic data to classify congestion. Comparisons are then made with an existing internal control centre system used by Greater Manchester Transport authority in the UK and other known classification algorithms.

7.3.1 Methodology: Binary fuzzy decision-making system for predicting urban road congestion

This section describes the methodology, which was used to develop a fuzzy system for road congestion on an urban city network. The model utilises real-world data from Bluetooth sensors and inductive loop counters provided by TfGM for Manchester, UK. These data sources will provide data, which is equivalent to what CAVs and RSUs would provide. Moreover, experts in road congestion management from TfGM and a road congestion ontology (Abberley et al., 2017; Gould and Abberley, 2017) were used to help define the fuzzy sets to ensure a thorough domain coverage.

The road congestion ontology which was used to support the development of a fuzzy system capable of classifying road congestion was presented in (Abberley et al., 2017). The road congestion ontology states that congestion can be measured using multiple dimensions, such as journey time and volume. Furthermore, congestion is often the consequence of an event, such as rush hour, a road accident, a concert, a football match, and roadworks. Finally, depending on the severity of congestion the magnitude can vary from very low to very high. Therefore, in this study, the magnitude ranges defined in the urban road congestion ontology (Abberley et al., 2017) will be used to determine the membership functions: Very low, VL, low, L, medium, M, high, H, and very high, VH which will ensure coverage of the domain.

7.3.2 Data sources and variables

A subset of the real-world spatial-temporal dataset, known as the Manchester Urban Congestion Data (MUCD) dataset was used. This subset consists of both journey time and traffic volume dimensions, which were collected from Bluetooth sensors and inductive loop counters and the subset has a total of 17376 records. Each record consists of two attributes and a classification that was created using the Red, R, Amber, A, and Green, G, (RAG) method implemented by TfGM, UK. Where red (Equation 10) and amber (Equation 11) are both congested and green (Equation 12) is non-congested.

The problems associated with the MUCD dataset has been discussed in section 4.6 and can be summarised as:

- Due to the limited number of inductive loops Traffic counters, the ability to calculate the volume of traffic for each link is limited.
- The data quality of the Bluetooth sensors has many issues. For example, capture rates; during the night periods or a period where no vehicle with a Bluetooth device passes the sensors cause the sensors to provide an incorrect average journey time when being observed.
- In bad weather, the sensors which use a mobile network to transmit the data to a central location, can fail and cause the dataset to have missing data.
- One class out significantly outweighs the other, causing the MUCD dataset classed as imbalanced, which cause challenges for machine learning classification algorithms. Since classification algorithms are often biased towards the majority class, which in this study is non-congestion.

7.3.3 Methodology for determining the membership functions and fuzzy rules

Stages 1 and 2 of the Mamdani methodology discussed in section 7.2.1.1 requires a set of rules to be determined and the data inputs to be fuzzified according to the determined membership functions. Therefore, it is important to set out a methodology for determining the membership functions and fuzzy rule set.

To assist with the initial determination of membership functions (section 7.3.3.1) and rules (section 7.3.3.2) an empirical approach was taken with the support of the urban road congestion ontology (Abberley et al., 2017), (Abberley et al., 2017), which is part of the URCC model discussed in chapter three. During the discussions with domain experts at TfGM, it was agreed the formal terms defined within the URCC model should be used to define the membership functions for journey time and traffic volume (due to their simplistic wording). This was achieved by using the magnitude concepts, such as very low, low, medium, high, and very high presented in the urban road congestion ontology.

Figure 68 shows the methodology for determining the membership functions and fuzzy rules which consists of nine stages and are described as follows:

1) Collect, process relevant data, and store the data in a database for ease of access.
2) The URCC model was initially used to determine the number of memberships and their names, however, empirical experimentation was conducted to determine the best approach for creating the membership functions.
    a. The approaches taken are: equal size memberships, using mean and +/- standard deviation one and two, and a unsupervised learning algorithm called K-means++.
3) Initially to determine the rule set, all combinations of inputs-outputs were used and through several iterations of empirical experimentation and feedback from TfGM a final rule set was determined.
4) Using the memberships and fuzzy rules, a subsample of data is used to evaluate the functionality of the fuzzy system.
5) Analyses of the subsample data.
6) Conducted empirical optimisation based on the analysis conducted in (5) to refine the membership functions over several iterations.
7) Review the membership functions, rules, and the outcome produced using the subsample data with TfGM using TIM for visualisation support.
8) Using the feedback from TfGM, a second stage of empirical optimisation is conducted to refine the rules over several iterations.
9) Once the membership functions and rules are optimised, perform predictions against the test data.

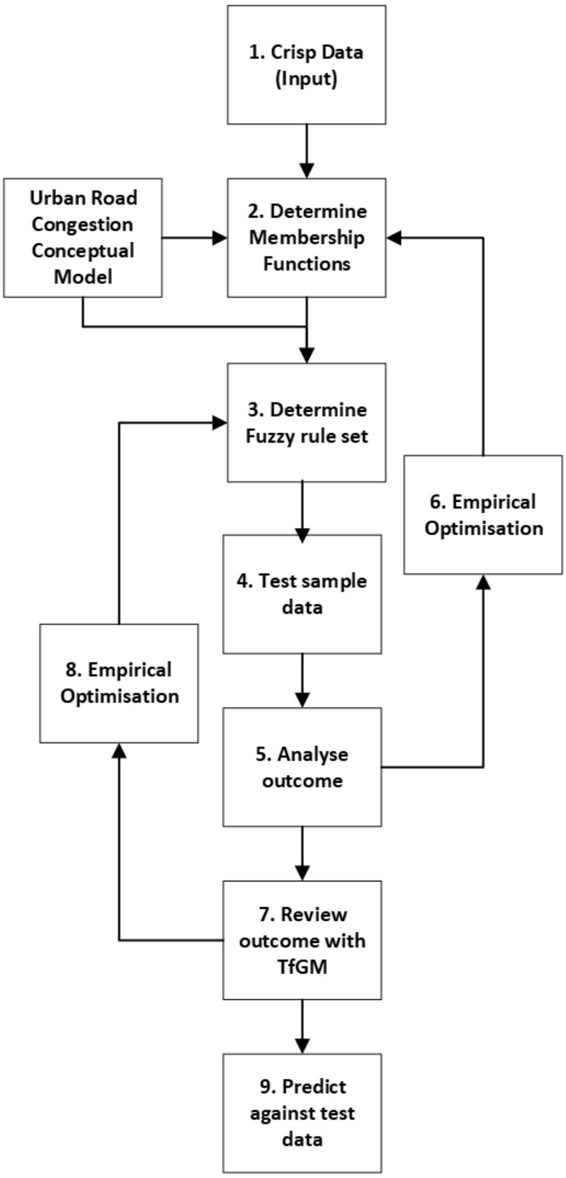

1. Crisp Data (Input)
Urban Road Congestion Conceptual Model
2. Determine Membership Functions
3. Determine Fuzzy rule set
6. Empirical Optimisation
4. Test sample data
8. Empirical Optimisation
5. Analyse outcome
7. Review outcome with TfGM
9. Predict against test data

Figure 68. Methodology for determining membership functions and rules [Source: Author]

7.3.3.1 Membership function determination

Table 11 shows the dimensions, data sources, and the linguistic values determined from the urban road congestion ontology (Abberley et al., 2017), (Abberley et al., 2017). The linguistic values of the membership functions representing journey time and traffic volume are also shown.

Table 11: Dimensions and their linguistic values

Dimension (Variables)	Data sources	Linguistic values (Membership functions)
Journey time	Bluetooth remote sensors	Very Low (*VL*) Low (*L*) Medium (*M*) High (*H*) Very High (*VH*)
Volume	Inductive loop counters	Very Low (*VL*) Low (*L*) Medium (*M*) High (*H*) Very High (*VH*)

Using the linguistic values identified in Table 11, the creation of the fuzzy membership functions can be performed using three steps:

- **Step 1**: Perform k-means++ clustering discussed in (Abberley et al., 2017). Section 5.4.3.2 provides the algorithm used to conduct k-means++ on both journey time and volume data.
- **Step 2**: Identify the final boundary values for a set of groups where they connect and define this value as dt.
- **Step 3**: Using the dt value, determine membership function domain coverage using one of three membership functions: linear up, linear down, and trapezoidal shape.

The primary objective of machine learning is to discover patterns within large datasets, such as the MUCD dataset used within this study. k-means++ clustering is an unsupervised algorithm used within machine learning to find a cluster of patterns in data. k-means++ uses the inherent structures in the data to best organise the data into groups of maximum commonalities (Aggarwal, 2013). This is achieved by partitioning n observations into k (in this study $k=5$) clusters. The use of five clusters was chosen based upon early empirical experiments, which found that five clusters provided sufficient resolution (Abberley et al., 2017).

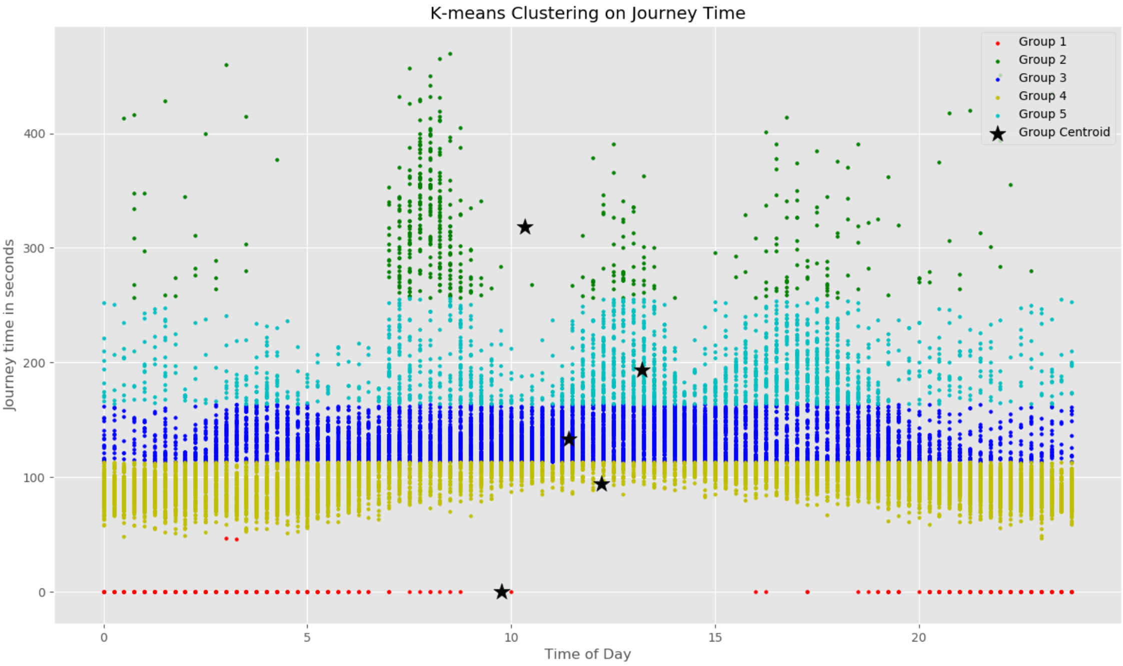


Figure 69: Example of k clusters where k=5 being performed on 6 months of journey time data

Figure 69 shows 17376 journey time records plotted on a 24-hour scale. Each observation within Figure 69 belongs to the cluster with the nearest mean value. Once k-means++ has been conducted, it becomes possible to identify the boundary values between each cluster, which will be used to create the membership functions in the fuzzy system.

Figure 70 shows an example pair of linear opposing membership functions, which will be used for the VL (very low) and VH (very high) memberships for both journey time and traffic volume. The two pairs Equation 15 and Equation 16 are both linear increasing and decreasing membership functions L, can be defined as (K. Crockett et al., 2006):

$$L \uparrow (x, dm, dn) = \begin{cases} 0, & x \leq dm \\ \dfrac{x - dm}{dn - dm}, & dm < x < dn \\ 1, & x \geq dn \end{cases}$$

Equation 15: Linear increasing membership function

$$L \downarrow (x, dm, dn) = \begin{cases} 1, & x \leq dm \\ 1 - \dfrac{x - dm}{dn - dm}, & dm < x < dn \\ 0, & x \geq dn \end{cases}$$

Equation 16: Linear decreasing membership function

Where dm is defined as $dm{=}dt{-}n\sigma$ and dt is the value generated by K-means clustering on all variable i records. n is a real number $n \rightarrow$ [0.0, ∞], σ is the standard deviation, and x is the value of the variable i. n is empirically determined. Additionally, dn is defined as $dn{=}dt{+}n\sigma$.

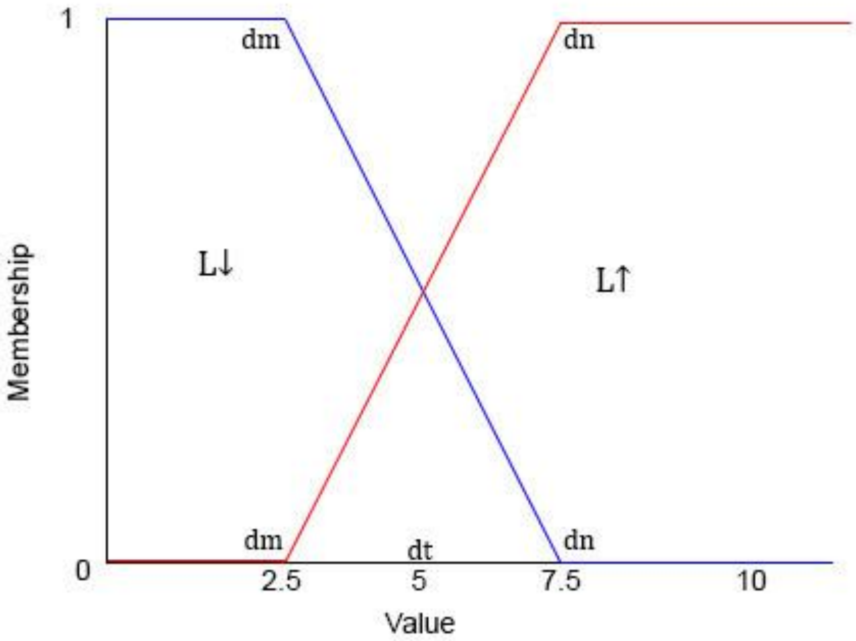


Figure 70: Example of a linear pair opposing fuzzy memberships functions

Figure 71 shows an example of a trapezoidal-shaped membership function, which will be used for the *L*, *M*, and *H* memberships. The trapezoidal-shaped membership function *T*, Equation 17, may be defined as:

$$T(x, dm^1, dn^1, dm^2, dn^2) = \begin{cases} 0, & x \leq dm^1 \\ \dfrac{x - dm^1}{dn^1 - dm^1}, & dm^1 < x < dn^1 \\ 1, & dn^1 \leq x \leq dm^2 \\ 1 - \dfrac{x - dm^2}{dn^2 - dm^2}, & dm^2 < x < dn^2 \\ 0, & x \geq dn^2 \end{cases}$$

Equation 17: trapezoidal-shaped membership function

Where *dm1*, *dn1*, *dm2*, and *dm2* are defined using the same method as *dm* and *dn*.

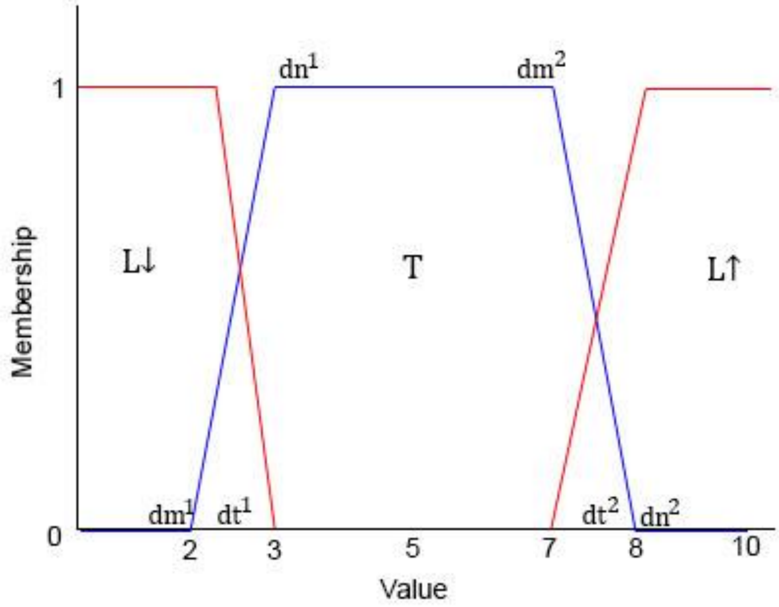


Figure 71: Example of a trapezoidal-shaped membership function

7.3.3.2 Fuzzy Rules Determination (manual and expert)

The fuzzy rules were initially created with every possible variant for each of the five membership functions, such as *VL*, *L*, *M*, *H*, and *VH* for journey time and volume. A total of 25 rules were created. However, the consequences of using this approach were observed in early analysis of the subsample data. It was observed from the early predictions that the results were not optimal due to more than expected false positives producing an overall weak performance. Therefore, with the support of the urban road congestion ontology (Abberley et al., 2017) and domain experts, TfGM (TfGM, n.d.), the rules were manually optimised down to just six. As a result of empirical optimisation, it was discovered that several rules were not firing correctly due to overlapping of rules and it was determined many rules were not relevant. For example, if journey time was *VH* then the output is congested regardless of the volume.

Algorithm 2 uses both antecedents and consequents membership functions to fire six unique rules to acquire each rule strength ready for fuzzy inference.

Algorithm 2
Rules for congestion.

Antecedents: Journey Time, JT. Traffic Volume, V.
Antecedents memberships: Very Low, VL. Low, L. Medium, M. High, H. Very High, VH.
Consequents: Congestion, C.
Consequents memberships: Congested, Con. Non-congested, Non.

```
$\boldsymbol{if}\ JT\ is\ VH\ \boldsymbol{then}$
    $C \leftarrow Con$
$\boldsymbol{if}\ JT\ is\ H\ \boldsymbol{then}$
    $C \leftarrow Con$
$\boldsymbol{if}\ JT\ is\ M\ \boldsymbol{and}\ V\ is\ VH\ \boldsymbol{then}$
    $C \leftarrow Con$
$\boldsymbol{if}\ JT\ is\ M\ \boldsymbol{and}\ V\ is\ not\ VH\ \boldsymbol{then}$
    $C \leftarrow Non$
$\boldsymbol{if}\ JT\ is\ L\ \boldsymbol{then}$
    $C \leftarrow Non$
$\boldsymbol{if}\ JT\ is\ VL\ \boldsymbol{then}$
    $C \leftarrow Non$
$return\ C$
```

7.3.4 Fuzzy inference

One of the first control systems and most commonly implemented methods for computing fuzzy inference is Mamdani (Mamdani and Assilian, 1975). Furthermore, Mamdani was first implemented within the transport domain, where it was used in an attempt to control a steam engine and boiler combination (Mamdani and Assilian, 1975). In this exploratory work on fuzzy systems, Mamdani inference was therefore selected.

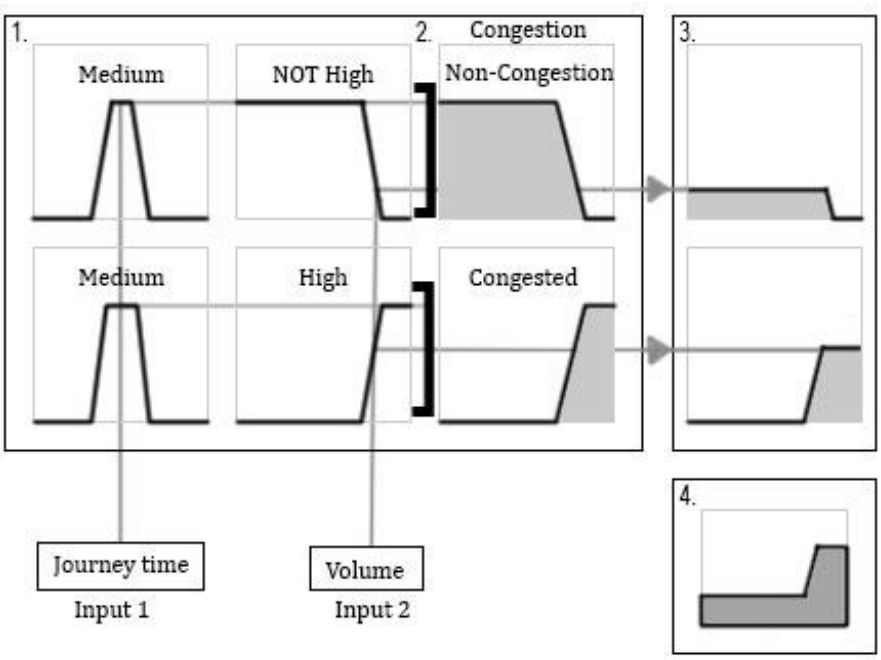


Figure 72: An example of how Mamdani fuzzy inferences works

Figure 72 shows the composition of fuzzy inference, the four stages are:

Stage 1. Fuzzification of the non-fuzzy inputs, which are crisp, numerical, and specific to the attribute domain. The inputs are fuzzified according to membership functions.

Stage 2. If the antecedent of a given rule has more than one part, the application of a fuzzy operator is required to obtain a single value that represents the individual rule. For instance, the top rule within Figure 72 has two parts in the antecedent, so a AND operator is used to identify the minimum value as the result.

Stage 3. Using the single value acquired in stage 2, the consequent is reshaped to provide the result of implication which is weighted depending on the linguistic characteristics that are attributed to it.

Stage 4. Aggregation is the combination of the fuzzy sets that represent the outputs of each rule into a single fuzzy set (fuzzy output distribution).

7.3.5 Defuzzification

The method centroid of area (COA), also known as the centre of gravity (COG) (Equation 18) is one of the most commonly used methods to defuzzify a fuzzy set (the output distribution membership in Figure 72) and output a crisp numeric value, which in this study is the probability of congestion. To achieve this, the total area of the output distribution membership is divided into a number of sub-areas and then the COA is calculated for each sub-area. Finally, all sub-areas COA are summed to find the defuzzied value (probability of congestion). The defuzzification using COA, Z^*, is the defuzzied value of the fuzzy sets, Z. Where $\mu_{\bar{A}}(z)$ is the degree of membership for the fuzzy set, where for all $z \in Z$.

$$Z^* = \frac{\int \mu_{\bar{A}}(z).zdz}{\int \mu_{\bar{A}}(z)dz}$$

Equation 18: Centre of gravity

7.3.6 Experimental Methodology

The aim of the experiment is to determine whether a fuzzy system can be used to analyse traffic data to classify congestion.

The following hypothesis will be evaluated.

Hypothesis

HA0: Using journey time and volume data, it is possible to classify congestion using a fuzzy system.

HA1: Using journey time and volume data, it is not possible to classify congestion using a fuzzy system.

To evaluate the performance of the fuzzy system, the binary fuzzy decision-making system was compared against two alternative machine-learning algorithms: The decision tree C4.5 (using the Weka implementation J48) (Weka, 2018) and naïve bayes. The decision tree C4.5 was chosen because C4.5 is explainable, and it would be useful to compare tree rules against fuzzy rules. Naïve bayes, which is a probabilistic classifier, was chosen because it is intuitive and simple, however, the performance is strong in many cases, and it manages all values independently. The statistical measurement to compare the three models are: True Positive Rate, False Positive Rate, Precision, F-score, and overall efficiency. All three models used the same MUCD subset, which was split into two parts: training that contains 8688 records of which 6665 were classified as non-congestion and 2023 were classified as congestion (accounting for only 23% of records). The test dataset contains the remainder of the dataset. Datasets were mutually exclusive.

In order to evaluate the three methods using an unbalanced dataset, five statistical measurements were chosen, which are: True Positive Rate, TPR, also known as recall and sensitivity. TPR measures the proportion of actual positives that are correctly identified. TPR is defined in Equation 19 where TP is a true positive, and FN is a false negative.

$$TPR = \frac{TP}{TP + FN}$$

Equation 19: True positive rate

False Positive Rate, FPR, measures the negative instance that is wrongly classified as positive. FPR is defined in Equation 20 where FP is false positive, and TN is a true negative.

$$FPR = \frac{FP}{FP + TN}$$

Equation 20: False positive rate

Precision, also known as a positive predictive value, PPV, measures the number of positive predictions divided by the total number of positive class values predicted. Precision is defined in Equation 21.

$$PPV = \frac{TP}{TP + FP}$$

Equation 21: Positive predictive value

F-measure, also known as F1 Score, $F1$, measures the balance between the precision and TPR. F-measure is defined in Equation 22.

$$F1 = \frac{2 * TP}{2 * TP + FP + FN}$$

Equation 22: F-measure

Overall efficiency, also known as accuracy, measures the amount of correctly classified instances. Overall efficiency is defined in Equation 23.

$$Overall\ Efficiency = \frac{TP + TN}{TP + TN + FP + FN}$$

Equation 23: Overall efficiency

However, due to the class imbalance as mentioned above, it is important to provide a single value that represents the performance of both classes for TPR, precision, and F-measure. To achieve this a weighted average will be used and is defined in Equation 24. Where C_{non} represents the statistical measurement being weighted for the class non-congestion. C_{con} represents the statistical measurement being weighted for the class congested.

$$W = \frac{C_{non} * (TP + FN) + C_{con} * (TN + FP)}{TP + FN + TN + FP}$$

Equation 24: Binary weighted average

7.3.7 Results and Discussion

The purpose of this study was to determine whether it is possible to classify road congestion using a fuzzy system and real-world traffic data. Table 12 shows the results for each statistical measurement for three machine-learning algorithms and their classes: Non-congestion (Figure 73), congested (Figure 74) and the weighted average of both classes (Figure 75).

Table 12: Results for Fuzzy System, J48, and Naïve Bayes

Experiment	Class	TP Rate / Recall	FP Rate	Precision	F-Measure	Overall Efficiency (%)
Binary Fuzzy System	Non	94.4	32.9	90.4	92.3	88
	Congested	67	5.5	78.4	72.2	
	Weighted Avg.	88	26.5	87.6	87.6	
Decision tree (J48)	Non	95.2	59	84.2	89.3	82.5
	Congested	41	4.8	72.1	52.3	
	Weighted Avg.	82.6	46.4	81.4	80.7	
Naïve Bayes	Non	99.8	57.8	85	91.8	86.3
	Congested	42.2	0.2	98.5	59.1	
	Weighted Avg.	86.4	44.4	88.2	84.2	

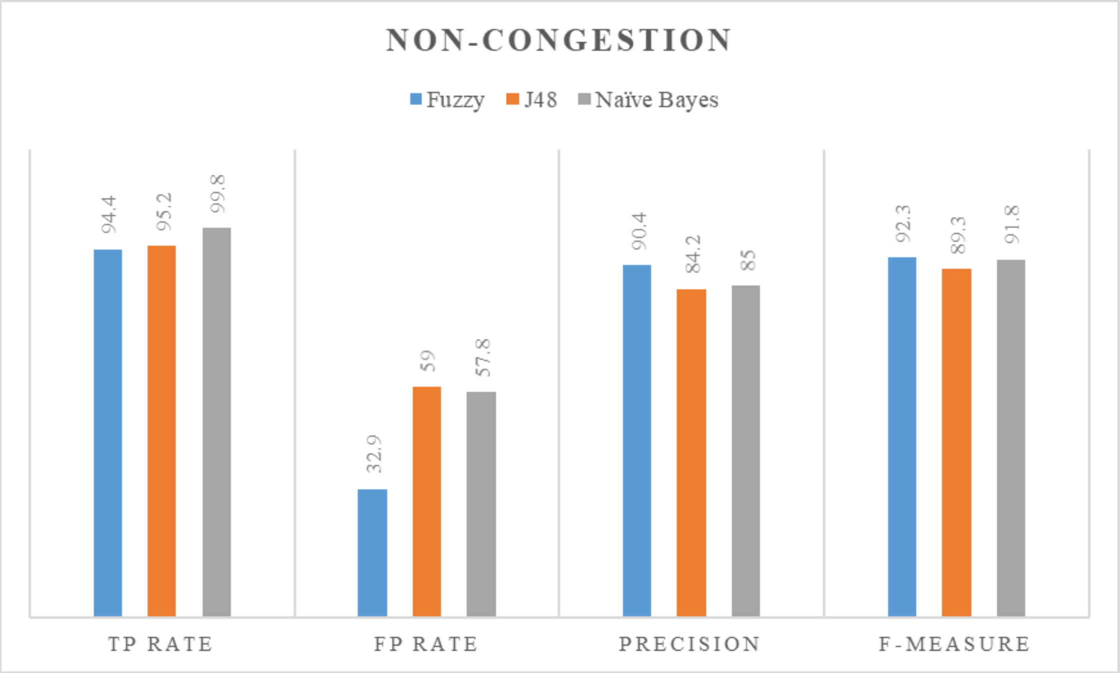


Figure 73 : TP rate, FP rate, precision, F-measure, and overall efficiency for non-congestion

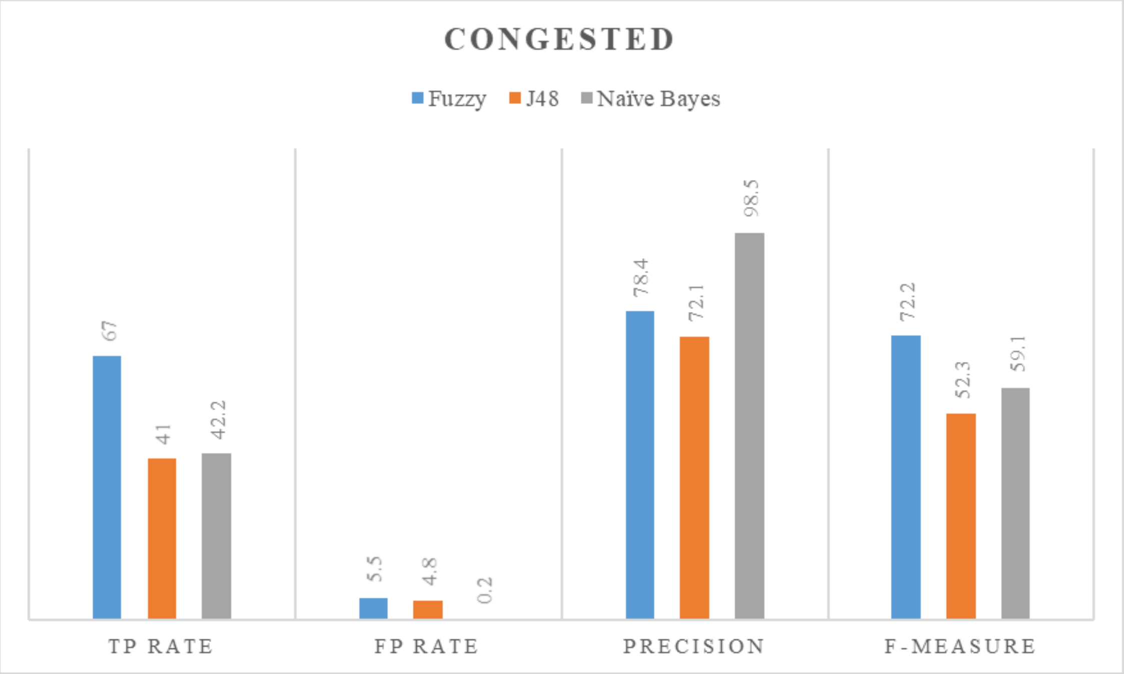


Figure 74: TP rate, FP rate, precision, F-measure, and overall efficiency for congested

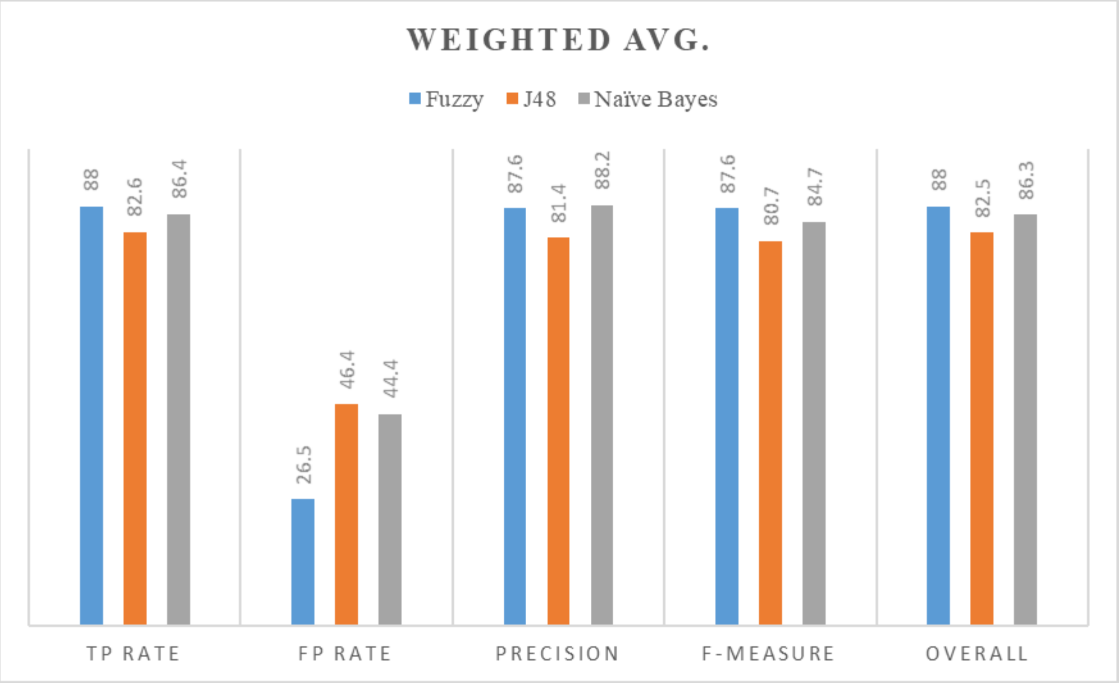


Figure 75: Weighted average of TP rate, FP rate, precision, F-measure, and overall efficiency

Before discussing the results, it is important to reiterate the challenges of performing classification on an imbalanced subset of data. Global performance measurements, such as overall efficiency, provides an advantage to the majority class and can be misleading. For example, the overall efficiency of the fuzzy system is 88 per cent, which seems good.

However, assume the dataset had 100 instances, with a split of 80 for non-congestion and 20 for congestion. Again, assume the system classifies non-congestion as 92 instances and congested as eight instances. This means the class, congested is only 40 per cent efficient/accurate and not 88 per cent. Therefore, the discussion will focus on TPR, FPR, precision, F-measure.

In addition to the challenges mentioned above, it should be noted that due to the data considerations and concerns mention in section 4.6, such as:

- A lack of consistent distance between the Bluetooth sensors causing each link to have its own heterogeneous characteristics.
- The data quality of Bluetooth and inductive loop counter sensors are poor. Therefore, sensors are providing an incorrect observation value.
- In bad weather, the sensors which use a mobile network to transmit the data to a central location, can fail and cause the dataset to have missing data.

Initial empirical experimentation was conducted to ascertain how to manage the incorrect and missing data values. Therefore, two approaches where taken, the first approach was to set all missing data values to zero (This approach relates to the results in Table 12) and aligns with the manual approach taken with TfGM. The second approach was to replace the missing and incorrect values with the last reliable measurement.

Table 13: Empirical Experiment Results for Fuzzy System: Comparing Handling of Missing and Incorrect Data

APPROACH	PRECISION	RECALL	F-MEASURE	OVERALL EFFICIENCY (%)
ZERO VALUES	78.4	67	72.2	88
LAST RELIABLE VALUES	63.1	62.7	62.9	45
DIFFERENCE	15.3	4.3	9.3	43

Table 13 shows the two different approaches and demonstrates that using the last reliable value was 43% less efficient overall and had performed worst across the other three measurements: Precision, Recall, and F-measure. Therefore, it was decided that the best approach to handling the missing and incorrect values were to set the values to zero as the performance is better and it was truer to the expected behaviour for TfGM.

The results in Table 12 show naïve bayes achieved a TPR of 99.8 per cent for non-congestion, which is the highest TPR across all algorithms and both classes. However, it also achieved the second highest FPR of 57.8 per cent. This is attributed to the paradox of imbalanced datasets. The FPRs for the minority class across all three algorithms are significantly low, for instance, the fuzzy system is 5.5 per cent, the decision tree is 4.8 per cent, and the naïve bayes is 0.2 per cent. The FPRs for the majority class across all three algorithms are noticeably higher, for instance, the fuzzy system is 32.9 per cent, the decision tree is 59 per cent, and the naïve bayes is 57.8 per cent. Because of these noticeable differences, it has been decided from this point to only compare the weighted averages of both classes.

The TPR weighted average for the fuzzy system is 88 per cent, which is higher than both, decision tree by ≈6 per cent and naïve bayes by ≈2 per cent. The FPR weighted average for the fuzzy system is 26.5 per cent, which is lower than both, decision tree by ≈20 per cent and naïve bayes by ≈18 per cent. The precision weighted average for the fuzzy system is 87.6 per cent, which is higher than the decision tree by ≈6 per cent, however, it was lower than the naïve bayes by ≈1 per cent. The F-measure weighted average for the fuzzy system was 87.6 per cent and is higher than both the decision tree by ≈7 per cent and Naïve Bayes by ≈3 per cent. Furthermore, the fuzzy system overall, efficiency was the highest of all three machine-learning algorithms.

Although all algorithms perform to a similar level with the fuzzy system performing the best overall, it should be noted that each algorithm has its own level of complexity, which some stakeholders may struggle to understand based on the complex explainability. For instance, the easiest of the three algorithms to implement and understand is the fuzzy system. As the system is built using linguistic values that all stakeholders are able to understand, and the output is a single defuzzified value (probability of congestion) where anything above 95 per cent is congestion compared to the RAG (Section 5.4.2) method, which requires the stakeholder to compare the journey time to three equations to identify congestion.

The second easiest to understand is the decision tree, J48, where a branch of the tree is split based on a value of the variable being used and this is repeated until the leaves are reached and an outcome is decided. It should be noted the bigger the tree and the more leaves it has the hard it is to understand the decision transparency and may make it harder for stakeholders to follow. The decision tree model in this experiment has a tree size of 17 and a total of 9 leaves. The 9 rules are transparent and could be understood by a transport expert. The most complex algorithm for stakeholders to understand is Naïve Bayes because it is a probabilistic classifier, which uses a probability distribution over a set of classes, instead of only outputting the most likely class that an observation should belong to.

