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.

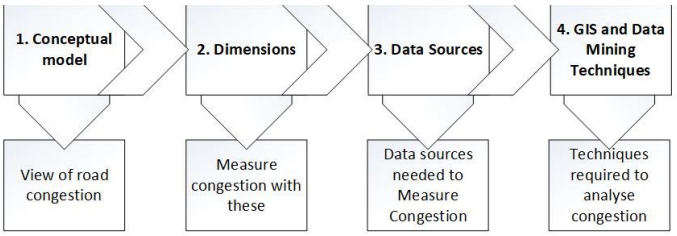

Figure 1: Research methodology

- **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.

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.

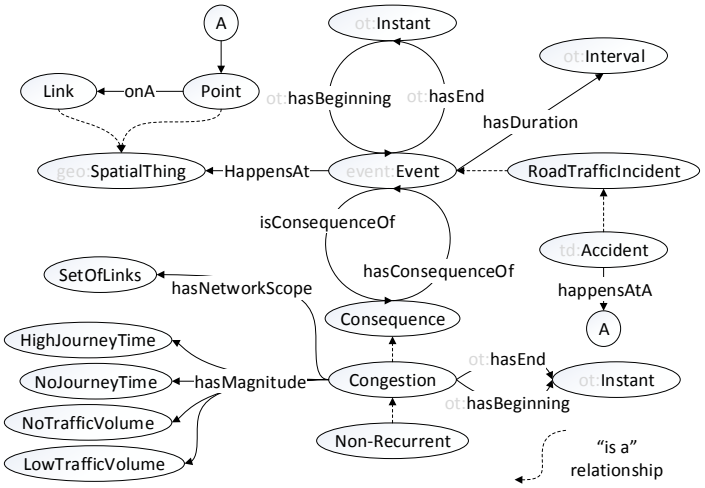

Figure 2: Ontology: Road Accident

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**.

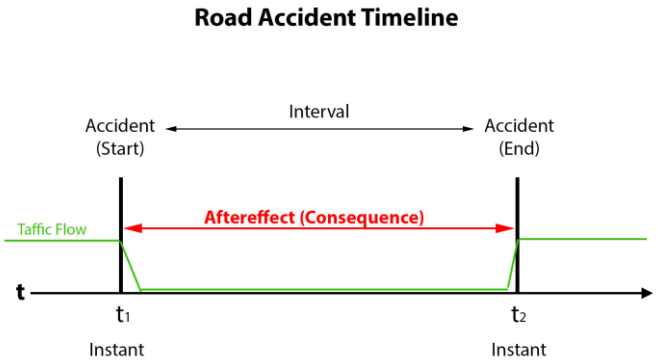

Figure 3: Accident temporal aspect example

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.

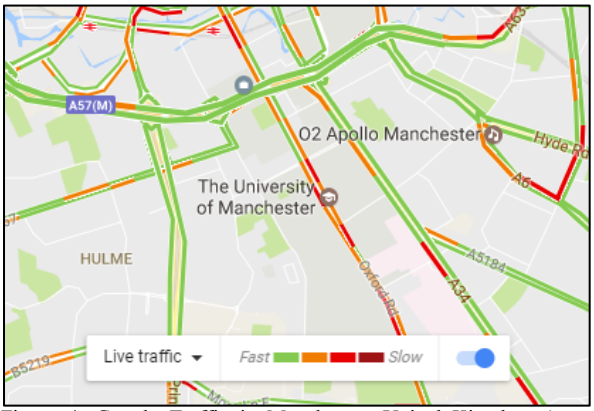

Figure 4: Google Traffic in Manchester, United Kingdom (copyright Google 2017)

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.

