USING DATA MINING TO PREDICT ROAD CRASH COUNT WITH A FOCUS ON SKID RESISTANCE VALUES

Daniel Emerson, Computer Science Discipline, Queensland University of Technology, Brisbane, Queensland, Australia.

Richi Nayak, Computer Science Discipline, Queensland University of Technology , Brisbane, Australia.

Justin Weligamage, Department of Transport and Main Roads Queensland , Australia.

ABSTRACT

Road crashes cost world and Australian society a significant proportion of GDP, affecting productivity and causing significant suffering for communities and individuals. This paper presents a case study that generates data mining models that contribute to understanding of road crashes by allowing examination of the role of skid resistance (F60) and other road attributes in road crashes. Predictive data mining algorithms, primarily regression trees, were used to produce road segment crash count models from the road and traffic attributes of crash scenarios. The rules derived from the regression trees provide evidence of the significance of road attributes in contributing to crash, with a focus on the evaluation of skid resistance.

---

1. INTRODUCTION

Road safety is a major concern worldwide with road crashes costing countries between one and three percent of annual Gross Domestic Product. WHO predicts road traffic crashes emerging as the 3rd leading cause of disease or injury burden in 2020 [1]. The annual economic cost of road crashes in Australia is conservatively estimated at $18 billion per annum, and the social impacts are devastating [2]. While the cause of road crashes is a multi-facetted problem including factors such as driver demography [3], vehicle speed and characteristics, weather, traffic and so on, an argument by Shankar, Mannering & Barfield in 1995 [4] proposed that road characteristics, i.e. poor road geometry & poor signage, make roads more crash prone. They argued that positive road characteristics can assist in reducing crashes. World-wide, road authorities reflect this assumption by following strict design codes described in their road design manuals and applying known engineering principles within the contexts of safety, cost, driver expectation, and economic and environmental parameters [5]. Cairney [6] stated that skid resistance (F60) has the best established relationship with crash cause. Inspired by these studies, this paper focuses on the relationships among road crash characteristics, skid resistance and other road attributes.

This paper aims to contribute to the knowledge of management of skid resistance by using data mining models [7] to identify crash risk associated with road attributes, with a focus on skid resistance. The study ignores vehicle, driver and crash attributes, and seeks to find risk relationships in road-related attributes only. This paper utilises the regression tree [7] to successfully model *road segment crash count* using the road characteristics of surface measures, surface type and age, road geometry and traffic. The resulting models, based on historic data, can be deployed to predict the *road segment crash count* in unknown road segments from the network and will be used in decision support systems for the identification of road segments where intervention resulting in reduction in skid resistance will have a high probability of reducing crashes.

The dataset, spanning between the years 2004 and 2007 for all Queensland main roads, was provided by the Queensland Department of Transport and Main Roads (QDTMR). The attributes

provided the necessary road and traffic information, with results from data mining showing a high level of model accuracy for the various models. Successful instance classification rates were between 70% to 93% of the road segments, depending on the configuration of the model.

The paper traces the development and outcomes of the study, with section 2 focusing on background and related work, section 3 on data pre-processing and model development, section 4 on providing the data mining methodology and section 5 on discussing the results. The conclusion provides a summary of the outcome from the models and proposes future directions.

2. BACKGROUND AND RELATED WORK

Data mining methods are increasingly in demand in industrial situations [8] and are being applied in many innovative ways in road crash studies [9-11]. This study contributes to meeting this demand by producing an analytical process required for developing a road asset manager's skid resistance decision support system. In this section we discuss the studies that led to the evolution of the proposed road crash study, and discuss general road crash studies of domain interest, and how the models relate to this long-term goal.

The formative influences of our work came from the following studies. The statistical study of Shankar, Milton & Mannering [12] developed the concept of road segment crash proneness, and suggested that road characteristics contributed to the crash risk. Our preliminary data mining modelling supported that concept [23] and provided an important building block on our journey to modelling crash count. Statistical methods were the main method of analysis until the late 1990s; however Chang and Chen [13] establish the that data mining, specifically decision trees, is suitable for the study of traffic crashes and road variables in a comparative study between Poisson regression and data mining methods. Wong & Chung [14] utilised data mining models for modelling crash counts; using the same method as ours for measuring crash rates. Anderson [15] used clustering in a road crash study to identify crash hotspots.

The paper utilizes road characteristics as input variables and the outputs from the models help in understanding their role in road crashes. Many studies including Qin & Ivan [16], Haynes et al. [17] and McCartt [18] demonstrate that road factors influence crashes, with factors including speed limit, traffic rates, time of day, volume/capacity ratio, percent of passing lanes, shoulder width, number of intersections and driveways. These characteristics coincide with many of the variables in our dataset. Our earlier findings on skid resistance correlate with the general consensus of findings for example, a general reduction of crashes is found with an increase in skid resistance [19] and wet pavement is more dangerous at higher skid resistance values [20]. In addition Mayora et al. [20] quote Yerpez and Ferrández [21] in proposing high crash rates occur where the friction demand is high. This is a concept of which we will be mindful, and will be seeking evidence for later analysis of our results.

Predicting crash count was a novel task for data mining and the models presented in this paper are the culmination of progressive development and improvement through a number of stages. The first models explored the prediction of a road segment having crash [22]. In this study, road segments with crash and those without crash were assembled and decision trees were found capable of identifying over 85% of road instances successfully. A comparative wet and dry road surface study was conducted using this method [23]. These earlier studies concluded that:

- A strong inverse relationship exists between skid resistance and crash rates showing reducing crash probability with increasing skid resistance.
- The relationship between crashes and traffic (AADT) is not linear, with the highest crash rate hotspot found at a skid resistance (F60) of 0.2 and a traffic rate of 20,000 vehicles per day.
- On wet surface road segments, the crash rates increased in regions with a combination of higher skid resistance (F60), and higher traffic rate (AADT).

We also explored the grouping of characteristics of roads that have crashes above certain crash rate thresholds with crash proneness data mining models [24] and provided evidence that:

- Many groups of roads, particularly those with low crashes per year have more in common with non-crash roads than roads with crashes.

- Road segments, grouped into clusters based on their common attributes using data mining clustering methods, have a disposition towards a certain crash range e.g. low range , mid-range or high range.

Inspired by the strong relationship between crashes and road / traffic attributes in the clustering model, we sought a model that could provide a higher resolution view of crash risk in this paper. Road segment *crash count* was chosen as the target, and for a DM algorithm the regression tree, with its ability to predict numerical outcomes was selected to develop the models.

