UNIVERSITY OF CAMBRIDGE
Department of Engineering

CONSTRUCTION ENGINEERING MASTERS DISSERTATION ABSTRACT

Vegetables and vehicles, the value of secondary use data. A cross-sectional correlation case study approach

The aim of this research is to understand if increased highways safety could be developed using statistical and machine learning approaches applied to secondary analysis of crowd sourced data. The results of which will be correlated to another disparate open data source. What is clear from the data is that accident rates are constantly changing over time, which requires interpretation to fully understand and this has provided the impetus for this study.

Across the world there were approximately 1.3 million fatalities on the highways in 2013. This number has been steadily rising as more countries enter stages of economic growth which allows more of the population to become vehicle owners. In the United Kingdom the number of fatalities on the highways has fallen to 1713 in 2013, although 1713 remains too high a number.

A wide range of industries have moved towards the collection and use of Big Data for the purposes of increased productivity, efficiency and safety, yet the construction and engineering sectors continue to lag. In the more technology savvy companies such as Facebook, Amazon or Tesco there has been a significant growth in crowd sourced data collection and use. However, crowd sourced data is yet to find real traction in the construction and engineering sectors around the globe (Bilal *et al.*, 2016).

This lack of penetration associated with crowd sourced data and its use in the construction sector provides an interesting hypothesis to test, which is: Statistical analysis and machine learning has utility to predict potential accident locations through secondary analysis of crowd sourced data.

This hypothesis may be helped by looking at several research questions which aim to; provide a methodology for secondary analysis of crowd sourced data for improved highway safety, unlocking of additional value which may be discovered in the wide-ranging data sets already being collected by the industry and how this value could be unlocked through the application of data mining techniques.

This quantitative research will look to use a cross-sectional correlation case study, specifically, the research will analyse crowd sourced vehicle telemetry black box data in the county of East Sussex and open sourced accident data. The data will be analysed using statistical techniques to understand locations of hard braking events, before machine learning approaches are applied to predict new braking event locations. When overlaid with open sourced accident data locations, this may show correlation of locations that can be recognised as being part of a network event. In doing so, any correlation could provide locations of predicted heavy braking and accident locations, so these events where an accident has yet to happen, could be investigated as a theoretical near miss or provide risk assessment criteria for sections of highway.

This research illustrates it has been possible to justify a correlation between the two disparate data sets for the purposes of improving highways safety. The use of statistical

UNIVERSITY OF CAMBRIDGE
Department of Engineering

and machine learning approaches on vehicle telemetry black box data to predict braking event locations, overlaid with the open sourced data for accident locations may provide an early indication of potential future accident locations or increased risk on sections of highway. Heinrich's triangle theory has led to trending of unsafe acts, which lead to near misses, on to prevention of more serious accidents and ultimately fatalities, the same principles may apply to predicted braking locations when tied to a recognisable highway network.

This study identifies further areas for research using secondary analysis of crowd sourced data such as; prediction of accident locations and correlation to departure from standards, use of variable speed limits in location of predicted braking events, highways defect identification, potential locations of poor air quality, and driver behavioural issues that would warrant further research to gain better understanding around the use of crowd sourced data for these purposes.

Dan Rennison

April 2019