7.3.8 Conclusion

This study has proven the hypothesis, *H_A0: Using journey time and volume data, it is possible to classify congestion using a fuzzy system* and has demonstrated the proof of concept. The initial results have demonstrated the binary fuzzy systems ability to predict congestion using volume and journey time, outperforming both the decision tree and Naïve Bayes. Moreover, the fuzzy system using only six rules was able to manage an unbalanced dataset. Additionally, it would be possible to implement this model other urban road networks.

The next step was to develop a multi-classification Fuzzy decision-making system that capable of recognise one of three types of congestion (Abberley et al., 2017): non-recurrent congestion, recurrent congestion, and semi-recurrent congestion plus non-congestion for when the traffic flow is good. This is an important requirement for TfGM who would benefit from not only being able to identify congestion but the type of congestion, which would allow for different mitigation strategies to be put in place. Additionally, they will be able to measure how much of the network is, at a given time, exhibiting signs of non-congestion, recurrent, non-recurrent, and semi-recurrent congestion. To achieve this goal, a multi-classification fuzzy decision-making system will be developed and discussed in section 7.4. The next fuzzy decision-making system will focus on having multi-classifications and will expand the linguistic variables to add times of day, days of the week, bank holidays, distance from an attraction, and direction of traffic flow.

7.4 Multi-classification Fuzzy decision-making system

Following on from the results of the binary fuzzy decision-making system, a multi-class fuzzy decision-making system was designed and developed.

The main differences are as follows

- Instead of only being able to predict whether congestion has occurred or not. The multi-classification model looks to predict the type of congestion, recurrent congestion, non-recurrent, and semi recurrent as well as non-congestion.
- The system extracts a subset of the real-world spatial-temporal data from MUCD dataset. The extra data fields being extracted are, Time-of-Day, Day-of-Week, distance from attractor
- Extract a subset of the real-world spatial-temporal dataset, known as the Manchester Urban Congestion Data (MUCD) dataset. This subset of data is then processed and store within a database. (Chapter four).
- The binary fuzzy decision-making system was analysing single links however, the multi-classification fuzzy decision-making system is designed to work with all links on the network.
- As the multi-classification system is predicting against all 64 links where each link has its own characteristics, a percentile model will be created to standardise each link, allowing it to replace the k-means++ algorithm used to determine the memberships.

The experiments described in this chapter contribute towards answer the following research question (RQ4) - "*Can a fuzzy rule-based system be designed to predict road congestion through validation of the Urban Road Congestion Conceptual (URCC) model?"*

This section aims to create a fuzzy Decision-making system that can model the complex nature of urban road congestion using a real-world dataset. To visualise the multi-classification fuzzy decision-making system, an extension to TIM was created (Figure 76). This extension allows the user additional functionality, such as performing an ad-hoc prediction of the type of congestion occurring by allowing the user to set the parameters and instantly see the outcome.

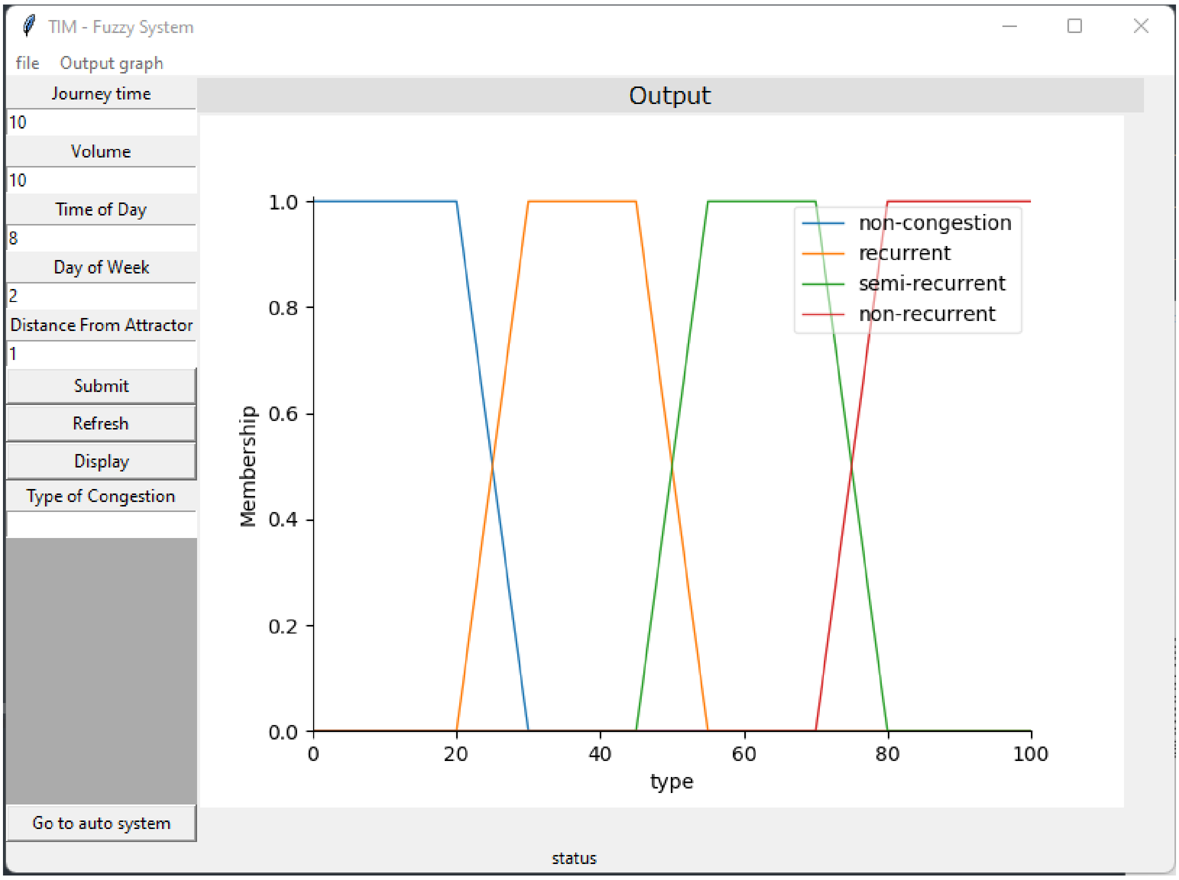


Figure 76: TIM - Fuzzy system

Additional automated functionalities are shown in Figure 77. These functions are as follows: automatically classify the subset of data based on the label definitions discussed in section 7.4.1.2. Perform the automatic predictions using the multi-classification fuzzy decision-making system against the training data. Automatically, statistically analysis the results using the following statistical measurements: Recall, Precision, F-measure, and weighted average for all 64 links individually and combined.

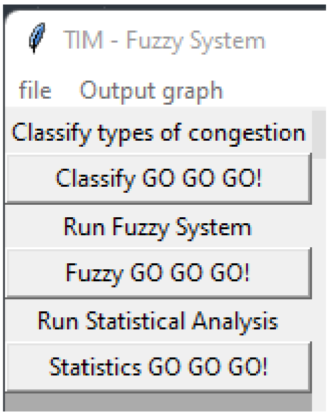


Figure 77: Additional functionality

7.4.1 Methodology: Multi-classification fuzzy decision-making system for predicting urban road congestion type

This section describes the methodology that was used to develop the second iteration of the fuzzy decision-making system for road congestion on an urban city network. The model is similar to its predecessor with several differences which have been mentioned in section 7.4.

The method for creating the multi-classification fuzzy decision-making system is as followed:

- The subset of data utilised to predict the type of urban road congestion are from the MUCD Dataset and consists of the following, journey time from Bluetooth sensors, traffic volume from inductive loop counters, event information from o2 Apollo and the Etihad Stadium, distance from attractor, road traffic accident injury statistic data (Stats19) from GOV.UK, and local school term times from GOV.UK (Section 7.4.1.1).
- Each link within the subset of data at every 15-minute intervals requires a classification to be allocated to it. As the data is unsupervised knowledge gain from the URCC model and with support of domain experts, such as TfGM, definitions for each classification are specified in section 7.4.1.2.
- Due to the multifaceted nature of the individual links, a new approach was chosen for standardising each links behaviour which can then be used to assist in the creation of the membership functions was implemented, referred to as the percentile model (Section 7.4.1.3).
- Once the relevant data had been standardised and the classification for each link at ever 15-minute intervals for the 6-month period has been calculated, the next step is to determine the membership functions, see section 7.4.1.4.
- After the fuzzy memberships functions have been determined, the next step is to determine the fuzzy rules, see section 7.4.1.5.
- Now the memberships and rules have been determined. A fuzzy inference method needs to be selected in section 7.4.1.6.
- The same method for performing defuzzification was discussed in section 7.3.5. the method used was centroid of area (COA), also known as the centre of gravity (COG).

7.4.1.1 Data sources and dataset

For this study, a subset of the real-world spatial-temporal data from MUCD dataset, which was discussed in chapter four was used. This subset of data is then processed and stored within a database. The subset is representative of a sub-network of 64 links, data for each link is collected every 15 minutes for a total of six months and in its current form the MUCD is unsupervised.

The MUCD consists of several types of data, such as average journey time between two sensors, traffic volume count at a single point, and event information from two attractors etc. This data is provided by Transport for Greater Manchester (TfGM), the Etihad Stadium, and the O2 Apollo. The data is then modified to create two new datasets of equal size, one for training and one for testing. Both datasets have a total of 555876 tuples and each tuple contains five attributes and one label. The attributes are as follows: Distance from Attractor (DfA), Day of Week (DoW), Journey Time (JT), Time of Day (ToD), and Volume (V). The label is one of four classifications: Non-Congestion (NC), Recurrent Congestion (RC), Semi-Recurrent Congestion (SRC), Non-Recurrent Congestion (NRC). Each type of congestion is clearly defined in section 7.4.1.2.

7.4.1.2 Labelling multi-classification MUCD Dataset

To be able to predict the type of congestion, a label is required. As this research introduces new concepts of congestion, such as semi-recurrent. It is important to define how each label is calculated. Non-congestion (section 7.4.1.2.1), recurrent congestion (section 7.4.1.2.2), semi-recurrent congestion (section 7.4.1.2.3), and non-recurrent congestion (section 7.4.1.2.4).

7.4.1.2.1 Non-congestion

Non-congestion, NC, (Equation 25) will be defined using a one of the three-methods implemented by the domain experts at TfGM. The method for labelling is called RAG which stands for Red, Amber, and Green. This research will only use the green (Equation 12) to define non-congestion.

$$NC = JT \leq \widetilde{JT} * 1.25$$

Equation 25: Non-congestion

Where $JT$is the average journey time for all Bluetooth enabled vehicles travelling between two sensors on each link. The $\widetilde{JT}$ is the 50th percentile of journey time for a single link within the MUCD. 1.25 is the congestion factors boundary used by TfGM.

7.4.1.2.2 Recurrent congestion

The definition of recurrent congestion, RC, (Equation 26 and Equation 27) can be summarised from literature as occurring when significant amounts of vehicles simultaneously use a limited space on a road network on the same day and at the same time (Verhoef, 1999; Hendricks et al., 2001; Arnott, 2013; Fosgerau and Small, 2013). Using this description and the semantic knowledge gained from the urban road congestion ontology (Abberley et al., 2017; Gould and Abberley, 2017) it is possible to provide a semantic description of recurrent congestion as: Recurrent congestion is caused by an event that is predictable and cyclical, such as rush hour which always occurs on a weekday between 6 am and 10 am or 3 pm and 7 pm causing worst that expected journey time on a city-scale.

$$RC = TOD_{am} + DOW_{wd} + \ JT_{worst} + V_{Worst}$$

Equation 26: Recurrent congestion (AM)

Or

$$RC = TOD_{pm} + DOW_{wd} + JT_{worst} + V_{Worst}$$

Equation 27: Recurrent congestion (PM)

Where t, time-of-day, for a specific 15-minute slot on road link x, where x is all road links on the urban network being modelled has a range of $t=\{0,\dots,24\}$ and is an element of Time of Day, TOD, (Equation 28) which is defined as:

$$TOD(t) = \begin{cases} em, & t < 6 \\ am, & 6 \leq t < 10 \\ day, & 10 \leq t < 15 \\ pm, & 15 \leq t < 19 \\ le, & t \geq 19 \end{cases}$$

Equation 28: Time of day

And d, day-of-week, for a specific 15-minute slot on road link x, where x is all road links on the urban network being modelled has a range of $d=\{1,\dots,7\}$ and is an element of Day of Week, DOW, (Equation 29) which is defined as:

$$DOW(d) = \begin{cases} wd, & d \leq 5 \\ we, & d > 5 \end{cases}$$

Equation 29: Day of week

And j, journey time, (Equation 30) for a specific 15-minute slot on road link x, where x is all road links on the urban network being modelled has a range of $j=\{0,\dots,\infty\}$ and is an element of Journey time, JT, which is defined as:

$$JT(j) = \begin{cases} expected, & j \leq \tilde{j} * 1.25 \\ worst, & j > \tilde{j} * 1.25 \end{cases}$$

Equation 30: Journey time

And v, volume, (Equation 31) for a specific 15-minute slot on road link x, where x is all road links on the urban network being modelled has a range $v=\{0,\dots,\infty\}$ and is an element of Volume, V, which is defined as:

$$V(v) = \begin{cases} expected, & v \leq \tilde{v} * 1.25 \\ worst, & v > v * 1.25 \end{cases}$$

Equation 31: Traffic volume

7.4.1.2.3 Semi-recurrent congestion

Semi-recurrent congestion, SRC, (Equation 32 and Equation 33) was coined by the author and is described as being predictable and non-cyclical unlike recurrent congestion, which is predictable and cyclical and non-recurrent that is non-predicable and non-cyclical. Semi-recurrent congestion is caused by scheduled events, such as a football match and concerts, which are not cyclical because they do not happen at the same time or on the same day.

$$SRC = TOD_{le} + DOW_{wd} + JT_{worst} + V_{worst} + D_{Near}$$

Equation 32: Semi-recurrent congestion (on a weekday)

Or

$$SRC = TOD_{day/pm/le} + DOW_{we} + JT_{worst} + V_{worst} + D_{Near}$$

Equation 33: Semi-recurrent congestion (on a weekend)

Where dis, distance (Equation 34) from attractor for road link x, where x is all road links on the urban network being modelled has a range $v=\{0.0,\ldots,\infty\}$ and is an element of Distance, D, which is defined as:

$$D(dis) = \begin{cases} near, & dis \leq Z \\ far, & dis > Z \end{cases}$$

Equation 34: Distance

Where Z is determined as an empirical variable which is the distance from the nearest attractor on a given urban network. For the purpose of this work and for the network shown above. Following empirical experimentation, the value Z is set at 2.5, which represents 2.5 km from the links nearest attractor.

7.4.1.2.4 Non-recurrent congestion

The definition of non-recurrent congestion, NRC, (Equation 35) can be summarised from the literature as occurring due to a non-predicable and non-cyclical event, such as a traffic accident and unplanned road works (Cassidy and Bertini, 1999; Verhoef and Rouwendal, 2004; Djahel et al., 2015), which can cause expected journey times and volumes to increase around the event. Non-recurrent congestion is defined as:

$$NRC = JT_{worst} + V_{Worst}$$

Equation 35: Non-recurrent congestion

7.4.1.3 Percentile Model

To standardise the performance of each link and to allow a single membership function to be determined for both the journey time and traffic volume, a percentile method was applied. See Equation 36, where each percentile group is represented by P, and n is the nth percentile. For this research the nth percentiles are broken down into 10 groups (0-10th, 10th-20th, 20th-30th, …, 90th-100th). X is the total number of observations.

Equation 36: Percentile method

$$P(n) = \left(\frac{n}{100}\right) * X$$

To calculate the percentile classification for journey time, use the following algorithm 3.

Algorithm 3
Standardising **journey time** for each link at each 15-minute observation.

Variables: *I* the set of observations, L the set of links, x_i: the observation, O: An array of outcomes.

```
for $l \in L$ do
    for $i \in I$ do
        if $x_i = 0$ then
            $O_i \leftarrow 10$
        end if
        if $x_i > 0$ and $x_i \leq P(10)$ then
            $O_i \leftarrow 1$
        end if
        if $x_i > P(10)$ and $x_i \leq P(20)$ then
            $O_i \leftarrow 2$
        end if
        if $x_i > P(20)$ and $x_i \leq P(30)$ then
            $O_i \leftarrow 3$
        end if
        if $x_i > P(30)$ and $x_i \leq P(40)$ then
            $O_i \leftarrow 4$
        end if
        if $x_i > P(40)$ and $x_i \leq P(50)$ then
            $O_i \leftarrow 5$
        end if
        if $x_i > P(50)$ and $x_i \leq P(60)$ then
            $O_i \leftarrow 6$
        end if
        if $x_i > P(60)$ and $x_i \leq P(70)$ then
            $O_i \leftarrow 7$
        end if
        if $x_i > P(70)$ and $x_i \leq P(80)$ then
            $O_i \leftarrow 8$
        end if
        if $x_i > P(80)$ and $x_i \leq P(90)$ then
            $O_i \leftarrow 9$
        end if
        if $x_i > P(90)$ and $x_i \leq P(100)$ then
            $O_i \leftarrow 10$
        end if
    end for
end for
return O
```

To calculate the percentile classification for traffic volume, use the following algorithm 4.

Algorithm 4
standardising **traffic volume** for each link at each 15-minute observation.

Variables: *I* the set of observations, L the set of links, x_i: the observation, O: An array of outcomes.

```
for $l \in L$ do
    for $i \in I$ do
        if $x_i = 0$ then
            $O_i \leftarrow 10$
        end if
        if $x_i > 0$ and $x_i \leq P(10)$ then
            $O_i \leftarrow 1$
        end if
        if $x_i > P(10)$ and $x_i \leq P(20)$ then
            $O_i \leftarrow 2$
        end if
        if $x_i > P(20)$ and $x_i \leq P(30)$ then
            $O_i \leftarrow 3$
        end if
        if $x_i > P(30)$ and $x_i \leq P(40)$ then
            $O_i \leftarrow 4$
        end if
        if $x_i > P(40)$ and $x_i \leq P(50)$ then
            $O_i \leftarrow 5$
        end if
        if $x_i > P(50)$ and $x_i \leq P(60)$ then
            $O_i \leftarrow 6$
        end if
        if $x_i > P(60)$ and $x_i \leq P(70)$ then
            $O_i \leftarrow 7$
        end if
        if $x_i > P(70)$ and $x_i \leq P(80)$ then
            $O_i \leftarrow 8$
        end if
        if $x_i > P(80)$ and $x_i \leq P(90)$ then
            $O_i \leftarrow 9$
        end if
        if $x_i > P(90)$ and $x_i \leq P(100)$ then
            $O_i \leftarrow 10$
        end if
    end for
end for
return O
```

7.4.1.4 Membership Function Determination

Table 14 shows the dimensions, data sources, and the linguistic values determined from the urban road congestion ontology. The linguistic values of the membership functions representing journey time, traffic volume, distance from attractor, time of day, and day of the week are also shown.

Table 14: Dimensions and their linguistically values.

Dimension (Variables)	Data sources	Linguistic values (Membership functions)
Journey time	Bluetooth remote sensors	Very Low (*VL*) Low (*L*) Medium (*M*) High (*H*) Very High (*VH*)
Volume	Inductive loop counters	Very Low (*VL*) Low (*L*) Medium (*M*) High (*H*) Very High (*VH*)
Distance from Attractor	GIS analysis	Very Near (*VN*) Near (*N*) Far (*F*) Very Far (*VF*)
Time of day	Temporal value of instance	Early Morning (*EM*) AM Peak (*AM*) Day (*D*) PM Peak (*PM*) Late Evening (*LE*)
Day of the week	Recorded day of instance	Weekday (*WD*) Weekend (*WE*)

Using the linguistic values identified in Table 14, the creation of the fuzzy membership functions can be performed using three steps:

Step 1: Perform the percentile model algorithms discussed in section 7.4.1.3 on the journey time and traffic volume data.
Step 2: Identify the final boundary values for a set of groups where they connect and define this value as dt.
Step 3: Using the dt value, determine membership function domain coverage using one of three membership functions: linear up, linear down, and trapezoidal shape.

The primary objective of using the percentile model (step 1) was to standardise the observed journey times and traffic volume for each link and sensors, regardless of the behavioural characteristics. Therefore, the journey time for the 10th percentile on one link could be 200 seconds and on another it could be 2000 seconds. By standardising these boundaries, it will allow for a single membership function to be created that covers all links instead of requiring a membership function for every link.

Using the fuzzy system extension to TIM, it is possible to visualise the membership functions. These use the same three fuzzy membership function as discussed in section 7.3.3.1. The three types of functions are: linear increasing membership, linear decreasing membership, and trapezoidal-shaped membership.

Figure 78 and Figure 79 shows the memberships for both journey time and traffic volume, where very low is calculated using linear decreasing (Equation 16). Low, medium, and high is calculated using trapezoidal-shaped (Equation 17). Very high is calculated using linear increasing (Equation 15).

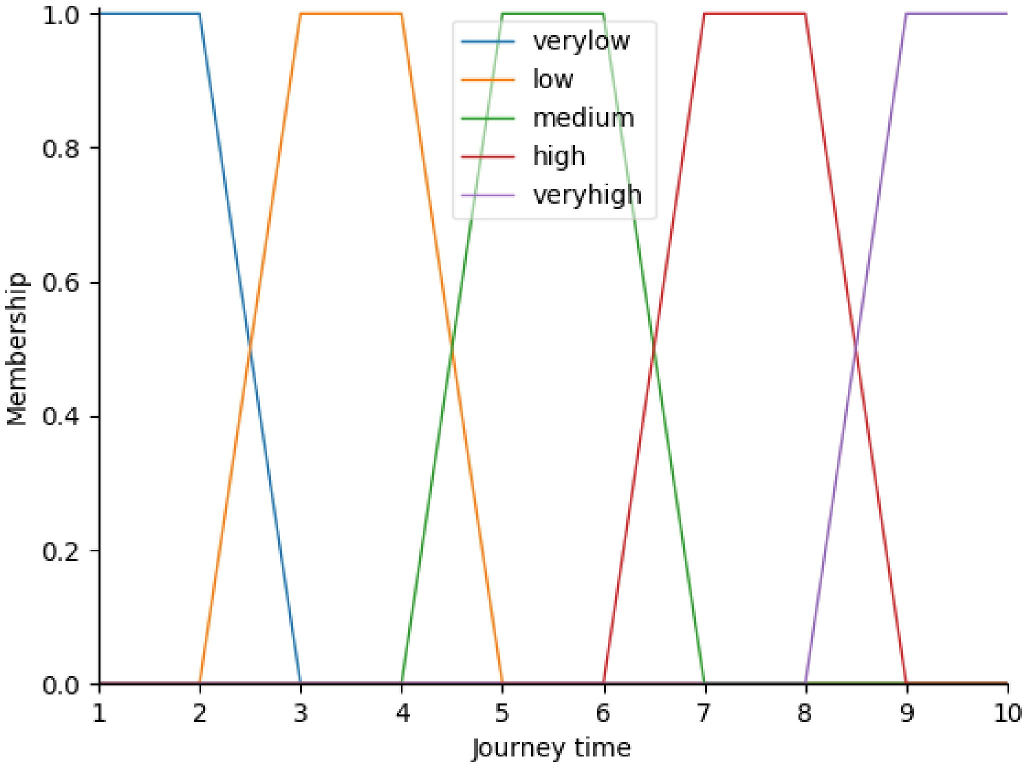


Figure 78: Journey time membership function

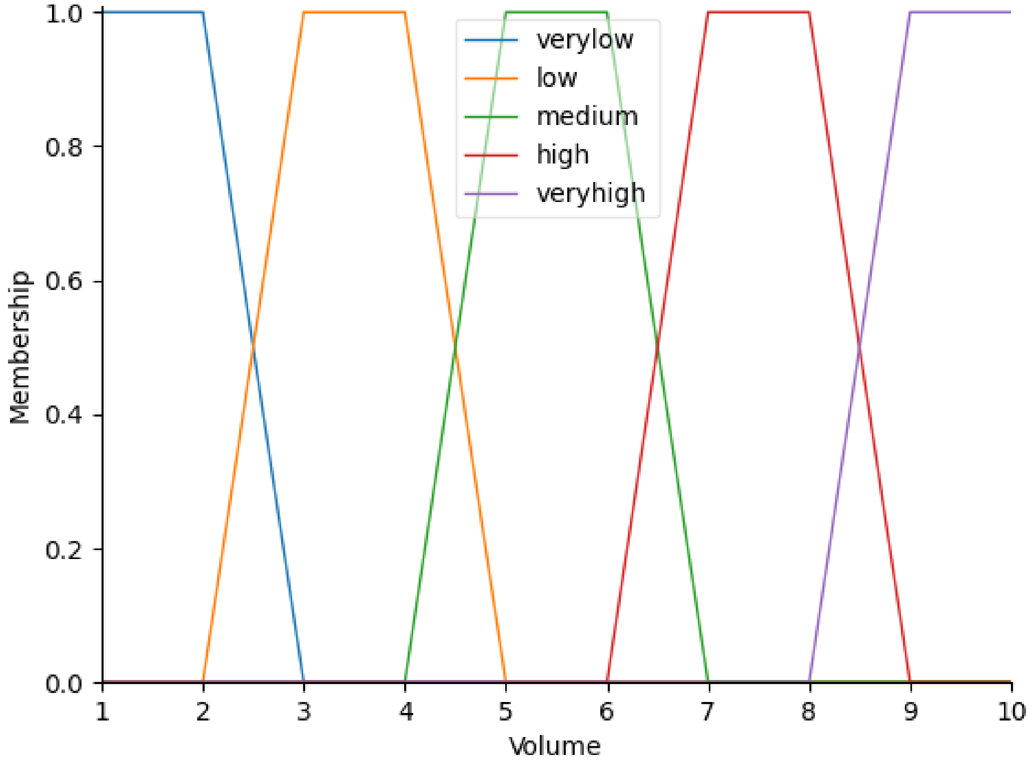


Figure 79: Traffic volume membership function

Figure 80 shows the membership for school bank holidays, where 'no' is calculated using linear decreasing (Equation 16) and 'yes' is calculated using linear increasing (Equation 15).

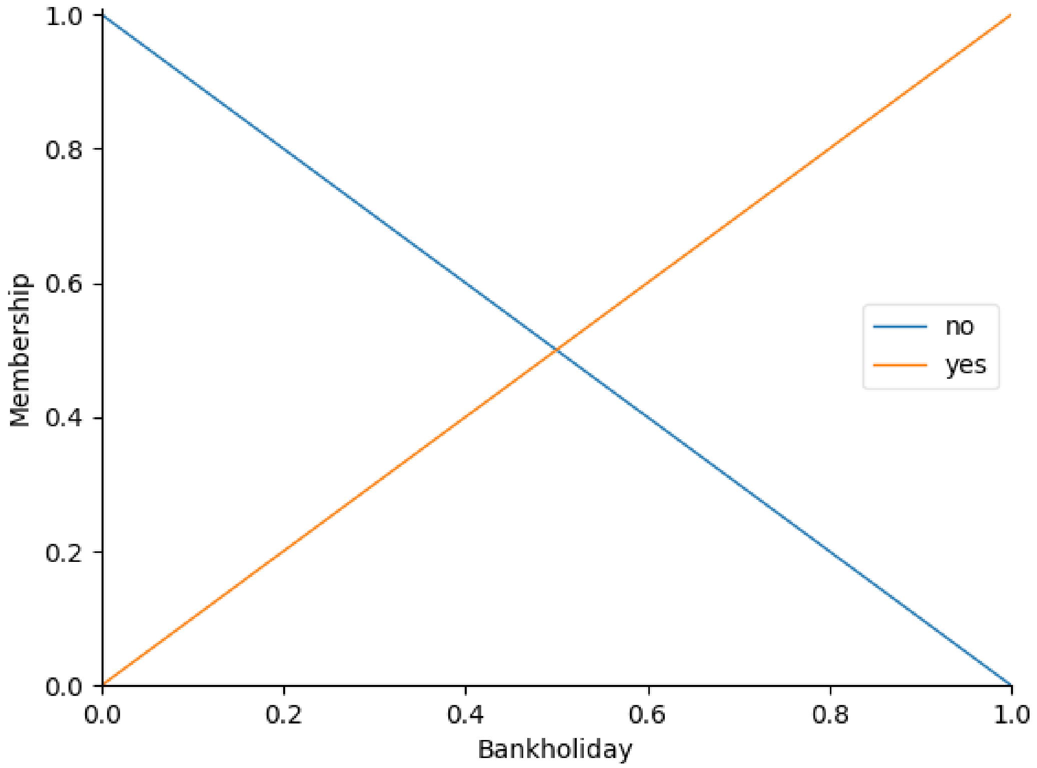


Figure 80: School term (bank holiday) membership function

Figure 81 shows the memberships for day of the week, where weekday is calculated using linear decreasing (Equation 16) and weekend is calculated using linear increasing (Equation 15).

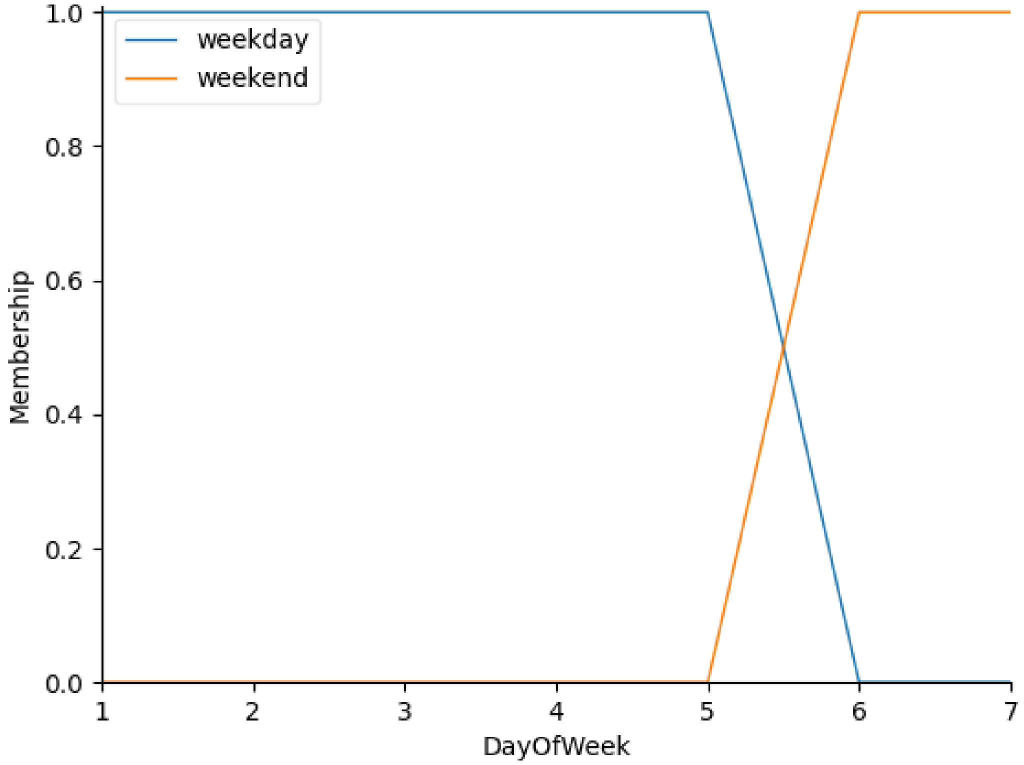


Figure 81: Day of week membership function

Figure 82 shows the memberships for time of the day, where early morning (em) is calculated using linear decreasing (Equation 16). am, day, and pm is calculated using trapezoidal-shaped (Equation 17). Late evening (le) is calculated using linear increasing (Equation 15).

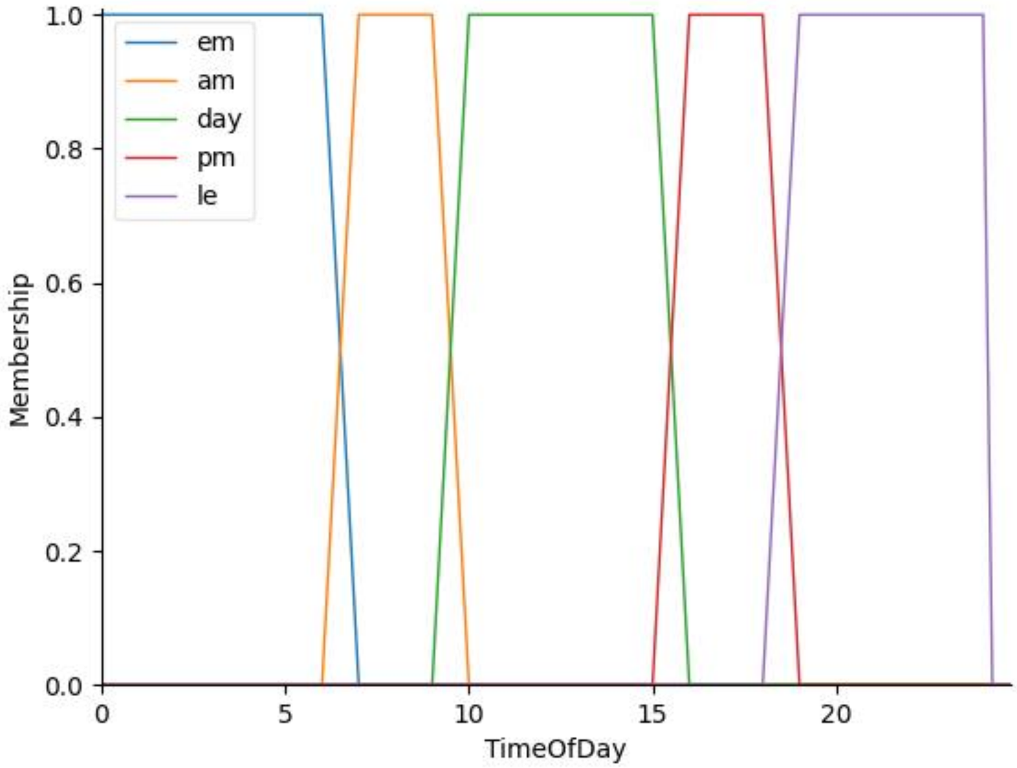


Figure 82: Time of day membership function

Figure 83 shows the memberships for distance from attractor, where very near is calculated using linear decreasing (Equation 16). Near and far are calculated using trapezoidal-shaped (Equation 17). Very Far is calculated using linear increasing (Equation 15).

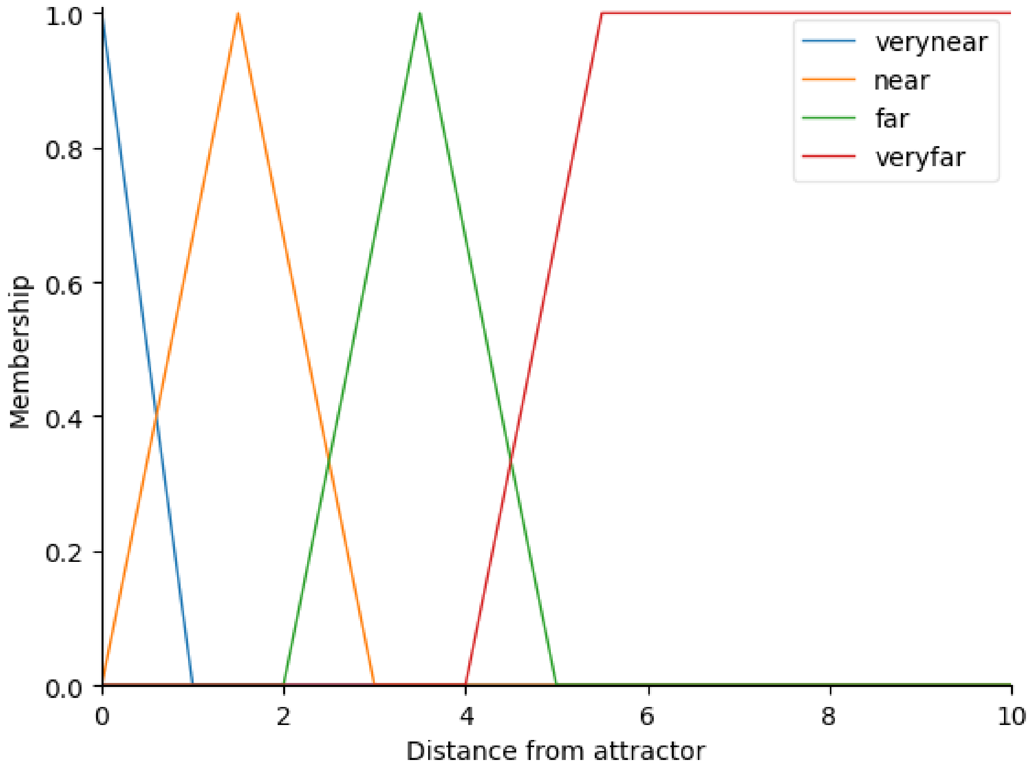


Figure 83: Distance from Attractor

7.4.1.5 Fuzzy Rules Determination (manual and expert)

The fuzzy rules were created with the knowledge gained from the URCC model in chapter three, by visualising the data using TIM in chapter four, and using expert knowledge (TfGM). A series of empirical experiments was conducted with feedback from TfGM to ascertain that a total of 12 rules were required for the initial multi-classification decision-making system.

Algorithm 5 uses both antecedents and consequents membership functions to fire 12 unique rules to acquire each rule strength ready for fuzzy inference.

Algorithm 5
Rules for predicting congestion type.