3. DATASET CHARACTERISTICS

The data mining dataset was crash-centred where crash instances were populated with the road data, with 42,388 crashes reported in the period between the beginning of January 2004 and the end of December 2007. The road data was obtained from the end-of-financial year 1 km road segment snapshots.

While skid resistance was a major focus of the paper, the paucity of survey records created a challenge by limiting the crash training set. Skid resistance surveys were available for only about 25% of the road segments across the network and only a few segments were measured more than once. Characteristically, each road segment had 10 readings, sometimes fewer, or more with up to 20 readings for dual lane roads. The readings were processed to provide a single 1 km value. The processes and the resulting dataset is outlined below

The 100 metre skid readings (F60) for each road segment were aggregated into a single 1 kilometric average identified by the road section, carriageway and starting distance. The standard deviation of the readings was calculated and included. The crash readings did not have lane readings data and therefore lane information was not included in F60 calculations. Of the 34 thousand one kilometre road segments, skid resistance values were available for 9866 segments. Of these road segments, 4,362 had crashes and when joined with the crash instances yielded 16,750 crashes, making around one third of the crashes available for analysis as training data. Crash records were included if the crash incident happened after a known skid resistance survey, or prior to a recent survey provided the seal intervention event was known. Time in months between the survey and crash was noted. Testing, to determine which F60 measure type was most successful for modelling, found that 1km averages were slightly favoured above the 100metre averages. While models for both types were maintained, results from models with the 1km F60 values are reported in this paper.

The following road, traffic and crash attributes were included with each crash record.

- ***Roadway Surface:*** Average Friction_at_60 (or F60) over1 km road segments, Average Friction_at_60 (100metre road segments), Texture Depth, Seal age, Seal type.
- ***Road Surface Wear:*** rutting average
- ***Road Surface Damage:*** Roughness average
- ***Roadway Geometry:*** Horizontal Alignment (curved-open view, straight etc), Vertical Alignment (level, grade, dip, crest), General Terrain (level, rolling, mountainous)
- ***Roadway Design:*** Crash Speed Limit, Road Speed Limit, Divided Road, Road Type (highway, urban arterial etc) , Carriageway Type (single or dual), Lane Count
- ***Roadway Features:*** Roadway Features (roundabouts, bridges, intersections etc)
- ***Roadway Furniture:*** Traffic Control (lights, signs etc)
- ***Traffic:*** AADT (Annualized Average Daily Traffic), traffic percent heavy
- ***Road Surface Wetness:*** Wet Road Surface, was Raining or not
- ***Crash Type:*** Crash Nature, Crash Description (head on, rear end etc.)
- ***Crash Severity:*** Crash severity index (fatal (1) through to property damage only(5))
- ***Crash Characteristics:*** Crash Time, Crash Day of the Week, Atmospheric, Lighting.

The most significant limitation in the data is the absence the friction demand categories [21] and gaps in the data relating to the classification of the demand categories. Curve radii, which are an integral component of the road segment *"investigatory levels"* and used in the process to define high demand road segments and the skid resistance were not available. However many significant characteristics were available, such as roadway features, traffic control, traffic rates and speed limits and were included in the data.

Data pre-processing methods were trialled extensively; however generally variables performed the best as provided and their original values were retained for mining. A small percentage of attributes had missing values; however this was not of major concern for this study that uses tree algorithms for modelling the data. Trees are known to be not sensitive to missing values [7]. In this study missing values were allowed to remain in their valid null state.

A stated above, crash risk was measured by 1 km road segment *crash counts* for annual and four year periods. The maximum road segment crash count was 32 crashes per km per year, and 100 crashes per km for the four year period. *Figure 1* shows the distribution of crash Instances by crash count for road segments, illustrating the similarity between years and an exponential drop in the number of instances of crashes as the crash count increases with only a few roads with many crashes.

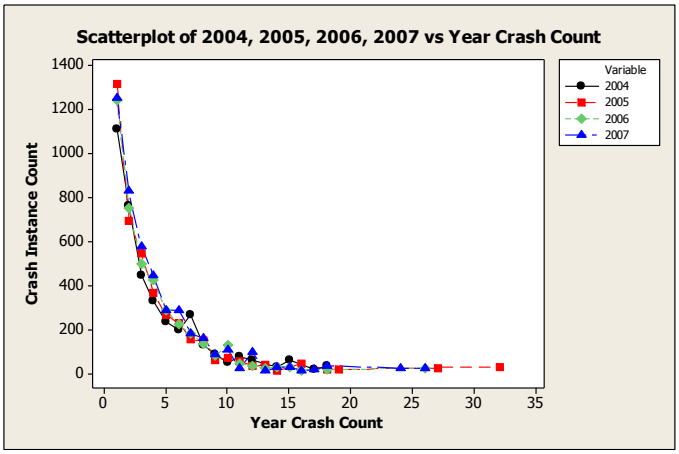

Figure 1: Distribution of Annual Crash Counts for years 2004 -2007.

Figure 2 shows the variation of the range of annual crash count values for each four year crash count value, providing further evidence supporting our inference that road segments maintained a characteristic crash count range from year to year. It shows that roads with high crash counts maintained a high level and those with low counts remained within their characteristic low range over time.

The objective of this study is to identify the attributes and their values that are critical in moving road segments from a character of high crash to low crash, and use this information in road asset management decision support.

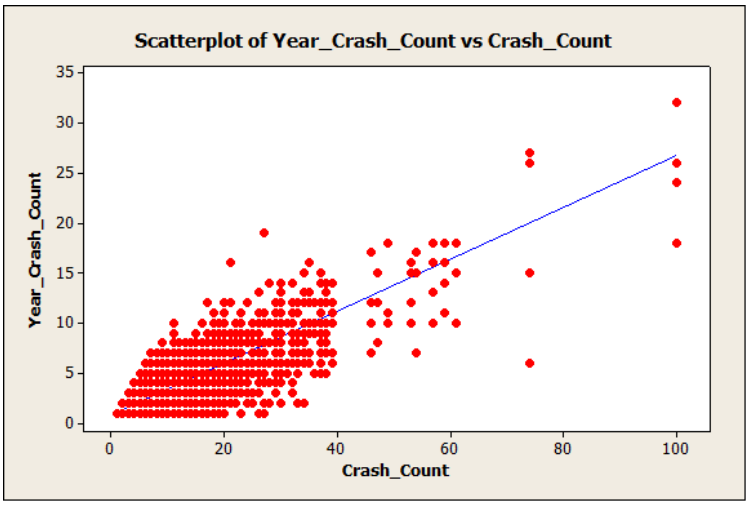

Figure 2: Distribution of annual crash counts for each four year crash count.

