Traffic Prediction and Analysis using a Big Data and Visualisation Approach

Declan McHugh*[1]

[1]Department of Computer Science, Institute of Technology Blanchardstown

February 4, 2015

Summary

This abstract illustrates an approach of using big data, visualization and data mining techniques used to predict and analyse traffic. The objective is to understand Traffic patterns in Dublin City. The prediction model was used as an estimator to identify unusual traffic patterns. The generic model was designed using data mining techniques, multivariate regression algorithms, ARIMA and visually correlated with real-time traffic tweets. Using the prediction model and tweet event detection. The result is a high-performance web application containing over 500,000,000,000 traffic observations that produce analytical dashboard providing traffic prediction and analysis.

KEYWORDS: Big Data, Data Mining, Visualisation, Traffic Analysis, Twitter Analysis.

*declan.mchugh@gmail.com

1 Introduction

The aim of this paper is to analyze traffic patterns for an urban city, Dublin and to provide a visual dashboard for analyzing traffic patterns. The data sets used vary from remotely sensed data and social media information from open data sources.

One of the challenges of this work is the Big Data (Four V's), Sheth (2014). Using data mining techniques for resolving issues around data quality and performing complex aggregation tasks in the form of Map Reduce enabled high-performance computation executions.

When visualizing the correlation between traffic-related tweets and adverse Traffic conditions, it is necessary obtain a prediction model. The prediction model provides an expected travel time. The actual travel time compared to the predicted travel time as a key performance indicator (KPI).

Variables from weather data, moving average and spatially related data, multivariate regression and statistical models is implemented. For each monitored traffic segment along with datasets to perform statistical and visual analysis such as the impact of weather conditions and spatial patterns.

The models were used to perform analysis on the volatility, the effect weather and prediction on travel times. Using visualization, the interpretation of the analysis is made simple using an analytic dashboard. The features of the dashboard allow the user compare inbound and outbound travel time, weather analysis, volatility analysis, twitter analysis for any hour of any day.

Using a classification model on Twitter data, traffic-related tweets were plotted on Google-Maps to identify Traffic events. The events can be visually correlated against the spurious Traffic events from the traffic prediction models. Tweets classified as traffic-related provided insights into the predictions the varied away from actual travel times.

The following sections provide more detail on the data sources, visualizations and conclusions.

2 Data Sources

Open data sources DubLinked (2014), Wunderground (2014) and Twitter (2014) were used in this work for the visualizations. The following section is a summary of the data used in the data collection and creation of the visualizations.

2.1 Traffic Data Sets

The prediction model was generated from data accumulated from open source portal known as DubLinked.

Figure 1: Dublin Traffic Data

2.2 Weather Data Sets

The Wunderground API is used to access three of the available historical weather data points 2.

