# 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.

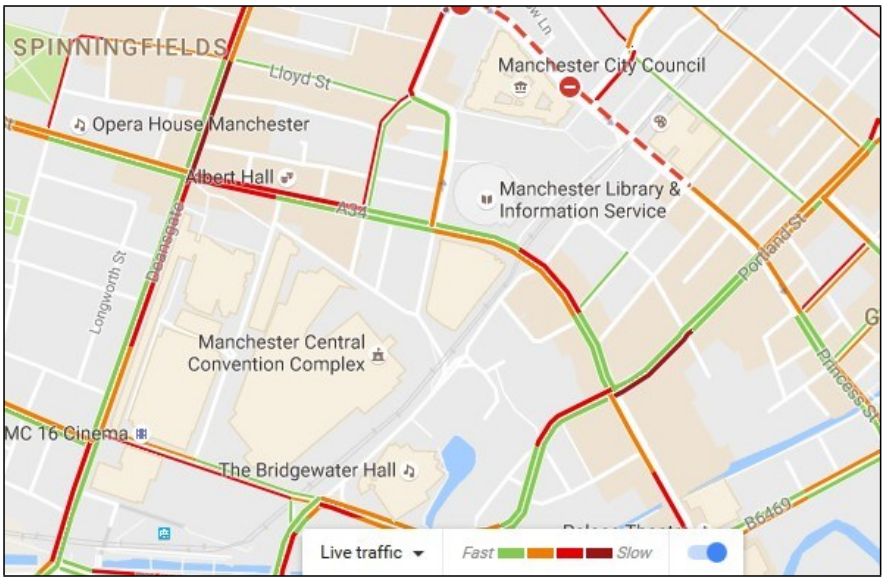


**Figure 1 Google Traffic in Manchester City Centre (copyright Google 2016)**

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.

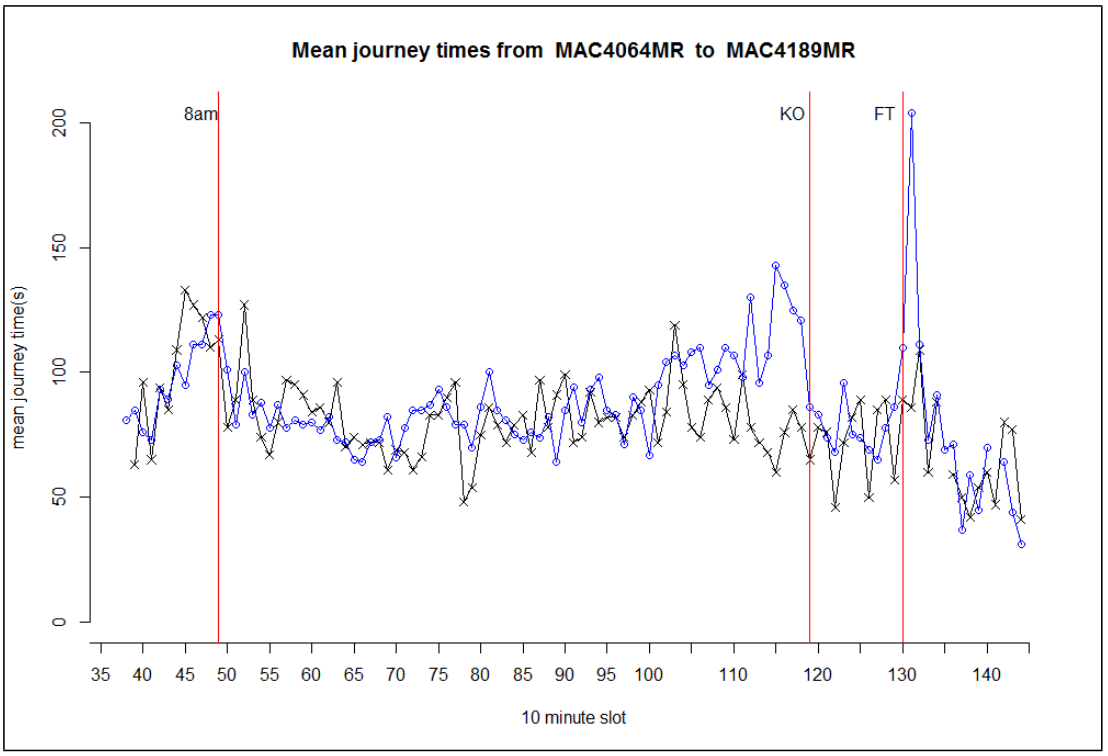


**Figure 2 Mean journey time between two Bluetooth sensors on 13th January (o) and 20th January 2016 (x) from 6am to midnight**

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).