Crash rate is frequently calculated and discussed in road asset management circles with a traffic component, e.g. crashes per million vehicle kilometres travelled (MVKT). Initially, our study deployed the crashes per million vehicle kilometres travelled measure calculated by the method shown, but it was found to be unsatisfactory.

$$Crashes/MVKT = Crashes\ per\ year * 10^6 / road\ segment\ length\ (km) * AADT * days/\ year$$

Thus with reluctance our study selected the absolute road segment crash counts as crash-risk measure, and found it to be used in other studies [25,16]. The reason for the change from the commonly accepted unit was that models using *crashes/MVKT* yielded poor results in comparison to absolute road segment crash counts.

The MVKT unit makes the assumption that crash rate has a linear relationship with crashes; i.e. as traffic increases on a road segment the crashes will also increase. This assumption belies the road design principle that a road segment is designed to carry a given volume of traffic at a designated speed, and crash rates can be expected to increase once either parameter is exceeded. The empirical evidence does not support the assumption of a linear relationship between skid resistance and traffic rates as shown by the distribution of crashes with respect to AADT, illustrated on a chart developed from QDTMR data (*Figure 4* below). The chart shows peak road segment crash values at AADT values of 20k vehicles per day, while road segments with higher traffic rates of 40k and 60k have quite low annual crash counts.

A distinction needs to be made between two measures: the traffic vs. crash on a single or homogeneous set of roads where crash rate may increases with traffic volumes, compared to the traffic averages (AADT) measured against road segment crashes for all road segments over the network. Our data and discussion focused on the latter.

Further, for data mining, a far more serious issue exists for developing models with the MVKT measure: the loss of information. Aggregating crashes with different crash counts combines crashes which most likely have different causes. In the following examples, both road segment crash measures of 60 crashes p.a. at 60k vehicles/day and 15 crashes p.a. at 15k vehicles/day (and many values in between) produce a single rating of 2.73 crashes/MVKT, and in each case the actual crash count and AADT values which hold important information about the system are lost from analysis. Causes of crashes on road segments at 60k vehicles per day are different from crashes at 15k. Different dynamics are at work. Grouping the attribute sets under a single MKVT value severely degrades the predictive capability of the value of the model.

Our earlier data mining investigations demonstrated that with the QDTMR dataset, the non-aggregated crash values were superior for data mining analysis [23], with models using the MVKT model performing worse by a factor of between 2 and 10 times depending on the input variables involved. A current test model using *crash count per MKVT, using the standard configuration from the study,* returned a correlation coefficient of 0.64, compared to the *crash count* only model which returned 0.93. Consequently, in this paper, road segment crash count was maintained as the crash rate target variable and AADT was included as an input variable.

Profiling the road segment crash count variable against other road and crash attributes further supported the benefit of using road segment crash counts. The crash count relationships between road & traffic attributes supported domain knowledge findings, e.g. that crash risk increased with lowering of skid resistance, whereas MVKT models failed to show this relationship. The chart of *skid resistance* vs. *annual road segment crash count* (*Figure 3*) demonstrates a general increase in annual crash count with a drop of average skid resistance in the range of 1 to 7 crashes per road segment per year, accounting for a very high proportion of all crashes (*Figure 1*).

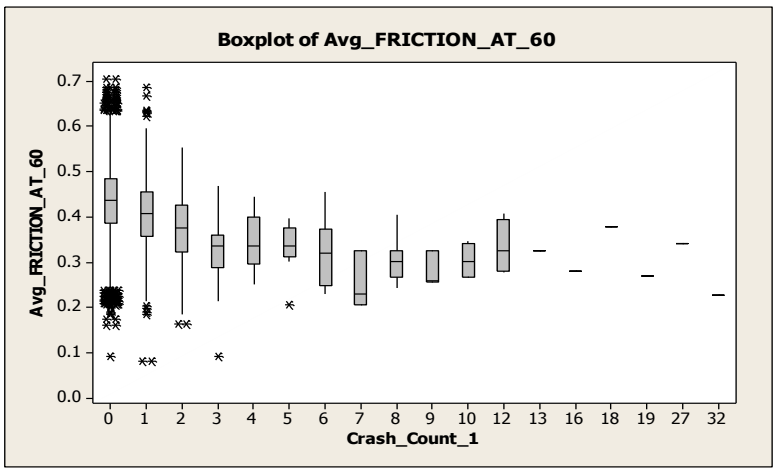

Figure 3: Distribution of average skid resistance with annual road segment crash count.

Characteristics of the data provided reasons for selecting data mining tree models for model development. Methods such as linear regression that rely on linear relationships between the road feature predictors and the crash rate response produce models of low accuracy. The distribution of road segment crash counts with respect to skid resistance (F60) and traffic rates (AADT) shown in *Figure 4* developed from the QDTMR dataset, demonstrates a non-linear pattern of crash count between these two most significant attributes.

Road segment crash counts, shown by the contour lines, peak in the F60 range of 0.2 and AADT band centred on 20,000 vehicles per day, whereas roads with much higher traffic rates have lower crash rates. This situation demonstrates the folly of proposing the use of linear modelling when using the heterogeneous road set. The ability of data mining tree algorithms to develop models in data exhibiting non-linear relationships and to select and report common patterns from data from different road types makes tree models an ideal application for this problem.

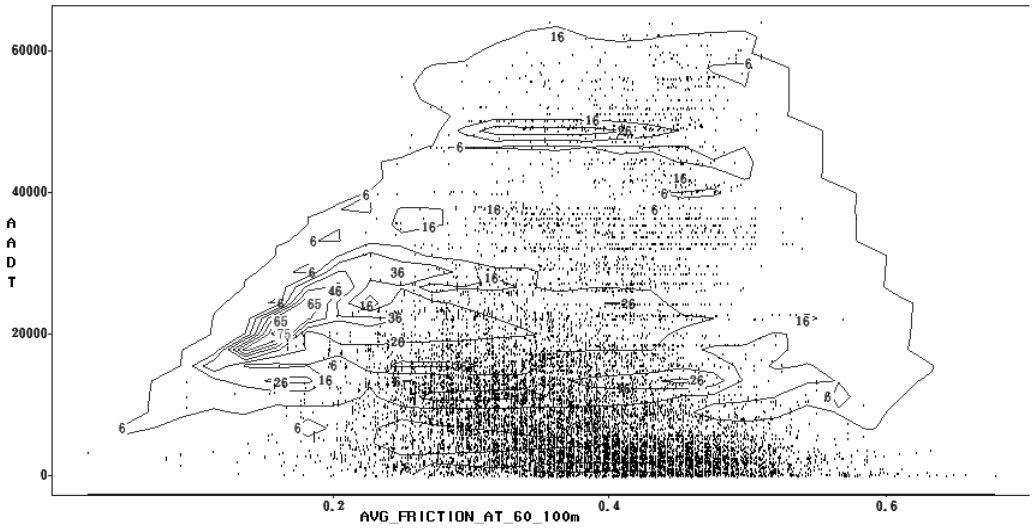

Figure 4: Distribution of road segment crash count by skid resistance and AADT for all crashes.