Antecedents: Journey Time, JT. Traffic Volume, V. Distance from attractor, Dis. Bank holidays, BH. Day of week, DoW, Time of day, ToD.
Antecedents memberships: Very Low, VL. Low, L. Medium, M. High, H. Very High, VH. Very Near, VN. Near, N. Far, F. Very Far, VR. No, NO, Yes, Y. Weekday, WD. Weekend, WE. Early Morning, EM. AM, A, Day, D, PM, P. Late Evening, LE.
Consequents: Congestion Type, CT.
Consequents memberships: Non-congestion, NC. Recurrent Congestion, RC. Semi-Recurrent Congestion, SRC. Non-Recurrent Congestion, NRC.

```
$\textbf{if}\ JT\ is\ M\ \textbf{and}\ V\ is\ NOT\ VH\ \textbf{then}$
    $CT \leftarrow NC$
$\textbf{if}\ JT\ is\ L\ \textbf{or}\ JT\ is\ VL\ \textbf{then}$
    $CT \leftarrow NC$
$\textbf{if}\ (JT\ is\ VH\ \textbf{or}\ JT\ is\ H)\ \textbf{and}\ (ToD\ is\ A\ \textbf{or}\ ToD\ is\ P)$
                                    $\textbf{and}\ DoW\ is\ WD\ \textbf{then}$
    $CT \leftarrow RC$
$\textbf{if}\ JT\ is\ M\ \textbf{and}\ V\ is\ VH\ \textbf{and}\ (ToD\ is\ A\ \textbf{or}\ ToD\ is\ P)$
                                    $\textbf{and}\ DoW\ is\ WD\ \textbf{then}$
    $CT \leftarrow RC$
$\textbf{if}\ (JT\ is\ VH\ \textbf{or}\ JT\ is\ H)\ \textbf{and}\ (ToD\ is\ NOT\ A\ \textbf{and}\ ToD\ is\ NOT\ P)$
        $\textbf{and}\ DoW\ is\ WD\ \textbf{and}\ (Dis\ is\ NOT\ N\ \textbf{or}\ Dis\ is\ NOT\ VN)\ \textbf{then}$
    $CT \leftarrow NRC$
$\textbf{if}\ JT\ is\ M\ \textbf{and}\ V\ is\ VH\ \textbf{and}\ (ToD\ is\ NOT\ A\ \textbf{or}\ ToD\ is\ NOT\ P)$
    $\textbf{and}\ DoW\ is\ WD\ \textbf{and}\ (Dis\ is\ NOT\ N\ \textbf{or}\ Dis\ is\ NOT\ VN)\ \textbf{then}$
    $CT \leftarrow NRC$
$\textbf{if}\ (JT\ is\ VH\ \textbf{or}\ JT\ is\ H)\ \textbf{and}\ DoW\ is\ WE$
    $\textbf{and}\ (Dis\ is\ NOT\ N\ \textbf{or}\ Dis\ is\ NOT\ VN)\ \textbf{then}$
    $CT \leftarrow NRC$
$\textbf{if}\ JT\ is\ M\ \textbf{and}\ V\ is\ VH\ \textbf{and}\ DoW\ is\ WE$
                $\textbf{and}\ (Dis\ is\ NOT\ N\ \textbf{or}\ Dis\ is\ NOT\ VN)\ \textbf{then}$
    $CT \leftarrow NRC$
$\textbf{if}\ (JT\ is\ VH\ \textbf{or}\ JT\ is\ H)\ \textbf{and}\ ToD\ is\ LE\ \textbf{and}\ DoW\ is\ WD$
                                    $\textbf{and}\ (Dis\ is\ N\ \textbf{or}\ Dis\ is\ VN)\ \textbf{then}$
    $CT \leftarrow SRC$
$\textbf{if}\ JT\ is\ M\ \textbf{and}\ V\ is\ VH\ \textbf{and}\ ToD\ is\ LE\ \textbf{and}\ DoW\ is\ WD$
                                    $\textbf{and}\ (Dis\ is\ N\ or\ Dis\ is\ VN)\ \textbf{then}$
    $CT \leftarrow SRC$
$\textbf{if}\ (JT\ is\ VH\ or\ JT\ is\ H)\ \textbf{and}\ (ToD\ is\ D\ \textbf{or}\ ToD\ is\ LE)\ \textbf{and}\ DoW\ is\ WE$
                                    $\textbf{and}\ (Dis\ is\ N\ \textbf{or}\ Dis\ is\ VN)\ \textbf{then}$
    $CT \leftarrow SRC$
$\textbf{if}\ JT\ is\ M\ \textbf{and}\ V\ is\ VH\ \textbf{and}\ (ToD\ is\ D\ \textbf{or}\ ToD\ is\ LE)\ \textbf{and}\ DoW\ is\ WE$
                                    $\textbf{and}\ (Dis\ is\ N\ \textbf{or}\ Dis\ is\ VN)\ \textbf{then}$
    $CT \leftarrow SRC$
$return\ CT$
```

7.4.1.6 Fuzzy inference

The same fuzzy inference as the binary Fuzzy decision-making system was used. The implemented methods for computing fuzzy inference known as Mamdani (Mamdani and Assilian, 1975) was discussed in section 7.3.4.

7.4.1.7 Defuzzification

The same method for performing defuzzification was discussed in section 7.3.5. the method used was centroid of area (COA), also known as the centre of gravity (COG).

7.4.1.8 Experimental methodology

An empirical study was undertaken to evaluate the multi-classification fuzzy decision-making system. The aim of the experiment is to determine whether a fuzzy system can be used to analyse traffic data to classify congestion.

The following hypothesis will be evaluated.

Hypothesis

H_B0: It is possible to accurately identify the type of road traffic congestion using a Fuzzy system.

H_B1: It is not possible to accurately identify the type of road traffic congestion using a Fuzzy system.

To evaluate the performance of the fuzzy system, it was compared against two alternative machine-learning algorithms: decision tree C4.5 (using the Weka implementation J48) (Weka, 2018) and naïve bayes, which both algorithms used the same subset of data as the fuzzy system. The justification for their selection was given in section 7.3.6.

In order to evaluate the three methods using the MUCD dataset, three statistical measurements were chosen, which are: Precision (Equation 21), Recall (Equation 19), and F-score (Equation 22). In addition to these measurements a weighted variables will be calculated for all three statistics which takes into consideration all four classifications: non-congestion, recurrent congestion, semi-recurrent congestion, and non-recurrent congestion.

Equation 37 shows how the weighted average for recall value is calculated. Where TP is the diagonal value in the confusion matrix presented in Table 15. For example, where actual and prediction equal the same value, such as 'NC' or 'NRC'. FP is the sum of the column minus the TP value. FN is the sum of the row minus the TP.

$$W_{recall} = \frac{\begin{array}{c} C_{NC}(TPR) * C_{NC}(TP + FN) + C_{NRC}(TPR) * C_{NRC}(TP + FN) \\ + C_{SRC}(TPR) * C_{SRC}(TP + FN) + C_{RC}(TPR) * C_{RC}(TP + FN) \end{array}}{TP + FN + TN + FP}$$

Equation 37: Multi-classification weighted average (recall)

Equation 38 shows how the weighted average for the precision value is calculated.

$$W_{precision} = \frac{\begin{array}{c} C_{NC}(PPV) * C_{NC}(TP + FN) + C_{NRC}(PPV) * C_{NRC}(TP + FN) \\ + C_{SRC}(PPV) * C_{SRC}(TP + FN) + C_{RC}(PPV) * C_{RC}(TP + FN) \end{array}}{TP + FN + TN + FP}$$

Equation 38: Multi-classification weighted average (precision)

Equation 39 shows how the weighted average for the F-score value is calculated.

$$W_{Fscore} = \frac{\begin{array}{c} C_{NC}(Fscore) * C_{NC}(TP + FN) + C_{NRC}(Fscore) * C_{NRC}(TP + FN) \\ + C_{SRC}(Fscore) * C_{SRC}(TP + FN) + C_{RC}(Fscore) * C_{RC}(TP + FN) \end{array}}{TP + FN + TN + FP}$$

Equation 39: Multi-classification weighted average (F-score)

7.4.1.9 Results and Discussion

Table 15 presents the confusion matrix for the classification of the types of congestion. In the confusion matrix, the four classifications are: Non-congestion (NC), non-recurrent congestion (NRC), semi-recurrent congestion (SRC), and recurrent congestion (RC).

Table 15: Multiclassification confusion matrix

		Prediction			
		NC	**NRC**	**SRC**	**RC**
Actual	**NC**	247094	102202	24352	33282
	NRC	1752	65821	1296	113
	SRC	216	4097	9318	19
	RC	708	9020	5481	51106

Table 21 in Appendix 2 shows the performance of each individual link which each of the link's locations are plotted on Figure 60. Each link in Figure 60 have two directions, therefore, link 'a' relates to both 'AU' (upstream) and 'AD' (downstream) in Table 21 in Appendix 2. In addition to presenting the results of everything, it is important to present the results for individual links as well because this helps to demonstrate the impact of the data concerns on the overall perform. Taking into consideration these concerns some links and types were easier to predict that others.

To validate the URCC model a fuzzy system was developed, using the subset of data extracted from the MUCD dataset which contains the dimensions identified in the URCC. The subset of data was used to predict the types of congestion. Some links and types were easier to predict than others. This is because of several contributing factors, such as quality of the data, location of Bluetooth and Inductive Loop Counters sensors. Using Table 16 and the graph presented in Figure 84.

Table 16: Each link predicted accuracy

Link	TP	# Of Obvs	Accuracy	Link	TP	# Of Obvs	Accuracy
au	12150	17376	69.92%	**ad**	11899	17376	68.48%
bu	9660	17376	55.59%	**bd**	9710	17376	55.88%
cu	12861	17376	74.02%	**cd**	12750	17376	73.38%
du	9228	17376	53.11%	**dd**	9299	17376	53.52%
eu	12756	17376	73.41%	**ed**	12492	17376	71.89%
fu	10148	17376	58.40%	**fd**	10714	17376	61.66%
gu	13842	17376	79.66%	**gd**	13597	17376	78.25%
hu	11756	17376	67.66%	**hd**	12157	17376	69.96%
iu	14308	17376	82.34%	**id**	14360	17376	82.64%
ju	11697	17376	67.32%	**jd**	11751	17376	67.63%
ku	9282	17376	53.42%	**kd**	9692	17376	55.78%
lu	11228	17376	64.62%	**ld**	11376	17376	65.47%
mu	11992	17376	69.01%	**md**	12200	17376	70.21%
nu	13259	17376	76.31%	**nd**	13243	17376	76.21%
ou	11136	17376	64.09%	**od**	11643	17376	67.01%
pu	9646	17376	55.51%	**pd**	9778	17376	56.27%
qu	11763	17376	67.70%	**qd**	11565	17376	66.56%
ru	13348	17376	76.82%	**rd**	13312	17376	76.61%
su	11807	17376	67.95%	**sd**	11655	17376	67.08%
tu	11381	17376	65.50%	**td**	11355	17376	65.35%
uu	13894	17376	79.96%	**ud**	14019	17376	80.68%
vu	13615	17376	78.36%	**vd**	13656	17376	78.59%
wu	11987	17376	68.99%	**wd**	12118	17376	69.74%
xu	11655	17376	67.08%	**xd**	11949	17376	68.77%
yu	9922	17376	57.10%	**yd**	9869	17376	56.80%
zu	10052	17376	57.85%	**zd**	10252	17376	59.00%
aau	13084	17376	75.30%	**aad**	12871	17376	74.07%
abu	11669	17376	67.16%	**abd**	11838	17376	68.13%
acu	9787	17376	56.32%	**acd**	9707	17376	55.86%
adu	11563	17376	66.55%	**add**	11209	17376	64.51%
aeu	12745	17376	73.35%	**aed**	12316	17376	70.88%
afu	12516	17376	72.03%	**afd**	12844	17376	73.92%

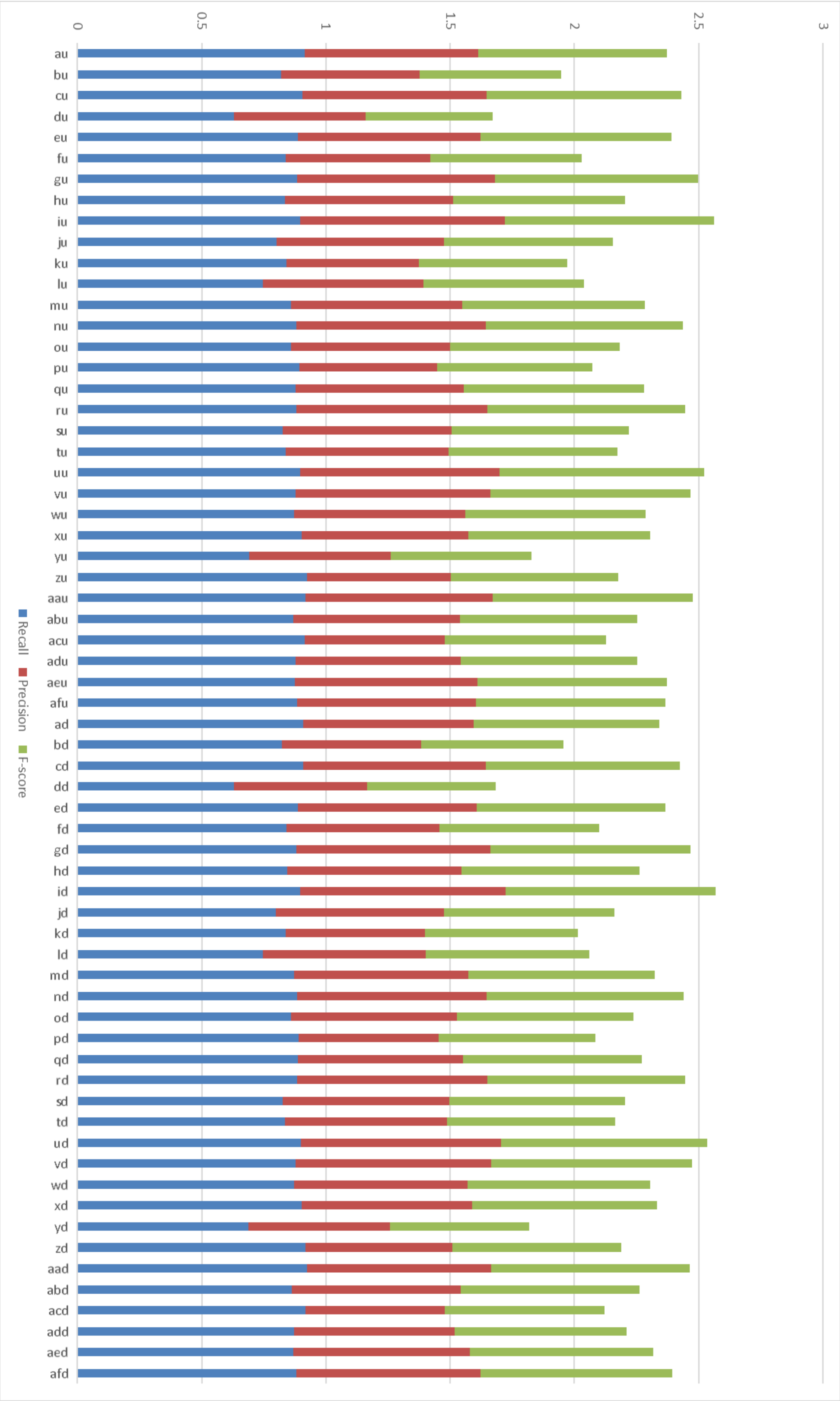


Figure 84: Graph of recall, precision, F-score for all 64 links

The first thing to notice when looking at the accuracy of all 64 links in Table 16 is that upstream and downstream are relatively similar to each other. Link ‘u’ and ‘v’ which connects the two attractors both had high accuracy and the weighted recall, precision and f-scores were predominantly in the ~0.80 which suggests predicting the type of congestion on these links was relatively precise. However, links, such as ‘b’ and ‘d’, which are connecting links and are not main roads have a lower level of accuracy and weighted precision, recall and f-score. This could be caused by the characteristics of the roads not being main routes and also the lack of nearby inductive loop counters is likely to of impacted the accuracy of the predictions, which unlike u and v that achieved ~79%, b and d only achieved ~54% accuracy.

7.4.1.10 Comparison of multi class fuzzy decision-system against other methods

The purpose of this experiment was to determine whether it is possible to classify they types of urban road congestion, recurrent, semi-recurrent, and non-recurrent using a fuzzy system and real-world data extracted from the MUCD dataset. Table 17 shows the results for each statistical measurement, recall, precision, F-score (F1) for three different types of machine-learning algorithms rule-based system (fuzzy), decision tree (J48) and a probabilistic (naïve bayes (NB)) and their classes: non-congestion (NC), recurrent congestion (RC), semi-recurrent congestion (SRC), and non-recurrent congestion (NRC) the weighted average of all classes.

It is important to compare the fuzzy decision-making system against other machine learning algorithms because it allows the performance from one model to be compared again others to identify similarities or extreme differences which could demonstrate a model over or under performing. Performing the analysis against other type of models may give inspiration to future work, for instance, a ‘fuzzy decision-making decision tree’.

As the aim of the comparison was to identify how the fuzzy system compared against traditional machine learning algorithms, the Decision tree C4.5 (using the Weka implementation J48) (Weka, 2018) and naïve bayes algorithms were used implemented using the same subset of data as the fuzzy system.

Table 17: Results for Fuzzy System, J48, and Naïve Bayes (multi-classification)

		NC	RC	SRC	NRC	wAvg
Fuzzy	**Recall**	**99**	64	24	35.8	85.1
	Precision	61	77	68.2	**95.7**	67.7
	F1	76	70	35.4	52.1	71.1
J48	**Recall**	94	**93**	**78.7**	**81.4**	**92**
	Precision	**95**	**87**	**70.3**	84.7	**92.1**
	F1	**95**	**90**	**74.3**	**83**	**92**
NB	**Recall**	90	79	47.6	16.6	78.5
	Precision	89	50	42.8	48	78
	F1	89	61	45.1	24.7	76.9

The results in Table 17 shows the three algorithms all perform to an adequate level. However, the Decision tree C4.5 (J48) performs the most consistently in predicting the types of congestion by achieving an overall weighted average of ~92 per cent for all three recall, precision, and F-score (F1) compared to the multi-classification fuzzy decision-making system which achieved 85.1 per cent for the weighted recall, 67.7 per cent for the weighted precision, and 71.1 per cent for the weighted F-score. Furthermore, Naïve Bayes achieved 78.5 per cent for weighted recall, 78 per cent for precision, and 76.9 for the F-score. Additionally, it was noticed that the multi-classification Fuzzy decision-making system achieved the highest recall (99 per cent) for non-congestion, meaning it identified the majority of data points within the relevant class. Furthermore, the fuzzy model was able to achieve 95.7 per cent for non-recurrent congestion in regard to precision, meaning it was able to predict the most accurately within the relevant class.

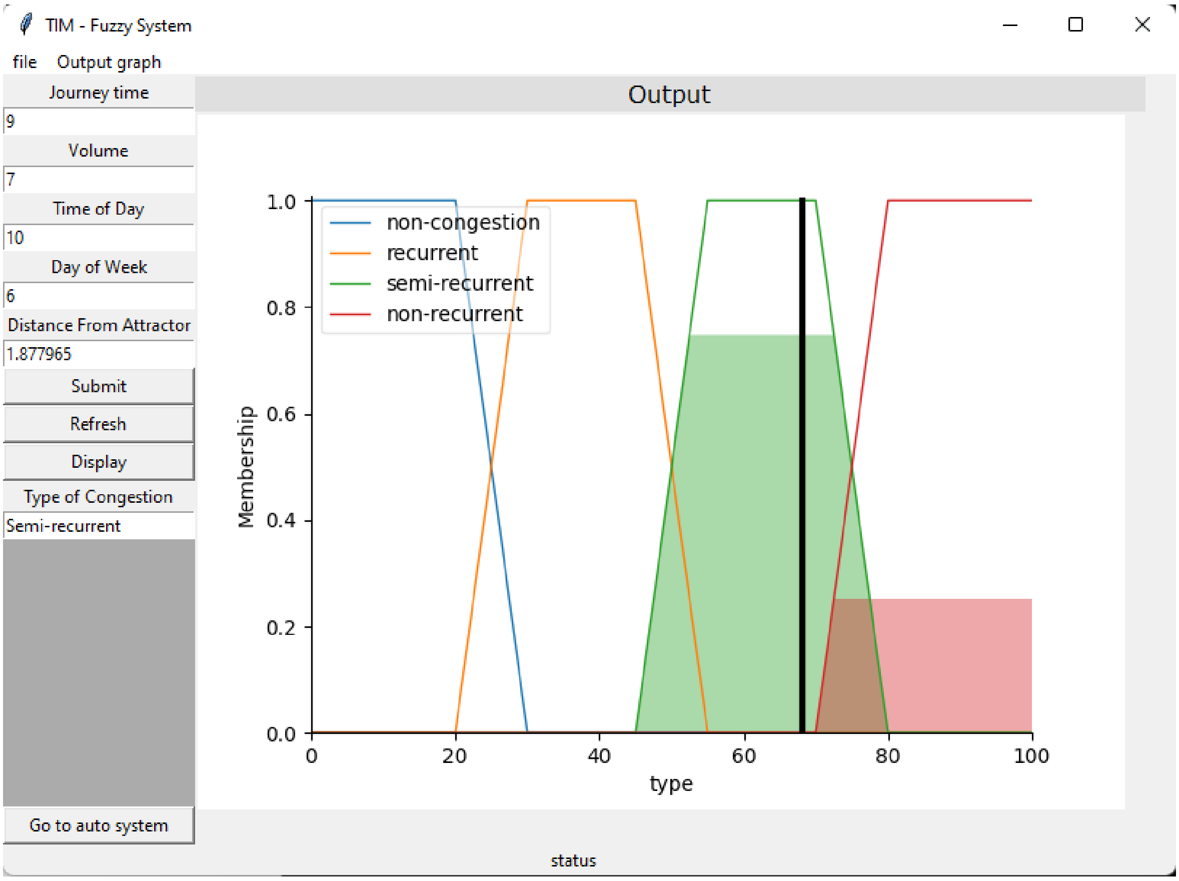
7.4.1.11 Examples of misclassifications

The observations shown in Table 18 demonstrate six instances when the multiclassification fuzzy decision-making system misclassified the observations.

Table 18: Observations of misclassification

Observation	*Class*	*DfA*	*DoW*	*JT*	*Time*	*Volume*	*Result*
1	NRC	1.877965492	6	9	10	7	SRC
2	NRC	5.140836974	7	6	13	8	NC
3	RC	1.877965492	2	9	15.25	9	SRC
4	RC	1.877965492	2	10	18.25	8	SRC
5	SRC	1.07647663	7	9	16.75	8	NRC
6	SRC	1.641062935	3	6	21.75	5	NC

Figure 85 shows a visualisation in TIM of an observation that was expected to be identified as non-recurrent congestion, however, instead the multiclassification fuzzy system misclassified the observation as semi-recurrent congestion. Furthermore, Figure 85 shows that the fuzzy system was able to identify the observation as both semi-recurrent and non-recurrent congestion, however, due to the degree of membership being stronger for semi-recurrent congestion the final crisp outcome was semi-recurrent congestion. The control rule for this outcome was “when journey time is high or very high, time of day is daytime or late evening, day of the week is the weekend, and distance from the attractor is near or very near then semi-recurrent congestion”.



Journey time
9

Volume
7

Time of Day
10

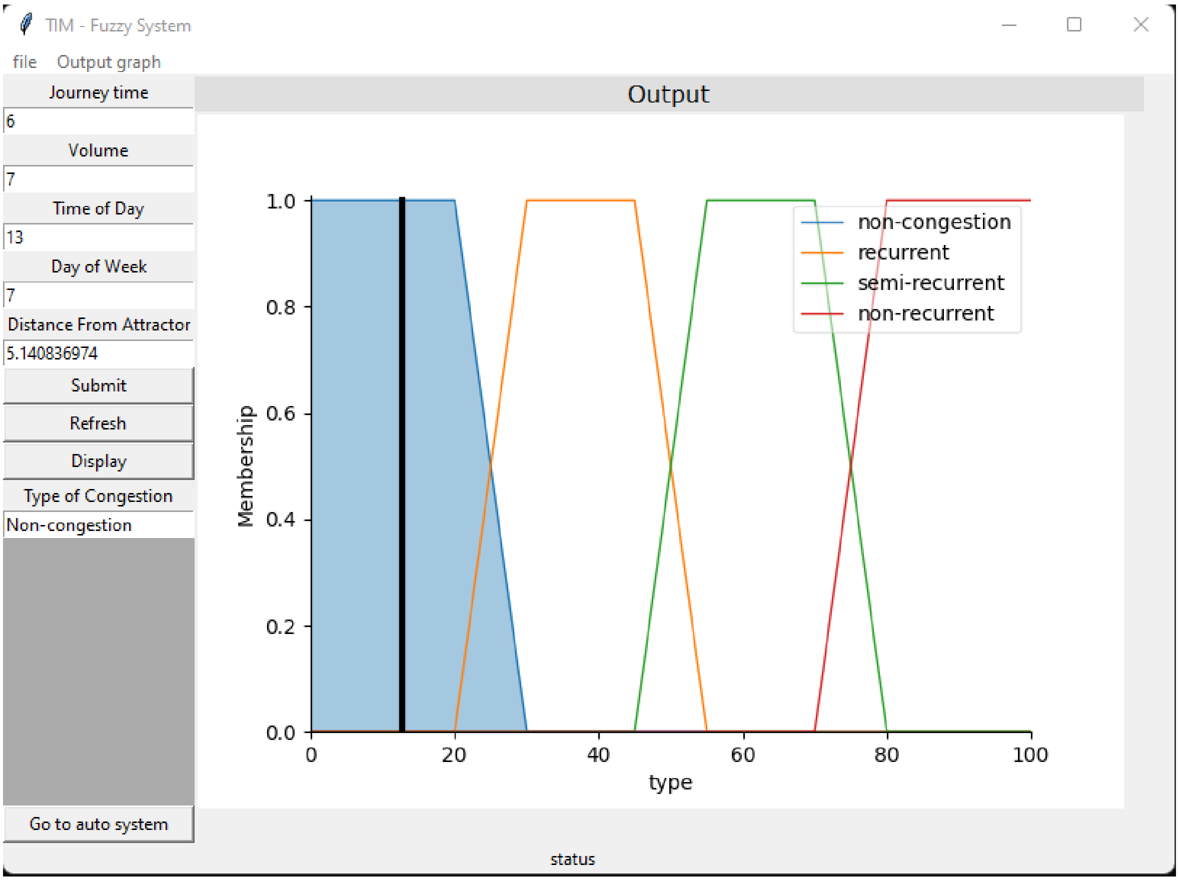
Day of Week
6

Distance From Attractor
1.877965

Type of Congestion
Semi-recurrent

Figure 85: Observation 1: NRC classified as SRC

Figure 86 shows a visualisation in TIM of an observation that was expected to be identified as non-recurrent congestion, however, instead the multiclassification fuzzy system misclassified the observation as non-congestion. The reason for the misclassification was due to one of the control rules being overly dominant causing this outcome, which was “when journey time is medium, and volume is very high then non-congestion”.



Journey time: 6

Volume: 7

Time of Day: 13

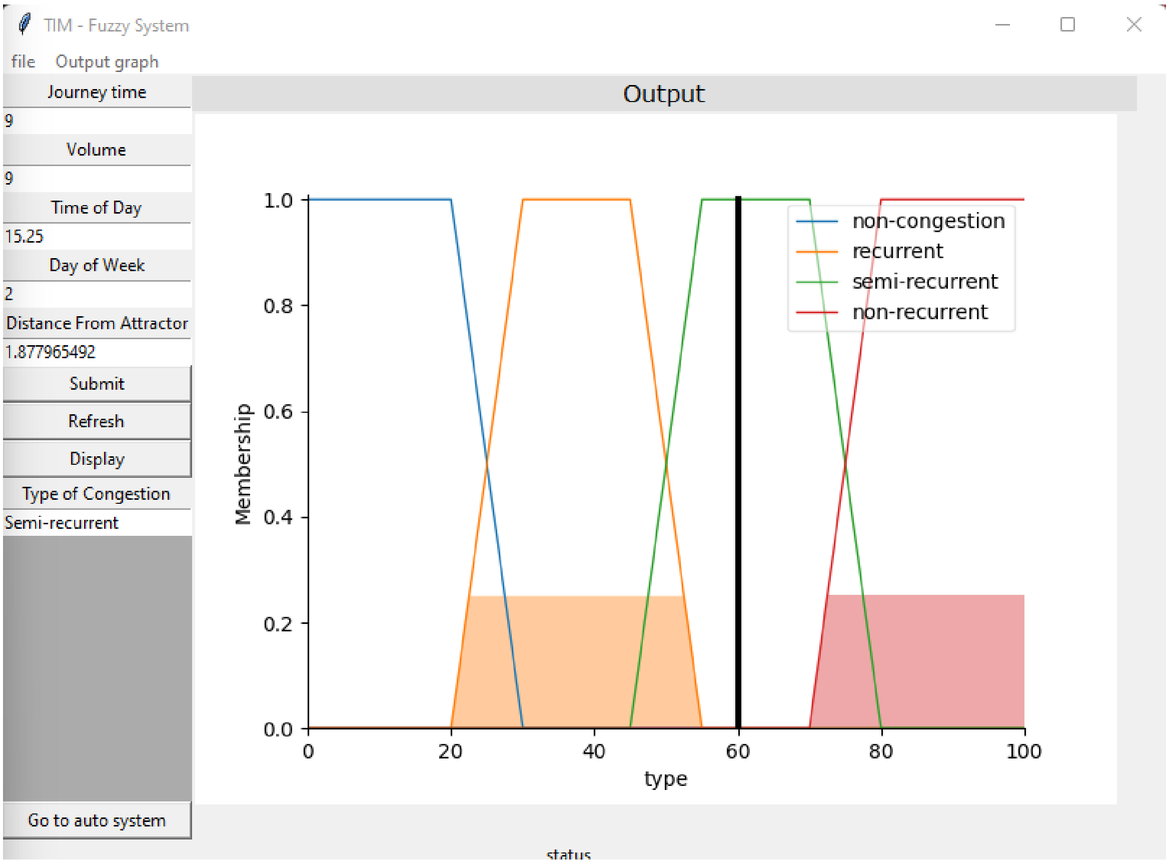
Day of Week: 7

Distance From Attractor: 5.140836974

Type of Congestion: Non-congestion

Figure 86: Observation 2: NRC classified as NC

Figure 87 shows a visualisation in TIM of an observation that was expected to be identified as recurrent congestion, however, instead the multiclassification fuzzy system misclassified the observation as semi-recurrent congestion. The reason for the misclassification was due to multiple fuzzy rules having an equal degree of membership and the defuzzification step that uses centroid of area to create a single crisp output value, in this instance the centre of the aggregation of the consequences is non-recurrent even though it only has a zero of degree of membership.



Journey time: 9

Volume: 9

Time of Day: 15.25

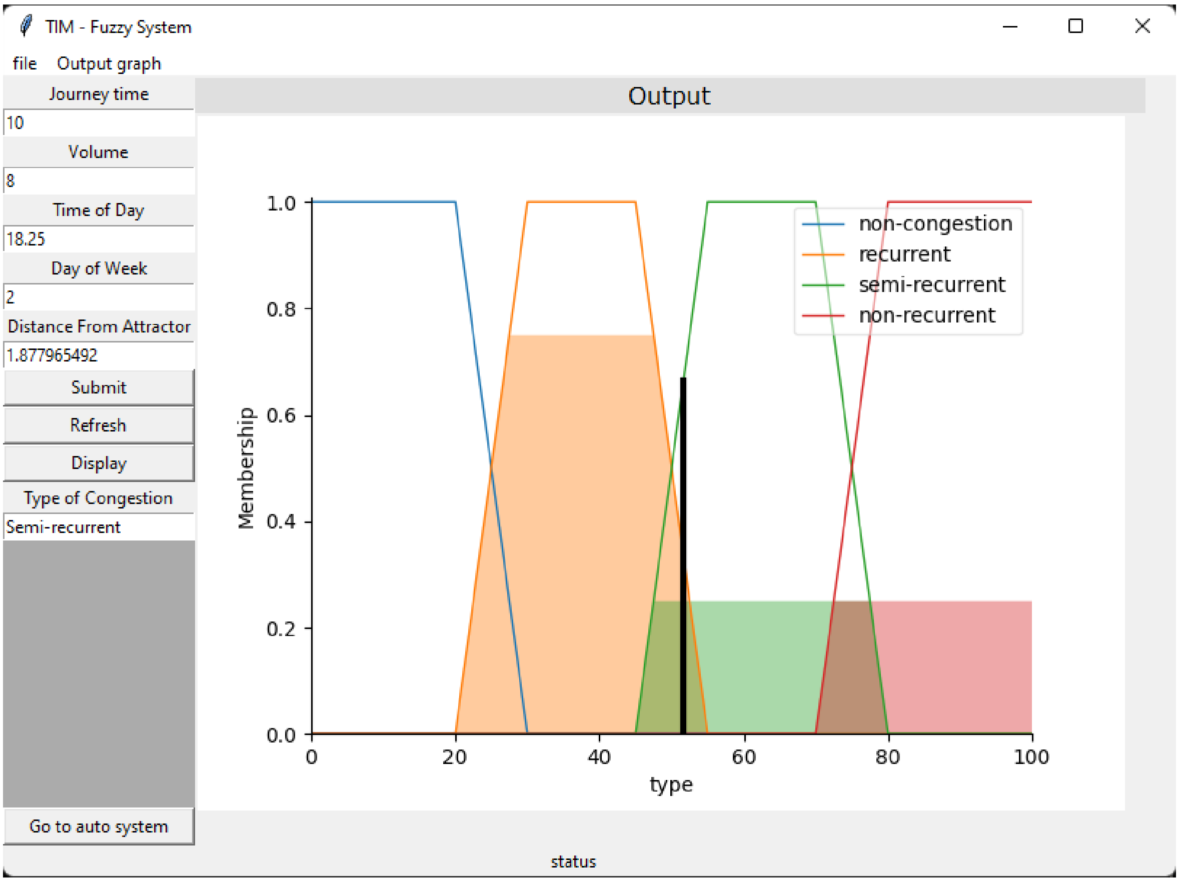
Day of Week: 2

Distance From Attractor: 1.877965492

Type of Congestion: Semi-recurrent

Figure 87: Observation 3: RC classified as SRC

Figure 88 shows a visualisation in TIM of an observation that was expected to be identified as recurrent congestion, however, instead the multiclassification fuzzy system misclassified the observation as semi-recurrent congestion. The reason for the misclassification was due to multiple fuzzy rules firing and creating a consequence for all three classifications: recurrent, semi-recurrent, and non-recurrent congestion with various degree of memberships. Although, the expected classification of recurrent congestion has the highest degree of membership, due to using the centroid defuzzification method, this observation was misclassified.



Journey time: 10

Volume: 8

Time of Day: 18.25

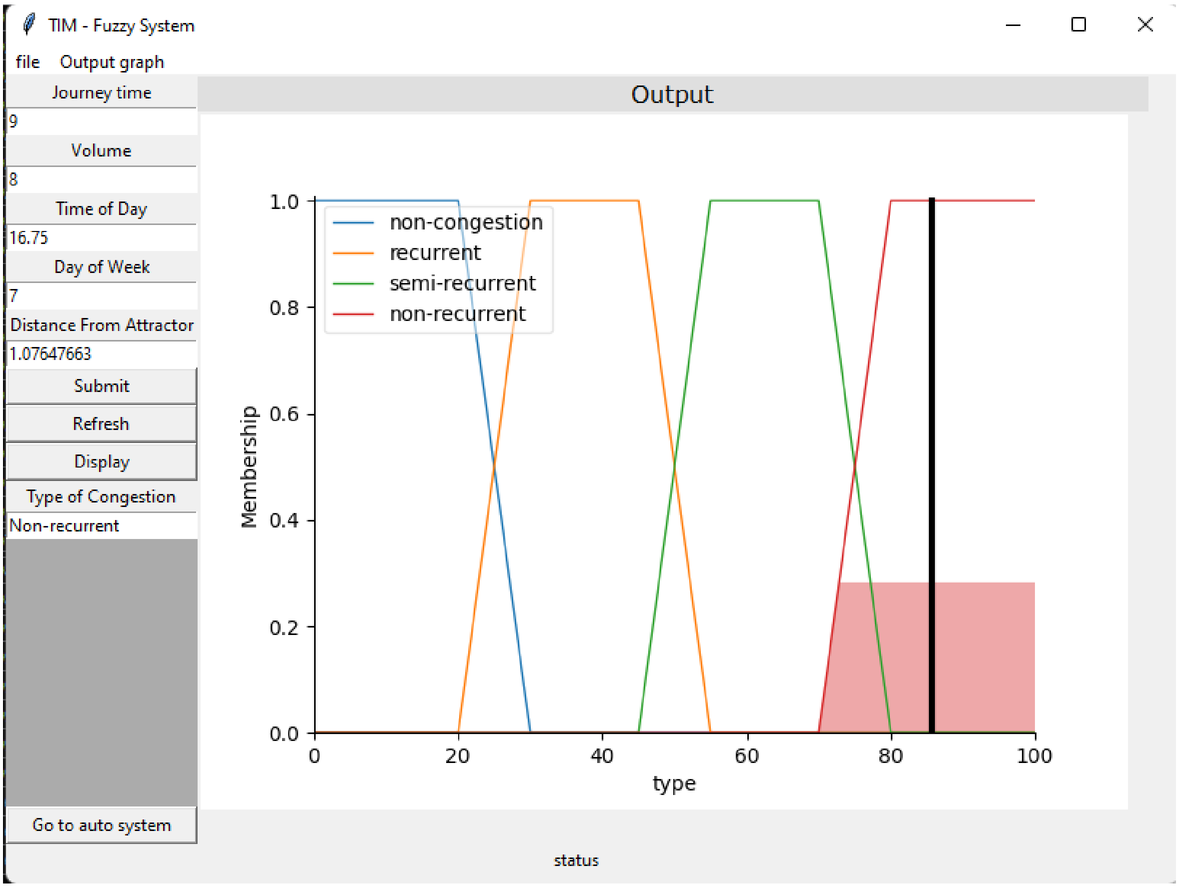
Day of Week: 2

Distance From Attractor: 1.877965492

Type of Congestion: Semi-recurrent

Figure 88: Observation 4: RC classified as SRC

Figure 89 shows a visualisation in TIM of an observation that was expected to be identified as semi-recurrent congestion, however, instead the multiclassification fuzzy system misclassified the observation as non-recurrent congestion. The reason for this misclassification is due to the following control rule which identified the observation as non-recurrent congestion with a ~0.3 degree of membership. The rule is “If journey time is high or very high, day of week is weekend, and distance from attractor is not near or very near then non-recurrent congestion”.