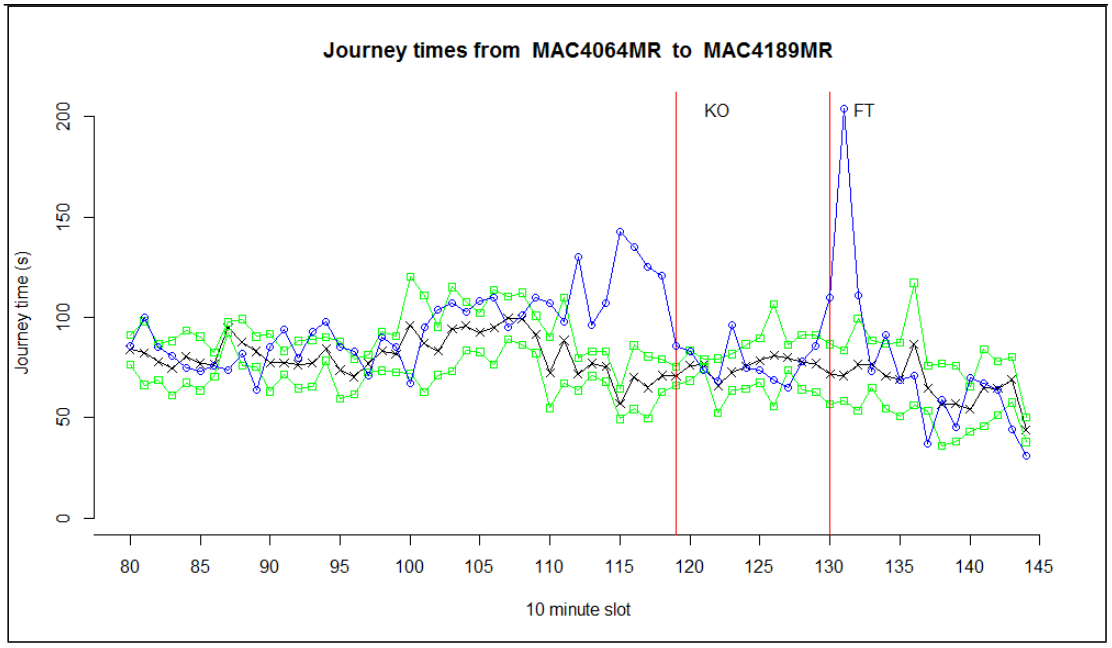


**Figure 3 Journey times on a match day compared to typical days from 1pm**

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.

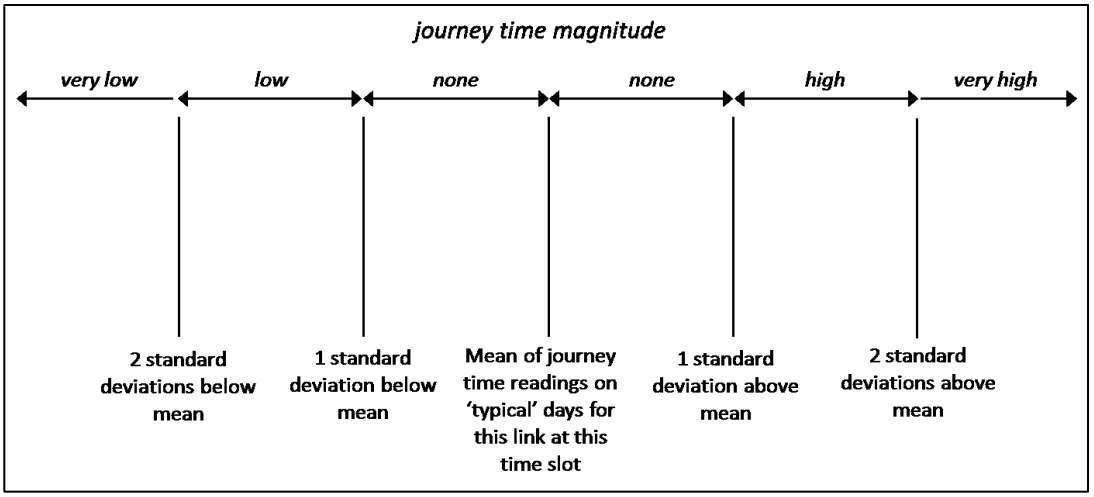


**Figure 4 Classifying journey time magnitude**

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.