While the models in this study have focused on all crashes to date, investigation into serious crashes is a priority of road authorities and road associations. *Figure 5*, plotting serious crashes, i.e. those leading to hospitalization or fatality, shows that the distribution of serious crashes has a relatively more even coverage than the distribution of all crashes, with an absence of the very high density of road segments in low AADT areas and an absence of the pronounced ridge in the order of 50k vehicles per day present in the model depicting all crashes. Serious crash models will evolve from the models presented in this study.

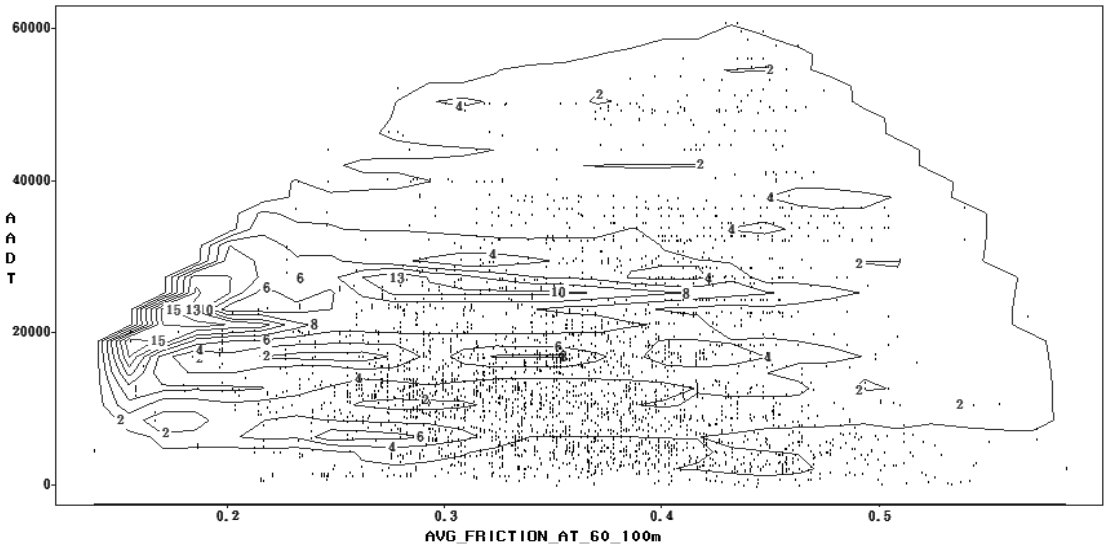

Figure 5: Distribution of road segment crash count by skid resistance and AADT for serious crash instances.

4. DATA MODELS: CONFIGURATION AND EVALUATION

The objective of the study was to develop a data mining model for the generation of a rules set that could accurately classify road segments by crash count. These rules would have a dual purpose: (1) to help understand the cause of high crash rates in some road segments; and (2) to be processed by an inference engine in a decision-support application that would carry out targeted road asset management tasks to reduce crash risk.

Predictive modelling required the fundamentals of a target, input variables and the evaluation method discussed above. Road segment crash count was selected as the target value. Road and traffic attributes provided the input data for the algorithm. A range of algorithms from the numeric prediction family was applied to develop benchmarks, with the models including the regression tree, support vector machines, linear regression and neural networks. The successful algorithm candidate was the regression tree. Selection was made on the usefulness of the output and quality of classification. Outcomes are discussed in the results in section 5.

The goal of this paper was to demonstrate that road segment crash count could be successfully modelled, and to produce models that would apply to the wider QDTMR dataset. The models were configured to optimise the rules so that they represent many instances while retaining the ability to meaningfully predict the target: in this case to produce rules with relevance to reasonable number of roads while making reasonable crash count predictions. These two configuration methods were utilized for controlling the models and rule quality;

- selection of the best attributes (feature selection)
- controlling the size of the tree by controlling parameters, and thus managing how specific the resulting rules will be and the number of instances that each rule describes.

Controlling the tree size is important to ensure that rules are general enough to apply to novel data sets, while maintaining the accuracy of the model. While preliminary models have been selected, our goal of testing the models with the whole dataset is beyond the scope of this stage and is for future development.

In datasets where the attributes of interest have a strong relationship to the target, models will produce compact trees with fewer rules and predictions will have a high levels of accuracy. Rules will apply well to large groups of instances. Where the attributes of interest do not relate well, the trees will be large and sometimes leaf count approaches the number of instances [7], and these rules perform poorly when applied in other datasets. Thus evaluation methods in this study involved an examination of rules, along with the model statistics. Rule assessment included:

- the measures of number of rules in comparison to the number of instances
- the proportion of instances that apply to a rule
- the proportion of instances that are correctly classified.

Model statistics included the c*orrelation coefficient (r-squared). R-squared* is the aggregated value measuring how near the target values were to their real values. Other *whole-of-model* indicators cited in our model assessments are *absolute mean error* and *root mean square error* giving indicators of the average of all predicted values and magnifying small changes in model accuracy. A summary of the nature and use of these measures is shown in Table 1.

Table 1: Evaluation measures used in regression tree models assessment.

Measure	**Statistic Definition**	**Performance**
Correlation Coefficient (R-squared)	A result of regression trees and interval targets, **subtracts the sum of observation variance squared value from the predicted value from 1.** Provides a valuable decimal result between 0 and 1 for the model and individual leaf nodes indicating the purity of the instance collection. 1-SS(err)/SS(total) (SS : sum of the squares)	R-squared is a whole of model nearness of classification measure and alternatively called the *coefficient of determination* effectively predicts how well the model will apply to like datasets. Higher the value better the model is.
Mean Absolute Error (MAE)	Measures the average of cumulative absolute value of the error(predicted value minus the true value) sum abs(predicted value - true value)/Total number of Instances	It is compared to the expected average value and useful for monitoring trends over a number of models, but exaggerates the effect of outliers [7]. Lower the value better the model is.
Root Mean Square Error (RMSE)	Root of the average error squared Sq root of (sum (predicted value - true value)^2/Total number of Instances)	RMSE is more sensitive to changes in models and will show changing trends between models that are more difficult to perceive in MAE. The lower the value better the model .