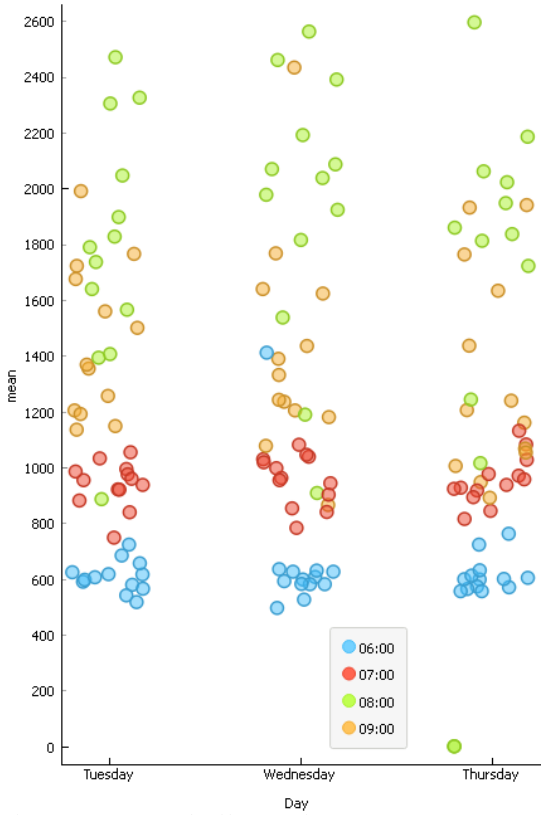

Figure 5: Scatter graph of journey times

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.

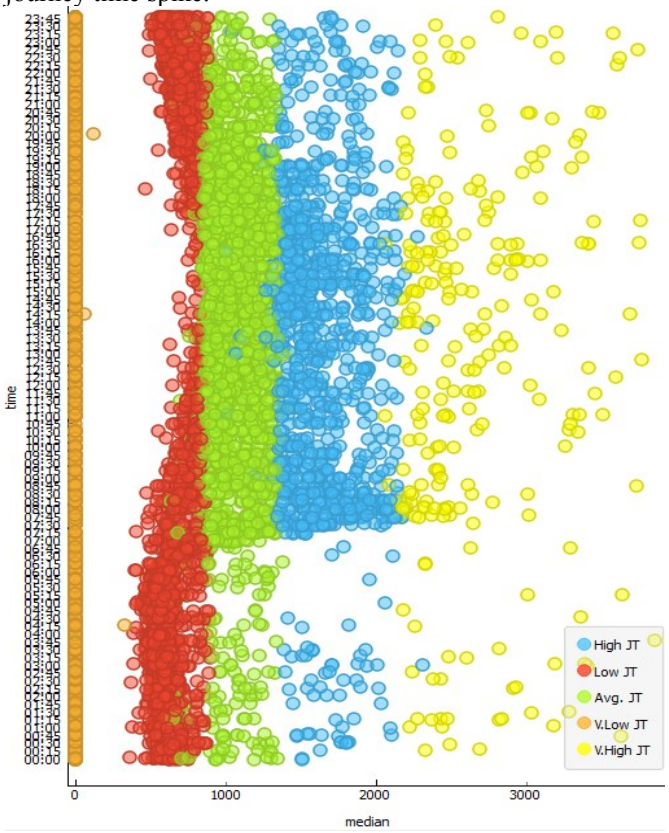

Figure 6: Clustered journey time into five categories 1) V. High JT 2) High JT 3) Avg. JT 4) Low JT 5) V. Low JT

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.