TIM - Fuzzy System

Journey time: 9

Volume: 8

Time of Day: 16.75

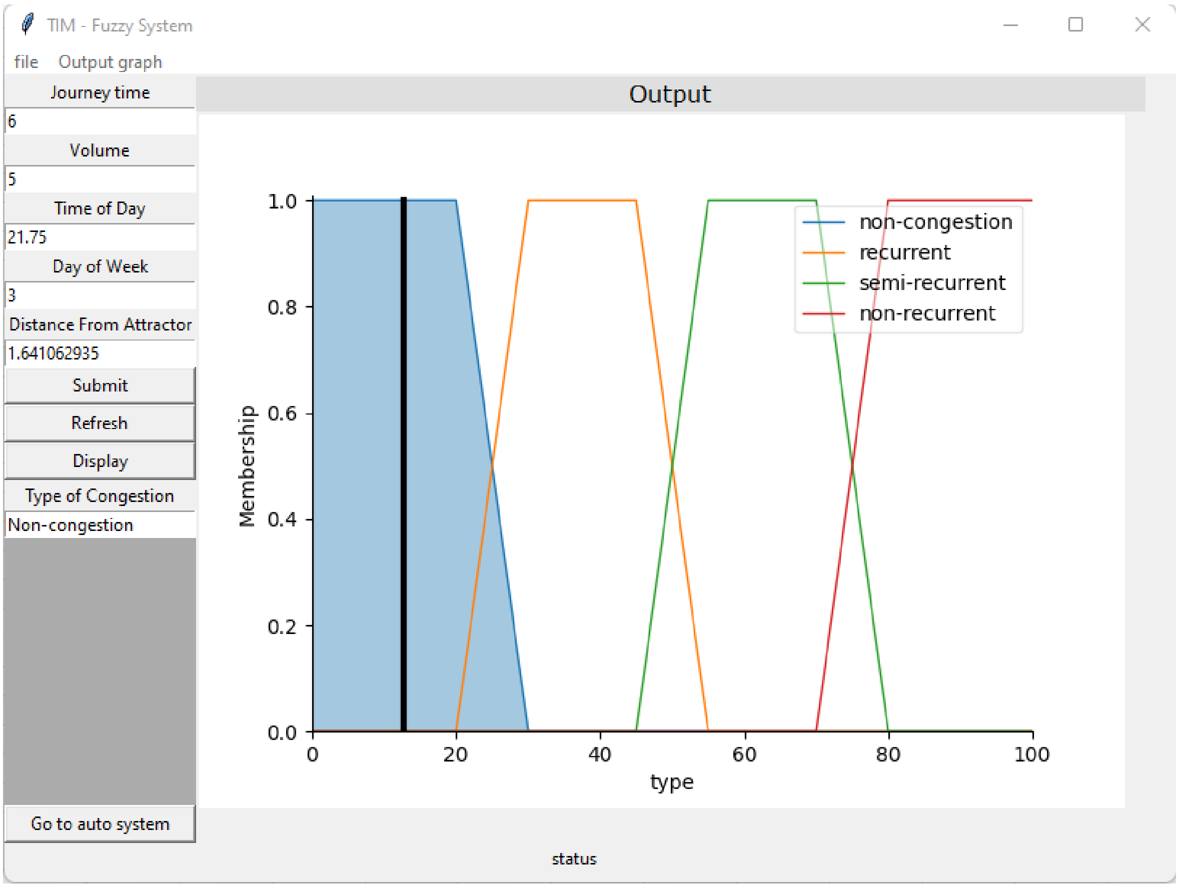
Day of Week: 7

Distance From Attractor: 1.07647663

Type of Congestion: Non-recurrent

Figure 89: Observation 5: SRC classified as NRC

Figure 90 shows a visualisation in TIM of an observation that was expected to be identified as recurrent congestion, however, instead the multiclassification fuzzy system misclassified the observation as non-congestion. The reason for this misclassification is due to the ambiguous non-congestion rule being too loose and taking dominance over the semi-recurrent rules. The non-congestion rule that caused this misclassification is “if journey time is medium and volume is not very high then non-congestion”.



Journey time: 6

Volume: 5

Time of Day: 21.75

Day of Week: 3

Distance From Attractor: 1.641062935

Type of Congestion: Non-congestion

Figure 90: Observation 6: SRC classified as NC

Therefore, after observing the misclassification of the six observations mentioned in Table 18. It was observed that to improve the performance of the multiclassification fuzzy system would require additional fuzzy control rules and optimisations of the membership function boundaries. One of the observations that would benefit from extra rules would be observation 5 which requires a rule similar to a non-recurrent congestion rule, however with the addition of distance from attractor and time of day to reduce the false positives when predicting semi-recurrent. Furthermore, it should be noted that additional rules would reduce the fuzzy system efficiency and explainability, making it more multifaceted and harder for the layperson to understand.

In addition to adding extra control rules to the multiclassification fuzzy system, a more efficient way to determine the membership functions would be to employ a search-based optimization technique known as a Genetic Algorithm (GA) over the manual approached currently used. Alternative approaches for defuzzification may also resolve some of these problems and will be considered in future work. The addition of extra defined rules and GA will be explored in the future (further work) in the hopes of creating a better performing multiclassification fuzzy system at predicting the type of congestion.

7.4.1.12 Conclusion

In conclusion, the multi-classification fuzzy decision-making system did not achieve the same level of performance as the J48 algorithm and performed similarly to the naïve bayes model, even outperforming it some areas, such as predicting recurrent congestion and non-recurrent congestion. However, although, the J48 algorithm tends to be easy to interpret, the overall size of the tree is 2367 and the tree contains 1184 leaves. Consequently, this has generated 1184 rules which create an extra level of complexity when it comes to understand the outcome and lack explainability which is key for a stakeholder to understands the outcome of the model compared with the 12 rules used in the multi-classification fuzzy decision-making system discussed in section 7.4.1.5.

Therefore, although the multi-classification fuzzy decision-making system did not outperform both the J48 and naïve bayes, it is easier for explainability, interpretation, and providing useful qualitive context back to stakeholders which naïve bayes is known to struggle with due to characteristic of probabilistic model which tend to struggle with big datasets. It is also important to note that the fuzzy system was only manually optimised and further work would employ techniques such as genetic algorithms to optimise membership functions.

7.5 Chapter conclusion

In conclusion, this chapter has demonstrated it is possible to use knowledge gained from the URCC model and the creation of a non-optimised multi-classification fuzzy decision-making system to predict urban road congestion validating concepts defined in the universal road congestion ontology. This was achieved used a combination of expert knowledge, an unsupervised learning technique known as clustering, and a percentile model to construct two fuzzy decision-making systems.

The outcome of both fuzzy decision-making systems has proven both hypotheses true. The first hypothesis *H_A0: Using journey time and volume data, it is possible to classify congestion using a fuzzy system* was not only proven true, but it also demonstrated the initial proof of concept setting the groundwork for the second Fuzzy decision-making system. Although, the second system did not perform as strong as the J48 decision tree, it did however, perform at an acceptable level to prove the second hypothesis *H_B0: It is possible to accurately identify the type of road traffic congestion using a Fuzzy system* true. Furthermore, the multi-classification decision-making system is easier to interpret and provide meaningful context compared to the J48 and naïve bayes models. This is because the multi-classification fuzzy decision-making system only uses 12 rules compared to the J48 decision tree which has a total 1184 rules.

In further work, the author plans to increase the performance of the multi-classification fuzzy decision-making system by focusing two main areas, which are data quality and optimisation. To improve the data quality, it is important to address some of the data considerations mentioned in section 4.6. The main consideration the author would like to address is the lack of consistent distance between two Bluetooth sensors and the point the

Bluetooth sensors do not always align with the location of the inductive loop counters.

To achieve this a new topology will be developed with a more flexible alternative data source to the Bluetooth sensors used within this research will be explored. The alternative data source being considered is Googles traffic data (The Directions API) because although, it charges for each API request, it allows the user to define each point to create a link without being limited by physical hardware. Additionally, it should help to remove some of the noise created by pedestrians and cyclists with an active Bluetooth device being recorded as a journey time as Google collects data directly from apps such as google maps and Android auto.

Finally, to optimise the multi-classification fuzzy decision-making system, the techniques known as a genetic algorithm (GA) will be explored as it has previously been used to optimise fuzzy membership functions (Crockett et al., 2013) and the fuzzy inference parameters (K. A. Crockett et al., 2006). Additionally, GAs have been successfully used in other domains than fuzzy systems for the purpose of optimisation. For example, a GA was used within a scheduling-based system for medical treatment (Squires et al., 2022).

Chapter Eight: Conclusion and further work

The research in this thesis has developed a formal and explicit conceptualisation of urban road congestion, which has a multifaceted nature. Analogical and ontological methods were used to conceptualise urban road congestion and produce an Urban Road Congestion Conceptual (URCC) model. The research presented validates this conceptual model using a real-world big data dataset and a custom-built fuzzy decision-making system. In this chapter, a discussion of each research question is provided, and the overall contributions of this thesis are presented.

8.1 Overview

The research presented in this thesis aimed to answer the following four research questions:

RQ1: Is it possible to provide a clear conceptualisation of urban road traffic congestion using an ontological model?

RQ2: Can quantitative Big Data be used to provide qualitative information in conjunction with a road traffic ontology with the support of Machine Learning?

RQ3: Can quantifiable big data on urban road congestion be visualised to provide quasi-real-time insight?

RQ4: Can a Fuzzy rule-based system be designed to predict road congestion through validation of the Urban Road Congestion Conceptual (URCC) model?

How each question has been addressed will now be discussed.

8.1.1 RQ1: Is it possible to provide a clear conceptualisation of urban road traffic congestion using an ontological model?

The main problem with modelling urban road congestion is the lack of a clear and consistent definition of what is meant by '*road congestion*' in an increasingly multifaceted urban context and how it relates to the events that cause it. To address this problem, an Urban Road Congestion Conceptual (URCC) model was created, using four analogies and a universal road congestion ontology which is made up of five core ontologies (Dimensions of congestion, events, congestion, direction. and spatial things). One of the limitations of the current literature regarding urban road congestion became apparent with the development of the URCC model, which identified there to be a lack of granularity between the types of congestion being presented.

Therefore, this research introduced a third type of congestion coined as 'semi-recurrent congestion'. Another key finding was established through the comprehensive review of the literature, these seemingly simple questions, such as What is congestion? What is the cause of congestion? Where has congestion occurred? did not have a clear and consistent way to answer. Using the developed URCC model, these question can now be answered in the same manner every time in a formal and explicit way. Thus, the research question (RQ1) – "*Is it possible to provide a clear conceptualisation of urban* *road traffic congestion using an ontological model?*" was address in chapter three and four.

8.1.2 RQ2: Can quantitative Big Data be used to provide qualitative information in conjunction with a road traffic ontology with the support of Machine Learning?

To answer this question, it was vital that a real-world quasi-real-time big data dataset was created. The data was collected from several sources and merged into a dataset known as the Manchester Urban Congestion Data (MUCD) dataset and was introduced in chapter three. The MUCD dataset has typical data issues associated with big data, such as noise, data sparsity and missing values. However, there were other unique challenges, such as each link having its own different characteristics, such as length size, number of lanes, and different speed limits. Each of these characteristics will cause the expected journey times and traffic volume counts to differ dramatically. Furthermore, another challenge was trying to design a topology which had sufficient coverage of Bluetooth sensors and inductive loop counters whilst encompassing the requirements set out by TfGM, such as focusing on the A6 road and Etihad Stadium.

Once the MUCD dataset has been created, it was important to identify which unsupervised learning algorithm was going to be used. Therefore, in chapter five, the decision was taken to implement k-mean++. A series of empirical experiments were conducted in chapter six in conjunction with the URCC model to identify the characteristics of urban road congestion. The key finding was clustering an unsupervised dataset made it possible to predict expected journey time and identify the differences between a weekday and a weekend. Therefore, this demonstrated that it is possible to take *quantitative* data and extract *qualitative* information, which can be provided to the stakeholders, such as road users or transport managers. The stakeholders (in this case TfGM) could then use the meaningful information to make better decisions. Therefore, answering (RQ2) – "*Can quantitative Big Data be used to provide qualitative information in conjunction with a road traffic ontology with the support of Machine Learning?*"

8.1.3 RQ3: Can quantifiable big data on urban road congestion be visualised to provide quasi-real-time insight?

To answer RQ3, the development of a visualisation tool was needed. Therefore, chapter three describes the development of the visualisation tool called Transport Incident Manager (TIM). TIM is a tool developed by the author using SQL Server and Python to visualise the statistical performance of the urban road network within Manchester, UK. Some of the functionalities created were real-time view of individual links and overall network performance, spatial autocorrelation, classification, and the ability to look at the data in different temporal states. Therefore, TIM has managed to demonstrate to the stakeholders at TfGM it is possible to visualises their quasi-real-time data, such as journey time.

TIM included the implementation of several statistical functions to gain insight into the behaviour and characteristics of congestion and the events that cause it, such as rush hour, a road accident, a football match. The development of TIM and validation by the expert stakeholders at TfGM answers the research question (RQ3) – “*Can quantifiable big data on urban road congestion be visualised to provide quasi-real-time insight?*”. TIM is an adaptable system, which has impact beyond this project and be used to visually model road congestion in wider national / international locations.

8.1.4 RQ4: Can a fuzzy rule-based system be designed to predict road congestion through validation of the Urban Road Congestion Conceptual (URCC) model?

To answer RD4, two fuzzy systems were developed. The first fuzzy decision-making system was a binary classification system, which used the unsupervised learning classifications to assist with determining the membership function for journey time and traffic volume. This system focused purely on a single link, using only two data sources and two classifications (congestion or non-congestion). Once, this system was developed and was proven to be a success, the next step was to develop a second fuzzy decision-making system which is more complex and useful to the stakeholders at TfGM.

The second fuzzy system, incorporated data from multiple sources and predicted on the whole neighbourhood network to classify the road conditions, non-congestion, recurrent, congestion, semi-recurrent, non-recurrent congestion. This system is a great way for TfGM to analysis their network at link level and depending on the type of congestion being identified they can respond in a more meaningful manner. Making the network more resilient. The URCC model was used to create the memberships and rules ensuring the fuzzy system results are consistent with what is defined as urban road congestion. Therefore, the development of a non-optimised multi-classification Fuzzy decision-making system made it possible to answer the research question (RQ4) – “*Can a fuzzy rule-based system be designed to predict road congestion through validation of the Urban Road Congestion Conceptual (URCC) model?*”

8.2 Research Contributions

This research has produced some significant contributions in the field of transportation.

- Firstly, the development of a novel Urban Road Congestion Conceptual (URCC) model which conceptualises the three types of congestion: non-recurrent, semi-recurrent, and recurrent congestion. Being able to conceptualise the events that causes these types of congestion is an important contribution to the stakeholders. It will give them the ability to respond to semi-recurrent causing event such as planned roadworks differently to non-recurrent events, such as unplanned roadworks, which were previously all classified as the same type of congestion. (Chapter Three).

- Secondly, the development of the Manchester Urban Congestion Data (MUCD) Dataset which incorporates real-world data from several sources, such as Transport for Greater Manchester (TFGM) and the United Kingdom's Governments freely open data. The MUCD dataset is the first dataset to combine data from low costing devices, such as Bluetooth sensors with openly free data, such as accident and event data, and more expensive sources, such as inductive loop counters. Being able to integrate these extra data sources with TfGM current data, provides them better opportunity to gain greater knowledge with regards to their network performance. (Chapter Four)
- The third contribution is the development of a visualisation toolkit Graphical User Interface (GUI) called Transport Incident Manager (TIM) which will provide the stakeholders, such as TfGM the ability to visualise and perform statistical analysis on individual links or the whole network in quasi-real-time, this will allow them to respond in a timelier manner making the network more resilient. Additionally, TIM has the ability to feed data from any data source which has the capability to monitor the relevant dimensions, such as journey time and volume. (Chapter Five)
- The fourth contribution is the development of a binary fuzzy decision system to determine if a rule base system could identify congestion at a high level. It was found through empirical experimentation that using a fuzzy system was more efficient than traditional methods such as a decision tree or probabilistic model. Not only was a fuzzy system more efficient, but it also has better explainability for stakeholders to understand as it uses only six linguistical rules to make the prediction of either congestion or non-congestion. (Chapter Six and Seven)
- The fifth contribution is a continuation to the fourth as it involves developing another one-of-a-kind fuzzy decision-making system, however, this fuzzy system is developed to classify multiple types of congestion. The classifications are non-recurrent congestion, semi-recurrent congestion, recurrent congestion, and non-congestion. The novelty of this system is similar to the binary fuzzy decision-making system, as it doesn't require training data to teach the model what patterns to look for. The fuzzy systems are developed using expert knowledge and one of the main benefits of the multi-classification model is it uses only 12 linguistically rules making it easier to explain the outcome of the predictions compared to the decision tree which has 1184 leaves which would need to be explained to understand the prediction. (Chapter Seven)

The research in this thesis has led to the following peer-reviewed publications at the time of submission.

Gould, N. and Abberley, L. (2017) 'The semantics of road congestion.' *In UTSG*. Dublin.

L. Abberley, N. Gould, K. Crockett and J. Cheng, 'Modelling road congestion using ontologies for big data analytics in smart cities,' 2017 International Smart Cities Conference (ISC2), 2017, pp. 1-6, Doi: 10.1109/ISC2.2017.8090795

L. Abberley, K. Crockett and J. Cheng, 'Modelling Road Congestion Using a

Fuzzy System and Real-World Data for Connected and Autonomous Vehicles,' 2019 Wireless Days (WD), 2019, pp. 1-8, Doi: 10.1109/WD.2019.8734238.

The following paper is currently being resubmitted to Transportation Research Interdisciplinary Perspectives following corrections.

- L. Abberley, N. Gould, K. Crockett, J. Cheng (2022) "Development and validation of a conceptual model for different types of road congestion: recurrent, non-recurrent, and semi-recurrent congestion"

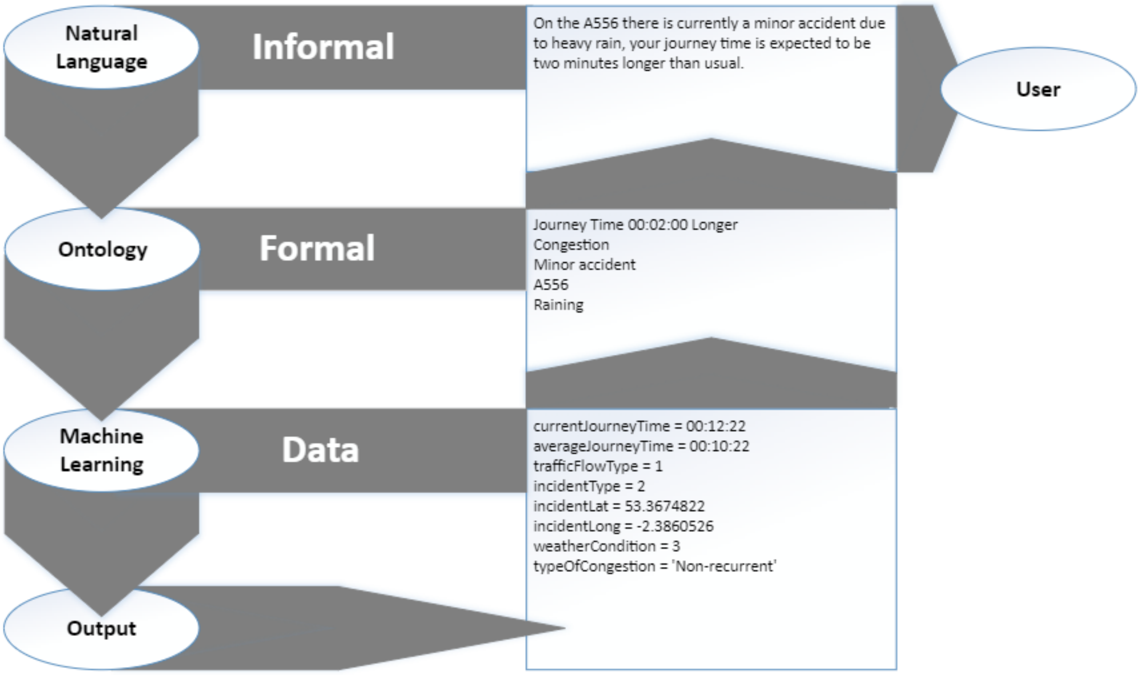
8.3 Future Work

8.3.1 Improve the Manchester Urban Congestion Data (MUCD) Dataset

The first focus with regards to future work is improving on the MUCD Dataset by firstly, increasing the number of links within the Manchester's neighbourhood network topology which is currently 64. The second improvement would be to incorporate more data sources that are in line with the relevant dimensions used to predict urban road congestion, such as traffic volume and journey time. Other data sources include Googles Directions API and Traffic master (https://www.basemap.co.uk/trafficmaster-data/), which both rely on GPS data source and can contribute three main dimensions: Speed, journey time, and traffic volume. The final improvement would be to gain access to more accurate weather in real-time, which will provide more meaningful data to predict the impact of severe weather on urban road congestion.

8.3.2 Extend the Urban Road Congestion Conceptual (URCC)

The second focus with regards to further work would be to advance the ontology. This will be achieved by incorporating new objects to implement prediction techniques and to explore agent-based modelling to simulate the interactions between the different stakeholder (agent) and how they will use URCC. Each stakeholder will have their own properties (attributes) and will use will interact with the URCC in different ways, such as a road user will be focused on a journey on several between A and B, but a traffic manager is likely to focus on the performance of the overall road network. Additional it would be beneficial to explore the relationship between natural language which is informal, the formal and explicit definitions presented in the universal ontology, and the data being used to predict urban road congestion. Figure 91 shows the concept of how informal information can be processed for the use of prediction an output urban road congestion and how that raw output data can then be translated back into a meaningful informal description of congestion which a stakeholder would find useful.



On the A556 there is currently a minor accident due to heavy rain, your journey time is expected to be two minutes longer than usual.

Journey Time 00:02:00 Longer
Congestion
Minor accident
A556
Raining

```
currentJourneyTime = 00:12:22
averageJourneyTime = 00:10:22
trafficFlowType = 1
incidentType = 2
incidentLat = 53.3674822
incidentLong = -2.3860526
weatherCondition = 3
typeOfCongestion = 'Non-recurrent'
```

Figure 91: Relationship between informal, formal, and data

8.3.3 Advancements to Transport Incident Manager (TIM)

The third focus with regards to further work would follow further improvements to the MUCD and would consist of incorporating additional data sources, such as Googles Directions API which can provide journey times for self-defined links into TIM. Other additions would be to allow users to save their session configurations and to incorporate additional statistics such as an average journey time over multiple self-defined links, for instance, link {a, b, c} to form a new larger 'temporary link' that incorporates several smaller links.

8.3.4 Optimisation of the membership functions

The fourth focus with regards to further work would be to research and incorporate a Genetic algorithm into the multi-classification decision-making system to optimise the membership functions to achieve better performance at prediction the type of congestion occurring. In the first instance, this could involve coding the lower and upper bounds of all membership functions in the system onto a chromosome and defining a fitness function which maximises the prediction accuracy. The challenges will lie in coding the problem and determining the most suitable fitness function.

8.4 Overall conclusion

In conclusion, this chapter has clearly stated the main research questions and explained how they were address (section 8.1), discussed the significant contributions in the field of transportation (section 8.3), and discussed the limitation of this research and made recommendation for future work (section 8.3). This has posed three concluding thoughts that need to be address.

The first thought is how can the proposed approach be generalised to address other situations?

The proposed method of conceptualising urban road congestion using ontologies to assist with the creation of a fuzzy system capable of predicting congestion and what type has occurred is scalable as long as the data is present, however, reliable data sources would need to be considered. Additionally, the proposed method already takes into consideration the different characteristics of each road (link), such as road length, speed limit, capacity, direction, distance from attractors, etc. Therefore, a percentile approach was taken to generalise journey time and volume. So regardless of each link characteristics, very high journey time means the same thing on each link, making it possible for this approach to be extended to highways and rural areas, however, more concepts may need to be added to account for different road behaviours.

The second thought is what is the model transferability?

The proposed fuzzy system has the ability to work in other countries and other major cities in the United Kingdom, such as Birmingham and London. However, a few considerations that need to be considered are: alternative data sources and city specific constraints, such as London’s congestion charge and Birmingham’s zero emissions zones would need to be modelled in the urban road congestion ontology and the fuzzy system to maintain performance.

Finally, what calibration is necessary to use the proposed method on other data?

The benefit of the proposed method is it has been developed around selected dimensions rather than specific data sources. For instance, the dimension known as journey time can use any of the several different data sources, such as Bluetooth sensors, ANPR cameras, and GPS. As part of early experimental exploration, Google API (GPS) data was explored, and it was noted that the journey time from Bluetooth sensors and Google API were both capable of working with the approach. However, the limitation of Google API is the expensive cost of each API request.

References

Abberley, L. (2016) 'Internal Technical report.'

Abberley, L., Crockett, K. and Cheng, J. (2019) 'Modelling Road Congestion Using a Fuzzy System and Real-World Data for Connected and Autonomous Vehicles.' *2019 Wireless Days (WD)*. IEEE pp. 1–8.

Abberley, L., Gould, N., Crockett, K. and Cheng, J. (2017) 'Modelling road congestion using ontologies for big data analytics in smart cities.' *In 2017 International Smart Cities Conference (ISC2)*. Wuxi, China: IEEE, pp. 1–6.

Acharya, B. K., Cao, C., Xu, M., Chen, W. and Pandit, S. (2018) 'Spatiotemporal Distribution and Geospatial Diffusion Patterns of 2013 Dengue Outbreak in Jhapa District, Nepal.' *Asia-Pacific Journal of Public Health*. SAGE PublicationsSage CA: Los Angeles, CA, April, p. 1010539518769809.

Agarwal, A. and Kickhöfer, B. (2015) 'Agent-based simultaneous optimization of congestion and air pollution: A real-world case study.' *Procedia Computer Science*, 52(1) pp. 914–919.

Aggarwal, C. C. (2013) 'A Survey of Uncertain Data Clustering Algorithms.' *Data Clustering: Algorithms and Applications*, 21(5) pp. 455–480.

Amini, M., Hatwagner, M. F., Mikulai, G. C. and Koczy, L. T. (2021) 'An intelligent traffic congestion detection approach based on fuzzy inference system.' *In SACI 2021 - IEEE 15th International Symposium on Applied Computational Intelligence and Informatics, Proceedings*. Institute of Electrical and Electronics Engineers Inc., pp. 97–104.

Anbaroğlu, B., Cheng, T. and Heydecker, B. (2015) 'Non-recurrent traffic congestion detection on heterogeneous urban road networks.' *Transportmetrica A: Transport Science*, 11(9) pp. 754–771.

Anbaroglu, B., Heydecker, B. and Cheng, T. (2014) 'Spatio-temporal clustering for non-recurrent traffic congestion detection on urban road networks.' *Transportation Research Part C: Emerging Technologies*. Elsevier Ltd, 48 pp. 47–65.

Arampatzis, G., Kiranoudis, C. T., Scaloubacas, P. and Assimacopoulos, D. (2004) 'A GIS-based decision support system for planning urban transportation policies.' *European Journal of Operational Research*, 152(2) pp. 465–475.

Arnott, R. (2013) 'A bathtub model of downtown traffic congestion.' *Journal of Urban Economics*. Elsevier Inc., 76(1) pp. 110–121.

Arnott, R. and Buli, J. (2018) 'Solving for equilibrium in the basic bathtub model.' *Transportation Research Part B: Methodological*. Pergamon, 109, March, pp. 150–175.

Arnott, R., Palma, A. De and Lindsey, R. (1993) 'A Structural Model of Peak-Period Congestion: A Traffic Bottleneck with Elastic Demand.' *The American Economic Review*, 83(1) pp. 161–179.

Arthur, D. and Vassilvitskii, S. (2007) ‘K-Means++: the Advantages of Careful Seeding.’ *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, 8 pp. 1027–1025.

Atkin, D. (2016) ‘Personal Communication.’

Bao, Y., Verhoef, E. T. and Koster, P. (2021) ‘Leaving the tub: The nature and dynamics of hypercongestion in a bathtub model with a restricted downstream exit.’ *Transportation Research Part E: Logistics and Transportation Review*. Pergamon, 152, August, p. 102389.

Bar-Gera, H. (2007) ‘Evaluation of a cellular phone-based system for measurements of traffic speeds and travel times: A case study from Israel.’ *Transportation Research Part C: Emerging Technologies*. Elsevier Ltd, 15(6) pp. 380–391.

Bauza, R. and Gozalvez, J. (2013) ‘Traffic congestion detection in large-scale scenarios using vehicle-to-vehicle communications.’ *Journal of Network and Computer Applications*. Elsevier, 36(5) pp. 1295–1307.

Bauza, R., Gozalvez, J. and Sanchez-Soriano, J. (2010) ‘Road traffic congestion detection through cooperative Vehicle-to-Vehicle communications.’ *Proceedings - Conference on Local Computer Networks, LCN*. IEEE pp. 606–612.

Bifulco, G. N., Cantarella, G. E., Simonelli, F. and Velonà, P. (2016) ‘Advanced traveller information systems under recurrent traffic conditions: Network equilibrium and stability.’ *Transportation Research Part B: Methodological*. Elsevier Ltd, 92 pp. 73–87.

Black, W. R. and Thomas, I. (1998) ‘Accidents on Belgium’s motorways: A network autocorrelation analysis.’ *Journal of Transport Geography*, 6(1) pp. 23–31.

Breitman, K. K., Barbosa, S. D. J., Casanova, M. a. and Furtado, A. L. (2007) ‘Conceptual modeling by analogy and metaphor.’ *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management - CIKM ’07* p. 865.

Cassidy, M. J. and Bertini, R. L. (1999) ‘Some traffic features at freeway bottlenecks.’ *Transportation Research Part B: Methodological*, 33(1) pp. 25–42.

Changnon, S. A. (1996) ‘Effects of summer precipitation on urban transportation.’ *Climatic Change*, 32(4) pp. 481–494.

Chen, C. L., Chen, P. C. and Chen, C. K. (1993) ‘Analysis and design of fuzzy control system.’ *Fuzzy Sets and Systems*. North-Holland, 57(2) pp. 125–140.

Chen, D. and Ahn, S. (2015) ‘Variable speed limit control for severe non-recurrent freeway bottlenecks.’ *Transportation Research Part C: Emerging Technologies*. Elsevier Ltd, 51 pp. 210–230.

Chen, H. and Rakha, H. A. (2016) ‘Multi-step prediction of experienced travel times using agent-based modeling.’ *Transportation Research Part C: Emerging Technologies*. Elsevier Ltd, 71 pp. 108–121.

Chen, P.-T., Chen, F. and Qian, Z. (2014) ‘Road Traffic Congestion Monitoring in Social Media with Hinge-Loss Markov Random Fields.’ *2014 IEEE International Conference on Data Mining* pp. 80–89.

Chen, Y. (2020) ‘An Analytical Process of Spatial Autocorrelation Functions Based on Moran’s Index’ pp. 1–27.

Chen, Y., Chen, C., Wu, Q., Ma, J., Zhang, G. and Milton, J. (2020) ‘Spatial-temporal traffic congestion identification and correlation extraction using floating car data.’ *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations*. Taylor and Francis Inc. pp. 1–18.

Chen, Y., Sabri, S., Rajabifard, A. and Agunbiade, M. E. (2018) ‘An ontology-based spatial data harmonisation for urban analytics.’ *Computers, Environment and Urban Systems*. Elsevier, 72(February) pp. 177–190.

Chen, Z., Liu, X. C. and Zhang, G. (2016) ‘Non-recurrent congestion analysis using data-driven spatiotemporal approach for information construction.’ *Transportation Research Part C: Emerging Technologies*, 71 pp. 19–31.

Cheng, T., Haworth, J. and Wang, J. (2012) ‘Spatio-temporal autocorrelation of road network data.’ *Journal of Geographical Systems*, 14(4) pp. 389–413.

Colak, S., Lima, A. and Gonz, M. C. (2016) ‘Understanding congested travel in urban areas.’

Corsar, D., Markovic, M., Edwards, P. and Nelson, J. D. (2015) *The Transport Disruption ontology*. [Online] [Accessed on 10th October 2015] https://transportdisruption.github.io/transportdisruption.html.

Couclelis, H. (1992) ‘People manipulate objects (but cultivate fields): Beyond the Raster-Vector Debate in GIS.’ *Theories and Methods of Spatiotemporal Reasoning in Geographic Space*, 639(716) pp. 65–77.

Cox, S. and Little, C. (2017) *OWL-Time*. [Online] https://www.w3.org/TR/owl-time/.

Crockett, K. A., Bandar, Z., Fowdar, J. and O’Shea, J. (2006) ‘Genetic tuning of fuzzy inference within fuzzy classifier systems.’ *Expert Systems*, 23(2) pp. 63–82.

Crockett, K., Bandar, Z., Mclean, D. and O’Shea, J. (2006) ‘On constructing a fuzzy inference framework using crisp decision trees.’ *Fuzzy Sets and Systems*, 157(21) pp. 2809–2832.

Crockett, K., Latham, A., Mclean, D. and O’Shea, J. (2013) ‘A fuzzy model for predicting learning styles using behavioral cues in a conversational intelligent tutoring system.’ *IEEE International Conference on Fuzzy Systems*.

Demchenko, Y., Grosso, P., De Laat, C. and Membrey, P. (2013) ‘Addressing big data issues in Scientific Data Infrastructure.’ *Proceedings of the 2013 International Conference on Collaboration Technologies and Systems, CTS 2013* pp. 48–55.

Department for Transport (2013) ‘An introduction to the Department for Transport’ s road congestion statistics,’ (August).

Department for Transport (2014) *Transport Resilience Review - A review of the resilience of the transport network to extreme weather events*.

Department for Transport (2018) 'Reported Road Casualties in Great Britain: notes, definitions, symbols and conventions.' *Department for Transport* pp. 1–6.

Department for Transport (2020) *Statistical Release Minor Road Traffic Estimates Revisions*.

Djahel, S., Doolan, R., Muntean, G.-M. and Murphy, J. (2015) 'A Communications-Oriented Perspective on Traffic Management Systems for Smart Cities: Challenges and Innovative Approaches.' *IEEE Communications Surveys & Tutorials*, 17(1) pp. 125–151.

Djahel, S., Jones, A., Hadjadj-Aoul, Y. and Khokhar, A. (2018) 'CRITIC: A cognitive radio inspired road traffic congestion reduction solution.' *IFIP Wireless Days*, 2018-April (February) pp. 151–157.

Djahel, S., Jones, A., Hadjadj-aoul, Y. and Khokhar, A. (2018) 'CRITIC: A Cognitive Radio Inspired Road Traffic Congestion Reduction Solution CRITIC : A Cognitive Radio Inspired Road Traffic Congestion Reduction Solution.' *In The 10th Wireless Days Conference (WD 2018)*.

Downs, A. (2005) *Still Stuck in Traffic: Coping with Peak-Hour Traffic Congestion*. Brookings Institution Press.

Economics, D., Arnott, R., Bates, J. J., Hall, F., Timothy, H., John, M. and Se-il, M. (2003) 'Kenneth A. Small and Xuehao Chu,' 37(May) pp. 319–352.

Emmerink, R. H. M., Axhausen, K. W., Nijkamp, P. and Rietveld, P. (1995) 'The potential of information provision in a simulated road transport network with non-recurrent congestion.' *Transportation Research Part C*, 3(5) pp. 293–309.

Faro, A. and Giordano, D. (2016) 'Algorithms to find shortest and alternative paths in free flow and congested traffic regimes.' *Transportation Research Part C: Emerging Technologies*. Elsevier Ltd, 73 pp. 24–28.

Fernandez, S. and Ito, T. (2015) 'Driver Behavior Model Based on Ontology for Intelligent Transportation Systems.' *2015 IEEE 8th International Conference on Service-Oriented Computing and Applications (SOCA)* pp. 227–231.

Fernandez-Caballero, A., Gomez, F. J. and Lopez-Lopez, J. (2008) 'Road-traffic monitoring by knowledge-driven static and dynamic image analysis.' *Expert Systems with Applications*, 35(3) pp. 701–719.

Fosgerau, M. and Small, K. A. (2013) 'Hypercongestion in downtown metropolis.' *Journal of Urban Economics*. Elsevier Inc., 76 pp. 122–134.

Fox, M. S. (2015) 'The role of ontologies in publishing and analyzing city indicators.' *Computers, Environment and Urban Systems*. Elsevier B.V., 54 pp. 266–279.

Gani, A. (2015) 'Sinkhole opens on busy Manchester Road.' *The Guardian*. August.

Golestan, K., Soua, R., Karray, F. and Kamel, M. S. (2015) 'Situation awareness within the context of connected cars: A comprehensive review and recent trends.' *Information Fusion*, 29, May, pp. 68–83.

Google (2016) *Google Traffic Maps*. [Online] [Accessed on 22nd November 2016] https://www.google.co.uk/maps/place/Manchester/@53.4712328,-2.242336,17z/data=!3m1!4b1!4m5!3m4!1s0x487bb1ec9277cf3d:0x737d31e8e02cefa2!8m2!3d53.4712328!4d-2.2401473!5m1!1e1.

Gould, N. and Abberley, L. (2017) 'The semantics of road congestion.' *In UTSG*. Dublin.

Gould, N. M., Mackaness, W. A., Touya, G. and Hart (2014) 'Collaboration on an Ontology for Generalisation,' (September).

Gov.uk (2017) *STATS19*. [Online] [Accessed on 15th October 2016] https://data.gov.uk/dataset/road-accidents-safety-data.

Grote, M., Williams, I., Preston, J. and Kemp, S. (2016) 'Including congestion effects in urban road traffic CO2 emissions modelling: Do Local Government Authorities have the right options?' *Transportation Research Part D: Transport and Environment*. Elsevier Ltd, 43 pp. 95–106.

Gualtieri, G. and Tartaglia, M. (1998) 'Predicting urban traffic air pollution: A gis framework.' *Transportation Research Part D: Transport and Environment*, 3(5) pp. 329–336.

GUO, R. and HUANG, H. (2009) 'Network Traffic Flow Evolution Model Considering OD Demand Mutation.' *Systems Engineering - Theory & Practice*. Systems Engineering Society of China, 29(1) pp. 118–123.

Hall, F. and Agyemang-Duah, K. (2000) 'Freeway Capacity Drop and the Definition of Capacity' pp. 91–98.

Hara, Y. (2015) 'Behaviour Analysis Using Tweet Data and geo-tag Data in a Natural Disaster.' *Transportation Research Procedia*, 11 pp. 399–412.

Hartgen, D. and Fields, G. (2009) 'Gridlock and Growth: The Effect of Traffic Congestion on Regional Economic Performance,' (371).