While comparing models, the model accuracy statistics are useful; however, an equally useful measure is comprehensibility to assess "model goodness" by comparing the number and quality of the rules, how well the rule set classifies the instances, and its suitability for the task at hand [7].

5.RESULTS AND DISCUSSION

Benchmarking of the candidate models was conducted through two stages. In the first stage, commonly used algorithms for numerical prediction were tested with the crash count dataset and default configurations. The results are shown in Table 2.

Table 2: Performance of commonly used numerical predictive algorithms with the crash model.

Model Type	**Implementation used**	**Comment**	**Correlation coefficient (R-squared)**	**Rule Count**
Regression Tree	M5	Provides a set of rules and a predicted numerical value for each, or a regression formula for each rule	93%	161
Linear Regression	Linear Regression	Provides a single regression formula for calculating crash count	70%	n.a
Neural Network	Multilayer perceptron	Provides a "black box" model which can be	83%	n.a
Support Vector Machine	SMOreg	Provides a single regression formula	68%	n.a

The regression tree provides the best results, with M5 having a reasonably low count of 182 rules for reasonable prediction of 93% of the instances. M5 has the benefit of optionally providing a regression formula which provides each significant attribute's contribution to the final

crash count prediction for the rule. Linear regression, neural network and the support vector machines all performed adequately, classifying the percentage of instances within the range of 68% to 83%

Automated attribute selection methods were used to improve the model. A list of the most significant attributes was compiled from the attributes selected from four ranking algorithms (Table 4). These attributes included most relevant road and traffic attributes. Additional attributes of interest will be tested during later modelling aimed at optimizing the models to perform with the whole dataset; however at this stage it is assumed that the model is of sufficient quality for evaluation of suitability of purpose.

Table 3: Ranking of attributes using algorithms.

Order	Attribute	Significance
1	AVG_FRICTION_AT_60_lkm	2.37121
2	AADT	1.99425
3	traffic_percent_heavy	0.97523
4	lane_count	0.96304
5	TEXT_DEPTH_SPTD_OWP_Avg	0.78307
6	roughness_average	0.72793
7	rutting_average	0.65875
8	seal_age	0.53282
9	seal_type	0.47272
10	CRASH_SPEED_LIMIT	0.37733
11	CWAY_TYPE	0.32681
12	CRAS_DIVIDED_ROAD	0.23064
13	ROAD_TYPE_STATE_LCN	0.23016
14	CRASH_NATURE	0.22575
15	ROADWAY_FEATURE	0.19661
16	TRAFFIC_CONTROL	0.17793
17	Intersection	0.12098
18	LIGHTING	0.10606
19	general_terrain	0.08244
20	HORIZONTAL_ALIGNMENT	0.05935
21	VERTICAL_ALIGNMENT	0.04596
22	CRASH_DAY_OF_WEEK	0.02203
23	ATMOSPHERIC	0.01145
24	WetRoadSurface	0.00522
25	wasRaining	0.00459
26	CRASH_TIME	0
27	CRASH_SEVERITY_MAX	0

Table 4: Selected *Best 14* road and traffic attributes.

Attribute
AVG_FRICTION_AT_60_lkm
AADT
traffic_percent_heavy
lane_count
Text Depth_SPTD_OWP_Avg
roughness_average
rutting_average
seal_age
seal_type
CRASH_SPEED_LIMIT
CWAY_TYPE
CRAS_DIVIDED_ROAD
ROAD_TYPE_STATE_LCN
Roadway Feature

In the second stage, a method of tree pruning was used to reduce the rule count and generalize the model. A series of models was generated for the M5 regression tree. Results are shown in Table 5.

Table 5: Reduction in classification accuracy of M5 model with size and rule count.

Model	Leaves & rules	R-squared	Mean absolute error	Root Mean Square Error
1	143	0.9279	3.7567	5.7576
2	159	0.9308	3.6246	5.2557
3	**161**	**0.9327**	**3.5208**	**5.1679**
4	163	0.9216	3.7069	5.5363
5	119	0.9098	3.9278	5.8754
6	88	0.8677	4.5369	6.9873

The results show that, in the regression tree M5 model (Table 5), a substantial reduction in rule count has only a minimal effect on the correlation coefficient (r-squared) from 0.93 to 0.86 even though the leaf count reduces almost by half from 161 leaves to 88 leaves. Plots were created for the extreme values shown in the table showing the distributions of the *predicted crash count* value in relation to the *actual crash count value* (*Figure 6* and *Figure 7*).

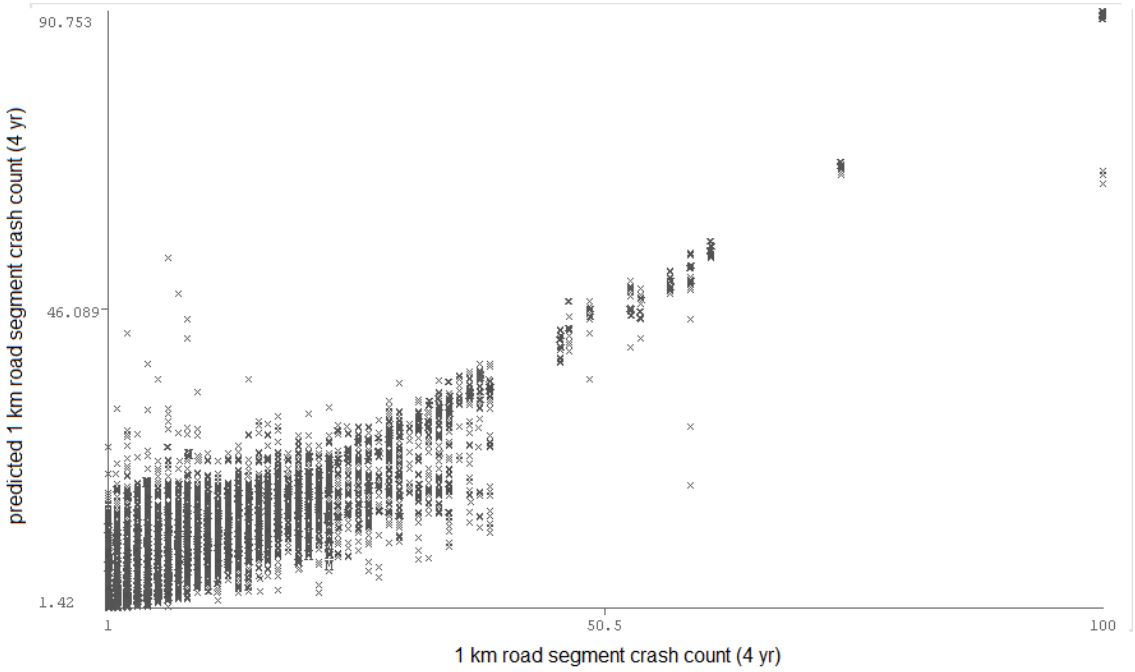

Figure 6: Crash count predictive accuracy plot of M5 regression tree with 143 leaves (93% correlation).