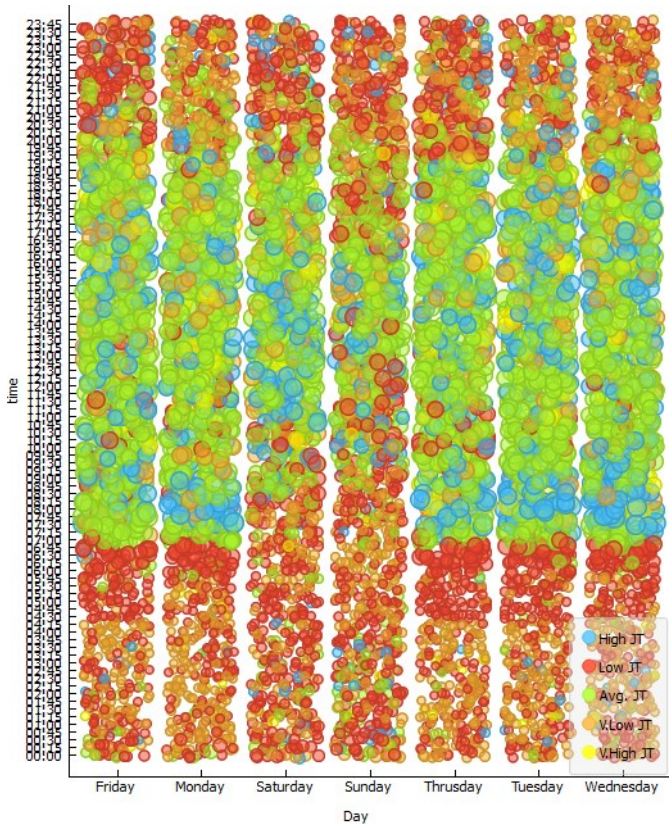

Figure 7: Clustered daily journey times into five categories 1) V. High JT 2) High JT 3) Avg. JT 4) Low JT 5) V. Low JT

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.

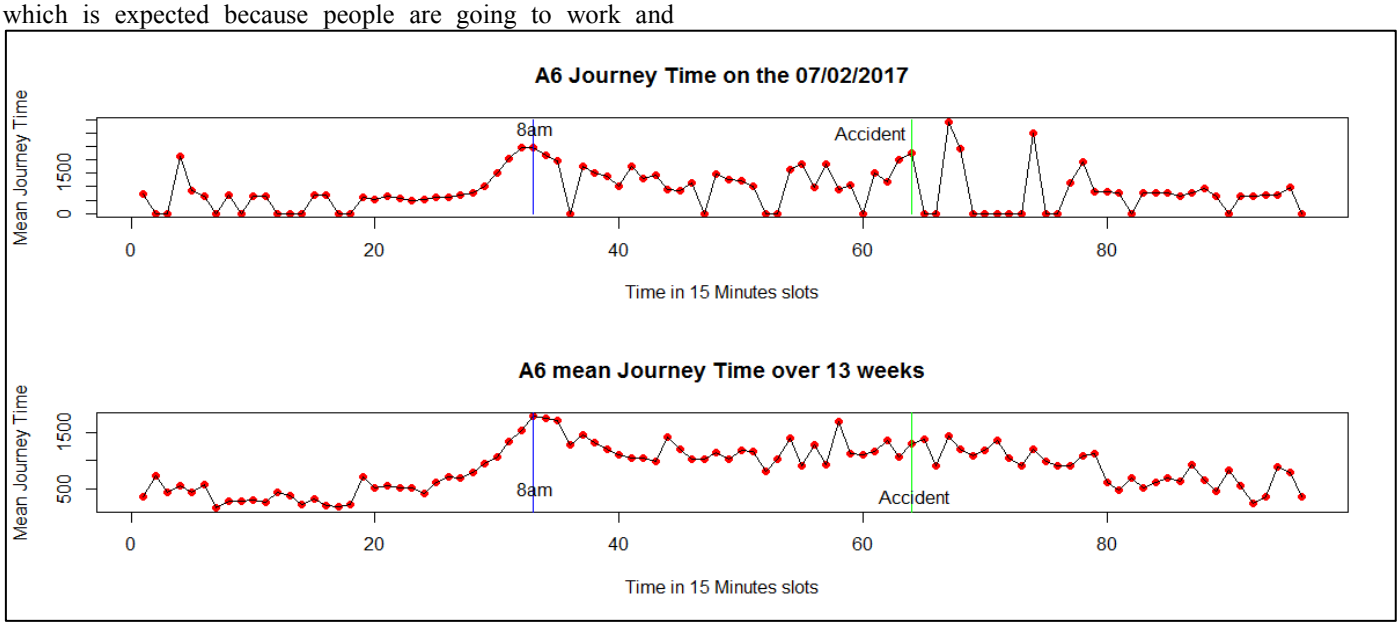

Figure 8: Journey time on the a) 7th February 2017 b) Over a 13 week period.

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. Lécué, 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.