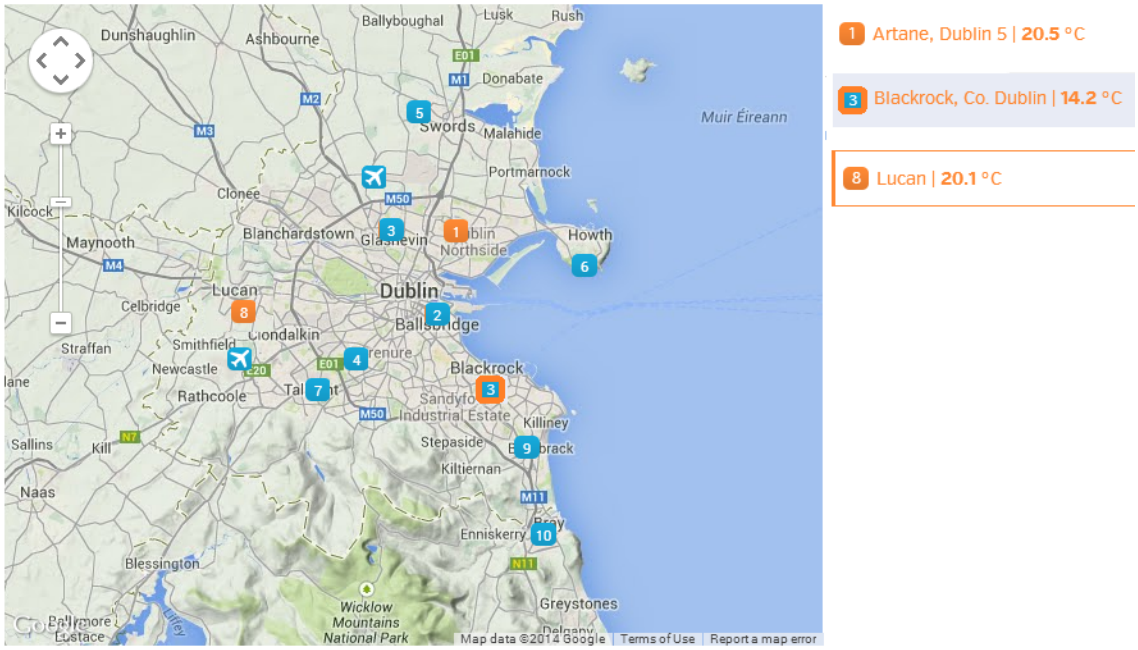


Figure 2: Open Data Weather Stations Dublin

2.3 Twitter Data Sets

Twitter provides an API for searching historical and real-time data. The historical data is used to as training for the classification model. Using traffic-related tweets from Twitter account, **#aaroadwatch** a training set is formed. The non-traffic tweets from the real-time data that contain geospatial data were used to make it possible to plot a tweet onto the map.