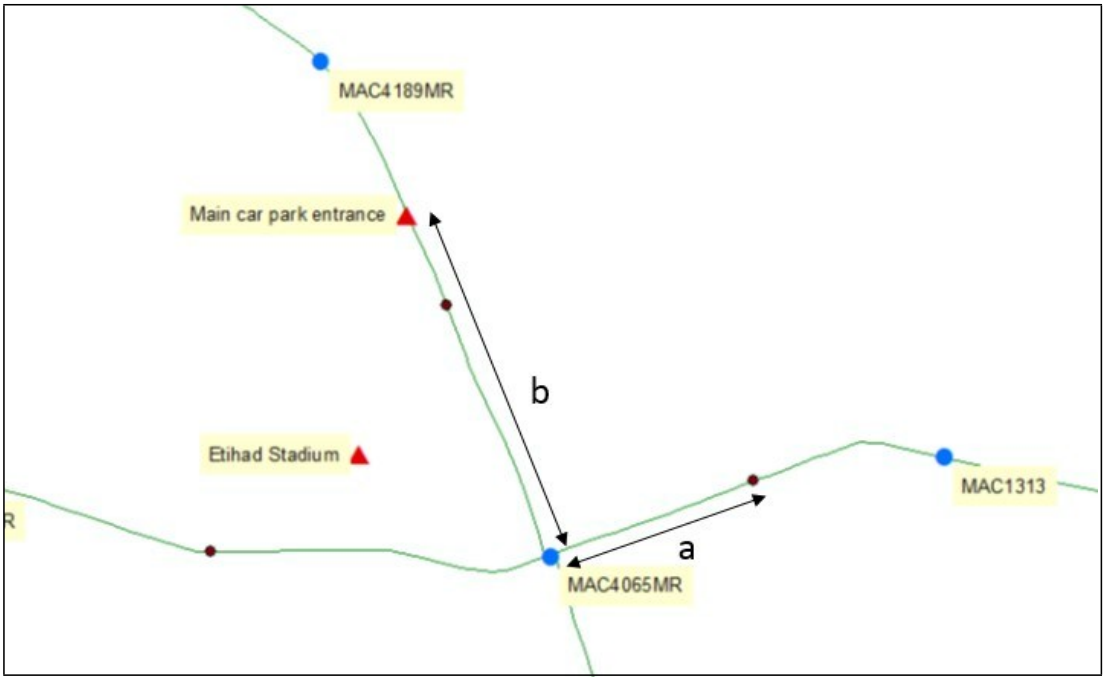


**Figure 5 Measuring the distance (a + b) of a link from the stadium (small dots are mid-points of links, large dots are sensor locations)**

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.