He, Z., Zheng, L., Song, L. and Zhu, N. (2016) 'A Jam-Absorption Driving Strategy for Mitigating Traffic Oscillations.' *IEEE Transactions on Intelligent Transportation Systems* pp. 1–12.

Hendricks, D. L., Engineering, V., Freedman, M., Zador, P. L., Fell, J. C., Mountain, S., Page, J. F., Bellis, E. S., Scheifflee, T. G., Hendricks, S. L., Steinberg, G. V and Lee, K. C. (2001) 'THE RELATIVE FREQUENCY OF UNSAFE Contract No . DTNH22-94-C-05020 Final Report Authors: Veridian Engineering , Inc . U. S. Department of Transportation National Highway Traffic Safety Administration Office of Program Development and Evaluation Washington.'

Herman, R. and Prigogine, I. (1979) 'A two-fluid approach to town traffic.' *Science (New York, N.Y.)*, 204(4389) pp. 148–151.

Highways England (2015) *Smart motorway*.

Hooke, A., Knox, J. and Portas, D. (1996) ‘Cost benefit analysis of traffic light & speed cameras.’

Iqbal, K., Khan, M. A., Abbas, S. and Hasan, Z. (2018) *Intelligent Transportation System (ITS) for Smart-Cities using Mamdani Fuzzy Inference System*. *IJACSA) International Journal of Advanced Computer Science and Applications*.

Isa, N., Yusoff, M. and Mohamed, A. (2014) ‘A Review on Recent Traffic Congestion Relief Approaches.’ *2014 4th International Conference on Artificial Intelligence with Applications in Engineering and Technology*, i(November) pp. 121–126.

Jagadish, H. V. (2015) ‘Big Data and Science: Myths and Reality.’ *Big Data Research*. Elsevier Inc., 2(2) pp. 49–52.

Jelokhani-Niaraki, M. (2018) ‘Knowledge sharing in Web-based collaborative multicriteria spatial decision analysis: An ontology-based multi-agent approach.’ *Computers, Environment and Urban Systems*. Elsevier, 72(March) pp. 104–123.

Jin, C. J., Wang, W., Jiang, R., Zhang, H. M., Wang, H. and Hu, M. Bin (2015) ‘Understanding the structure of hyper-congested traffic from empirical and experimental evidences.’ *Transportation Research Part C: Emerging Technologies*. Elsevier Ltd, 60 pp. 324–338.

Jin, W. L. (2020) ‘Generalized bathtub model of network trip flows.’ *Transportation Research Part B: Methodological*. Pergamon, 136, June, pp. 138–157.

Jones, P. (2016) ‘Congestion Reduction in Europe: Advancing Urban Congestion and Network Operation : Towards a Broader Set of Metrics for Assessing Performance’ pp. 1–44.

Jung, C. Te, Sun, C. H. and Yuan, M. (2013) ‘An ontology-enabled framework for a geospatial problem-solving environment.’ *Computers, Environment and Urban Systems*, 38(1) pp. 45–57.

Kaddoura, I. and Nagel, K. (2016) ‘Agent-based Congestion Pricing and Transport Routing with Heterogeneous Values of Travel Time Savings.’ *Procedia Computer Science*. Elsevier Masson SAS, 83 pp. 908–913.

Kianfar, J. and Edara, P. (2013) ‘A Data Mining Approach to Creating Fundamental Traffic Flow Diagram.’ *Procedia - Social and Behavioral Sciences*. Elsevier B.V., 104 pp. 430–439.

Kilpeläinen, M. and Summala, H. (2007) ‘Effects of weather and weather forecasts on driver behaviour.’ *Transportation Research Part F: Traffic Psychology and Behaviour*, 10(June 2015) pp. 288–299.

Knoop, V., Hoogendoorn, S. and van Zuylen, H. (2008) ‘Capacity Reduction at Incidents: Empirical Data Collected from a Helicopter.’ *Transportation Research Record: Journal of the Transportation Research Board*, 2071 pp. 19–25.

Koetse, M. J. and Rietveld, P. (2009) ‘The impact of climate change and weather on transport: An overview of empirical findings.’ *Transportation Research Part D: Transport and Environment*. Elsevier Ltd, 14(3) pp. 205–221.

Kohli, D., Sliuzas, R., Kerle, N. and Stein, A. (2012) ‘An ontology of slums for image-based classification.’ *Computers, Environment and Urban Systems*. Pergamon, 36(2) pp. 154–163.

Kurkcu, A. and Ozbay, K. (2017) ‘Estimating Pedestrian Densities, Wait Times, and Flows with Wi-Fi and Bluetooth Sensors:’ *https://doi.org/10.3141/2644-09*. SAGE PublicationsSage CA: Los Angeles, CA, 2644(1) pp. 72–82.

Kwon, J., Mauch, M. and Varaiya, P. (2006) ‘Components of Congestion: Delay from Incidents, Special Events, Lane Closures, Weather, Potential Ramp Metering Gain, and Excess Demand.’ *Transportation Research Record: Journal of the Transportation Research Board*, 1959(1959) pp. 84–91.

Latham, A. M. (2011) ‘Personalising Learning with Dynamic Prediction and Adaptation to Learning Styles in a Conversational Intelligent Tutoring System,’ (December).

Lécué, F., Schumann, A. and Sbodio, M. L. (2012) ‘Applying semantic web technologies for diagnosing road traffic congestions.’ *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 7650 LNCS (PART 2) pp. 114–130.

Lee, J. and Li, S. (2017) ‘Extending Moran’s Index for Measuring Spatiotemporal Clustering of Geographic Events.’ *Geographical Analysis*, 49(1) pp. 36–57.

Li, C. S. and Chen, M. C. (2014) ‘A data mining based approach for travel time prediction in freeway with non-recurrent congestion.’ *Neurocomputing*. Elsevier, 133 pp. 74–83.

Li, L. (2015) ‘Research on traffic congestion mathematical model in traffic signal control system.’ *International Journal of Smart Home*, 9(12) pp. 279–288.

Li, Y., Guo, T., Xia, R. and Xie, W. (2018) ‘Road Traffic Anomaly Detection Based on Fuzzy Theory.’ *IEEE Access*. IEEE, 6 pp. 40281–40288.

Liang, Z. and Wakahara, Y. (2014) ‘Real-time urban traffic amount prediction models for dynamic route guidance systems.’ *EURASIP Journal on Wireless Communications and Networking*, 2014(1) p. 85.

Lieberman, J., Singh, R. and Goad, C. (2015) *Geospatial*. [Online] https://www.w3.org/2005/Incubator/geo/XGR-geo-ont-20071023/.

Liu, Y. Y., Wang, Y. Q., An, R. and Li, C. (2015) ‘The spatial distribution of commuting CO2 emissions and the influential factors: A case study in Xi’an, China.’ *Advances in Climate Change Research*. Elsevier Ltd, 6(1) pp. 46–55.

López, F. A., Páez, A., Carrasco, J. A. and Ruminot, N. A. (2017) ‘Vulnerability of nodes under controlled network topology and flow autocorrelation conditions.’ *Journal of Transport Geography*, 59 pp. 77–87.

Lozano, A., Manfredi, G. and Nieddu, L. (2009) ‘An algorithm for the recognition of levels of congestion in road traffic problems.’ *Mathematics and Computers in Simulation*, 79(6) pp. 1926–1934.

Luan, S., Ma, X., Li, M., Su, Y. and Dong, Z. (2021) ‘Detecting and interpreting non-recurrent congestion from traffic and social media data.’ *IET Intelligent Transport Systems*. John Wiley and Sons Inc, 15(12) pp. 1461–1477.

Mamdani, E. H. and Assilian, S. (1975) ‘An experiment in linguistic synthesis with a fuzzy logic controller.’ *International Journal of Man-Machine Studies*, 7(1) pp. 1–13.

Mandal, K., Sen, A., Chakraborty, A. and Roy, S. (2011) ‘Road Traffic Congestion Monitoring and Measurement using Active RFID and GSM Technology’ pp. 1375–1379.

Maruyama, T. and Sumalee, a (2007) ‘Efficiency and equity comparison of cordon-and area-based road pricing schemes using a trip-chain equilibrium model.’ *Transportation Research Part A*, 41(7) pp. 655–671.

Mathew, J. and Xavier, P. M. (2014) ‘A survey on using wireless signals for road traffic detection.’ *IJRET: International Journal of Research in Engineering and Technology*, 1163(2319) pp. 97–102.

Mudge, R., Montgomery, D., Groshen, E., Groshen, J. P., Helper, S. and Carson, C. (2018) ‘America’s Workforce and the Self-Driving Future Realizing Productivity Gains and Spurring Economic Growth,’ (June).

Nankervis, M. (1999) ‘The effect of weather and climate on bicycle commuting.’ *Transportation Research Part A: Policy and Practice*, 33 pp. 417–431.

News, M. E. (2015) *Huge hole opens in Mancunian Way after flooding causes major disruption*. [Online] [Accessed on 1st September 2015] http://www.manchestereveningnews.co.uk/news/greater-manchester-news/manchester-rain-flooding-travel-live-9856044.

News, M. E. (2016) *Mancunian Way to fully reopen after giant sinkhole appeared in road*. [Online] [Accessed on 24th August 2016] http://www.manchestereveningnews.co.uk/news/greater-manchester-news/mancunian-way-open-date-sinkhole-11416320.

Noy, N. F. and McGuinness, D. L. (2001) ‘Ontology Development 101: A Guide to Creating Your First Ontology.’ *Stanford Knowledge Systems Laboratory* p. 25.

OECD (2006) *Managing Urban Traffic Congestion*. *Managing*.

OnStar (2016). [Online] [Accessed on 10th August 2016] https://www.onstar.com/us/en/home.html.

Othman, N. Bin, Legara, E. F., Selvam, V. and Monterola, C. (2015) ‘A Data-Driven Agent-Based Model of Congestion and Scaling Dynamics of Rapid Transit Systems.’ *Journal of Computational Science*. Elsevier B.V., 10 pp. 338–350.

de Palma, A. and Lindsey, R. (2011) ‘Traffic congestion pricing methodologies and technologies.’ *Transportation Research Part C: Emerging Technologies*, 19(6) pp. 1377–1399.

Pan, B., Zheng, Y., Wilkie, D. and Shahabi, C. (2013) ‘Crowd Sensing of Traffic Anomalies Based on Human Mobility and Social Media.’ *Proceedings of the*

21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems pp. 344–353.

Patire, A. D., Wright, M., Prodhomme, B. and Bayen, A. M. (2015) ‘How much GPS data do we need?’ *Transportation Research Part C: Emerging Technologies*. Elsevier Ltd, 58 pp. 325–342.

Philip Chen, C. L. and Zhang, C. Y. (2014) ‘Data-intensive applications, challenges, techniques and technologies: A survey on Big Data.’ *Information Sciences*. Elsevier Inc., 275 pp. 314–347.

Pongpaibool, P., Tangamchit, P. and Noodwong, K. (2007) ‘Evaluation of road traffic congestion using fuzzy techniques.’ *IEEE Region 10 Annual International Conference, Proceedings/TENCON* pp. 1–4.

Pope, J. A., Rakes, T. R., Rees, L. P., Crouch, I. W. M., Pope, J. A., Rakes, T. R. and Rees, L. P. (1995) ‘A Network Simulation of High-Congestion Road-Traffic Flows in Cities with Marine Container Terminals Published by : Palgrave Macmillan Journals on behalf of the Operational Research Society Stable URL : http://www.jstor.org/stable/2584496 A Network Simula,’ 46(9) pp. 1090–1101.

Pourjavad, E. and Mayorga, R. v. (2019) ‘A comparative study and measuring performance of manufacturing systems with Mamdani fuzzy inference system.’ *Journal of Intelligent Manufacturing*. Springer New York LLC, 30(3) pp. 1085–1097.

Pourjavad, E. and Shahin, A. (2018) ‘The Application of Mamdani Fuzzy Inference System in Evaluating Green Supply Chain Management Performance.’ *International Journal of Fuzzy Systems*. Springer Berlin Heidelberg, 20(3) pp. 901–912.

Radak, J., Ducourthial, B., Cherfaoui, V. and Bonnet, S. (2015) ‘Detecting Road Events Using Distributed Data Fusion: Experimental Evaluation for the Icy Roads Case.’ *IEEE Transactions on Intelligent Transportation Systems* pp. 1–11.

Raimond, Y. and Abdallah, S. (2007) *The Event Ontology*. [Online] [Accessed on 10th October 2016] http://motools.sourceforge.net/event/event.html.

Reggiani, A., Nijkamp, P. and Lanzi, D. (2015) ‘Transport resilience and vulnerability: The role of connectivity.’ *Transportation Research Part A: Policy and Practice*. Elsevier Ltd, 81 pp. 4–15.

Riad, A. M. and Shabana, B. T. (2012) ‘Real Time Route for Dynamic Road Congestions,’ 9(3) pp. 423–428.

Romilly, P. (1999) ‘Substitution of bus for car travel in urban Britain: An economic evaluation of bus and car exhaust emission and other costs.’ *Transportation Research Part D: Transport and Environment*, 4(2) pp. 109–125.

Rui, L., Zhang, Y., Huang, H. and Qiu, X. (2018) ‘A new traffic congestion detection and quantification method based on comprehensive fuzzy assessment in VANET.’ *KSII Transactions on Internet and Information Systems*, 12(1) pp. 41–60.

Saberi, A., Student, N. P. H. D., Rezaei, M., Dashti Barmaki, M. and Student, P. H. D. (2012) *Analysis of Groundwater Quality using Mamdani Fuzzy Inference System (MFIS) in Yazd Province, Iran*. *International Journal of Computer Applications*.

Van Schijndel, W. J. and Dinwoodie, J. (2000) 'Congestion and multimodal transport: A survey of cargo transport operators in the Netherlands.' *Transport Policy*, 7(4) pp. 231–241.

Sen, R., Siriah, P. and Raman, B. (2011) 'RoadSoundSense: Acoustic sensing based road congestion monitoring in developing regions.' *2011 8th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks, SECON 2011* pp. 125–133.

Shao, L., Wang, C., Liu, L. and C., J. (2015) 'road topology-based scheme for traffic condition estimation via vehicular crowdsensing.' *Concurrency Computation Practice and Experience*, 22(6) pp. 685–701.

Shekhar, S., Jiang, Z., Ali, R. Y., Eftelioglu, E., Tang, X., Gunturi, V. M. V and Zhou, X. (2015) 'Spatiotemporal data mining: A computational perspective.' *ISPRS International Journal of Geo-Information*, 4(4) pp. 2306–2338.

Sheu, J. (1999) 'A stochastic modeling approach to dynamic prediction of section-wide inter-lane and intra-lane traffic variables using point detector data.' *Transportation Research*, 33 pp. 79–100.

Sheu, J.-B. and Ritchie, S. G. (1998) 'A new methodology for incident detection and characterization on surface streets.' *Transportation Research Part C: Emerging Technologies*, 6(5–6) pp. 315–335.

Silva, H. E., Verhoef, E. T. and van den Berg, V. a C. (2014) 'Airlines' strategic interactions and airport pricing in a dynamic bottleneck model of congestion.' *Journal of Urban Economics*. Elsevier Inc., 80 pp. 13–27.

Singh, A., Singh, S. and Aggarwal, A. (2021) 'Traffic Congestion Controller: A Fuzzy Based Approach.' *In 2021 International Conference on Disruptive Technologies for Multi-Disciplinary Research and Applications (CENTCON)*. IEEE, pp. 355–358.

Somuyiwa, A. O., Fadare, S. O. and Ayantoyinbo, B. B. (2015) 'Analysis of the Cost of Traffic Congestion on Worker's Productivity in a Mega City of a Developing Economy' pp. 644–656.

Squires, M., Tao, X., Elangovan, S., Gururajan, R., Zhou, X. and Acharya, U. R. (2022) 'A novel genetic algorithm based system for the scheduling of medical treatments.' *Expert Systems with Applications*. Pergamon, 195, June, p. 116464.

Staab, S. and Studer, R. (2007) 'What Is an Ontology?' *Decision Support Systems* p. 654.

Stanford University (2018) *Protégé*. [Online] http://protege.stanford.edu.

Steenbruggen, J., Tranos, E. and Rietveld, P. (2016) 'Traffic incidents in motorways: An empirical proposal for incident detection using data from mobile phone operators.' *Journal of Transport Geography*. Elsevier B.V., 54 pp. 81–90.

Stefanello, F., Buriol, L. S., Hirsch, M. J., Pardalos, P. M., Querido, T., Resende, M. G. C. and Ritt, M. (2015) ‘On the minimization of traffic congestion in road networks with tolls.’ *Annals of Operations Research*.

Studer, R., Benjamins, V. R. and Fensel, D. (1998) ‘Knowledge engineering: Principles and methods.’ *Data & Knowledge Engineering*, 25 pp. 161–197.

Sugeno, M. and Kang, G. T. (1988) ‘Structure identification of fuzzy model.’ *Fuzzy Sets and Systems*. North-Holland, 28(1) pp. 15–33.

Sun, Y., Hrušovský, M., Zhang, C. and Lang, M. (2018) ‘A Time-Dependent Fuzzy Programming Approach for the Green Multimodal Routing Problem with Rail Service Capacity Uncertainty and Road Traffic Congestion.’ *Complexity*, 2018 pp. 1–22.

Tadeusiak, M. (2014) ‘Traffic Flow Modelling conceptual model and specific implementations.’

Takagi, T. and Sugeno, M. (1985) ‘Fuzzy Identification of Systems and Its Applications to Modeling and Control.’ *IEEE Transactions on Systems, Man and Cybernetics*, SMC-15(1) pp. 116–132.

Tepanosyan, G., Sahakyan, L., Zhang, C. and Saghatelyan, A. (2019) ‘The application of Local Moran’s I to identify spatial clusters and hot spots of Pb, Mo and Ti in urban soils of Yerevan.’ *Applied Geochemistry*. Elsevier Ltd, 104, May, pp. 116–123.

TfGM (n.d.) *Transport for Greater Manchester*. [Online] https://www.tfgm.com/.

TfN (n.d.) *Transport for the North*. [Online] https://transportforthenorth.com/.

Thomas, N. E. (1998) ‘Multi-state and multi-sensor incident detection systems for arterial streets.’ *Transportation Research Part C: Emerging Technologies*, 6(5–6) pp. 337–357.

Toan, T. D. and Wong, Y. D. (2021) ‘Fuzzy logic-based methodology for quantification of traffic congestion.’ *Physica A: Statistical Mechanics and its Applications*. Elsevier B.V., 570, May.

Transport, D. for (2004) ‘Instructions for the Completion of Road Accident Reports,’ (October 2004) pp. 1–116.

Transport2020 (2016) *Transport2020*. [Online] [Accessed on 22nd February 2016] http://www.transport2020.org/.

Tsekeris, T. and Geroliminis, N. (2013) ‘City size, network structure and traffic congestion.’ *Journal of Urban Economics*. Elsevier Inc., 76 pp. 1–14.

U.S Department of Transportation (2018) *Traffic Congestion and Reliability: Trends and Advanced Strategies for Congestion Mitigation*.

Uschold, M., Bateman, J., Davis, M., Sowa, J., Bennett, C. M., Brooks, R., Dima, A., Gruninger, M., Guarino, N., Obrst, L., Ray, S., Schneider, T., Sriram, R., West, M. and Yim, P. (2011) ‘Making the Case for Ontology’ pp. 1–10.

Verhoef, E. T. (1999) ‘Time, speeds, flows and densities in static models of road traffic congestion and congestion pricing.’ *Regional Science and Urban Economics*, 29 pp. 341–369.

Verhoef, E. T. and Rouwendal, J. (2004) ‘A behavioural model of traffic congestion Endogenizing speed choice, traffic safety and time losses.’ *Journal of Urban Economics*, 56 pp. 408–434.

Vopham, T., Hart, J. E., Laden, F. and Chiang, Y.-Y. (2018) ‘Emerging trends in geospatial artificial intelligence (geoAI): Potential applications for environmental epidemiology.’ *Environmental Health: A Global Access Science Source*. Environmental Health, 17(1) pp. 1–6.

Waite, C. (2020) ‘2018 UK greenhouse gas emissions, provisional figures.’ *National Statistics*, (February 2020) p. 40.

Wang, C., Quddus, M. A. and Ison, S. G. (2009) ‘Impact of traffic congestion on road accidents: A spatial analysis of the M25 motorway in England.’ *Accident Analysis & Prevention*, 41(4) pp. 798–808.

Wang, Y., Peng, Z., Wang, K., Song, X., Yao, B. and Feng, T. (2015) ‘Research on Urban Road Congestion Pricing Strategy Considering Carbon Dioxide Emissions.’ *Sustainability*, 7(8) pp. 10534–10553.

Waters, N. M. (1999) ‘Transportation GIS: GIS-T.’ *Geographical information systems: Principles, techniques, management and applications* pp. 827–844.

Weka (2018) *Class J48*. [Online] [Accessed on 15th November 2018] http://weka.sourceforge.net/doc.dev/weka/classifiers/trees/J48.html.

Wen, W. (2008) ‘A dynamic and automatic traffic light control expert system for solving the road congestion problem.’ *Expert Systems with Applications*, 34(4) pp. 2370–2381.

Wilt, G. E., Adams, E. E., Thomas, E., Ekperi, L., LeBlanc, T. T., Dunn, I., Molinari, N. A. and Carbone, E. G. (2018) ‘A space time analysis evaluating the impact of hurricane sandy on HIV testing rates.’ *International Journal of Disaster Risk Reduction*. Elsevier, 28, June, pp. 839–844.

Wu, C., Thai, J., Yadlowsky, S., Pozdnoukhov, A. and Bayen, A. (2015) ‘Cellpath: Fusion of Cellular and Traffic Sensor Data for Route Flow Estimation via Convex Optimization.’ *Transportation Research Procedia*. Elsevier Ltd, 7 pp. 212–232.

Xiong, Y., Li, Y., Xiong, S., Wu, G. and Deng, O. (2021) ‘Multi-scale spatial correlation between vegetation index and terrain attributes in a small watershed of the upper Minjiang River.’ *Ecological Indicators*. Elsevier Ltd, 126 p. 107610.

Xuan, L. (2022) ‘Big data-driven fuzzy large-scale group decision making (LSGDM) in circular economy environment.’ *Technological Forecasting and Social Change*. North-Holland, 175, February, p. 121285.

Yang, Q. (1997) ‘A Simulation Laboratory for Evaluation of Dynamic Traffic Management Systems.’

Yang, W., Wang, Y., Webb, A. A., Li, Z., Tian, X., Han, Z., Wang, S. and Yu, P. (2018) ‘Influence of climatic and geographic factors on the spatial distribution of Qinghai spruce forests in the dryland Qilian Mountains of Northwest China.’ *Science of the Total Environment*. Elsevier B.V., 612, January, pp. 1007–1017.

Yao, W. and Qian, S. (2021) ‘From Twitter to traffic predictor: Next-day morning traffic prediction using social media data.’ *Transportation Research Part C: Emerging Technologies*. Pergamon, 124, March, p. 102938.

Yasdi, R. (1999) ‘Prediction of road traffic using a neural network approach.’ *Neural computing & applications*, 8 pp. 135–142.

Yuan, F. and Cheu, R. L. (2003) ‘Incident detection using support vector machines.’ *Transportation Research Part C: Emerging Technologies*, 11(3–4) pp. 309–328.

Zadeh, L. A. (1968) ‘Probability Measures of Fuzzy Events.’ *JOURNAL OF MATHEMATICAL ANALYSIS AND APPLICATIONS* pp. 421–427.

Zhang, C., Luo, L., Xu, W. and Ledwith, V. (2008) ‘Use of local Moran’s I and GIS to identify pollution hotspots of Pb in urban soils of Galway, Ireland.’ *Science of The Total Environment*. Elsevier, 398(1–3) pp. 212–221.

Zhang, Y., Ye, N., Wang, R. and Malekian, R. (2016) ‘A Method for Traffic Congestion Clustering Judgment Based on Grey Relational Analysis.’ *ISPRS International Journal of Geo-Information*, 5(5) p. 71.

Zheng, Y., Capra, L., Wolfson, O. and Yang, H. (2014) ‘Urban Computing.’ *ACM Transactions on Intelligent Systems and Technology*, 5(3) pp. 1–55.

Appendices

Appendix 1

This appendix has material related to chapter 3.

1.1 Enumerate important terms in the ontology

Table 19 shows a list of all the concepts being used and their associated descriptions that have been defined in the associated ontologies.

Table 19: Concepts and description of the road traffic congestion ontology

Concepts	Descriptions
Road Network	A network of **roads** that help vehicles to travel easily around a country.
Link	A **link** is a segment or segment of a **road** and can have several **lanes** going upstream and downstream.
Road	A **set of links** with the same name e.g. A6.
Lanes	A **lane** is a part of a **road** that is selected for use by a single row of vehicles.
City Centre	Area of a city where business, entertainment, shopping, and Political powers are concentrated. In addition, the **city centre** is also known as “downtown” in America or “Central Business District” in Australia.
Junction	**Junctions** are classified based on the number of **roads** that are involved. For example, a three-way intersection is known as a “T junction” or a “fork”. A four-way junction is known as a “crossroads”.
Highway	A **set of links** that has a minimum of 6 **lanes**.
Point	A **point**, typically described using a coordinate system **relative** to Earth, such as WGS84.
Spatial Thing	Anything with spatial extent, i.e., size, shape, or position. E.g., people, places, bowling balls, as well as abstract areas like cubes.
Event	An arbitrary classification of space/**time** region, by a cognitive agent. An **event** may have actively participating agents, passive factors, products, and a location in space and **time**.
Instant	A temporal entity with zero extents or **duration**.
Interval	A temporal entity with an extent or **duration**
Consequence	A result or effect, normally one that is unwanted.
Congestion	The state of a congested **road**.
Recurrent	Occurring often or repeatedly.

Concepts	Descriptions
Non-recurrent	Occurring at an unknown time.
Semi-recurrent	An **event** that occurs repeatedly but often at a different **time** and day or an **event** with an expected start and end **time**.
Dimensions	A way to measure.
Occupancy	The percentage of the **time** the detection zone of a detector is occupied by some vehicle
Density	A spatial measure that describes the number of vehicles occupying a section of a **road**.
Traffic Volume Count	The number of vehicles passing a point in a given period of **time**.
Average Speed	The rate at which someone or something moves or operates or can move or operate over a selected distance.
Speed	The rate at which someone or something moves or operates or can move or operate.
Speed At a Point	The rate at which someone or something moves or operates or can move or operate at a given **point**.
Velocity	A **speed** in a given **direction**.
Capacity	The maximum number of vehicles per unit of **time** that can be accommodated under given conditions with a reasonable expectation of occurrence.
Journey Time	The **time** it takes to go from origin to destination.
Time Frame	A specified period of **time** in which something occurs or is planned to take place.
Time	The indefinite continued progress of existence and events in the past, present, and future are regarded as a whole.
Public Events	The organised **public event**, which could disrupt traffic.
Roadworks	**Road** maintenance or improvement activity of an unspecified nature, which may potentially cause disruption to travel.
Terrorist Incident	A situation related to a perceived or actual threat of terrorism, which could disrupt traffic.
Road Traffic Incident	An event that causes disruption to the **road** network.
Concert	**Concert event** that could disrupt traffic.
Football Match	**Football match** that could disrupt traffic
Parade	Formal display of organized procession, which could disrupt traffic.
Marathon	**Marathon**, cross-country or **road** running event that could disrupt traffic.
Accident	**Accidents** are situations in which one or more vehicles lose control and do not recover.
Direction	A course along which someone or something moves
Absolute	Location of a fixed **point** on earth.

Concepts	Descriptions
Relative To Event	A location that is **relative to an event** location.
Relative To Traveller	A location that is relative to the traveller.
Relative To a Functional Site	A location that is relative to a functional site.
Magnitude	The severity of something.
Region	An area, especially the part of a country or the world having definable characteristics but not always fixed boundaries.

1.2 Define the classes and the class hierarchy

- Time
  - Instant
  - Interval
- Consequence
  - Congestion
    - Recurrent
    - Non-recurrent
    - Semi-recurrent
- Dimensions
  - Traffic Volume Count
  - Capacity
  - Density
  - Journey Time
  - Occupancy
  - Speed
    - Average Speed
    - Speed at A Point
  - Velocity
- Direction
  - Absolute
  - Relative
    - Relative to Event
    - Relative to Traveller
    - Relative to Functional Site
- Event
  - Public Events
    - Concert
    - Football Match
    - Marathon
    - Parade
  - Road Traffic Incident
    - Accident
  - Roadworks
  - Terrorist Incident
- Highway
- Junction
- Lanes

- Link
- Magnitude
  - Very Low
  - Low
  - Average
  - High
  - Very High
- Road
- Set of Links
- Spatial Thing
  - Point
  - Region
    - City Centre
  - Road Network
- Time Frame

1.3 Define the Classes-Properties

Table 20: Class-Properties (Domain, Properties, and Range)

Domain	Property	Range
Events	Happens at a	Point
Traffic Volume Count	Has a	Capacity
Velocity	Has a	Speed
Velocity	Has a	Direction
Lane	Has a	Capacity
Congestion	Has a beginning	Instant
Event	Has a consequence of	Consequence
Congestion	Has a duration	Interval
Congestion	Has an end	Instant
Link	Has numerous	Lanes
Set Of Links	Has multiple	Link
Congestion	Has a network scope	Set Of Links
Congestion	Analysed using	Dimensions
Consequence	Is a consequence of	Events
Highway	Is a part of	Road Network
Junction	Is a part of	Road Network
Link	Is a part of	Road
Set Of Links	Is a part of	Road Network
Time Frame	Is a part of	Time
Dimensions	Measured by	Magnitude
Journey Time	Measured by	Time
Occupancy	Measured by	Time

Appendix 2

Table 21 shows the performance of the multi-classification Fuzzy decision-making system and presents the prediction per link, per direction.

Table 21: Individual links performance, Precision, Recall, and F-score.

Link		NC	RC	SRC	NRC	wAvg
AU	**Recall**	1	0.65808	0	0.2052	0.91409
	Precision	0.6557	0.8602	0	1	0.69924
	F-score	0.792	0.74569	0	0.3406	0.7599
BU	**Recall**	1	0.62312	0	0.3554	0.82058
	Precision	0.3963	0.68075	0	1	0.55594
	F-score	0.5677	0.65066	0	0.5244	0.56889
CU	**Recall**	1	0.7803	0	0.3018	0.90608
	Precision	0.6962	0.81348	0	1	0.74016
	F-score	0.8209	0.79654	0	0.4637	0.78546
DU	**Recall**	0.7317	0.6535	0	0.4583	0.6293
	Precision	0.2997	0.57629	0	0.8761	0.53108
	F-score	0.4252	0.61247	0	0.6018	0.51033
EU	**Recall**	1	0.58884	0	0.3702	0.8865
	Precision	0.6929	0.74717	0	1	0.73412
	F-score	0.8186	0.65863	0	0.5404	0.77056
FU	**Recall**	1	0.65852	0	0.3439	0.83736
	Precision	0.4453	0.74344	0	1	0.58402
	F-score	0.6162	0.69841	0	0.5118	0.60771
GU	**Recall**	1	0.65813	0	0.5385	0.88324
	Precision	0.7654	0.68424	0	1	0.79662
	F-score	0.8671	0.67093	0	0.7001	0.81684
HU	**Recall**	1	0.70111	0	0.4756	0.83391
	Precision	0.5586	0.65284	0	1	0.67657
	F-score	0.7168	0.67612	0	0.6446	0.694
IU	**Recall**	1	0.75998	0	0.5736	0.89662
	Precision	0.8022	0.70599	0	1	0.82344
	F-score	0.8903	0.73199	0	0.729	0.84247
JU	**Recall**	0.9594	0.71125	0	0.5242	0.80145
	Precision	0.5261	0.66381	0	0.9724	0.67317
	F-score	0.6795	0.68671	0	0.6812	0.68093
KU	**Recall**	1	0.52305	0	0.1728	0.84026
	Precision	0.493	0.73111	0	1	0.53419
	F-score	0.6604	0.60982	0	0.2947	0.5963
LU	**Recall**	0.8966	0.57341	0	0.5497	0.74615
	Precision	0.4852	0.60334	0	0.9322	0.64618
	F-score	0.6296	0.588	0	0.6916	0.64556

Link		NC	RC	SRC	NRC	wAvg
MU	**Recall**	1	0.63989	0	0.2452	0.85899
	Precision	0.6904	0.73372	0	1	0.69015
	F-score	0.8169	0.6836	0	0.3938	0.73439
NU	**Recall**	1	0.7735	0.35159	0.3996	0.88107
	Precision	0.7273	0.78987	0.84091	0.9542	0.76306
	F-score	0.8421	0.7816	0.49586	0.5633	0.79174
OU	**Recall**	1	0.5639	0.26137	0.3012	0.85876
	Precision	0.5708	0.8075	0.84953	0.9347	0.64088
	F-score	0.7268	0.66406	0.39975	0.4556	0.68248
PU	**Recall**	1	0.56082	0.15083	0.1514	0.89189
	Precision	0.4811	0.89127	0.89461	0.9108	0.55513
	F-score	0.6497	0.68844	0.25813	0.2597	0.62516
QU	**Recall**	1	0.41577	0.29626	0.2839	0.87906
	Precision	0.6361	0.81708	0.83967	0.9196	0.67697
	F-score	0.7776	0.55111	0.43799	0.4338	0.72356
RU	**Recall**	1	0.66768	0.36392	0.4166	0.88209
	Precision	0.7339	0.84451	0.84477	0.9267	0.76819
	F-score	0.8465	0.74575	0.5087	0.5748	0.79528
SU	**Recall**	1	0.75304	0	0.2592	0.82712
	Precision	0.692	0.76508	0	1	0.6795
	F-score	0.818	0.75901	0	0.4117	0.71316
TU	**Recall**	1	0.60758	0.41404	0.3477	0.8379
	Precision	0.5524	0.86525	0.8812	0.9192	0.65498
	F-score	0.7117	0.71388	0.56337	0.5045	0.67899
UU	**Recall**	1	0.71147	0.28863	0.4619	0.89758
	Precision	0.7716	0.88052	0.85714	0.918	0.79961
	F-score	0.8711	0.78702	0.43184	0.6146	0.82554
VU	**Recall**	1	0.7222	0.49907	0.4238	0.87872
	Precision	0.7428	0.8699	0.87184	0.9225	0.78355
	F-score	0.8524	0.7892	0.63477	0.5808	0.80467
WU	**Recall**	1	0.6436	0.2845	0.3224	0.87155
	Precision	0.6194	0.86635	0.91774	0.9497	0.68986
	F-score	0.7649	0.73854	0.43435	0.4814	0.72631
XU	**Recall**	1	0.53571	0.16677	0.2043	0.90245
	Precision	0.6269	0.8802	0.83333	0.96	0.67075
	F-score	0.7707	0.66605	0.27793	0.3369	0.7308
YU	**Recall**	0.841	0.68269	0.44374	0.4036	0.69101
	Precision	0.3843	0.79806	0.80496	0.7994	0.57102
	F-score	0.5275	0.73588	0.57211	0.5364	0.56543
ZU	**Recall**	1	0.43722	0.07543	0.0678	0.92446
	Precision	0.5391	0.87161	0.82741	0.9422	0.5785
	F-score	0.7006	0.58233	0.13825	0.1265	0.67451

Link		NC	RC	SRC	NRC	wAvg
AAU	Recall	1	0.64644	0.17371	0.1966	0.91855
	Precision	0.7248	0.88606	0.79692	0.9805	0.75299
	F-score	0.8405	0.74752	0.28524	0.3275	0.80383
ABU	Recall	1	0.61913	0.22253	0.3288	0.86761
	Precision	0.6098	0.79037	0.8277	0.9779	0.67156
	F-score	0.7576	0.69435	0.35075	0.4922	0.71334
ACU	Recall	1	0.48811	0.06689	0.1113	0.91579
	Precision	0.5019	0.95078	0.93143	0.9615	0.56325
	F-score	0.6684	0.64506	0.12481	0.1995	0.64796
ADU	Recall	1	0.74505	0	0.3099	0.878
	Precision	0.5925	0.74969	0	1	0.66546
	F-score	0.7441	0.74736	0	0.4732	0.71054
AEU	Recall	1	0.77352	0	0.4289	0.87497
	Precision	0.6726	0.72412	0	1	0.73348
	F-score	0.8043	0.74801	0	0.6003	0.76301
AFU	Recall	1	0.52965	0	0.3519	0.88389
	Precision	0.6866	0.68541	0	1	0.7203
	F-score	0.8142	0.59755	0	0.5206	0.76125
AD	Recall	1	0.59077	0	0.2044	0.91005
	Precision	0.6421	0.84832	0	1	0.6848
	F-score	0.782	0.69649	0	0.3394	0.74653
BD	Recall	1	0.64968	0	0.3549	0.82355
	Precision	0.3964	0.70232	0	1	0.55882
	F-score	0.5678	0.67497	0	0.5239	0.57208
CD	Recall	1	0.79434	0	0.2929	0.90815
	Precision	0.6852	0.83196	0	1	0.73377
	F-score	0.8132	0.81271	0	0.4531	0.78096
DD	Recall	0.7262	0.67784	0	0.459	0.63019
	Precision	0.3115	0.57189	0	0.8706	0.53516
	F-score	0.436	0.62037	0	0.6011	0.51737
ED	Recall	1	0.55897	0	0.3481	0.88643
	Precision	0.6774	0.73518	0	1	0.71892
	F-score	0.8077	0.63508	0	0.5164	0.75915
FD	Recall	1	0.67877	0	0.3672	0.84144
	Precision	0.4932	0.72227	0	1	0.6166
	F-score	0.6606	0.69985	0	0.5371	0.64255
GD	Recall	1	0.63571	0	0.5066	0.88044
	Precision	0.7508	0.66649	0	1	0.78252
	F-score	0.8577	0.65074	0	0.6725	0.80531
HD	Recall	1	0.75513	0	0.4878	0.84494
	Precision	0.5929	0.66901	0	1	0.69964
	F-score	0.7444	0.70947	0	0.6557	0.71862

Link		NC	RC	SRC	NRC	wAvg
ID	Recall	1	0.75692	0	0.5857	0.89718
	Precision	0.8052	0.70308	0	1	0.82643
	F-score	0.8921	0.729	0	0.7387	0.84487
JD	Recall	0.9457	0.76082	0	0.521	0.79807
	Precision	0.5396	0.65329	0	0.9596	0.67628
	F-score	0.6871	0.70296	0	0.6753	0.68586
KD	Recall	1	0.58031	0	0.187	0.83919
	Precision	0.5205	0.72035	0	1	0.55778
	F-score	0.6846	0.64279	0	0.3151	0.61727
LD	Recall	0.8897	0.63684	0	0.5537	0.7478
	Precision	0.5003	0.61694	0	0.9198	0.6547
	F-score	0.6404	0.62673	0	0.6912	0.65595
MD	Recall	1	0.64333	0	0.24	0.87194
	Precision	0.7036	0.70349	0	1	0.70212
	F-score	0.826	0.67207	0	0.3872	0.74973
ND	Recall	1	0.77823	0.3157	0.3983	0.88343
	Precision	0.7264	0.7997	0.82681	0.9468	0.76214
	F-score	0.8415	0.78882	0.45693	0.5607	0.79235
OD	Recall	1	0.62495	0.27803	0.3411	0.85845
	Precision	0.604	0.79269	0.84111	0.9419	0.67006
	F-score	0.7531	0.6989	0.41792	0.5008	0.70735
PD	Recall	1	0.53211	0.15782	0.1567	0.89111
	Precision	0.4918	0.88275	0.91029	0.924	0.56273
	F-score	0.6593	0.66398	0.26901	0.268	0.63144
QD	Recall	1	0.42429	0.28073	0.2364	0.88701
	Precision	0.6262	0.82935	0.84572	0.9117	0.66557
	F-score	0.7701	0.56138	0.42153	0.3754	0.71889
RD	Recall	1	0.65725	0.34965	0.3951	0.8839
	Precision	0.7351	0.83503	0.82023	0.9357	0.76611
	F-score	0.8473	0.73555	0.4903	0.5556	0.79509
SD	Recall	1	0.70488	0	0.2556	0.82577
	Precision	0.6787	0.75455	0	1	0.67075
	F-score	0.8086	0.72887	0	0.4071	0.7061
TD	Recall	1	0.63657	0.4302	0.3505	0.83447
	Precision	0.5434	0.84998	0.88702	0.9239	0.65349
	F-score	0.7042	0.72796	0.5794	0.5082	0.67573
UD	Recall	1	0.73377	0.32413	0.4793	0.89794
	Precision	0.7773	0.86774	0.88939	0.9453	0.8068
	F-score	0.8747	0.79515	0.47511	0.6361	0.83015
VD	Recall	1	0.68877	0.4908	0.4221	0.87932
	Precision	0.7514	0.86895	0.86815	0.9014	0.78591
	F-score	0.8581	0.76844	0.62709	0.575	0.80738

Link		NC	RC	SRC	NRC	wAvg
WD	**Recall**	1	0.66278	0.30781	0.3193	0.87314
	Precision	0.6286	0.87134	0.90405	0.9489	0.6974
	F-score	0.7719	0.75288	0.45925	0.4778	0.73325
XD	**Recall**	1	0.57028	0.17659	0.2223	0.90196
	Precision	0.642	0.89747	0.87755	0.9478	0.68767
	F-score	0.7819	0.6974	0.29402	0.3602	0.74291
YD	**Recall**	0.8397	0.68192	0.44166	0.4016	0.68905
	Precision	0.3761	0.8061	0.78729	0.8057	0.56797
	F-score	0.5195	0.73883	0.56587	0.536	0.56125
ZD	**Recall**	1	0.4482	0.12644	0.0798	0.91786
	Precision	0.5523	0.83849	0.83898	0.9475	0.59001
	F-score	0.7116	0.58415	0.21976	0.1473	0.68017
AAD	**Recall**	1	0.67282	0.15917	0.1559	0.92424
	Precision	0.7094	0.9032	0.78457	0.9703	0.74073
	F-score	0.83	0.77117	0.26464	0.2687	0.79869
ABD	**Recall**	1	0.6121	0.24944	0.3587	0.86162
	Precision	0.6189	0.79034	0.82836	0.9687	0.68129
	F-score	0.7646	0.68989	0.38342	0.5235	0.71823
ACD	**Recall**	1	0.50895	0.05669	0.1035	0.91788
	Precision	0.4949	0.9559	0.9085	0.9725	0.55864
	F-score	0.6621	0.66424	0.10672	0.187	0.64547
ADD	**Recall**	1	0.73651	0	0.3037	0.87305
	Precision	0.5644	0.73867	0	1	0.64509
	F-score	0.7215	0.73759	0	0.4659	0.69039
AED	**Recall**	1	0.76037	0	0.4079	0.8697
	Precision	0.6361	0.72618	0	1	0.70879
	F-score	0.7776	0.74288	0	0.5795	0.73986
AFD	**Recall**	1	0.57198	0	0.4049	0.8814
	Precision	0.7057	0.66931	0	1	0.73918
	F-score	0.8275	0.61683	0	0.5764	0.77437

Appendix 3

This appendix has a copy of each published paper related to work within this thesis.