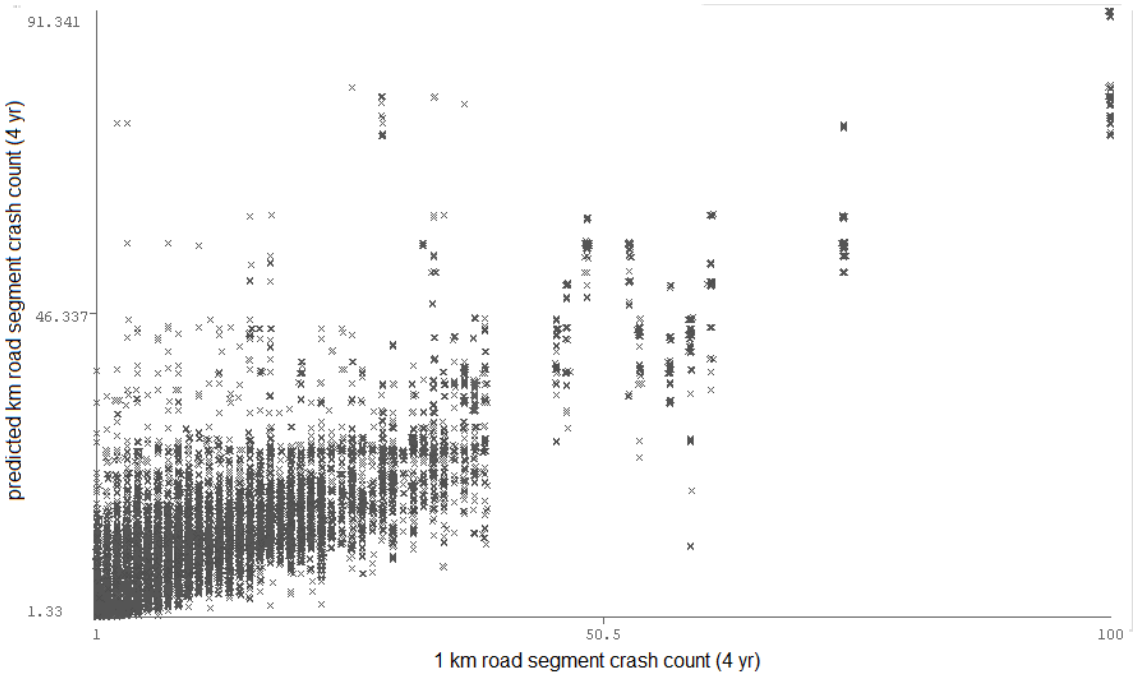

Figure 7: Crash count predictive accuracy plot of M5 regression tree with 88 leaves (87% correlation).

While the plot in the model with the lower rule count of 88 rules (*Figure 7*) shows a wider distribution of predicted values than the model of higher accuracy with 143 leaves (*Figure 6*), the predictive distributions remain in bands meeting our purpose of identifying road segments that have low crash rates, medium crash rates and high crash rates. Further work will assess the range of models in Table 5 to find the best configuration required to apply the models to the

whole dataset. Best practice dictates that conclusions be drawn from a number of models and models using alternative algorithms provided similar results.

Sample rules generated by tree models are examined. A rule has two parts, a condition composed of attribute range values making the rule into a propositional rule [7], and a result showing the node number (NODE), number of instances (N), the predicted rule average (AVE) and the standard deviation of the actual value of instances within the rule (SD).

Sample rule 1 shows a road collection with skid resistance (F60) value of below 0.4059 on high speed single lane roads in the low range of traffic, with the 1 km road segment having an average crash count of 4 crashes per 4 years.

Sample Rule 1.

IF AVG_FRICTION_AT_60 < 0.4095
AND CRASH_SPEED_LIMIT IS ONE OF: 90 100 110
AND 3987 <= AADT < 6105
AND CWAY_TYPE EQUALS SINGLE
THEN
NODE : 48
N : 315
AVE : 4.04444
SD : 2.5357

Sample rule 2 belongs to road segments with reasonably high crash rate having low speed, mid traffic density roads and texture depth is above 1.625 mm threshold. the lower boundary of skid resistance can reach low values of 0.228, and the number of lanes is high. Based on the skid resistance and AADT values, this set or roads approaches the crash the region of highest crash rate shown on *Figure 4*

Sample Rule 2.

IF 1.625 <= TEXT_DEPTH_SPTD_OWP_Avg
AND roughness_average < 98.5
AND 19320 <= AADT
AND lane_count < 3.455
AND CRASH_SPEED_LIMIT IS ONE OF: 10 20 30 40 50 60
AND 0.228 <= AVG_FRICTION_AT_60
THEN
NODE : 93
N : 69
AVE : 42.2899
SD : 19.3938

Sample rule 3 has a high instance membership of low crash roads, representing single lane, low traffic density roads with low crash rates.

Sample Rule 3.

IF 358 <= AADT < 1025
AND 3.65 <= traffic_percent_heavy
AND CWAY_TYPE EQUALS SINGLE
THEN
NODE : 107
N : 446
AVE : 1.51794
SD : 0.70131

Each of sample rules 4,5,and 6 represents road segments with low skid resistance and demonstrates a drop in 4 year road segment crash count from an average of 73.2 crashes to 5.9 crashes, and shows a correlating drop in traffic rated (AADT).

Sample Rule 4.

IF AVG_FRICTION_AT_60 < 0.228
AND 8904 <= AADT
THEN
NODE : 6
N : 92
AVE : 73.2065
SD : 38.1774

Sample Rule 5.

IF AVG_FRICTION_AT_60 < 0.277
AND 6105 <= AADT < 8904
THEN
NODE : 10
N : 111
AVE : 22.9279
SD : 12.8096

Sample Rule 6.