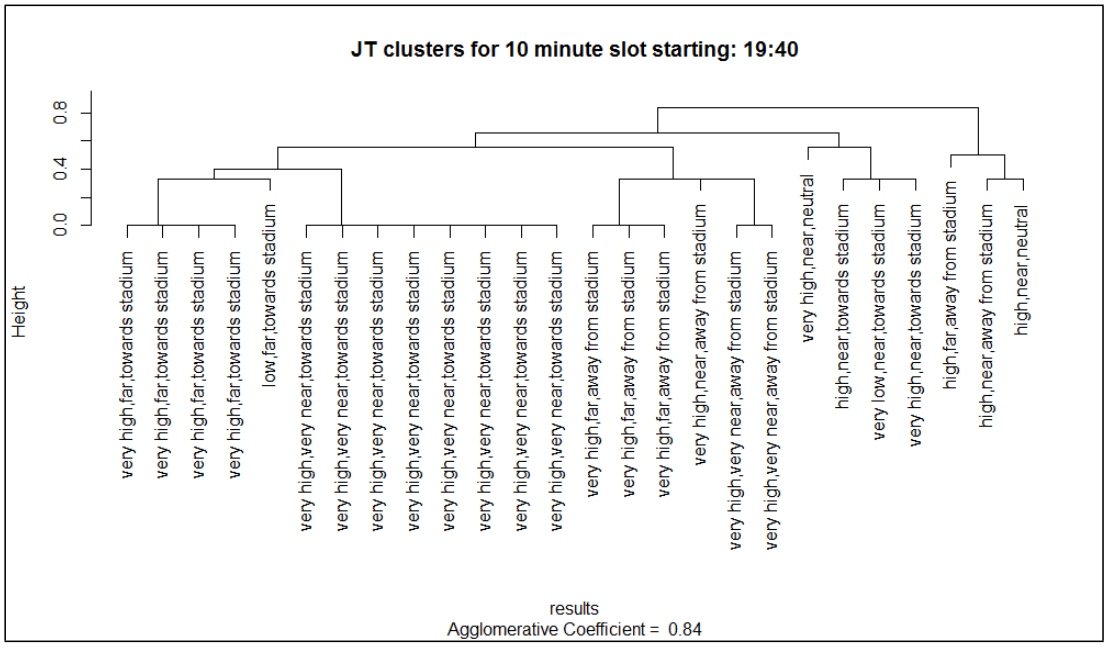


**Figure 6 Vehicle journey time clusters prior to kick-off on 13th January 2016**

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.

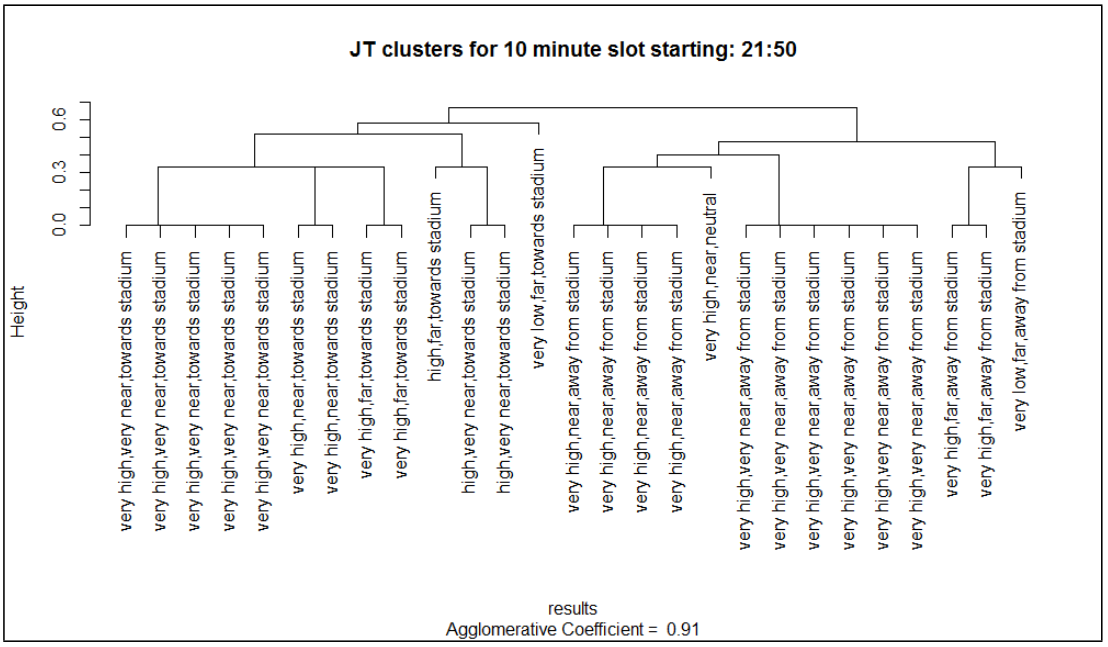


**Figure 7 Vehicle journey time clusters after full-time on 13th January 2016**

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).