Gould, N. and Abberley, L. (2017) ‘The semantics of road congestion.’ *In UTSG*. Dublin.

L. Abberley, N. Gould, K. Crockett and J. Cheng, ‘Modelling road congestion using ontologies for big data analytics in smart cities,’ 2017 International Smart Cities Conference (ISC2), 2017, pp. 1-6, Doi: 10.1109/ISC2.2017.8090795

L. Abberley, K. Crockett and J. Cheng, "Modelling Road Congestion Using a

Fuzzy System and Real-World Data for Connected and Autonomous Vehicles," 2019 Wireless Days (WD), 2019, pp. 1-8, Doi: 10.1109/WD.2019.8734238.

Modelling Road Congestion Using a Fuzzy System and Real-World Data for Connected and Autonomous Vehicles

Luke Abberley Member, IEEE, Keeley Crockett SMIEEE, Jianquan Cheng
Science and Engineering, Manchester Metropolitan University, UK
{Luke.Abberley@mmu.ac.uk, K.Crockett@mmu.ac.uk, J.Cheng@mmu.ac.uk}

***Abstract* - Road congestion is estimated to cost the United Kingdom £307 billion by 2030. Furthermore, congestion contributes enormously to damaging the environment and people's health. In an attempt to combat the damage congestion is causing, new technologies are being developed, such as intelligent infrastructures and smart vehicles. The aim of this study is to develop a fuzzy system that can classify congestion using a real-world dataset referred to as Manchester Urban Congestion Dataset, which contains data similar to that collected by connected and autonomous vehicles. A set of fuzzy membership functions and rules were developed using a road congestion ontology and in conjunction with domain experts. Experiments are conducted to evaluate the fuzzy system in terms of its precision and recall in classifying congestion. Comparisons are made in terms of performance with traditional classification algorithms decision trees and Naïve Bayes using the Red, Amber, and Green classification methods currently implemented by Transport for Greater Manchester to label the dataset. The results have shown the fuzzy system has the ability to predict road congestion using volume and journey time, outperforming both decision trees and Naïve Bayes.**

Keywords - Intelligent Transport Systems; Big Data; Fuzzy System; Urban Road Network; Congestion

I. INTRODUCTION

For centuries, people have naturally been migrating from rural to urban areas causing the natural occurrence of urbanization, which has contributed to one of the biggest challenges' society faces each day, which is road congestion. Road congestion in urban areas is estimated to cost the UK economy a total of £307 billion by 2030 [1]. Furthermore, road congestion contributes enormously to damaging the environment, due to air pollution which has an impact on peoples well-being [2], [3].

In an attempt to reduce the impact of road congestion, many large corporations, such as Google, Tesla, and Uber are developing *'smart vehicles'*, such as connected and autonomous vehicles (CAVs) that will be implemented as part of an Intelligent Transport System (ITS) of the future. Smart vehicles are expected to reduce congestion levels and the number of fatal accidents on the roads, with an estimated 37,000 lives a year predicted as being saved in the United States (U.S.) alone [4]. This is due to smart vehicles being able to communicate faster than a human and make better decisions based on information collected by sensors embedded within vehicles with other vehicles and infrastructure [5]. However, due to the limited access to these smart vehicles and their associated infrastructure, this study will use alternative data sources, which comprise of data similar to what is collected by CAVs and Road Side Units (RSUs) that will be used within ITS of the future, such as a VANETs. Furthermore, these types of ITS will provide data from vehicle to vehicle (V2V) and vehicle to infrastructure (V2I) providing a constant stream of big data [5]–[7], which can be used to provide different information, such as volume, journey time, speed, and weather conditions, which are also known as dimensions.

Little work has been conducted using fuzzy systems to model road congestion [8]–[10]. However, this limited work has indicated that fuzzy models of road congestion are better for a stakeholder, such as a domain expert to understand that the conventional quantitative models previously implemented, such as the probability model [11] and the spatial-temporal model [12]. Fuzzy sets are the ideal choice for modelling road congestion because of their ability to handle the ambiguity, multifaceted nature, and uncertainty within traffic data, such as journey time, speed, traffic flow, accidents, road works etc. They have the ability to capture such variables through the use of linguistic variables and hedges which are easier for a domain expert to understand [13].

The contribution of this paper is taking an unbalanced real-world big dataset with the support of an ontology and domain experts to construct a fuzzy model of road congestion. This is achieved through the construction of fuzzy systems comprises of a set of fuzzy membership functions and fuzzy rules that can be used to identify road congestion. An experiment is conducted to determine whether the fuzzy model can be used to analyse traffic data to classify congestion. Comparisons are made with an existing system used by Greater Manchester Transport authority in the UK and other known classification algorithms.

This paper is organised as follows: Section II provides a brief overview of recent studies of fuzzy systems within the field of transportation. Section III provides an in-depth explanation for developing a fuzzy system capable of capturing the levels of congestion on an urban road network. Section IV presents the experimental methodology and Section V provides the results and a comparative discussion. Section VI concludes this study and provides insight into further work.

II. Fuzzy Systems in Transportation

A. What is a fuzzy system?

A fuzzy system is typically a control system based on fuzzy logic. The term "fuzzy" refers to the system's ability to deal with terms that are not binary or predefined- often referred to as linguistic variables [13]. For instance, a humans' understanding of the phrase, near or far, could imply: very near, near, not near, far, and very far. Hence, fuzzy terms are subjective and mean different things to different people. The main advantage of a fuzzy system is that the model itself is made up of a number of fuzzy rules, which can model a problem, such as congestion that can be expressed in terms a human operator can understand.

B. Transport applications

The approach to use fuzzy systems within the discipline of transportation to classify road traffic congestion is a relatively new field. For instance, a study was presented in [14] into a cooperative V2V road traffic detection congestion on freeways. The study uses a level of service metric created by a third party that collected aerial surveys to define the levels of congestion: slight, moderate, and severe. The author then created a new metric that uses four membership function: VerySlow, Slow, Medium, and Fast, two inputs: Speed and Density, and sixteen rules to define an output for one of three levels of congestion. However, this study does not consider non-congestion as an output and has described only testing the model in a simulation with simulated data, furthermore, the focus of the study is on highways and does not reflect an urban road network, which has very different characteristics. Another study [9], examined road traffic anomalies that contribute to congestion at a single junction using a one-way traffic video sequence. This study uses two data inputs: Traffic flow and traffic density. Traffic flow has three membership functions called low, medium, and high. Traffic density also has three membership functions (sparse, normal, and dense) which are calculated using a statistical analysis of the pixels. This study uses nine rules, which were obtained through experts and empirical experiments and has an output of either: normal traffic, slight congestion, and heavy congestion. However, one of the limitations of this experiment is it was only tested on 3 different scenes and in total had 142 observations. Although the use of fuzzy systems is very new, despite limitations in current work, fuzzy set representation of the variables that model congestion encapsulate a greater human understanding of a multifaceted and dynamic environment.

III. Developing a Fuzzy System for Congestion

This section describes the methodology that was used to develop a fuzzy system for road congestion on an urban city network. The model utilises real-world data from Bluetooth sensors and inductive loop counters provided by Transport for Greater Manchester (TfGM) for Manchester, UK. These data sources will provide data that is equivalent to what will be provided by CAVs and RSUs. Moreover, experts in road congestion management (TfGM) and a road congestion ontology [3], [15] was used to help define the fuzzy sets to ensure thorough domain coverage. The road congestion ontology which was used to support the development of a fuzzy system capable of classifying road congestion was presented in [15]. The road congestion ontology states that congestion can be measured using multiple dimensions, such as journey time and volume. Furthermore, congestion is often the consequence of an event, such as rush hour, a road accident, a concert, a football match and roadworks. Finally, depending on the severity of congestion the magnitude can vary from very low to very high. Therefore, in this study, the magnitude ranges defined in the urban road congestion ontology [15] will be used to determine the membership functions: Very low, *VL*, low, *L*, medium, *M*, high, *H*, and very high, *VH* which will ensure coverage of the domain.

A. Data Sources and Variables

For this study, a real-world spatial-temporal dataset, known as the Manchester Urban Congestion dataset (MUCD) was used. The dataset consists of journey time, volume, weather, bank holidays, and event information. However, due to the nature of this study, only journey time and volume data collected from Bluetooth sensors and IDC will be used to simulate data collected by CAVs and RSUs. The MUCD has 17376 records and each record consists of 126 attributes. Furthermore, the MUCD dataset is labelled using the Red, *R*, Amber, *A*, and Green, *G*, (RAG) method implemented by TfGM, UK. Where (G)reen is non-congestion (1), (A)mber is slight congestion (2), and (R)ed is major congestion (3).

$$G = JT \leq median * 1.25 \tag{1}$$

$$A = Median * 1.25 < JT \leq Median * 1.5 \tag{2}$$

$$R = JT > Median * 1.5 \tag{3}$$

Where *JT* is the average journey time for all Bluetooth enabled vehicles travelling between two sensors on each link. The *Median* is the 50th percentile of journey time for a single link within the MUCD. 1.25 and 1.5 are the congestion factors that TfGM experts use to measure network performance. The problems associated with the MUCD can be summarised as:

- Due to the limited number of inductive loops counters, the ability to calculate the volume of traffic for each link in the network is limited.
- The data quality of the Bluetooth sensors has many issues. For example, capture rates; during the night periods or a period where no vehicle with a Bluetooth device passes the sensors cause the sensors to provide an incorrect average journey time when being observed.
- In bad weather, the sensors which use a mobile network to transmit the data to a central location, can fail and cause the dataset to have missing data.
- The non-congestion class significantly outweighs the other, causing the MUCD dataset to be imbalanced, which imposes challenges for machine learning classification algorithms that is a problem because classification algorithms are often biased towards the majority class, which in this study is non-congestion.

TABLE I. DIMENSIONS AND THEIR LINGUISTIC VALUES.

Dimension (Variables)	Data sources	Linguistic values (Membership functions)
Journey time	Bluetooth remote sensors	Very Low (*VL*) Low (*L*) Medium (*M*) High (*H*) Very High (*VH*)
Volume	Inductive loop counters	Very Low (*VL*) Low (*L*) Medium (*M*) High (*H*) Very High (*VH*)

B. Membership Function Determination

Table I shows the dimensions, data sources, and linguistic values determined from the urban road congestion ontology [15]. The linguistic values of the membership functions representing journey time and volume are also shown.

Using the linguistic values identified in Table I, the creation of the fuzzy membership functions can be performed using three steps:

- **Step 1**: Perform K-means clustering [16] on both journey time and volume data.
- **Step 2**: Identify the final boundary values for a set of clusters (referred to in this work as groups) where they connect and define this value as *dt (boundary threshold).*
- **Step 3**: Using the *dt* value, determine membership function domain coverage using one of three membership functions: linear increasing, linear decreasing, and trapezoidal.

The primary objective of data mining is to discover patterns within large datasets, such as the MUCD dataset used within this study. K-means clustering is an unsupervised algorithm used within data mining to find a cluster of patterns in data. K-means uses the inherent structures in the data to best organise the data into groups of maximum commonalities [17]. This is achieved by partitioning *n* observations into *k* (in this study *k=5*) clusters. Five clusters were used based upon early empirical experiments which found that five clusters provided sufficient resolution [15]. Figure 1 shows, as an example, 17376 journey time records plotted on a 24-hour scale. Each observation within Figure 1 belongs to the cluster with the nearest mean value. Once K-means has been performed, it becomes possible to identify the boundary values between each cluster, which will be used to create the membership functions in the fuzzy system.

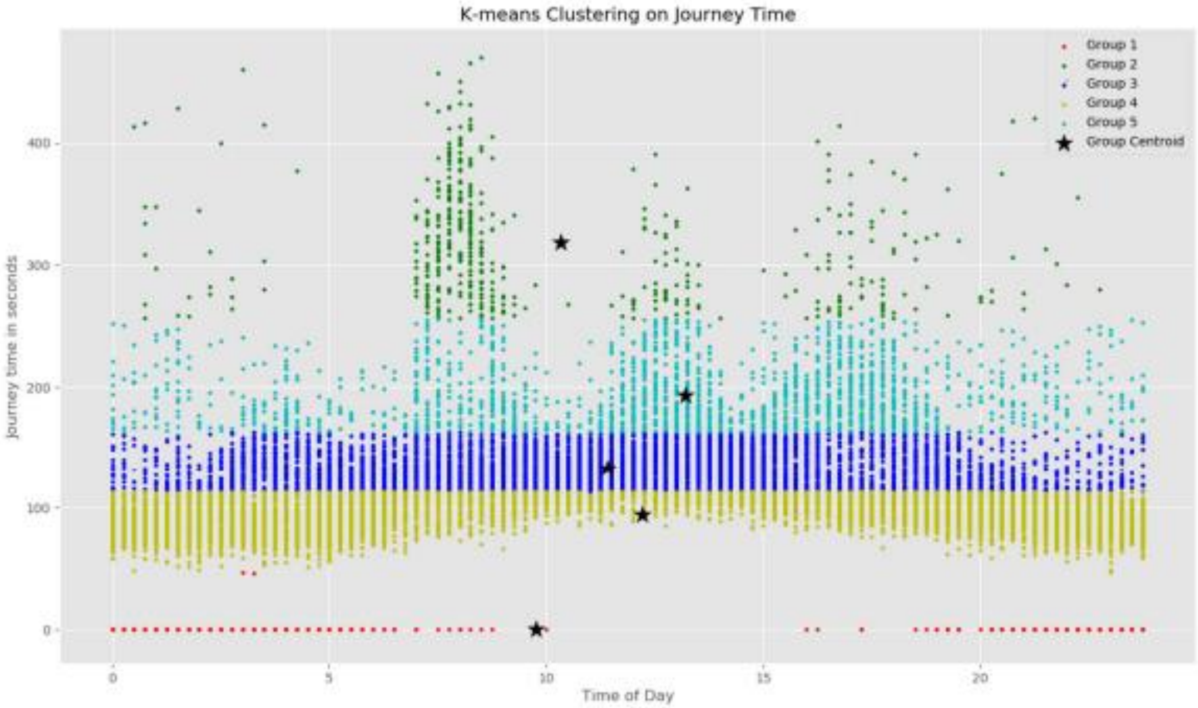


Figure 1: Example of k-means clusters on 6 months' worth of journey time data where k=5.

Figure 2 shows an example pair of linear opposing membership functions, which will be used for the *VL* and *VH* memberships. The two pairs (4) and (5) are both linear increasing and decreasing membership functions *L*, can be defined as [18]:

$$L\uparrow(x, dm, dn) = \begin{cases} 0, & x \le dm \\ \dfrac{x - dm}{dn - dm}, & dm \le x \le dn \\ 1, & x \ge dn \end{cases} \qquad (4)$$

$$L\downarrow(x, dm, dn) = \begin{cases} 1, & x \le dm \\ 1 - \dfrac{x - dm}{dn - dm}, & dm \le x \le dn \\ 0, & x \ge dn \end{cases} \qquad (5)$$

Where dm is defined as $dm{=}dt{-}n\sigma$ and dt is the value generated by K-means clustering on all variable *i* records. n is a real number $n \rightarrow [0.0, \infty]$, σ is the standard deviation, and x is the value of the variable *i*. n is empirically determined. Additionally, dn is defined as $dn{=}dt{+}n\sigma$.

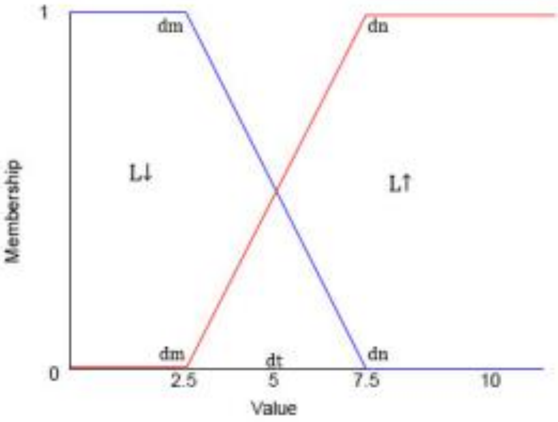


Figure 2: Example of a linear pair opposing fuzzy memberships functions.

Figure 3 shows an example of a trapezoidal-shaped membership function, which will be used for the *L*, *M*, and *H* memberships. The trapezoidal-shaped membership function *T* (6), may be defined as:

$$T(x, dm^1, dn^1, dm^2, dn^2) = \begin{cases} 0, & x \le dm^1 \\ \dfrac{x - dm^1}{dn^1 - dm^1}, & dm^1 \le x \le dn^1 \\ 1, & dn^1 \le x \le dm^2 \\ 1 - \dfrac{x - dm^2}{dn^2 - dm^2}, & dm^2 \le x \le dn^2 \\ 0, & x \ge dn^2 \end{cases} \qquad (6)$$

Where *dm1*, *dn1*, *dm2*, and *dm2* are defined using the same method as *dm* and *dn*

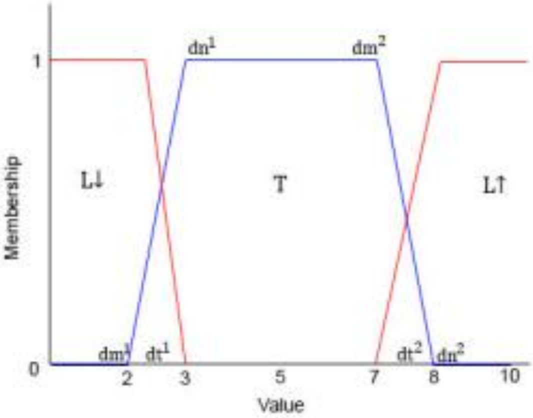


Figure 3: Example of a trapezoidal-shaped membership function.

C. *Fuzzy Rules Determination (manual and expert)*

The fuzzy rules were initially created with every possible variation of each five membership functions, such as *VL*, *L*, *M*, *H*, and *VH* for journey time and volume, which gave a total of 25 rules. However, with the support of the urban road congestion ontology [15] and domain experts, TfGM [19], the rules were humanly optimised down to just six. This manual optimisation revealed that several rules were not firing so therefore, they were not relevant. For example, if journey time was *VH* then the output is congested regardless of the volume.

Algorithm 1 uses both antecedents and consequents membership functions to fire six unique rules to acquire each rule strength ready for fuzzy inference.

Algorithm 1
Rules for congestion

Antecedents: Journey time, *JT*. Volume, *V*.
Antecedents memberships: Very low, *VL*. Low, *L*. Medium, *M*, High, *H*. Very high, *VH*.
Consequents: Congestion, *C*.
Consequents memberships: Congested, Con. Non-congested, *Non*.

```
IF JT is VH THEN Con
IF JT is H THEN Con
IF JT is M AND V is VH THEN Con
IF JT is M AND V is NOT VH THEN Non
IF JT is L THEN Non
IF JT is VL THEN Non
```

D. *Fuzzy inference*

One of the first control systems and most commonly implemented methods for computing fuzzy inference is Mamdani [20]. Furthermore, Mamdani was first implemented within the transport domain, where it was used in an attempt to control a steam engine and boiler combination [20].

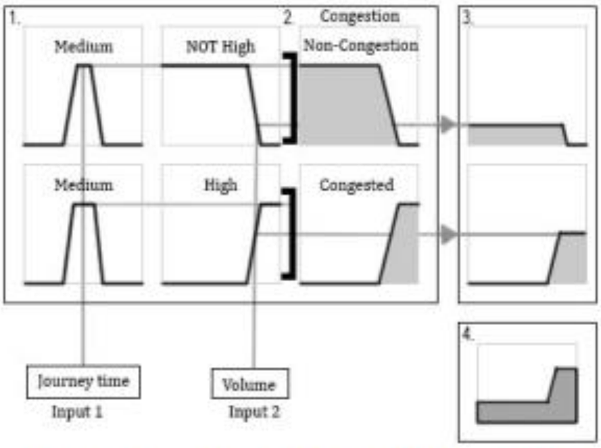


Figure 4: A example of how Mamdani fuzzy inferences works.

Figure 4 shows the composition of fuzzy inference, the four stages are:

Stage 1. Fuzzification of the non-fuzzy inputs (the average journey time and volume over a 15-minute slot for one link in the network), which are crisp, numerical, and specific to the attribute domain. The inputs are fuzzified according to membership functions.

Stage 2. If the antecedent of a given rule has more than one part, the application of a fuzzy operator is required to obtain a single value that represents the individual rule. For instance, the top rule within Figure 4 has two parts in the antecedent, so a AND operator is used to identify the minimum value as the result.

Stage 3. Using the single value acquired in stage 2, the consequent is reshaped to provide the result of implication which is weighted depending on the linguistic characteristics that are attributed to it.

Stage 4. Aggregation is the combination of the fuzzy sets that represent the outputs of each rule into a single fuzzy set (fuzzy output distribution).

Stage 5. The input for defuzzification is the single aggregated fuzzy set and the output is a single value. This is discussed in more detail in section E.

E. Defuzzification

The method centroid of area (COA), also known as the centre of gravity (COG) (7) is used to defuzzify the final output fuzzy set (Figure 4) and output a crisp numeric value, which in this study is the probability of congestion. To achieve this, the total area of the output distribution membership is divided into a number of sub-areas and then the COA is calculated for each sub-area. Finally, all sub-areas COA are summed together to find the defuzzied value (probability of congestion).

$$Z^* = \frac{\int \mu_{A}(z).zdz}{\int \mu_{\bar{A}}(z)dz} \tag{7}$$

Where μ is defined as the degree of membership (y-axis), *z* is defined as the value on the x-axis, $\bar{A}$ is the fuzzy set, and *dz* is the derivative of *z*.

IV. Experimental Methodology

The aim of the experiment is to determine whether a fuzzy system can be used to analyse traffic data to classify congestion. The hypothesis for this study is H_1: Using journey time and volume data, it is possible to classify congestion using a fuzzy system. To evaluate the performance of the fuzzy system, it was compared against two alternative machine-learning algorithms: The decision tree C4.5 (using the Weka implementation J48) [21] and Naïve Bayes, on the MUCD dataset. The training and testing strategy is described as follows: the MUCD dataset was split into two parts: Training Set containing 8688 records of which 6665 were classified as non-congestion and 2023 were classified as congestion, which accounts for only 23% of records. The test set containing the remainder of the dataset. Datasets were mutually exclusive.

In order to evaluate the three methods using an unbalanced dataset, five statistical measurements were chosen, which are: True Positive Rate, *TPR*, also known as recall and sensitivity. TPR measures the proportion of actual positives that are correctly identified. TPR is defined in equation (8) where *TP* is true positive, and *FN* is false negative.

$$TPR = \frac{TP}{TP + FN} \tag{8}$$

False Positive Rate, *FPR,* measures the negative instance that is wrongly classified as positive. FPR is defined in equation (9) where *FP* is false positive, and *TN* is true negative.

$$FPR = \frac{FP}{FP + TN} \tag{9}$$

Precision, also known as positive predictive value, *PPV*, measures the number of positive predictions divided by the total number of positive class values predicted. Precision is defined in equation (10).

$$PPV = \frac{TP}{TP + FP} \tag{10}$$

F-measure, also known as F1 Score, *F1*, measures the balance between the precision and TPR. F-measure is defined in equation (11).

$$F1 = \frac{2 * TP}{2 * TP + FP + FN} \tag{11}$$

Overall efficiency, also known as accuracy measures the amount of correctly classified instances. Overall efficiency is defined in equation (12)

$$Overall\ Efficiency = \frac{TP + TN}{TP + TN + FP + FN} \tag{12}$$

However, due to the class imbalance as mentioned above, it is important to provide a single value that represents the performance of both classes for TPR, False Positive Rate, precision, and F-measure. To achieve this a weighted average will be used and is defined in equation (13). Where C_{non} represents the statistical measurement being weighted for the class non-congestion. C_{con} represents the statistical measurement being weighted for the class congested.

$$W = \frac{C_{non} * (TP + FN) + C_{con} * (TN + FP)}{TP + FN + TN + FP} \tag{13}$$

V. Results and Discussion

The purpose of this study was to determine whether it is possible to classify road congestion using a fuzzy system and real-world traffic data. Table II shows the results for each statistical measurement for the three machine-learning algorithms and their classes: Non-congestion (Figure 5) and Congested (Figure 6) and the weighted average of both classes (Figure 7) defined in equation (13).

Before discussing the results, the authors would like to reiterate the challenges of performing classification on an imbalanced dataset. Global performance measurements, such as overall efficiency, provides an advantage to the majority class and can be misleading. For example, the overall efficiency of the fuzzy system is 88 per cent, which seems good. However, assume the dataset had 100 instances, with a split of 80 for non-congestion and 20 for congestion. Assume the system classifies non-congestion as 92 instances and congested as eight instances. This means the class, congested is only 40 per cent efficient/accurate and not 88 per cent. Therefore, the discussion will focus on TPR, FPR, precision, F-measure.

The results show Naïve Bayes achieved a TPR of 99.8 per cent for non-congestion, which is the higher TPR across all algorithms and both classes. However, it achieved the second highest FPR of 57.8 per cent. This is attributed to the paradox of imbalanced datasets. The FPRs for the minority class across all three algorithms are significantly low, for instance, the fuzzy system is 5.5 per cent, the decision tree is 4.8 per cent, and the Naïve Bayes is 0.2 per cent. The FPRs for the majority class across all three algorithms are noticeably higher, for instance, the fuzzy system is 32.9 per cent, the decision tree is 59 per cent, and the Naïve Bayes is 57.8 per cent. Because of these noticeable differences, it has been decided from this point to only compare the weighted averages of both classes. The TPR weighted average for the fuzzy system is 88 per cent, which is higher than both, the decision tree by ≈6 per cent and Naïve Bayes by ≈2 per cent. The FPR weighted average for the fuzzy system is 26.5 per cent, which is lower than both, the decision tree by ≈20 per cent and Naïve Bayes by ≈18 per cent. The precision weighted average for the fuzzy system is 87.6 per cent, which is higher than the decision tree by ≈6 per cent, however, it was lower than the Naïve Bayes by ≈1 per cent. The F-measure weighted average for the fuzzy system was 87.6 per cent and is higher than both the decision tree by ≈7 per cent and Naïve Bayes by ≈3 per cent. Furthermore, the fuzzy system overall efficiency was the highest of all three machine-learning algorithms.

Although all algorithms perform to a similar level with the fuzzy system performing the best overall, it should be noted that each algorithm has its own level of complexity with some stakeholders possibly struggling to understand how the model produces an explainable decision. For instance, the easiest of the three algorithms for a stakeholder to understand is the fuzzy system. This is because the rules are comprised of linguistic variables, which are easier to understand and interpret by stakeholders. The single defuzzied output of the fuzzy system gives a measure of the probability that congestion occurs in a specific 15-minute slot on road link *x*, where *x* is a road link on the urban network being modelled. The second easiest to understand is the decision tree, J48, where a branch of the tree is split based on a value of the variable being used and this is repeated until the leaves are reached and an outcome is decided. It should be noted the bigger the tree and the more leaves (and hence rules) the harder it is to understand the decision transparency and hence, may become harder for stakeholders to follow. The decision tree model in this experiment has a tree size of 17 and a total of 9 leaves. The 9 rules are transparent and could be understood by a transport expert. The most complex algorithm for a stakeholder to understand is Naïve Bayes because it is a probabilistic classifier, which uses a probability distribution over a set of classes, instead of only outputting the most likely class that the observation should belong to.

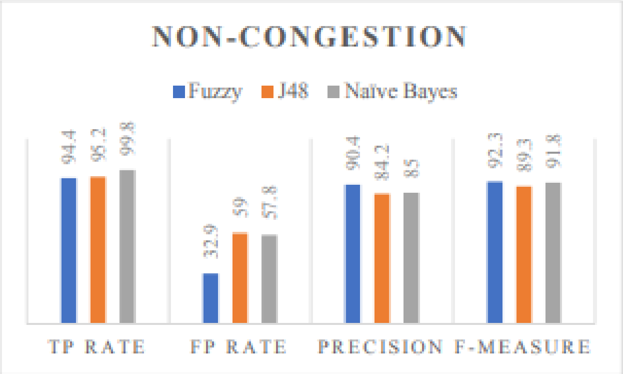


Figure 5: TP rate, FP rate, precision, F-measure, and overall efficiency for non-congestion.

TABLE II. Results for Fuzzy System, J48, and Naïve Bayes.

Experiment	Class	TP Rate / Recall (%)	FP Rate (%)	Precision (%)	F-Measure (%)	Overall Efficiency (%)
Fuzzy System	*Non*	94.4	32.9	90.4	92.3	88.0
	Congested	67.0	5.5	78.4	72.2	
	Weighted Avg.	88.0	26.5	87.6	87.6	
Decision tree (J48)	*Non*	95.2	59.0	84.2	89.3	82.5
	Congested	41.0	4.8	72.1	52.3	
	Weighted Avg.	82.6	46.4	81.4	80.7	
Naïve Bayes	*Non*	99.8	57.8	85.0	91.8	86.3
	Congested	42.2	0.2	98.5	59.1	
	Weighted Avg.	86.4	44.4	88.2	84.2	

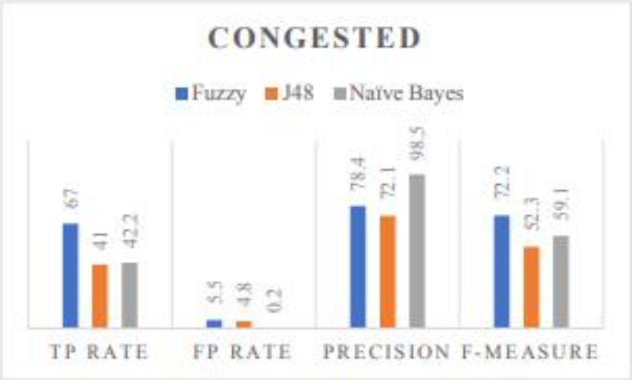


Figure 6: TP rate, FP rate, precision, F-measure, and overall efficiency for congested.

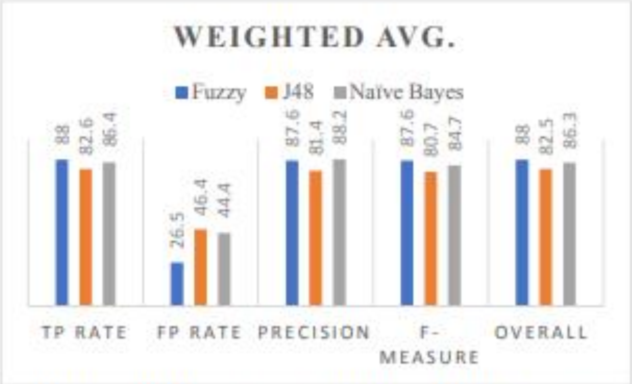


Figure 7: Weighted average of TP rate, FP rate, precision, F-measure, and overall efficiency.

VI. Conclusion and Further Work

This study has proven the hypothesis, H_1: Using journey time and volume data, it is possible to classify congestion using a fuzzy system and has demonstrated a proof of concept fuzzy model. The initial results have demonstrated the fuzzy systems ability to predict congestion using volume and journey time, outperforming both the decision tree and Naïve Bayes. Moreover, the fuzzy system using, only six rules was able to handle an unbalanced dataset. Additionally, the author believes it would be possible to implement this model on other urban road networks. To further this study, the authors are currently working on expanding the system to classify the three types of congestion [15]: Non-recurrent, Recurrent, and Semi-recurrent. This is an important requirement for TfGM who would benefit from not only being able to identify congestion but the type of congestion, which would allow for different mitigation strategies to be put in place. Additionally, they will be able to measure how much of the network is, at a given time, exhibiting signs of non-congestion, recurrent, non-recurrent, and semi-recurrent congestion. To achieve this goal, the fuzzy system will be expanded to add linguistic variables for different times of day, different days of the week, bank holidays, distance from an attraction, and direction of traffic flow.

Acknowledgement

The author would like to thank Transport for Greater Manchester for part funding this research. Additionally, this research contains Transport for Greater Manchester data licenced under the Open Government Licence v3.0.

References

[1] S. Djahel, A. Jones, Y. Hadjadj-Aoul, and A. Khokhar, "CRITIC: A cognitive radio inspired road traffic congestion reduction solution," *IFIP Wirel. Days*, vol. 2018–April, no. February, pp. 151–157, 2018.

[2] L. Rui, Y. Zhang, H. Huang, and X. Qiu, "A new traffic congestion detection and quantification method based on comprehensive fuzzy assessment in VANET," *KSII Trans. Internet Inf. Syst.*, vol. 12, no. 1, pp. 41–60, 2018.

[3] N. Gould and L. Abberley, "The semantics of road congestion," in *UTSG*, 2017.

[4] R. Mudge, D. Montgomery, E. Groshen, J. P. Groshen, S. Helper, and C. Carson, "America's Workforce and the Self-Driving Future Realizing Productivity Gains and Spurring Economic Growth," no. june, 2018.

[5] S. Djahel, R. Doolan, G.-M. Muntean, and J. Murphy, "A Communications-Oriented Perspective on Traffic Management Systems for Smart Cities: Challenges and Innovative Approaches," *IEEE Commun. Surv. Tutorials*, vol. 17, no. 1, pp. 125–151, 2015.

[6] K. Golestan, R. Soua, F. Karray, and M. S. Kamel, "Situation awareness within the context of connected cars: A comprehensive review and recent trends," *Inf. Fusion*, vol. 29, pp. 68–83, May 2015.

[7] N. Isa, M. Yusoff, and A. Mohamed, "A Review on Recent Traffic Congestion Relief Approaches," *2014 4th Int. Conf. Artif. Intell. with Appl. Eng. Technol.*, vol. i, no. November, pp. 121–126, 2014.

[8] P. Pongpaibool, P. Tangamchit, and K. Noodwong, "Evaluation of road traffic congestion using fuzzy techniques," *IEEE Reg. 10 Annu. Int. Conf. Proceedings/TENCON*, pp. 1–4, 2007.