IF AVG_FRICTION_AT_60 < 0.2735
AND 1464 <= AADT < 3987
AND CWAY_TYPE EQUALS SINGLE
THEN
NODE : 44
N : 78
AVE : 5.96154
SD : 3.9205

In studying these rules, the temptation is to jump to conclusions about the cause of crash rate being related to the traffic rate; however the rules merely define the range of certain attributes used in selecting the group of roads, and making the transition from correlation to causation needs to be carefully managed through collection of all evidence, inference management and testing over time. Nevertheless the classes and their defining rules are a powerful analytical components. Witten and Franks state that "*propositional rules are sufficiently expressive for concise, accurate concept description*" [7]. While the models have been developed for automated classification purposes in decision support, analysis of the rules will provide the understanding to evaluate the completeness of the models in their decision support role and contribute to our understanding of their quality. A model analysis stage is planned for future work.

Skid resistance is an active and prominent attribute in many rules. The chart (*Figure 8*) plots the predicted road segment crash count from the data mining model vs. the road skid resistance (F60) and shows that a strong relationship exists where crash count increases as skid resistance drops, for the road segments with up to seven crashes, and these make up a high proportion of crashes (*Figure 2*). Of the remaining categories the skid resistance is prone to fluctuation because of the low instance counts, as well as skid resistance fluctuations, where rules with low skid resistance may be those that will benefit from resurfacing, whereas groups of road segments with high skid resistance can be identified and other causes sought.

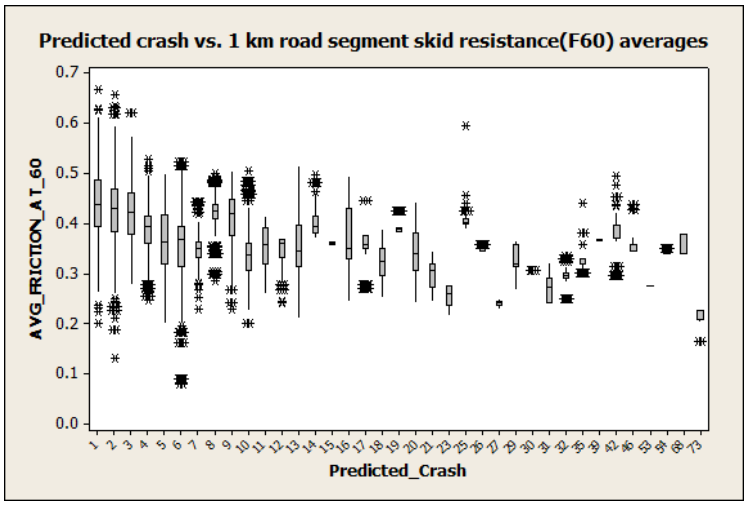

Figure 8: Predicted crash count vs. skid resistance (F60).

These regression tree models demonstrate that our study has met the objective of showing that crash count can be successfully modelled with road and traffic attributes. Regression trees are unshackled from the limitations of linear modelling and the information loss experienced by decision trees during the creation of classes through the discretization of continuous values. Our

models used road and traffic attributes as inputs to predict the range of road segment crash counts from the heterogeneous data sources of the whole of the road network. Examination of the annual variation of sample road segment crash count suggested that, as a measure, it was stable enough to model. From year to year, a given road segment varied in the number of crashes; however the crash numbers fell within a characteristic range over time, and therefore the aggregated four year crash count was selected as the target. The regression tree provides an elegant solution by providing output of an average value and a standard deviation for instances classified by a given rule. Having a model that can provide a prediction within narrow probability for a crash opens the floodgates for developing regression tree models to deploy in decision support in the road and crash management domain. Using the best of these regression tree models, a profiling method can be applied to a given road segment to estimate crash counts over a possible range of skid resistances (0.1 to 0.7) to provide a *skid resistance* vs. *crash count* performance curve. Crash counts will provide road asset managers with information to judge the current skid resistance value of a given road segment, and model the benefit of a proposed intervention.

6. CONCLUSION

The objective of the study was to demonstrate that a predictive data mining model could be developed to determine road segment crash count and crash risk using road and traffic attributes. Models, developed from the small proportion of road segments with skid resistance, provided the algorithms that would be used for analysis of road data for all roads on the network. These algorithms developed on the richest subset of data, will be applied to all road segments with the purpose of identifying those road segments that would benefit from road surface intervention to control skid resistance and reduce crash risk. In addition, the algorithms are expected to support ad hoc modelling with a combination of attributes and support filtering to produce a focus on problems such as serious crash. Further functionality such crash risk forecasting could be achieved by changing attribute values and reapplying the models, therefore allowing road asset managers to test hypotheses on their desktop with roads of interest.

The models developed in this study apply only to road networks with conditions similar to Queensland, however the data mining principles could be applied to produce models for any networks where data collection is representative of the whole network.

The study has found a number of candidate data mining models based on the predictive data mining method from the regression tree family. The regression tree produces rules comprehensible to humans enabling understanding and feedback by industry people, as well as providing machine-driven information processes to perform the required information analysis. Configuration of the models is important. Simply selecting models with the highest accuracy is not sufficient, because models with a high number of rules creates very specific propositions which relate to only a few instances and may not perform well with novel data. The rules must be representative enough to relate to many instances so that, once the model is operating on new data, the rules will classify the new instances with a high level of accuracy as well.

Among a different variety of data mining algorithms, the M5 regression tree showed regression trees to to be a good performers, producing high classification rates (93%) of instances with a low rule counts of 161 rules for 16,750 instances. In addition, when the regression tree was pre-pruned to generalise the rules, a high classification rate of 86% was maintained with a lower rule count of 88 rules.

Further work is required to test and configure the candidate models to ensure that the rules obtained from the data used in model development perform well when applied to the bulk of road instances. Once the models have been optimised, they can be utilized in decision support where all road segments can be risk profiled through the full skid resistance range. Those road found to have a critical skid resistance value will be prioritized for investigation on their predicted crash risk reduction potential.

7. ACKNOWLEDGMENTS

The study is part of an ongoing cooperative study of road surface and crash between Queensland University of Technology (QUT) and the Queensland Department of Transport and Main Roads (QDTMR), with sponsorship from the Cooperative Research Centre for Integrated Engineering Asset Management (CIEAM). Data mining operations were performed in SAS and WEKA and charts prepared in SAS, Weka and Minitab. The views presented in this paper are of the authors and not necessarily the views of the organizations.

REFERENCES

[1] World Health Organization: Global Health: today's challenges. *The World Health Organization,*(2002),Retrieved from http://www.who.int/whr/2003/en/Chapter1.pdf. Retrieved February 11, 2011.

[2]] Queensland-Fire-and-Rescue-Service. Firefighters called to record number of road crashes,. *Queensland Fire and Rescue Service, Queensland Government,* Brisbane, Last update: November 18, 2002), Retrieved from http://www.fire.qld.gov.au/news/view.asp?id=207 Retrieved October, 2008.