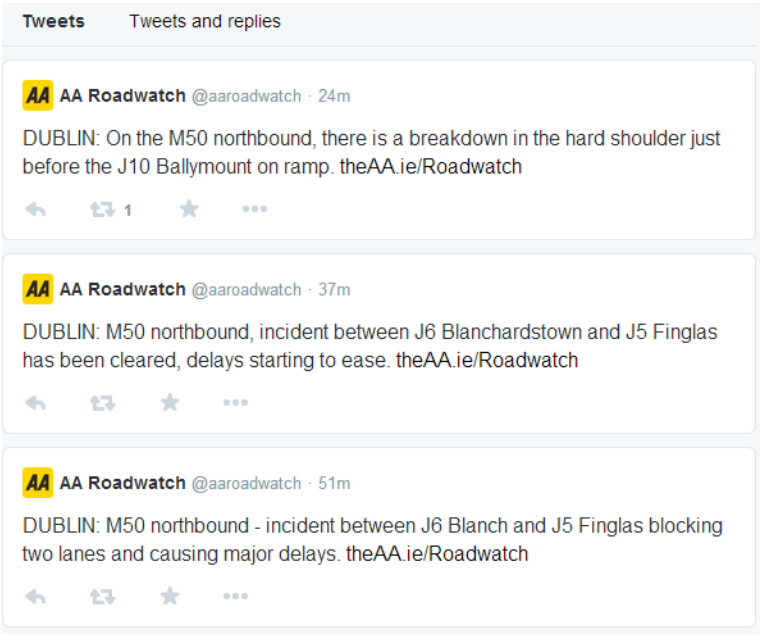


Figure 3: AA Roadwatch Tweets

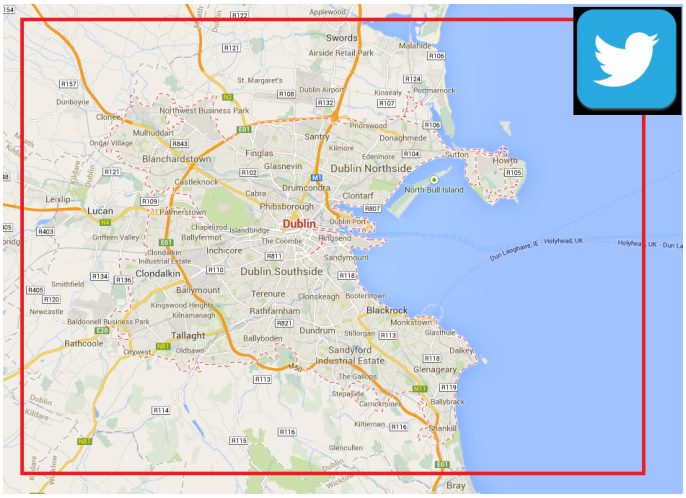


Figure 4: Monitering of Twitter Stream

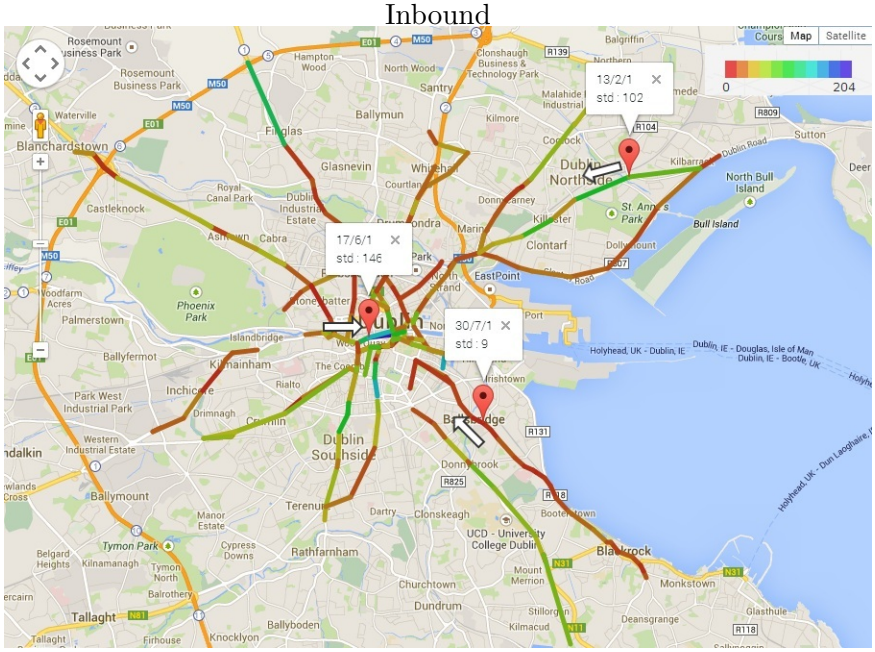
3 Visualizations

The analytics dashboard enables the users to identify the patterns of traffic visually as mentioned in the introduction 1. The visualizations are generated using Google Maps, JQuery along with a Python backend.

3.1 Volatility

Volatility is a way of identifying an inconsistency in the travel time. With this users can identify areas that are prone to delays. Standard deviation can be considered a way of measuring the volatility according to a paper from Tulloch (2012). A range of colors provides the result of the standard deviation from low red equal to 0 and high purple equal 200+ see Figure 5

Inbound



Outbound

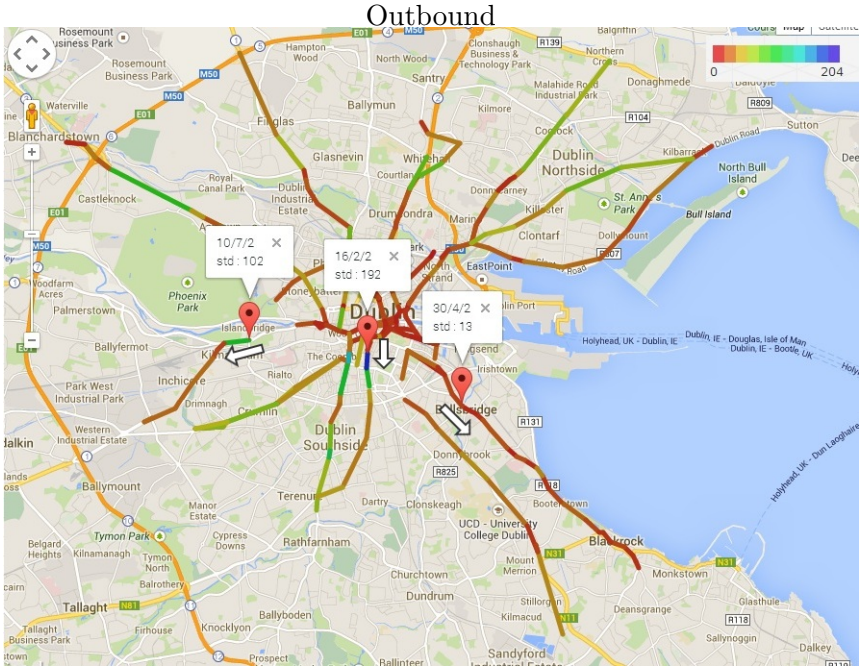


Figure 5: Inbound and Outbound Traffic Observations

3.2 Weather

The three weather stations selected for correlation analyses was performed (see section 2). The visualization demonstrates there is a spatial relationship between the observing weather station. The triangles in Figure 6 represent weather stations. The size circles represent correlation values from -1 to 1, for example, zero correlation is not visible on the map. The circle uses transparency as the negative correlation indicator.