[9] Y. Li, T. Guo, R. Xia, and W. Xie, "Road Traffic Anomaly Detection Based on Fuzzy Theory," *IEEE Access*, vol. 6, pp. 40281–40288, 2018.

[10] Y. Sun, M. Hrušovský, C. Zhang, and M. Lang, "A Time-Dependent Fuzzy Programming Approach for the Green Multimodal Routing Problem with Rail Service Capacity Uncertainty and Road Traffic Congestion," *Complexity*, vol. 2018, pp. 1–22, 2018.

[11] L. Li, "Research on traffic congestion mathematical model in traffic signal control system," *Int. J. Smart Home*, vol. 9, no. 12, pp. 279–288, 2015.

[12] B. Anbaroğlu, T. Cheng, and B. Heydecker, "Non-recurrent traffic congestion detection on heterogeneous urban road networks," *Transp. A Transp. Sci.*, vol. 11, no. 9, pp. 754–771, 2015.

[13] L. A. Zadeh, "Probability Measures of Fuzzy Events," *JOURNAL OF MATHEMATICAL ANALYSIS AND APPLICATIONS*, vol. 23. pp. 421–427, 1968.

[14] R. Bauza, J. Gozalvez, and J. Sanchez-Soriano, "Road traffic congestion detection through cooperative Vehicle-to-Vehicle communications," *Proc. - Conf. Local Comput. Networks, LCN*, pp. 606–612, 2010.

[15] L. Abberley, N. Gould, K. Crockett, and J. Cheng, "Modelling road congestion using ontologies for big data analytics in smart cities," in *2017 International Smart Cities Conference (ISC2)*, 2017, pp. 1–6.

[16] S. Khalifa, Y. Elshater, K. Sundaravarathan, A. Bhat, P. Martin, F. Imam, D. Rope, M. Mcroberts, and C. Statchuk, "The Six Pillars for Building Big Data Analytics Ecosystems," *ACM Comput. Surv.*, vol. 49, no. 2, pp. 1–36, 2016.

[17] C. C. Aggarwal, "A Survey of Uncertain Data Clustering Algorithms," *Data Clust. Algorithms Appl.*, vol. 21, no. 5, pp. 455–480, 2013.

[18] K. Crockett, Z. Bandar, D. Mclean, and J. O'Shea, "On constructing a fuzzy inference framework using crisp decision trees," *Fuzzy Sets Syst.*, vol. 157, no. 21, pp. 2809–2832, 2006.

[19] TfGM, "Transport for Greater Manchester." [Online]. Available: https://www.tfgm.com/.

[20] E. H. Mamdani and S. Assilian, "An experiment in linguistic synthesis with a fuzzy logic controller," *Int. J. Man. Mach. Stud.*, vol. 7, no. 1, pp. 1 13, Jan. 1975.

[21] Weka, "Class J48," 2018. [Online]. Available: http://weka.sourceforge.net/doc.dev/weka/classifiers/trees/J48.html. [Accessed: 15-Nov-2018].

Modelling Road Congestion using Ontologies for Big Data Analytics in Smart Cities

Luke Abberley Student Member, IEEE, Nicholas Gould, Keeley Crockett SMIEEE, Jianquan Cheng
Science and Engineering
Manchester Metropolitan University
Manchester, United Kingdom
Luke.Abberley@mmu.ac.uk, N.Gould@mmu.ac.uk, K.Crockett@mmu.ac.uk, J.Cheng@mmu.ac.uk

Abstract—**Intelligent Transport Systems are a vital component within Smart Cities but rarely provide the context that is required by the road user or network manager that will help support decision making. Such systems need to be able to collect data from multiple heterogeneous sources and analyse this information, providing it to stakeholders in a timely manner. The focus of this work is to use Big Data analytics to gain knowledge about road accidents, which are a major contributor to non-recurrent congestion. The aim is to develop a model capable of capturing the semantics of road accidents within an ontology. With the support of the ontology, selective dimensions and Big Data sources will be chosen to populate a model of non-recurrent congestion. Initial Big Data analysis will be performed on the data collected from two different sensor types in Greater Manchester, UK to determine whether it is possible to identify clusters based on journey time and traffic volumes.**

Keywords—Big Data; Intelligent Transport Systems; Clustering; Ontology;

I. INTRODUCTION

Currently one of the biggest challenges society faces each day is road congestion, which has an enormous impact on health because of pollutants being released from vehicles that are stuck in road congestion worldwide for a total of 4.8 billion hours [1]. In addition, road congestion costs the European Union an estimated 1-2% GDP (£100-200 billion) each year [1], [2]. However, the most crucial consequences of road congestion are the premature deaths caused by deadly chemicals being released and the delays caused to the emergency services using the road network. The road network is the linchpin that holds the other transport networks together [3]; making it vital to alleviate some of the high demand put on it. Two ways to achieve this would be to firstly, provide road users with better multimodal information allowing road users to make better choices such as taking an alternative transport mode. Secondly, it would be useful to develop an Intelligent Transport System (ITS) or a component of one, which is capable of handling multiple heterogeneous data sources. ITSs are an innovative application, which aims to provide traffic managers and road users with better information, allowing for 'smarter' use of transport networks. Current ITSs lack the capability of being dynamic by using multiple *heterogeneous* data sources in near real-time. Moreover, the most noticeable weakness of ITSs is the lack of context they provide to road users because of the *quantitative data* being processed and a lack of *qualitative information*. For example, road users driving on a highway currently would notice variable-message signs stating, "CONGESTION AHEAD EXPECT DELAYS" but this message lacks any useful context creating more questions than answers. For instance, what type of congestion? Where is the congestion? What is the cause? When did it start? When will it end? Are there any alternative routes? How will it influence the overall journey? A more informative message would be "CONGESTION AHEAD IN 2 MILES, DUE TO AN MINOR ACCIDENT AT 15:45 CAUSING INCREASED JOURNEY TIMES".

This research attempts to answer the question: "Can quantitative Big Data be used to provide qualitative information in conjunction with a road traffic ontology with the support of Machine Learning?"

Figure 1 shows the research methodology followed in order to attempt to answer this research question.

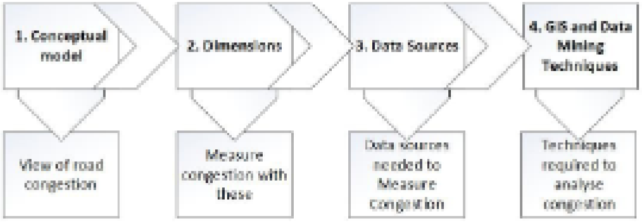

1. Conceptual model
2. Dimensions
3. Data Sources
4. GIS and Data Mining Techniques
View of road congestion
Measure congestion with these
Data sources needed to Measure congestion
Techniques required to analyse congestion

Figure 1: Research methodology

- **Stage 1** is the formulation of a conceptual model of congestion leading to the development of an ontology to provide a formal and explicit conceptualisation of congestion and in particular, the impact of road accidents.
- **Stage 2** From the ontology, the dimensions that describe the congestion caused by accidents are identified, in particular, journey time and traffic volume.
- **Stage 3** Now the dimensions have been identified through the development of the ontology, it is possible to identify which Big Data sources are relevant by reviewing which data sources have been previously used to calculate the journey time and traffic volume. Journey time has previously been calculated using Bluetooth sensors, Global Positioning Systems (GPS), cameras, and traffic volume with Radio-frequency Identification (RFID) and Inductive Loop Counters.
- **Stage 4** Utilising the relevant dimensions and their Big Data Sources, analytics is performed to identify patterns in

This research was partly funded by Transport for Greater Manchester, UK.

978-1-5386-2524-8/17/$31.00 ©2017 IEEE

the traffic volumes and journey times, which can be used to translate quantitative data into qualitative information.

The remainder of this paper is organised as follows. The concepts of congestion are introduced in Section II. In Section III, the road accident ontology will be presented. Section IV will discuss the Big Data sources this research uses. Section V will introduce the experimental design and analysis. Section VI will discuss the experimental results. Finally, we conclude and suggest further work in Section VII.

II. Concepts of Congestion

Although, congestion is not a new phenomenon, and it has been an outstanding problem for every civilisation including ancient Rome, which the Caesars noted [4] *'The passage of goods carts on narrow city streets so congested that they become impassable and unsafe for pedestrians to continue'*. The UK's Department for Transport (DfT) makes a distinction between *physical* congestion that can be characterised by considering average speeds on the network and *relative* congestion that is defined by the road user's expectation [5]. For example, a person who regularly drives a certain route, which is regularly congested, would consider this normal. However, a different person driving the same route for the first time may consider it to be severely congested [6]. A report into traffic congestion by the U.S Department of Transportation (DoT) focuses primarily on a *relative* approach to defining congestion using terms such as *'clog'*, *'impede'* and *'excessive fullness'* and adds *'For anyone who has ever sat in congested traffic, those words should sound familiar.'* [7]. The same report noted how congestion is typically related to an excess of vehicles on a portion of roadway or pedestrians on a sidewalk. There is still an apparent *absence* of consistency of how congestion is defined. This is partly due to the multifaceted nature of congestion and how it is perceived.

In this research, road traffic congestion is distinguished between two *vague* types: non-recurrent and recurrent congestion. Vague because, although the terms such as recurrent or non-recurrent are widely accepted by academics and transport management, the relative views and individual perspectives slightly differ. Table I shows a definition for each type of congestion.

TABLE I. Definition of Congestion

Congestion Type	Definition	References
Recurrent congestion	Occurs when significant amounts of vehicles simultaneously use a limited space of road. Such as weekday morning and afternoons peak hours' traffic jam situations.	[8]–[10]
Non-recurrent congestion	Occurs from a road traffic incident such as traffic accidents, work zones, extreme weather conditions and some special events like music concerts and important sports events.	[1], [11], [12]

III. An Ontology for Congestion

Ontologies have become an area of interest within many fields such as Computing [13], [14], Geography [15], [16] and Transportation [2], [17], [18]. The main motive for using an ontology is the way it allows data and algorithms to be described in a formal and explicit way forcing clarification and improving knowledge management and decision-making [19] whilst remaining accurate, conflict free and faithful to each individual domain [20].

The ontology shown in Figure 2 was developed by performing an extensive literature review into many concepts of congestion and road accidents and using data collected from Transport for Greater Manchester, UK (TfGM) to perform a data exploration of a road accident that happened on 1st November 2016 on the A5103 road in Manchester, UK.

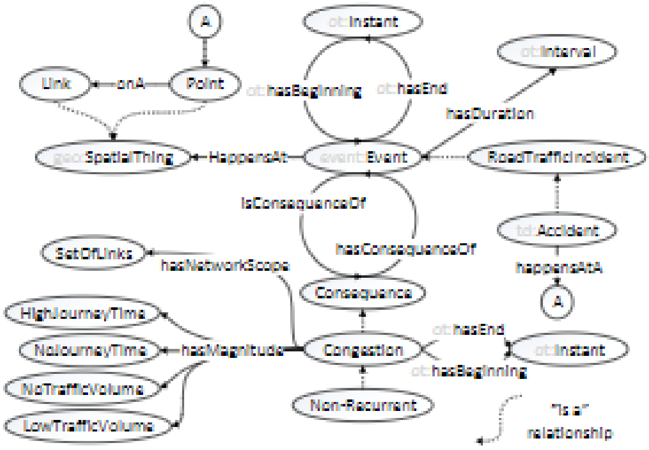


Figure 2: Ontology: Road Accident

The ontology (Figure 2), explains the relationship between **an event**, for example, a road **accident** and its **consequence**, which is **non-recurrent congestion**. From the ontology, we know an **accident** is a **road traffic incident**, which is a type of **event**. These types of **event** have temporal aspects, which are **instant**, and **interval**.

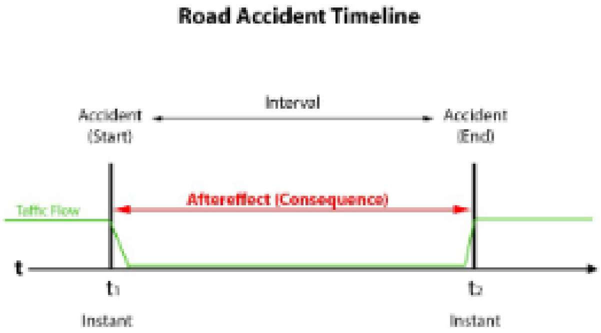


Figure 3: Accident temporal aspect example

Figure 3 provides a visual example of the temporal aspects of a road **accident**. *t1* is the **instant** of a vehicle impact another object, *t2* is the **instant** where the traffic flow returns to "normal" and the interval between these two instances is the **consequence** of impact on the **set of road network links**, which lasts an amount of time (**interval**). Additionally, **non-recurrent congestion** is the **consequence** of the **event** and has a network scope, which originates from a **spatial thing** such as a **point** on a **link** that the **accident** occurred. Finally, we define

congestion caused by a road **accident** as having magnitudes such as a **high journey time** and **low traffic volumes**.

.

IV. Big Data Sources

Many research projects and commercial tools such as Google Traffic provide a dimension of congestion with terms like 'free flow' and 'bound flow' [21] and road speeds in quasi-real-time by using data culled from mobile phone users.

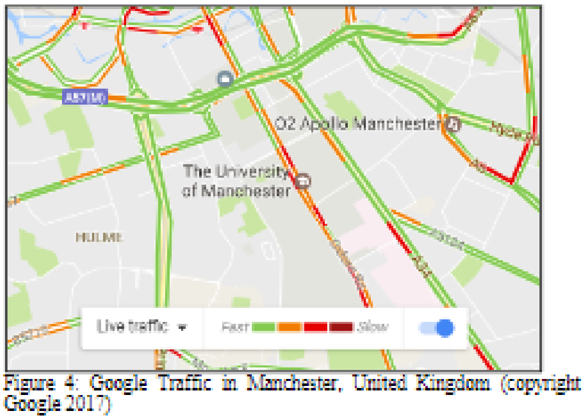


Figure 4: Google Traffic in Manchester, United Kingdom (copyright Google 2017)

However, Figure 4 shows the information that a Google traffic user would see where links are highlighted with one of four colours that relate to the average speed on that link. However, there is a clear absence of context. What speeds do the four colours refer too? Do slow speeds mean the link is congested? If congestion has occurred then what is the cause? When did it start and when will it end? Are these speeds normal for the day and time? According to [22], it is important to be able to identify the cause of congestion, e.g. A road accident.

This research will use the data presented in Table II and will focus on a 4.5-mile section of the A6 road, which connects Stockport to Manchester city centre, UK. Table II shows which data sources, where the data was acquired, the area covered, timeframe and dimension gained.

TABLE II. Table of Data

Data	From	Location	Timeframe	dimension
Bluetooth	TfGM	Manchester, UK	2016-Current	Journey Time
Inductive Loop Counter	TfGM	Manchester, UK	2015-Current	Traffic volume
Accident Data	STATS19 [34]	UK	2005-current	Casualty accidents only

These data sources have been discussed and previously used in research which aimed to improve ITS[1], [23]. Inductive Loop Counters have been discussed and used to save travel time and detect anomalies [9], [24]. The accident data is being used to provide an understanding of historical accidents to help identify new accidents in quasi-real-time. However, what makes this research novel is the combination of data from multiple sensor sources to identify the occurrence of road accidents, and providing this information to road users in a qualitative format. These data sources do come with their challenges. Bluetooth sensors are not 100% reliable since a zero second journey time could be due to several reasons. For example, there were no vehicles with a Bluetooth device that had driven past at least two sensors; also, the mobile network used to transmit sensor data to the central server could have been affected by bad weather; also, Bluetooth MAC address may have been allocated to multiple devices, which could cause an unexpected journey time. Inductive Loop Counters are sparsely deployed in the study area. An accident is only recorded if there are one or more casualties and a police officer has attended, which means that an accident that may have caused congestion might not be in the dataset. This is defined as an *incomplete* dataset.

V. Experimental Design and Analysis

For this research, a non-labelled dataset has been created using all the data sources mentioned in Table II. A non-labelled dataset is best suited to being analysed with an unsupervised learning algorithm such as clustering [25]. Clustering is a type of machine learning algorithm and is one of the most commonly used algorithms when a user has a non-labelled data problem that requires a solution [26]. Clustering models the relationship between variables using approaches such as centroid-based and hierarchical. All clustering methods use the inherent structures in the data to best organize the data into groups of maximum commonalities. Some of the most popular clustering algorithms are k-Means, k-Medians, Expectation Maximisation (EM) and Hierarchical Clustering [27].

Traditionally, congestion has been assessed by measuring speed, volume, and occupancy on the road network. However, these dimensions are not without limitations; for example, speed (as opposed to mean speed) is a measure at a single point on a link and cannot be used as a constant due to the possibility of a road block or incident which could cause a vehicle to reduce their speed before regaining speed before going passed another speed checkpoint. Volume and occupancy require frequently deployed 'expensive' equipment, for instance, Inductive Loop Counters. Therefore, the following hypotheses will use data from inexpensive technology that can be used to calculate journey times rather than speed and identify changes in journey time and traffic volume depending on day and time providing information that is more useful.

Hypothesis One

H0: Clustering an unsupervised dataset creates clusters that make it possible to predict journey time.

H1: Clustering an unsupervised dataset creates clusters that cannot be used to predict journey time.

Hypothesis Two

H0: Clustering an unsupervised dataset creates clusters that make it possible to identify differences between a weekday and a weekend.

H1: Clustering an unsupervised dataset creates clusters that cannot be used to identify differences between a weekday and a weekend.

A. Methodology

The first step is to collect the data, which is recorded when a vehicle or an occupant with a Bluetooth enabled device passes numerous sensors. The MAC address of the vehicle or a Bluetooth enabled device being carried by an occupant are recorded in a raw data file called Per Vehicle Record (PVR). These MAC addresses are then used to calculate the journey time of several users between an origin and destination in 15-minute intervals. Once sufficient data has been collected and processed; involving the conversion of mean journey times into seconds from a time stamp, the source file is imported into a database. Finally, modelling will be performed using the K-Means++ algorithm which is an unsupervised learning method with a non-labelled dataset. K-Means++ algorithm was chosen because according to [28] it has previously achieving functional values 20% better than K-Means and performed 70% faster.

VI. Results and Discussion

The purpose of this experiment was to discover patterns in the journey time and traffic volumes to help predict and classify journey time. Figure 5 displays 13 weeks of data in a scatter graph with Tuesday, Wednesday and across the x-axis, journey time along the y-axis and grouped into four time periods that are 6:00, 7:00, 8:00 and 9:00. Each group represents a 15 minutes slot. For example, 6:00 until 6:15.

From Figure 5, it is apparent that the group 6:00 and 7:00 are a lot more consistent with regards to journey time than 8:00 and 9:00 that appear to have a lot more variation ranging from 0 to 2600 seconds. 9:00 is positioned sparsely between 7:00 and 8:00 demonstrating a visible temporal pattern in the journey time data.

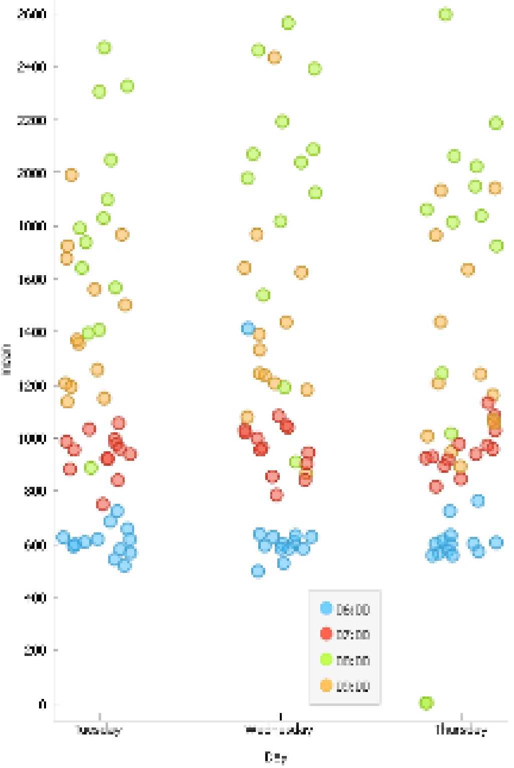


Figure 5: Scatter graph of journey times

Following the exploration of journey time in Figure 5, the next phase was to try to prove whether hypothesis one is true or not. To achieve this clustering using the K-Means++ algorithm was chosen.

Figure 6 was produced by using K-Means++ to choose the initial seeds and a euclidean distance was used. Five clusters were chosen for two reasons. Firstly, it achieves the second highest silhouette score with 0.609. Secondly, after an initial attempt with three clusters that did not provide sufficient resolution, it was decided to use five clusters instead. Resolution is vital to be able to prove hypothesis one true, because without the ability to classify journey time into meaningful classes it would be impossible to predict the level of journey time. In Figure 6 the five classifications are Very High, High, Average, Low and Very Low journey time. In addition, to the five classifications, Figure 6 has many interesting patterns, such as, journey time between 00:00 until 06:30 remained densely in the Very Low or Low journey time. In addition, during the remainder of the day Journey time becomes less Low journey time and more Average and High journey time. Finally, around 8:00 you can see a Very High journey time spike.

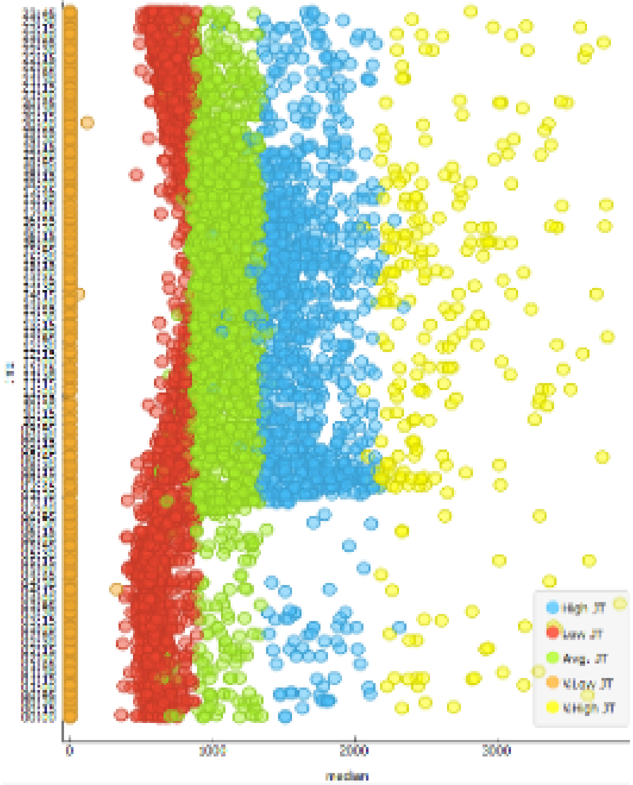


Figure 6: Clustered journey time into five categories 1) V. High JT 2) High JT 3) Avg. JT 4) Low JT 5) V. Low JT

To be able to prove hypothesis two either true or not a slightly different approach was used concerning how it was presented visually. In Figure 7, the x-axis is used for all 7 days of the week and the y-axis is used for time of day in 15-minute intervals. The size of each point is used to refer to the traffic volume and the five classifications remain the same.

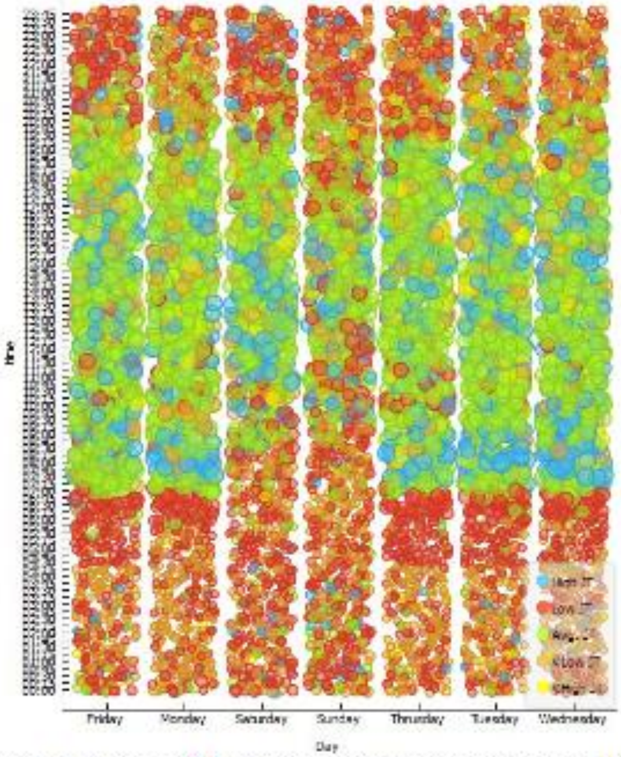


Figure 7: Clustered daily journey times into five categories 1) V. High JT 2) High JT 3) Avg. JT 4) Low JT 5) V. Low JT

Figure 7 shows it is possible to use clustering to identify differences between weekdays and weekend. For example, on Saturday and Sunday, there are long periods of low journey times and fewer vehicles using the road in the morning. In addition, on Monday, Tuesday, Wednesday, and Thursday there is noticeable High journey time at around 8:00 each morning, which is expected because people are going to work and dropping children off at school. Finally, it is worth noting the volume levels typically become high at 7 am during the week and does not reduce until around 8 pm proving hypothesis two true. Proving these hypotheses true is vital for when we attempt to identify the difference between a spike in journey time and a reduction in traffic volume caused by a road accident or a recurrent event such as morning rush hour.

A case study was chosen to attempt to answer the research question as to whether the impact of a road accident could be identified in the sensor data. The case study is from a fatal road accident on the A6 on the 7th of February 2017. Using the data sources mentioned in Table II, journey time and the time of the accident was plotted on two timelines, the first is the day of the accident and the second is the mean of 13 weeks (January until March 2017). Looking at Figure 8, there is a noticeable difference at the time of the fatal accident between the journey time average, which is around 2000 seconds (Average JT), and the day of the fatal accident that fluctuates between either 0 second (V. Low JT) or around 3500 seconds (V. High JT). For a road user, these values mean very little but after using the clusters created in the experimental analysis, we can say the journey time has changed from an average journey time to either no journey (road closed) or a very high journey time state, which lasts for around 3 hours overall before returning to the expected journey time. In addition, these measurements match up to what was proposed in the road accident ontology.

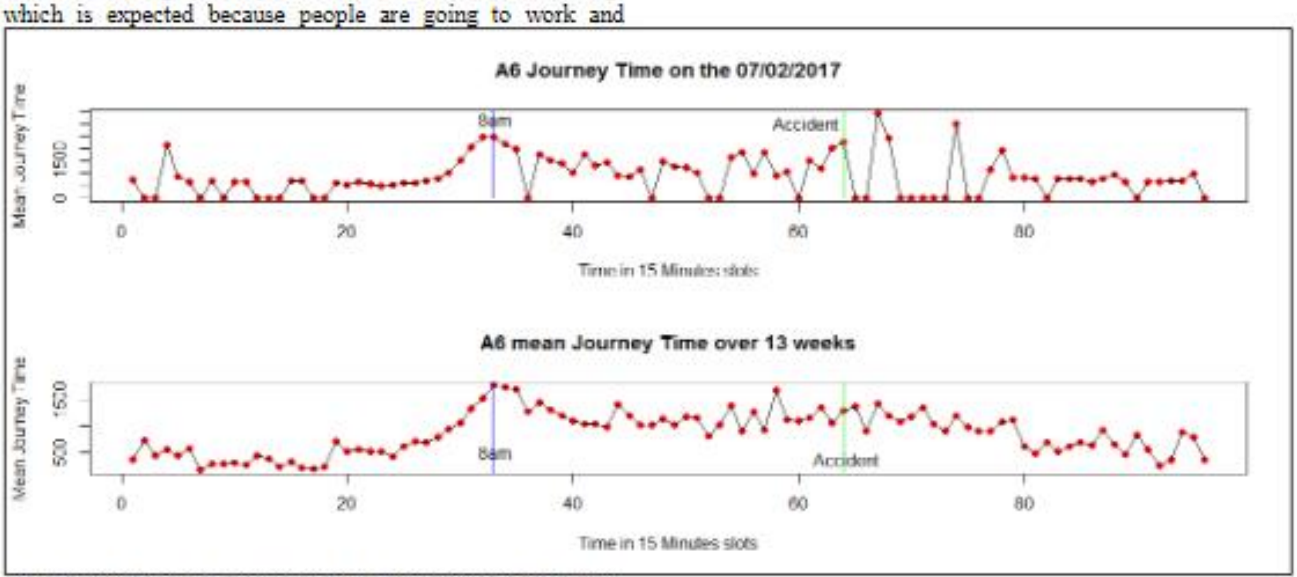


Figure 8: Journey time on the a) 7th February 2017 b) Over a 13 week period.

VII. Conclusion and Further Work

This paper has discussed the many concepts of congestion, which were used along with the support of TfGM to develop the road accident ontology. The two dimensions that were chosen with the support of the ontology were used to identify which data sources are best to be used to perform the experimental analysis that helped to prove both hypothesis and the research question. In addition, this research has demonstrated that it is possible to take *quantitative* data and extract *qualitative* information, which a road user or transport manager could use to help support decision-making. However, despite the promising results, further work is required to establish whether it is possible to identify similar patterns within a spatiotemporal dataset that can identify the shockwave caused by traffic events such as accidents. Then develop an early warning system that can detect such events, which cause non-recurrent congestion and predict the impact severity.

References

[1] S. Djahel, R. Doolan, G.-M. Muntean, and J. Murphy, "A Communications-Oriented Perspective on Traffic Management Systems for Smart Cities: Challenges and Innovative Approaches," *IEEE Commun. Surv. Tutorials*, vol. 17, no. 1, pp. 125–151, 2015.

[2] D. Corsar, M. Markovic, P. Edwards, and J. D. Nelson, "The Transport Disruption ontology," 2015. [Online]. Available: https://transportdisruption.github.io/transportdisruption.html#. [Accessed: 10-Oct-2015].

[3] A. O. Somuyiwa, S. O. Fadare, and B. B. Ayantoyinbo, "Analysis of the Cost of Traffic Congestion on Worker's Productivity in a Mega City of a Developing Economy," pp. 644–656, 2015.

[4] A. Downs, *Still stuck in traffic: Coping with peak-hour traffic congestion*. 2004.

[5] Department for Transport, "Reported Road Casualties in Great Britain: notes, definitions, symbols and conventions," *Dep. Transp.*, pp. 1–6, 2015.

[6] Department for Transport, "An introduction to the Department for Transport ' s road congestion statistics," no. August, 2013.

[7] U.S Department of Transportation, "Traffic Congestion and Reliability: Trends and Advanced Strategies for Congestion Mitigation," 2017.

[8] M. Fosgerau and K. A. Small, "Hypercongestion in downtown metropolis," *J. Urban Econ.*, vol. 76, pp. 122–134, 2013.

[9] E. T. Verhoef, "Time, speeds, flows and densities in static models of road traffic congestion and congestion pricing," *Reg. Sci. Urban Econ.*, vol. 29, pp. 341–369, 1999.

[10] R. Arnott, "A bathtub model of downtown traffic congestion," *J. Urban Econ.*, vol. 76, no. 1, pp. 110–121, 2013.

[11] E. T. Verhoef and J. Rouwendal, "A behavioural model of traffic congestion Endogenizing speed choice, traffic safety and time losses," *J. Urban Econ.*, vol. 56, pp. 408–434, 2004.

[12] M. J. Cassidy and R. L. Bertini, "Some traffic features at freeway bottlenecks," *Transp. Res. Part B Methodol.*, vol. 33, no. 1, pp. 25–42, 1999.

[13] D. Fensel, I. Horrocks, F. Van Harmelen, and D. Mcguinness, "OIL Ontology Infrastructure to Enable the Semantic Web," *Intell. Syst. IEEE*, vol. 16, no. 2, pp. 38–45, 2001.

[14] F. Bobillo and U. Straccia, "The fuzzy ontology reasoner fuzzyDL," *Knowledge-Based Syst.*, vol. 95, pp. 12–34, 2016.

[15] H. Couclelis, "People manipulate objects (but cultivate fields): Beyond the Raster-Vector Debate in GIS," *Theor. Methods Spat. Reason. Geogr. Sp.*, vol. 639, no. 716, pp. 65–77, 1992.

[16] M. Joronen and J. Häkli, "Politicizing ontology," *Prog. Hum. Geogr.*, pp. 1–19, 2016.

[17] K. Golestan, R. Soua, F. Karray, and M. S. Kamel, "Situation awareness within the context of connected cars: A comprehensive review and recent trends," *Inf. Fusion*, vol. 29, pp. 68–83, May 2015.

[18] F. Lecue, A. Schumann, and M. L. Sbodio, "Applying semantic web technologies for diagnosing road traffic congestions," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 7650 LNCS, no. PART 2, pp. 114–130, 2012.

[19] M. Uschold, J. Bateman, M. Davis, J. Sowa, C. M. Bennett, R. Brooks, A. Dima, M. Gruninger, N. Guarino, L. Obrst, S. Ray, T. Schneider, R. Sriram, M. West, and P. Yim, "Making the Case for Ontology," pp. 1–10, 2011.

[20] G. Shanks, E. Tansley, and R. Weber, "Using ontology to validate conceptual models," *Commun. ACM*, vol. 46, no. 10, pp. 85–89, 2003.

[21] J. Kianfar and P. Edara, "A Data Mining Approach to Creating Fundamental Traffic Flow Diagram," *Procedia - Soc. Behav. Sci.*, vol. 104, pp. 430–439, 2013.

[22] N. Gould and L. Abberley, "The semantics of road congestion," in *UTSG*, 2017.

[23] A. D. Patire, M. Wright, B. Prodhomme, and A. M. Bayen, "How much GPS data do we need?," *Transp. Res. Part C Emerg. Technol.*, vol. 58, pp. 325–342, 2015.

[24] F. Yuan and R. L. Cheu, "Incident detection using support vector machines," *Transp. Res. Part C Emerg. Technol.*, vol. 11, no. 3–4, pp. 309–328, 2003.

[25] Y. Zhang, N. Ye, R. Wang, and R. Malekian, "A Method for Traffic Congestion Clustering Judgment Based on Grey Relational Analysis," *ISPRS Int. J. Geo-Information*, vol. 5, no. 5, p. 71, 2016.

[26] C. L. Philip Chen and C. Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data," *Inf. Sci. (Ny).*, vol. 275, pp. 314–347, 2014.

[27] C. C. Aggarwal, "A Survey of Uncertain Data Clustering Algorithms," *Data Clust. Algorithms Appl.*, vol. 21, no. 5, pp. 455–480, 2013.

[28] D. Arthur and S. Vassilvitskii, "K-Means++: the Advantages of Careful Seeding," *Proc. eighteenth Annu. ACM-SIAM Symp. Discret. algorithms*, vol. 8, pp. 1027–1025, 2007.

The semantics of road congestion

Dr Nicholas Gould
Lecturer in Geographic Information Science

Luke Abberley
PhD candidate

Manchester Metropolitan University

Abstract

Most live road traffic information systems, such as Google Traffic, do not provide the user with the context of congestion. To usefully support decision making, by drivers and network managers, such systems need to provide information such as the probable cause of the congestion and its likely time span. The focus of this work is on non-recurrent congestion.

We aim to develop a system that captures the semantics of road congestion by interpreting sensor data collected in the Greater Manchester region. This data consists of journey time data (collected by Bluetooth sensors) and volume, or count, data collected by induction loops. Rather than supplying information such as the current journey time on a particular road link, which is meaningless without context, we aim to provide context sensitive information such as increasing, abnormal, journey times near the football stadium, in the direction of the football stadium.

Clusters of anomalous sensor readings are identified using an agglomerative hierarchical clustering algorithm in R. The main challenge is in determining which readings are anomalous. The characteristics of the largest clusters are then taken as typical of that kind of congestion causing event. Initial work has involved identifying the journey time and volume patterns of a known attractor, a football match and we aim to extend the work to automatically identify unplanned events such as road accidents, using the sensor data.

Introduction

The impact of road congestion on the economy, on air quality and on well-being (Office for National Statistics, 2014), is enormous. Congestion can be classed as recurrent (such as that experienced in the "rush hour") and non-recurrent, that caused by incidents such as road accidents. Traffic agencies define the two differently but the quantity of non-recurrent congestion has been estimated at between 40% and 70% of total congestion (Kwon et al., 2006). Furthermore, a reduction of recurrent congestion involves policy and the encouragement of behavioural change such as a modal shift to public transport. Could it be that the previous focus on recurrent congestion was based on the view that congestion was an urban planning problem and could be solved by planning and engineering approaches? Non-recurrent congestions now seems an easier target, especially with the availability of new near real-time data sources. Although that is not to say that solutions designed to reduce *recurrent* congestion will not influence *non-recurrent* congestion; a general reduction in road traffic will reduce the impact of unpredictable events and lead to a more *resilient* network (Reggiani, 2013).

To begin to solve the problem of non-recurrent congestion, however, still requires the identification of congestion, but this is difficult without a clear measure. Furthermore, the actuality of congestion is dependent on circumstances and the road user's perception. Low speeds on the road network near a football stadium will be perceived as expected by the match attendee but as congestion by the non-attendee. The UK's Department for Transport recognises this in its distinction between *physical* congestion that can be characterised by considering average speeds on the network, and *relative* congestion that is defined by the road user's expectation (Department for Transport, 2015).

Tools such as Google Traffic provide snapshots of road speeds in near real-time by using GPS data culled from mobile phone users (Figure 1). However, this information displays only average speeds on road links; there is a lack of context here. To what extent do slow speeds on particular links represent congestion? If there is congestion then what caused it? When did it start? When is it likely to finish? There is also no depiction of congestion as a relative phenomena. Figure 1 displays low speeds at major road junctions, but is that not just an expected downside of city centre driving?

Context can be provided by identifying the cause of the congestion. The road user stuck in heavy traffic would benefit from the knowledge that the congestion is caused by a football match that will kick off in five minutes time and after that, the congestion will reduce.