[3] Willamson, A.). Why are young drivers over represented in crashes. *National Library of Australia*, Sydney, N.S.W. Australia, 2003. Retrieved from http://catalogue.nla.gov.au/Record/3123551, Retrieved May19, 2009.

[4] Shankar, V., Mannering, F. and Barfield, W. Effect of roadway geometrics and environmental factors on rural freeway accident frequencies. *Accident Analysis & Prevention*, 27 (3),1995, pp 371-389.

[5] Queensland Department of Transport Main Roads. Road planning and design manual, Design Philosophy. *Queensland Government, Transport and Main Roads.* Brisbane, Queensland. Last Update: February 11, 2011 Accessed from http://www.tmr.qld.gov.au/Business-and-industry/Technical-standards-and-publications/Road-planning-and-design-manual.aspx, Retrieved August 02, 2010.

[6] Cairney, P. Road surface characteristics and crash occurrence : a literature review. *Austroads Publication No. AP-T96/08* Austroads, Sydney, 2008. Retrieved from http://catalogue.nla.gov.au/Record/4406507, Retrieved June 15, 2009.

[7] Witten, I. H. and Frank, E. Data Mining: Practical Machine Learning Tools and Techniques. *Morgan Kaufmann*, San Francisco, 2005.

[8] Alzghoul, A. and Löfstrand, M. Increasing availability of industrial systems through data stream mining. *Computers & Industrial Engineering*, 60 (2), 2011, pp.195-205.

[9] Pande, A. and Abdel-Aty, M. Market basket analysis of crash data from large jurisdictions and its potential as a decision support tool. *Safety Science*, 47 (1), 2009, pp.145-154.

[10] Zhao, Z., Jin, X., Cao, Y. and Wang, J. Data mining application on crash simulation data of occupant restraint system. *Expert Systems with Applications*, 37 (8), 2010, pp.5788-5794.

[11] Liu, P. A self-organizing feature maps and data mining based decision support system for liability authentications of traffic crashes. *Neurocomputing*, 72 (13-15), 2009, pp.2902-2908.

[12] Shankar, V., Milton, J. and Mannering, F. Modeling accident frequencies as zero-altered probability processes: An empirical inquiry. Accident Analysis & Prevention, 29 (6), 1997, pp. 829-837.

[13] Chang, L.-Y. and Chen, W.-C. Data mining of tree-based models to analyze freeway accident frequency. Journal of Safety Research, 36(4), 2005, pp. 365-375
.
[14] Wong, J.T. and Chung, Y.-S Analyzing heterogeneous accident data from the perspective of accident occurrence. Accident Analysis & Prevention, 40 (1), 2008, pp. 357-367.

[15] Anderson, T. K. Kernel density estimation and K-means clustering to profile road accident hotspots. Accident Analysis & Prevention, 41 (3), 2009, pp. 359-364.

[16] Qin, X., Ivan, J. N. and Ravishanker, N. Selecting exposure measures in crash rate prediction for two-lane highway segments. *Accident Analysis & Prevention*, 36 (2), 2004, pp. 183-191.

[17] Haynes, R., Lake, I. R., Kingham, S., Sabel, C. E., Pearce, J. and Barnett, R. The influence of road curvature on fatal crashes in New Zealand. *Accident Analysis & Prevention*, 40 (3), 2008, pp.843-850.

[18] McCartt, A. T., Northrup, V. S. and Retting, R. A. Types and characteristics of ramp-related motor vehicle crashes on urban interstate roadways in Northern Virginia. *Journal of Safety Research,* 35 (1), 2004, pp.107-114.

[19] Olsen, R. A. The driver as cause or victim in vehicle skidding accidents. *Accident Analysis & Prevention*, 10 (1), 1978, pp. 61-67.

[20] Mayora, J. and Piña, R. An assessment of the skid resistance effect on traffic safety under wet-pavement conditions. *Accident Analysis & Prevention,* 41, (4), 2009, pp.881-886.

[21] Yerpez, J. and Ferrández, F. (1986). Road characteristics and safety: identification of the part played by road factors in accident generation. *INRETS Synthesis No 2*. Paris, France. 1986.

[22] Nayak, R., Emerson, D., Weligamage, J. and Piyatrapoomi, N. Using Data Mining on Road Asset Management Data in Analysing Road Crashes. In Proceedings of the 16th Annual TMR Engineering & Technology Forum, Brisbane, July 20, 2010.

[23] Emerson, D., Nayak, R., Weligamage, J. and Piyatrapoomi, N. Identifying differences in wet and dry road crashes using data mining. Proceedings of the Fifth World Congress on Engineering Asset Management, WCEAM 2010, Brisbane, Australia.

[24] Nayak, R., Emerson, D., Weligamage, J. and Piyatrapoomi, N. Road Crash Proneness Prediction using Data Mining. *Proceedings of the Extending Database Technology,* EDBT 2011, Uppsala, Sweden.

[25] Kam, B. H. A disaggregate approach to crash rate analysis. Accident Analysis & Prevention, 35 (5), 2003, pp. 693-709.

AUTHOR BIOGRAPHIES

Daniel Emerson, research assistant to Dr Nayak, Master of Information Technology & enrolled in Master of Research at Queensland University of Technology (QUT), developed the data mining datasets from QDMR data and the data mining models. Daniel has been engaged in the project for over two years and has been working with Justin Weligamage and Dr Nappadol Piyatrapoomi, senior engineer at QDTMR. He has immersed himself in skid resistance lore through research and contact with QDTMR staff and other road engineers during meetings and conferences.

Dr Richi Nayak, the data mining supervisor for the study, is senior lecturer in the Computer Science Discipline of the Faculty of Science and Technology, Queensland University of Technology. Her research interests are data mining and Web intelligence. She has published about 100 articles related to data mining in international journals, conferences and books. She has successfully applied her data mining expertise in a number of application domains such as active ageing, road asset management, building construction, software engineering and Web services. She is editor-in-chief of the International Journal of Knowledge and Web Intelligence. Richi provided a guiding hand over model development.

Justin Weligamage is currently Manager (Road Asset Strategy) with the Department of Transport and Main Roads, Queensland, Australia. He has over 25 years of consulting, research and industry experience in the areas of road and civil infrastructure. Justin has been involved in several initiatives for developing and implementing road asset management initiatives, including the publication of "Asset Maintenance Guidelines" and “Skid Resistance Management Plan”, strategic application of the Highway Development and Management System, or HDM-4 within Queensland Main Roads, and road investment decision support research within the Cooperative Research Centre for Construction Innovation. He has written a number of research papers and technical reports, and has been published and presented at various refereed international conferences. Justin has been the leader, inspiration and support behind the project.