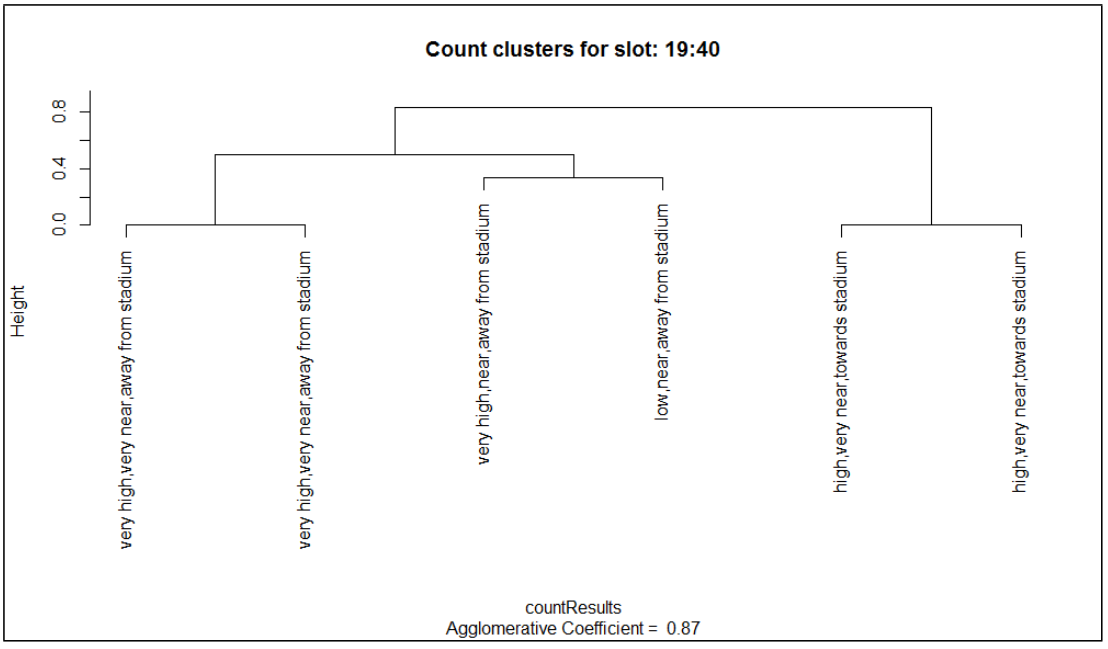


**Figure 8 Vehicle count clusters prior to kick-off on 13th January 2016**

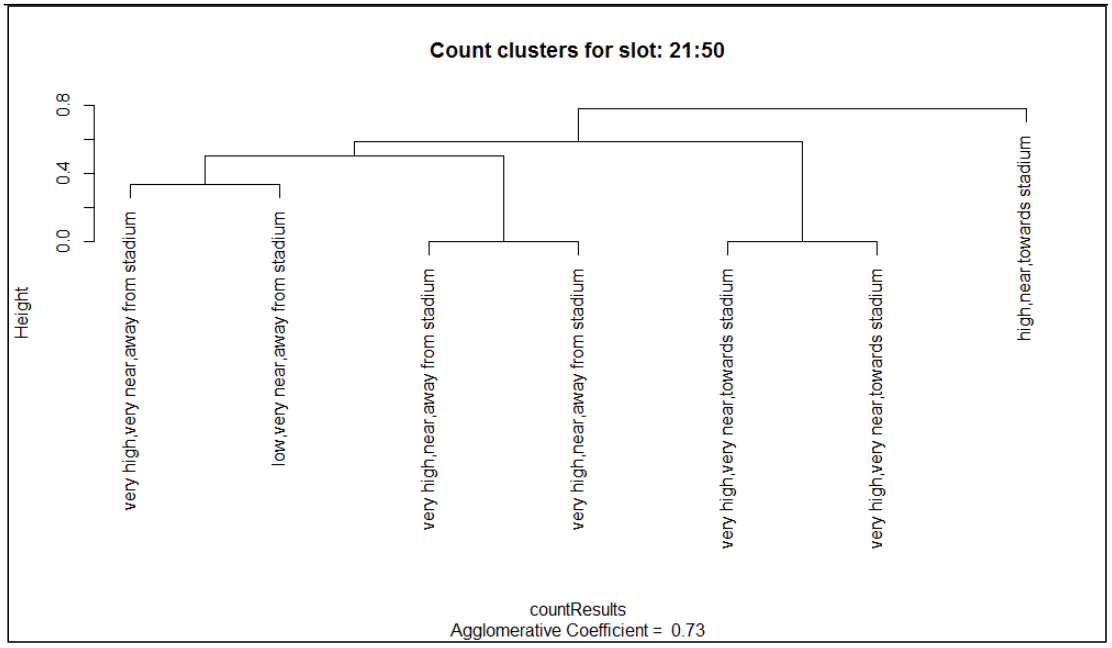


**Figure 9 Vehicle count clusters after full-time on 13th January 2016**

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.