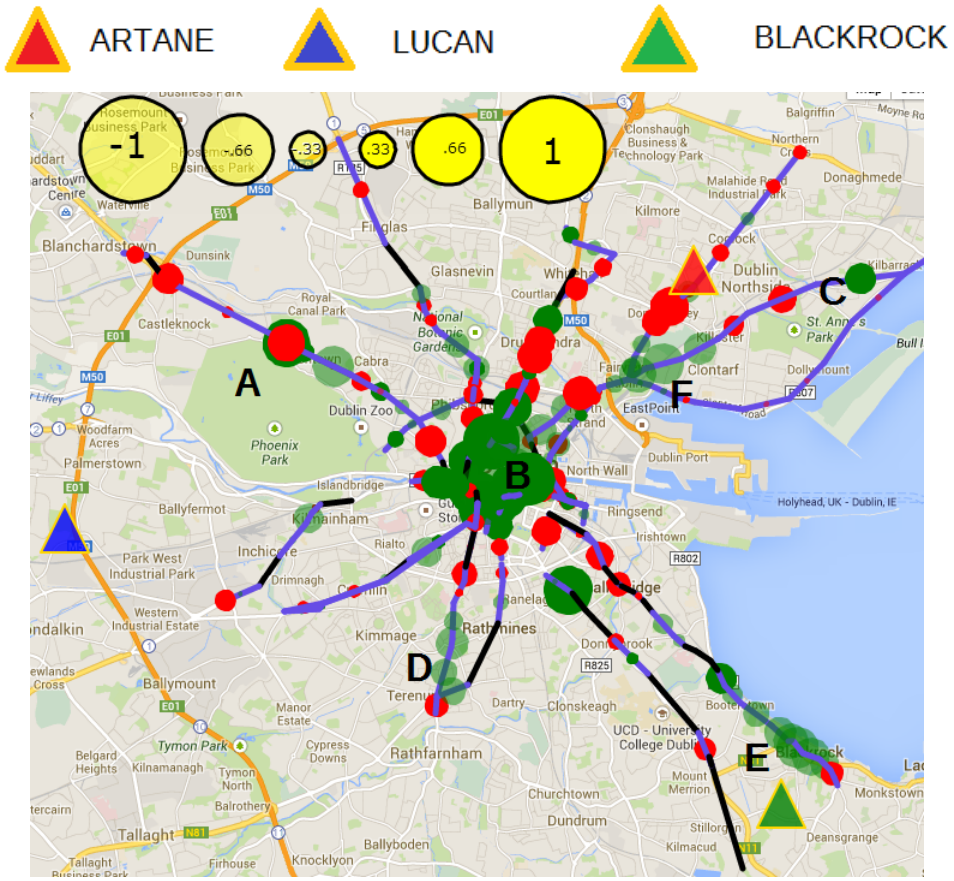


Figure 6: Correlation of Temperature Map Direction Inbound Peak Times

3.3 Prediction Model

During the exploration process, the highest correlated weather stations and spatial neighbour were used in the generic estimation data model. A generic data model provides the ability to reuse the process of building the data sets that would best fit the majority of the 512 observed locations while still keeping features that improve accuracy. As a result, the prediction algorithms behaved differently depending on the influence of features.

The best-performing algorithms for the least volatile road segments mentioned **??** are linear regression. Some road segments had little or no volatility. Other linear regressions, performed well that had volatility used a normalization of feature to improve accuracy.

Road segments with highly volatility with features of insignificant correlation resulted in a non-linear Support Vector Machine with Fourier transform with the highest accuracy.

Bayesian Ridge linear regression algorithm performed very well for the prediction. It demonstrates that when noise accounts for the more linear the data becomes.

In figures, **??** and **??** shows that the area of Finglas and Glasnevin is the least affected by weather and is highly volatile. Where the city centre and Clontarf are volatile and highly affected by weather conditions becomes a linear problem, see figure 7.