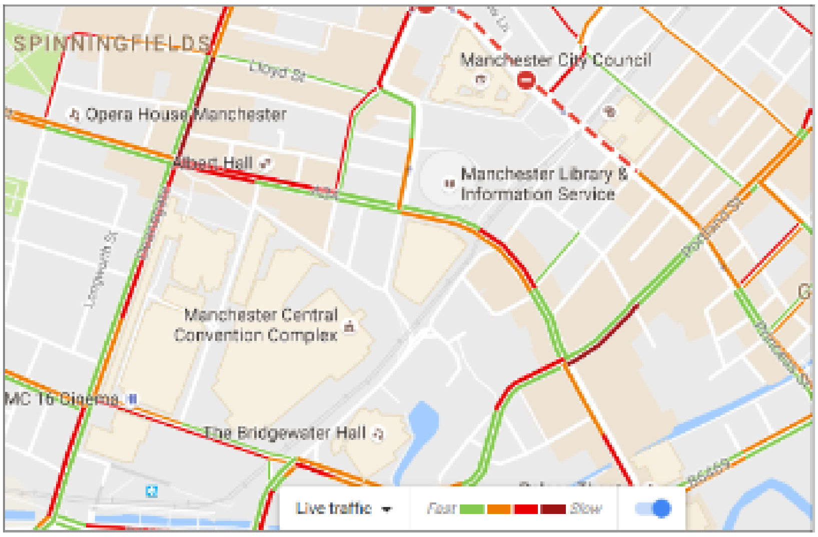


Figure 1 Google Traffic in Manchester City Centre (copyright Google 2016)

Many different sources can now be used to identify traffic congestion in addition to that data collected from sensors installed by municipalities: Google uses data from GPS enabled smart phones; fleet vehicles or high-end cars fitted with GPS can provide historical journey time data; Uber has started to make its GPS data available to city planners.

However, these sources may not persist. For example, data services may suffer temporary outages or be permanently withdrawn; sources that were once free to use may start extracting a charge or change their terms and conditions. It is therefore necessary to ensure that any model can embrace multiple data sources.

We suggest therefore that a purely numerical model is not sufficient to capture the complexities of road congestion, in particular the relative dimension. In order to understand congestion we require an open model that is neither reliant on opaque data sources, nor limited to road network sensors, but can be expanded to incorporate other data sources such as weather forecasts, air quality measurements and social media.

Ultimately more contextual information about road congestion can support multi-modal travel information systems; a driver might be informed that their route to the city centre is heavily congested owing to a serious road accident but five minutes drive away is a light-rail station with a car park that is 50% full and a service to the city centre due in fifteen minutes. Users of all modes of transport complain about the lack of detail in times of disruption; the more information provided to travellers will enable them to make appropriate decisions. If we are to respond to a congestion event effectively, we need to understand its cause as well as its nature.

To react appropriately in order to alleviate congestion we need *diagnosis* (Lécué et al., 2012), this requires an understanding of the causes and characteristics. Therefore, it is not sufficient simply to report the current state of the network (as for example Google maps can). To react to a storm that has been identified by sensors we need to know the characteristics of the storm - strength, size, direction - in order to mitigate against it, but we need not know its cause. We cannot avoid it. With congestion, if we know its cause we may be able to halt it or at least reduce its scope.

This leads us to conclude that a more nuanced description of congestion than current speeds on road links is necessary; in particular, there is a need to explain the context of road congestion. Ultimately, is it possible to use sensor data to allow traffic managers to alleviate congestion when it occurs, for example, by changing signal timings and priorities or by informing drivers using Variable Message Signs (VMS) and other tools?

A semantic approach to road congestion

Context is part of the semantics of a domain; we can define the concepts and the relationships between those concepts using semantics and we adopt the definition of Kuhn (2005) of semantics as the meaning of expressions in a language. The expression of a concept in a language aids understanding. We propose using an *ontology* to describe the characteristics and causes of congestion. An ontology can provide a formal, machine-readable, representation that makes intended meaning computable (Yim, 2015).

In our road network, we may have different sensor types that are influenced by road traffic. For example, Bluetooth sensors can be used to determine the mean journey time between two points on the network. This is an immediate and direction measure of congestion; the higher the journey time the worse the congestion. Induction loops, buried in the road can be used to accurately count the number of cars passing the loop. This count is not, however, a direct measure of congestion. Contrarily, a higher than normal volume can mean the opposite; that the traffic is flowing smoothly. However, it can be an indicator of future congestion if, for example, the flow is in the direction of an attractor. Other sensors such as rain gauges are not (directly) influenced by traffic but can be used as a predictor of possible congestion since weather conditions have an impact on demand (Creemers et al., 2015).

Lécué et al. (2012) use a semantic matching approach to compare the current road conditions with historic conditions. For example, if there is congestion on road x near event y and that has happened in the past then we can infer that the reoccurrence of the event is the cause of the congestion. However, they do not define patterns of congestion. Anicic et al. (2012) describe a semantic event processing system that tries to identify traffic bottlenecks in near real-time but describe congestion purely in terms of speeds on particular roads; there is no recognition of the relative nature of congestion.

Llaves and Kuhn (2014) separate event types and event patterns in the formalization of knowledge. Event patterns are not included in ontologies. For example, the type might be *heavy rainfall* and the pattern *rainfall above 4mm per hour*. This allows for flexibility; Transport for Greater Manchester and Transport for London can both have the conception of "high journey times" but can have different measures of them. This is the approach used by this research.

Method

Congestion before and after a football match at the Etihad stadium, East of Manchester city centre was used as the first case study. The football match represents a relatively predictable cause of non-recurrent congestion, with a known attractor (the stadium) and start and end times (kick-off and full-time). The aim was to identify and formalise patterns of congestion related to football matches in the sensor data. The intention is to model more unpredictable events, such as road accidents, in future work.

The data is supplied by TfGM and consists of journey time data on links collected from passive Bluetooth sensors and traffic volume data from permanent induction loops. For both data sources, the data was aggregated into 10-minute time slots[1]. This was a fairly arbitrary selection but any larger and the resolution would be too small to allow for real-time reactions by traffic managers to events, and any smaller and sample sizes would be too small.

[1] Thus, slot 1 will represent the 10 minutes between 12 midnight and 10 minutes past midnight, and slot 144 will represent the 10-minute slot prior to midnight.

Figure 2 shows the mean journey times between two different Bluetooth sensors on a section of road to the North East of the stadium on two different days - one a match day (circles) and one a non-match day (crosses). On both days the pattern in the early part of the day is similar, both exhibiting the morning rush hour, where journey times increase. On match day (13th January 2016), relatively high journey times over a relatively long time prior to kick-off can be seen, followed by a very high spike after the match finishes. This pattern is expected since some supporters make there way to the game early where others arrive just in time, whereas all supporters tend to leave at a similar time.

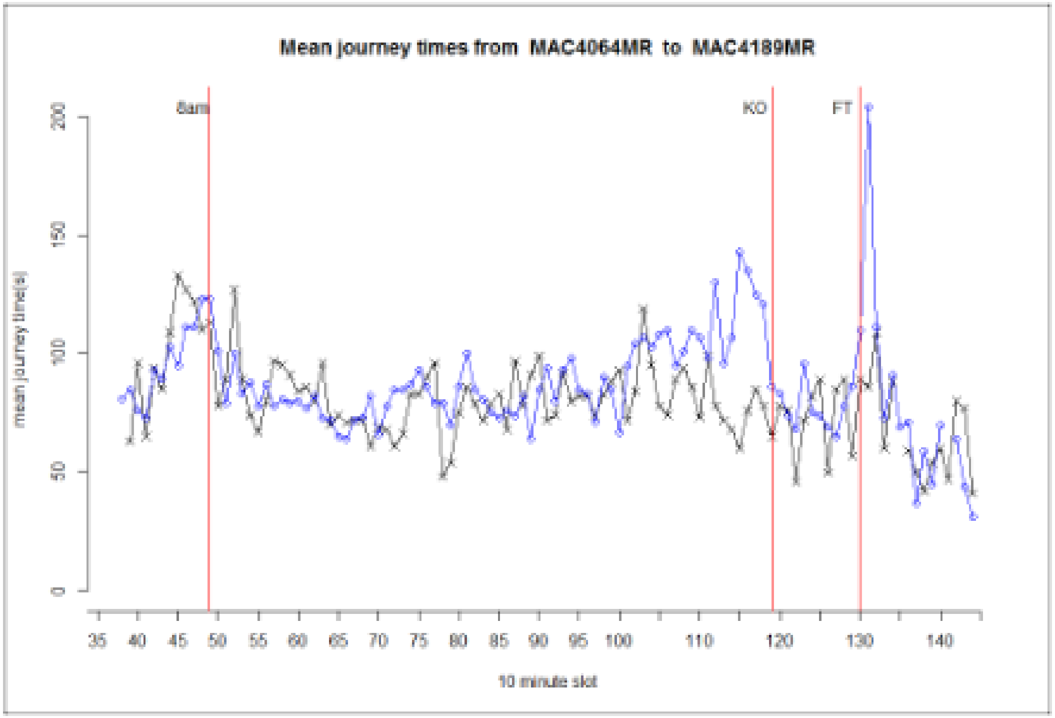


Figure 2 Mean journey time between two Bluetooth sensors on 13th January (o) and 20th January 2016 (x) from 6am to midnight

The mean journey times between pairs of sensors were analysed, following outlier removal. Since we are interested in non-recurrent, or atypical, congestion a measure of recurrent, or typical, congestion is required. As well as the time of day, road agencies typically allow for differences in demand on weekdays/weekends and holidays/non-holidays. Since this study focusses on Wednesday evening football matches, the data for typical traffic was based on four non-match day Wednesdays. This selection is relatively arbitrary, and the selection of "typical" road conditions is worthy of a study in itself. The more "typical" days used then the better the measure of "typical" conditions, however if we go too far back into the past then we will end up ignoring medium and longer term trends in the data. For example, we may end up including data from when a road link was controlled differently from when it was on the study date. Given these caveats, Figure 3 shows the journey times on a link on a match day (circle) compared to the mean of four "typical" days (cross) and one standard deviation either side of that mean (square).

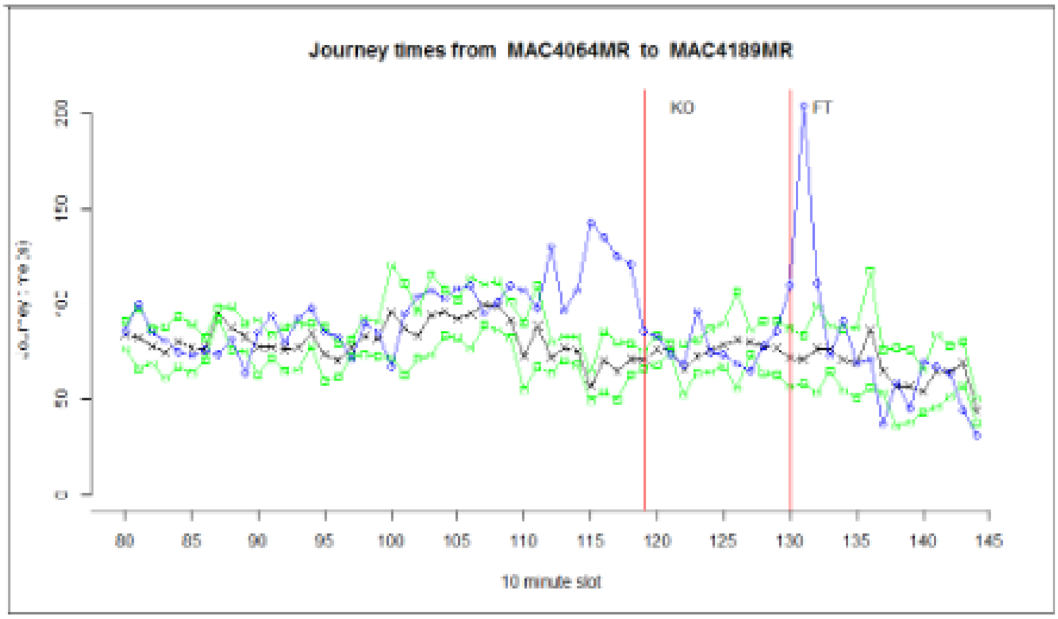


Figure 3 Journey times on a match day compared to typical days from 1pm

The next step is to classify the abnormal journey times using relative terms. The characteristics of the journey time on any road section between two sensors that are considered are *magnitude*, *direction* and *proximity* to the attractor, in this case the stadium. This approach allows for the generalisation of the approach; "high" journey times in Manchester city centre will have very different absolute values from a city such as London, say but with this approach we can use the same language.

Firstly, the magnitude of the journey times on the match day are classified using their differences from the mean value of the typical days in any particular time slot. For example, if the journey time is between one and two standard deviations from the typical day mean then that reading is classed as "high". This too is arbitrary but at least it allows for the relative nature of congestion.

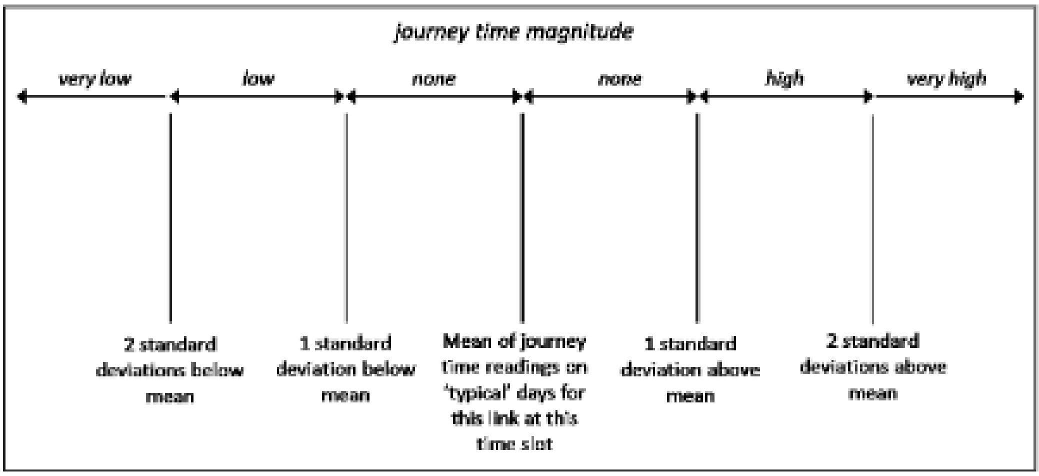

journey time magnitude
very low
low
none
none
high
very high
2 standard deviations below mean
1 standard deviation below mean
Mean of journey time readings on 'typical' days for this link at this time slot
1 standard deviation above mean
2 standard deviations above mean

Figure 4 Classifying journey time magnitude

The next step is to classify the distance of each road link from the football stadium, or more exactly the distance of the mid point of each road link to the entrance of the stadium car park. Therefore, for example, the distance of the link between sensors MAC4065MR and MAC1313 and the stadium is the sum of *a* and *b*. (Figure 5). The entrance of the main car park was used as a proxy for the centre of the attractor rather than the stadium itself. Obviously, this does not account for the fact that there are multiple car parks and informal street parking near the stadium.

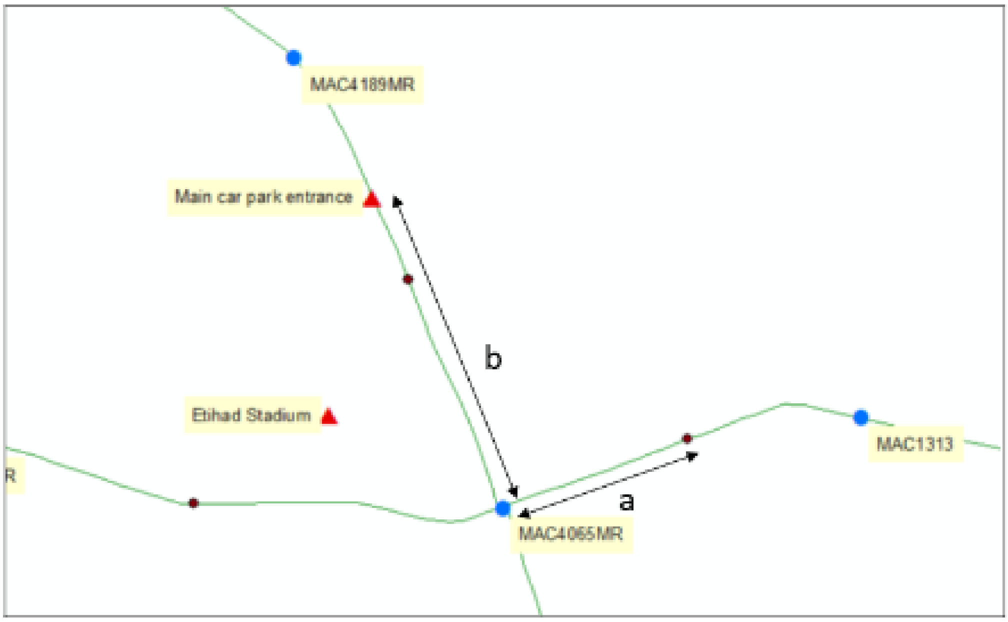


Figure 5 Measuring the distance (a + b) of a link from the stadium (small dots are mid-points of links, large dots are sensor locations)

The distances were placed in quantiles, based on thirds, and allocated a relative distance of *near*, *very near* and *far*. This allows for a richer, and scalable, description of distance than the calculating of Euclidean distance between link centre points.

Finally, each link was assigned a relative direction - towards the stadium or away from the stadium. For example, traffic traversing the link between sensors MAC4065MR and MAC1313 (Figure 5) is designated as *away from the stadium* (main car park) and in the reverse direction (MAC1313 to MAC4065MR) as *towards the stadium*. Again, this is a more semantically rich designation than using compass points, for example (*West* and *East* for this link).

A similar technique was used to identify anomalies in the vehicle count (volume) data. Abnormal volume magnitudes were classified in the same way (Figure 4) and a distance to the stadium was assigned to each counter location and a direction (towards or away from the stadium) was generated.

In any one ten minute time slot there are differences in the characteristics of each link even if they share the same distance and direction in relation to the stadium, given the unpredictable nature of traffic flow. The next step, therefore, is to identify clusters of journey times and volumes on links sharing the same characteristics in terms of magnitude, distance and direction. The DAISY algorithm (Kaufman and Rousseeuw, 2005), as implemented in R (Maechler et al., 2015), was used to create a dissimilarity matrix for the anomalous journey times based on magnitude, distance and direction. This matrix was used as input to the AGNES agglomerative hierarchical clustering algorithm (Kaufman and Rousseeuw, 2005) which generated the clusters and the tree. A cluster is categorised as the journey time readings in that time slot that share the same magnitude, distance from the stadium and relative direction.

Results

Dendrograms for each time slot were created from the clusters identified. Figure 6 shows an example dendrogram generated from hierarchical clustering for the anomalies in a 10-minute time slot starting just prior to kick off. Each item in the cluster represents a journey time on a road segment. The clusters at the lowest level (height = 0) are where there are exact matches of magnitude, relative distance from stadium, and relative direction. The largest cluster of links, of 8 readings are of the form *high journey times*, *very near* the stadium and in the direction *towards* the stadium, which matches expectations so near to kick off.

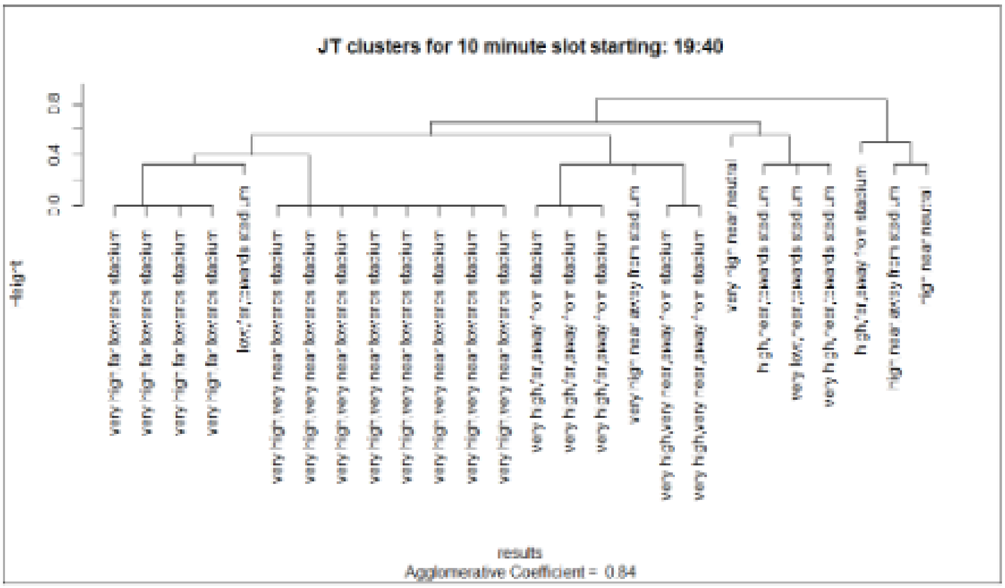


Figure 6 Vehicle journey time clusters prior to kick-off on 13th January 2016

Figure 7 shows a dendrogram for the 10-minute time slot starting at 21:50, some 15 minutes after full time. Here the largest cluster (6 members) is of the form very high journey times, very near to the stadium but this time travelling *away* from the stadium, which is, again, what would be expected following the end of the match. Note that the next most significant cluster is for high journey times, very near to the stadium but *towards* the stadium. This demonstrates that the area around the stadium is congested even for those heading towards the stadium.

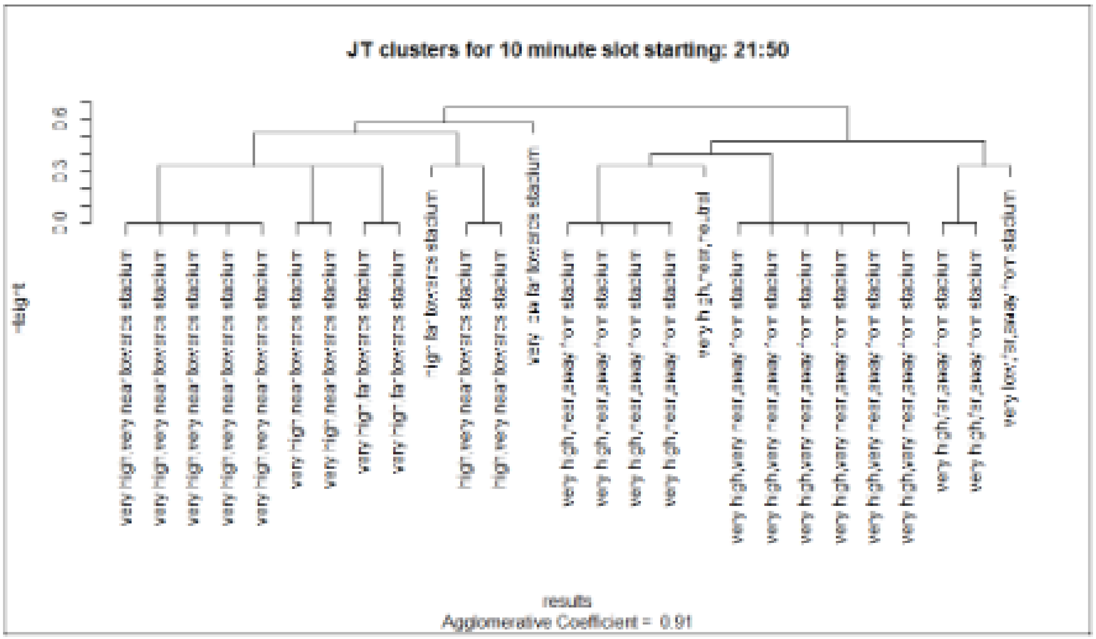


Figure 7 Vehicle journey time clusters after full-time on 13th January 2016

As with the journey time data clusters based on magnitude, direction and distance for traffic *volume* (count) were identified. Figure 8 and Figure 9 show the dendrograms for vehicle count clusters for the same time slots as Figure 6 and Figure 7. Traffic volume counters are relatively few in the study area compared to Bluetooth sensors and subsequently, relatively fewer clusters are identified in comparison to the journey time data and those clusters that are identified, have significantly fewer members. For the time slot displayed in Figure 9, for example, there are no perfect clusters (where Height = 0).

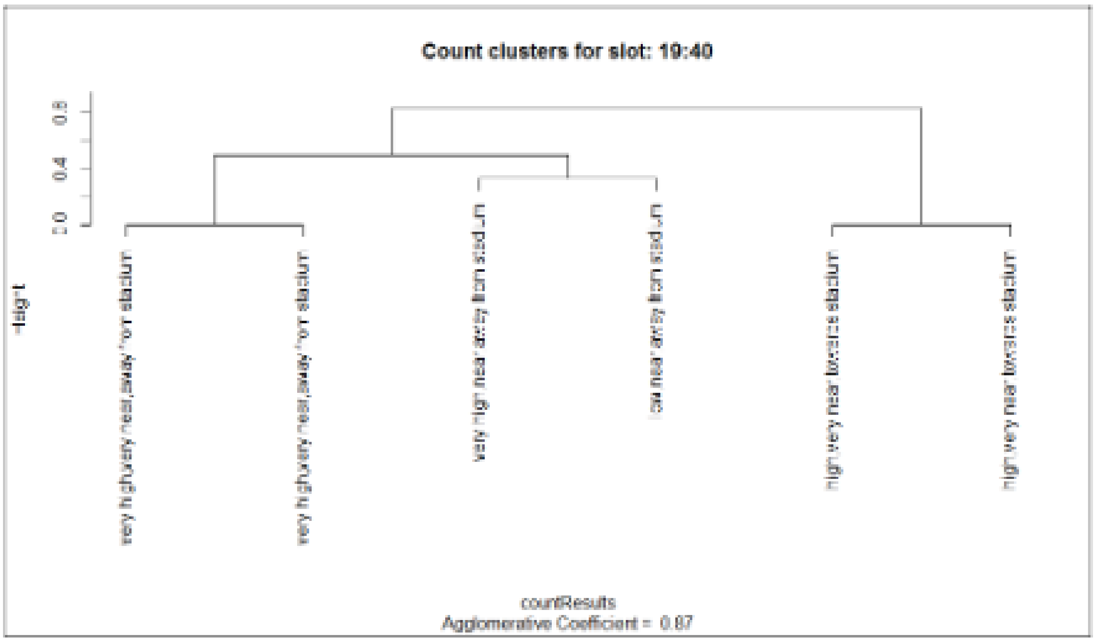


Figure 8 Vehicle count clusters prior to kick-off on 13th January 2016

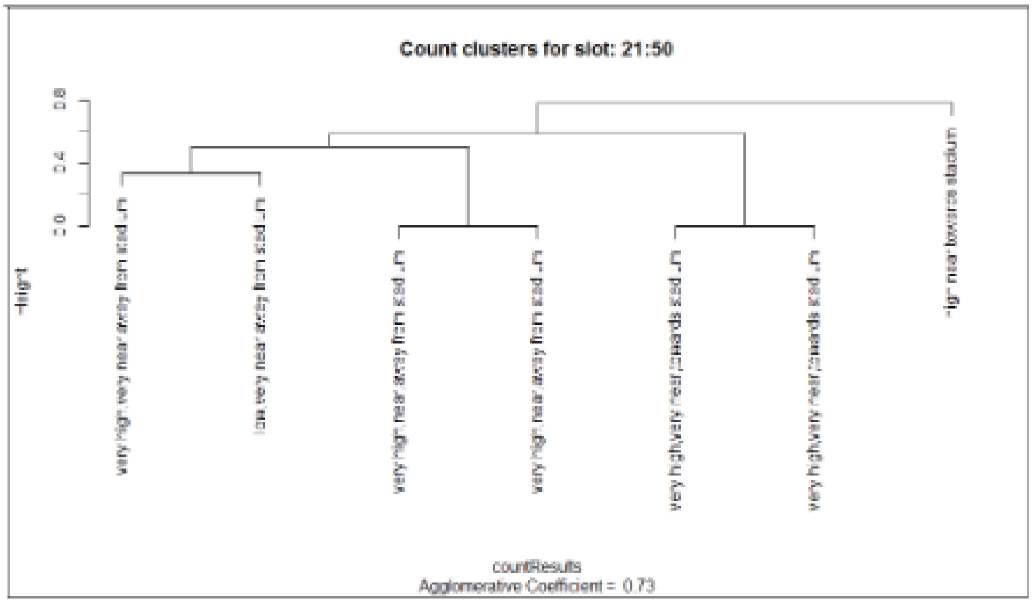


Figure 9 Vehicle count clusters after full-time on 13th January 2016

An analysis of the data on other Wednesday, evening kick-off match days, at the stadium (21st October 2015 and 27th January 2016) reveals similar patterns.

Now we have a better, although still simplified, understanding of road congestion caused by football matches, we can capture the semantics of congestion in an ontology. Formalisation, using an ontology, will eventually allow for automation of the response to congestion. A term such as "high congestion" on itself is meaningless. We need to use terms such as "high journey times" or "high volumes". We need not add absolute values to these terms when defining them in the ontology. Llaves and Kuhn (2014) make the distinction between *event types* and *event patterns*, where the latter has no place in the ontology. The same applies to our concept of "very high journey" times. We can include the concept in the common, shared ontology but the definition used above (Figure 4) would be part of a local implementation of a system that uses the ontology.

Some of the relevant concepts, such as Football Match are defined in the *Transport Disruption Ontology* (Corsar et al., 2015). The ontology lacks the concept of congestion but has the concepts of *Heavy Traffic*, *Queuing Traffic*, *Slow Traffic* and *Stationary Traffic* taken from the DATEX II specification (www.datex.eu). However, these concepts are defined in terms of a percentage of free-flow traffic; Stationary Traffic is defined as "average speed is less than 10% of its free-flow level", for example. These terms provide too simplified a view of congestion. There are also other gaps, for example the ontology has the concept Football *Match* but not Football *Stadium*. The latter is necessary in our case since we refer to the relative distance from the stadium. The football stadium has two roles, when it is hosting a football match it acts as an *Attractor* to traffic, in other times it serves as a *Landmark*, providing context. Elements of the OWL-Time ontology (W3C, 2006) were used to describe the temporal aspects of the football match and its resultant congestion (Figure 10). Concepts and relationships borrowed from the Transport Disruption and OWL-Time ontologies are prefixed *td* and *ot* respectively.

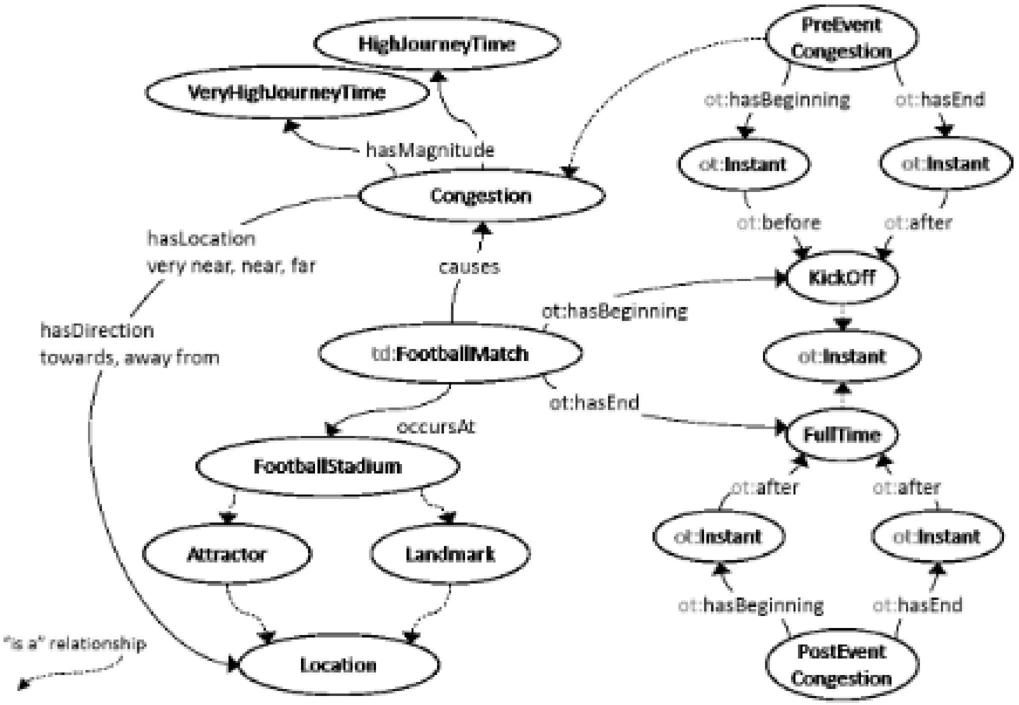


Figure 10 An ontology of the impact of a football match on road congestion

The ontology also describes the relationship between the football match and traffic count values (Figure 11). Here the post match relationship has been omitted for brevity. As stated earlier, high count values are not a direct indicator of congestion but more likely an indicator of future congestion, as drivers head for an attractor. Traffic counts play an entirely different role when the cause is a road accident; prior to the accident there will be no abnormal count, after the accident the count will reduce. These patterns, as identified by the sensors, can help distinguish between the causes of congestion providing diagnosis.

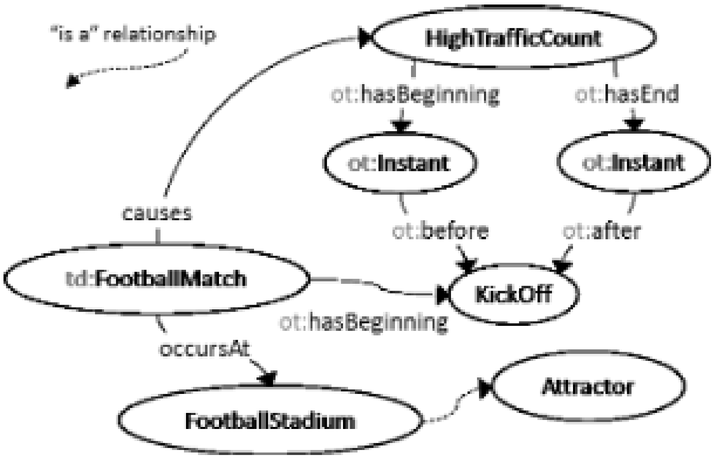


Figure 11 The relationship between traffic counts and an attractor

The start and end times of the congestion phenomena are defined using instances but could be represented using the OWL-Time *Interval* concept, to allow for a degree of fuzziness. The ontology should be extended and perhaps revised; is it the football match or the football stadium that is the attractor? Other congestion causing events could be described in a similar manner; an unpredictable event such as road accident would only have *PostEventCongestion*. If each type of event has a sufficiently distinct profile then the data sources could be potentially used to identify the cause of a congestion and thus help to alleviate it.

Discussion and further work

There is much work to be done on both the data analysis and the ontology. The definition of the magnitude of abnormal journey times - high, very high - (Figure 4) lacks the resolution to capture the difference between the significant differences in magnitude before and after the match (Figure 3). Rather than look at the data by time slot, it would be useful to include a temporal classification of each reading; for example, very near to the event start, a long time after the event end.

Also missing is a technique to describe the relative differences in the *duration* of the high journey times pre and post-match. Another consideration is whether the derivative of the magnitude of the journey times is more useful than the absolute values; i.e. is the journey time increasing or decreasing?

The classification of distance and direction presume that the location of the source of the congestion, in this case an attractor, is known. Further work is required to determine if it is possible to identify the source from sensor readings for events such as accidents and roadworks.

Ideally, the model should be able to infer the importance of the stadium and other reference points (e.g. motorway junctions) and include them where necessary. The stadium is only an *attractor* before and after a football fixture; however, it is a *landmark* at all times. We need to add other relevant features (attractors and landmarks) into the model and then determine the relative distance of the sensor sites from them and also the relative direction (towards/away) of the measured traffic.

Acknowledgements

This research was funded by the UK Department for Transport as part of the T-TRIG programme. Thanks are due to David Atkin at TfGM for supplying the data and identifying a suitable study area.

References

Anicic, D., Rudolph, S., Fodor, P. and Stojanovic, N. (2012) Stream reasoning and complex event processing in ETALIS. *Semantic Web*, 3(4) pp. pp. 397-407.

Corsar, D., Markovic, M., Edwards, P. and Nelson, J. (2015) The Transport Disruption Ontology. In *The 14th International Semantic Web Conference*. Bethlehem, Pennsylvania, 11th - 15th October 2015.

Creemers, L., Wets, G. and Cools, M. (2015) Meteorological variation in daily travel behaviour: evidence from revealed preference data from the Netherlands. *Theoretical and Applied Climatology*, 120(1) pp. 183-194.

Department for Transport. (2015) *An introduction to the Department for Transport's road congestion statistics*. [Online] [Accessed on 22nd November 2016] https://www.gov.uk/government/publications/road-congestion-and-travel-times-statistics-guidance

Kaufman, L. and Rousseeuw, P. J. (2005) *Finding groups in data: an introduction to cluster analysis*. Hoboken, New Jersey: John Wiley & Sons.

Kuhn, W. (2005) Geospatial Semantics: Why, of What, and How? In *Journal on Data Semantics III*. Vol. 3534. Berlin / Heidelberg: Springer

Kwon, J., Mauch, M. and Varaiya, P. (2006) Components of Congestion: Delay from Incidents, Special Events, Lane Closures, Weather, Potential Ramp Metering Gain, and

Excess Demand. *Transportation Research Record: Journal of the Transportation Research Board*, 1959 pp. 84-91.

Lécué, F., Schumann, A. and Sbodio, M. L. (2012) Applying Semantic Web Technologies for Diagnosing Road Traffic Congestions. In Cudré-Mauroux, P., Heflin, J., Sirin, E., Tudorache, T., Euzenat, J., Hauswirth, M., Parreira, J. X., Hendler, J., Schreiber, G., Bernstein, A. and Blomqvist, E. (eds.) *The Semantic Web – ISWC 2012: 11th International Semantic Web Conference*, Boston, MA, USA, November 11-15, 2012, Proceedings, Part II. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 114-130.

Llaves, A. and Kuhn, W. (2014) An event abstraction layer for the integration of geosensor data. *International Journal of Geographical Information Science*, 28(5) pp. 1085-1106.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M. and Hornik, K. (2015) *cluster: Cluster Analysis Basics and Extensions*. R package version 2.0.3.

Office for National Statistics. (2014) *Commuting and Personal Well-being, 2014*. Office for National Statistics.

Reggiani, A. (2013) Network resilience for transport security: Some methodological considerations. *Transport Policy*, 28 pp. 63-68.

W3C. (2006) *Time Ontology in OWL W3C*. [Online] [Accessed on 4th November 2016] http://www.w3.org/TR/owl-time/

Yim, P. (2015) Bootstrapping the applied ontology practice: Ontology communities, then and now. *Applied Ontology*, 10(3-4) pp. pp. 229-241.