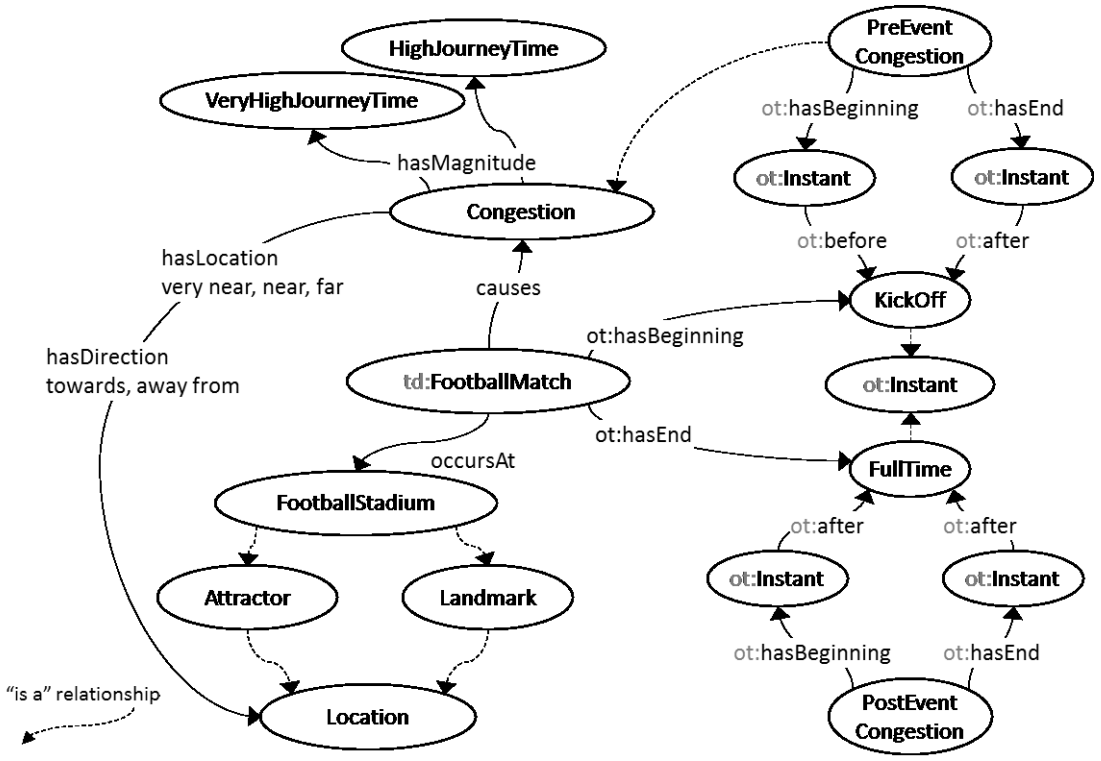


**Figure 10 An ontology of the impact of a football match on road congestion**

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.

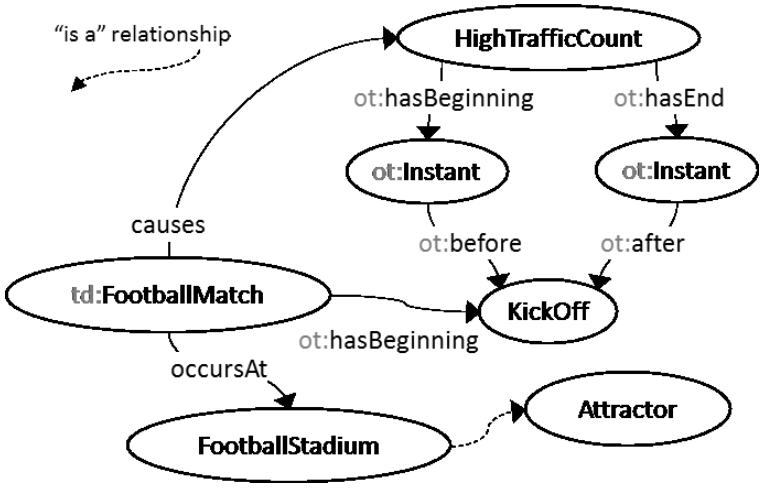


**Figure 11 The relationship between traffic counts and an attractor**

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.