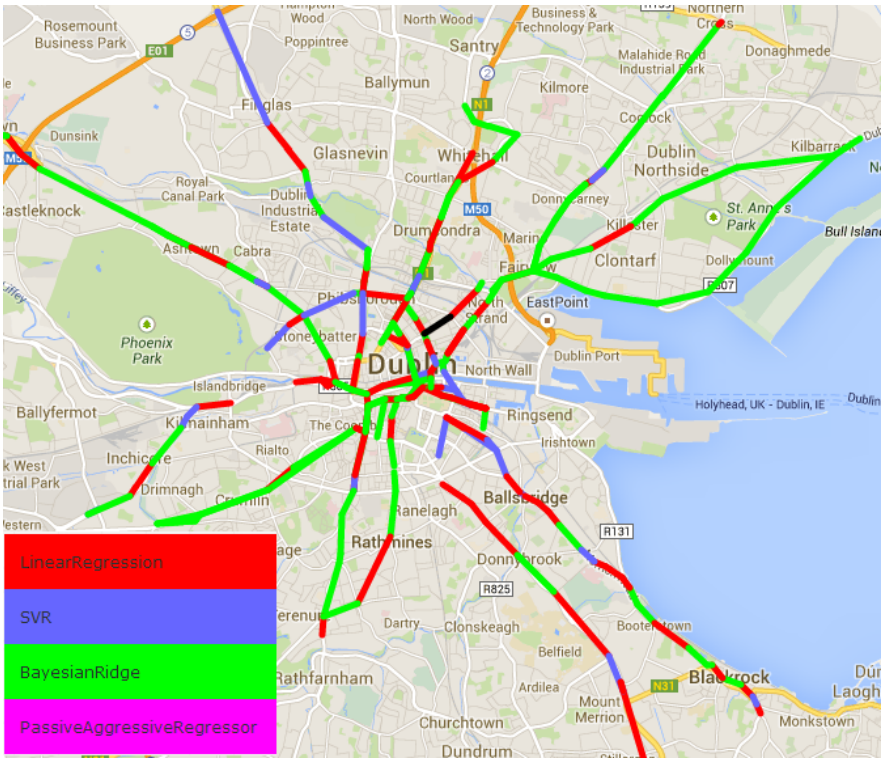


Figure 7: Off-peak and Inbound Algorithm Map

3.4 Twitter

The objective of twitter section is to analyse the approach of using the traffic domain tweets to extract tweets from real-time data that is related to the traffic domain. The tweets from the traffic domain do not contain geospatial information compared to the real-time tweet. Passive Aggressive Classifier is an example of one of the algorithms used to classify the real-time traffic tweets. Using

Tf-Idf scoring and Vectorizer algorithm. tweets are tokenised to fit the predictive model for the classifier. A sample of results fit on Table 1.

Table 1: Classified Tweets

Result	**Text**
True Positive	No better way to start your day with a car crash, and then forgetting about the banana in my pocket going through security...
False Positive	We're gonna crash vine if we keep doing this
False Positive	Lyndsay Lohan looks like a car crash.... She is wrote off #ChattyMan
True Positive	I bloody hate waiting #delays http://t.co/Yh55PrfQK3
True Positive	There's after been a crash outside my estate, 3 fire trucks and 3 ambulances

4 Results

The classification approach worked as a proof of concept. The real-time traffic tweet could be used to provide further analysis on traffic delays. In Figure 8 the dashboard demonstrates traffic related tweets as blue markers overlayed above the road segments and its estimated result. The red lines indicate delays, the green indicate better than expected while the grey is as expected. Each tweet marker is click-able to provide more informative details on the traffic conditions. Using the buttons on the left of the dashboard will display the different elements of the visualisations show in this abstract.

Using NoSQL to overcome the challenges of the four V's the approach stored volumes of data that on a single machine RDMS system would have been problematic 2.

Table 2: Data Volume

Data Source	*Items*	*No. of Documents*
Traffic Observations	501,402,840	8,356,714
Real-time Tweets	3,048,310	116
User Tweets	5,267	5,267
Weather Records	229,311	2,103

Issues such as in figure 8 the dashboard contains a some false positives, example *"Lyndey Lohan looks like a car crash.. she is wrote off #ChattyMan"* while true positive *"We hope our new display doesn't cause too many delays in Donnybrook .... http://t.co/sDKrey1pJf"* can be resolved in future work.

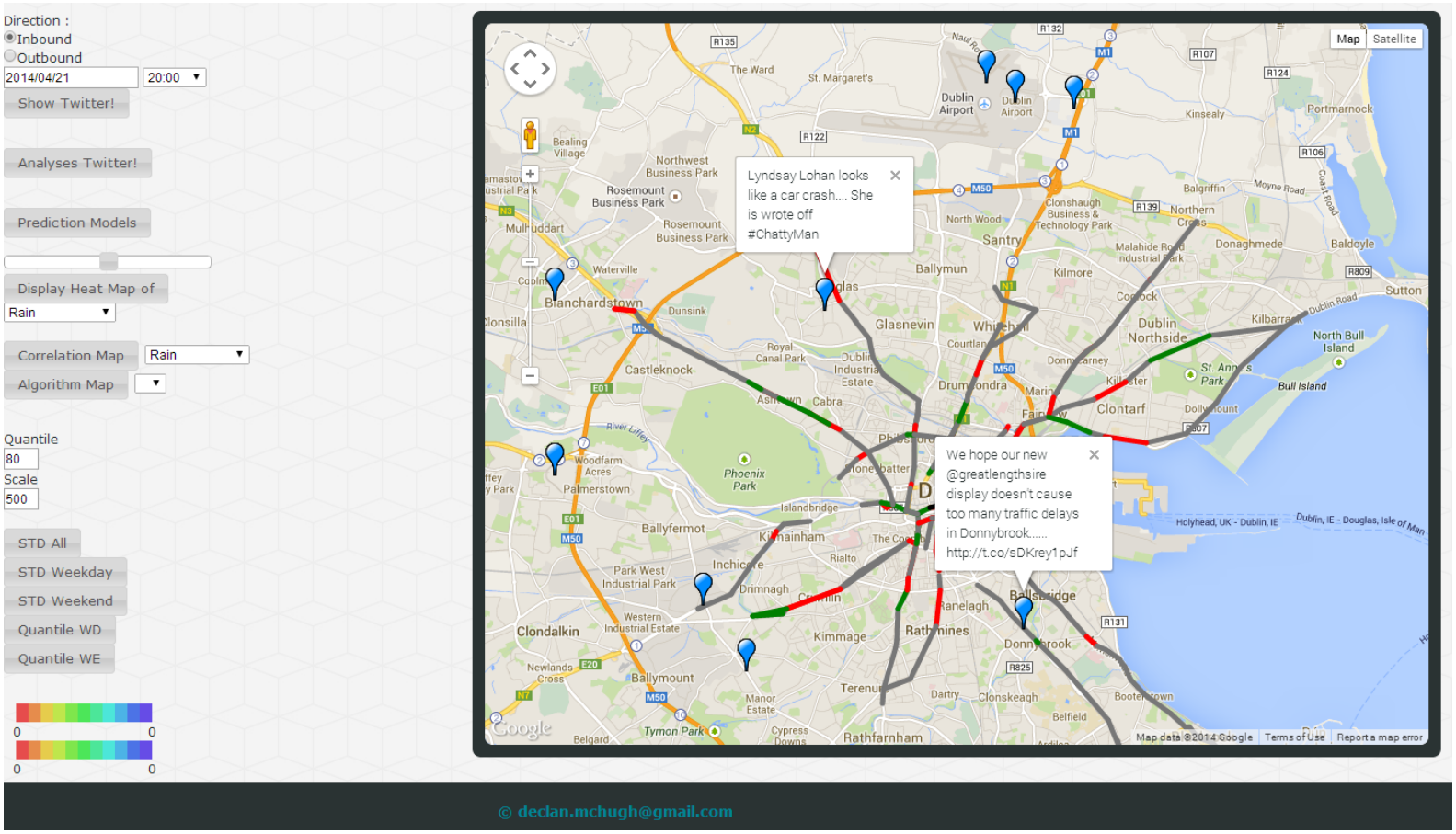


Figure 8: Dashboard Analysis for 21/04/2014 8pm to 9pm

References

DubLinked (2012-2014). Trips data.

Sheth, A. (2014). Transforming big data into smart data: Deriving value via harnessing volume, variety, and velocity using semantic techniques and technologies.

Tulloch, D. J. (2012). A garch analysis of the determinants of increased volatility of returns in the european energy utilities sector since liberalisation. *IEEE*.

Twitter (2014). Twitter.

Wunderground (2012 - 2014). Wunderground.

5 Acknowledgements

None of this work could have been achieved without data from Twitter (2014), DubLinked (2014) and Wunderground (2014)

6 Biography

Declan McHugh is a 12-year veteran of writing software. He has a broad exposure to developing in many technologies and is a keen advocate in all things analytical. Declan has recently obtained a 1st class honours Masters the Analytics is actively working projects with Big Data Analytics and Visualization.