A Data Mining Driven Risk Profiling Method for Road Asset Management

Daniel Emerson
Queensland University of Technology
Brisbane, Australia, 4000
d.emerson@connect.qut.edu.au

Justin Z. Weligamage
Roadway Engineering Consultant
Brisbane, Australia, 4000
jweligamage@gmail.com

Richi Nayak
Queensland University of Technology
Brisbane, Australia, 4000
r.nayak@qut.edu.au

ABSTRACT

Road surface skid resistance has been shown to have a strong relationship to road crash risk, however, applying the current method of using investigatory levels to identify crash prone roads is problematic as they may fail in identifying risky roads outside of the norm. The proposed method analyses a complex and formerly impenetrable volume of data from roads and crashes using data mining. This method rapidly identifies roads with elevated crash-rate, potentially due to skid resistance deficit, for investigation. A hypothetical *skid resistance/crash risk curve* is developed for each road segment, driven by the model deployed in a novel *regression tree extrapolation* method. The method potentially solves the problem of missing skid resistance values which occurs during network-wide crash analysis, and allows risk assessment of the major proportion of roads without skid resistance values.

Categories and Subject Descriptors

G.1.1 **[Mathematics of Computing]:** NUMERICAL ANALYSIS: Interpolation –*extrapolation,*/ G.4: MATHEMATICAL SOFTWARE: Algorithm design and analysis

Keywords

Risk management; data mining; model deployment; road asset management; missing data; skid resistance.

1. INTRODUCTION

Road asset managers have huge and diverse volumes of data at their disposal, including results from the significant road surface friction measure surveys. Skid resistance is a standardized measure of the road surface friction between the road surface and a skidding tire, and is recognized as the best established link between road attributes and crash risk (Cairney 2008, [1]). General world-wide practice relies strongly on historically developed skid resistance heuristics to identify risky roads for decision support in budget decisions [2]. This traditional method of relating roadway demand categories and corresponding skid resistance investigatory levels has the potential for over or under engineering of roads not conforming to the norm, thus approving unnecessary expenditure or creating latent safety issues.

In the data available, the presence of missing values and data quality issues favoured a scope of analysis converging towards local, homogeneous roadway segments, which is the norm for current roadway research [3]. However to solve the problem of finding all skid resistance problem roads, a scope of analysis of the whole road network was required. Thus a method was sought to overcome the limitations of the missing data.

This paper proposes a two-part solution. Based on the premise that historical roadway data can predict future crash rates [4], a regression tree model was trained on those data from the crash locations with skid resistance surveys available (40% of all crash locations) to predict the aggregated crash rate of each 1 km road segment. The bagged M5 algorithm [5], fitted with roadway features, road wear, crash and traffic variables, returned a coefficient of determination (r-sq) above 0.9 and was able to predict the full range of values in the crash range when deployed over the network. With the models developed, a deployment solution was sought to identify risky roads, including the 60% of crash locations without the key skid resistance survey results. To manage this situation, a deployment of the model used a *"what if" data table* framework. This method allowed the model to be deployed as a predictive engine in a *crash /skid resistance profile* for each crash location in the network. To provide the *x-axis* of the profile, each crash location instance was replicated and collectively populated with skid resistance (F60) values between the maximum and minimum at a default increment. To provide the corresponding *y-value* of the skid resistance/crash rate point for each replicate, the model was applied in the novel extrapolation process called *regression tree extrapolation*, and the crash prediction made.

The combined points from the replicates produced a curve showing a change in crash rate with increase in skid resistance. Each curve was independently interrogated to seek the existence of a skid resistance threshold where the rate dropped to a low crash rate plateau. In complying roads, comparison between the known crash rate of the roadway segment and the optimal predicted crash rate allowed identification of roads with elevated crash rate. Since the actual skid resistance of the road segment did not need to be known, the method could be applied to almost all roads. Thus the method provided a data-driven machine-learning method to identify risky roads from across the whole network.

The crash curves showed three patterns: (1) the pattern outlined above; (2) non-skid resistance sensitive roads showing little or no change found generally in low crash roads; and (3) erroneous patterns. The data was "observational" in nature, being sourced from other projects, and had many aggregated values and known quality issues. The emergence of the consistency of strong patterns of skid resistance thresholds in results reinforced the

KDD'13, August 11–14, 2013, Chicago, Illinois, USA.
Copyright © 2013 ACM 978-1-4503-2174-7/13/08…$15.00.

appropriateness of both the modelling method and the *"what if"* approach.

The method was evaluated by applying the principles of Coppi's *Informational Paradigm* [6]. Coppi provides the precursor of a formal data mining framework that allows acceptance of inductively derived knowledge using non-statistical data and methods, e.g. databases and data mining algorithms, when applied compliantly within the methodology. The high-level of confluence found between the components of the study was an indicator of goodness. Agreement was found among the domain knowledge, the statistical models and data mining models evaluated in the study's *Strategy of Analysis*.

In summary, (1) the problem of complexity has been solved by combining an extrapolation method with a regression tree model to analyze road crash data, and (2) the generation of a skid resistance/ crash rate profile sidestepped the problem of the missing values in the experimental variable *skid resistance*, thus allowing the whole of the roadway dataset to be processed.

The rest of the paper is as follows. Section 2 presents the related work. Section 3 introduces the dataset that we have for analysis. Section 4 discusses the proposed methodology. Section 5 evaluates the results and provides the related discussion.

2. PRIOR WORK

Road crash studies initially used statistical methods such as Analysis of Variance, Linear Regression, Poisson Regression, Negative Binomial Regression and Log Linear in the examination of homogenous or near homogeneous road sections. Investigations have examined roadway features, prevailing conditions and traffic factors related to the causes of crashes or to predict crash rate [3,4]. Analyses generally compared limited classes of roadways, with further limitations such as data quality, distribution assumptions and experimental design imposed by statistical methods [3].

A second wave of road and crash analysis, applying data mining (DM) on reasonably homogenous data, has shown skid resistance to be significant. A method, using Random Forest Trees to examine crash severity on arterial roads, concluded that high skid resistance has a correlation with severity of accident [7]. Outcomes from a study of skid resistance and texture depth on crash rate in an expressway tunnel found an inverse relationship between crash rate and skid resistance [8]. A similar inverse relationship between crash rate and skid resistance was recently described using clustering to evaluate the effects of skid resistance and texture depth in crash rates [9].

A third wave of road crash analysis progressed from historic aggregated crash rate to examination of conditions leading up to individual crashes, and fitted models with real-time traffic conditions to predict imminent crashes, thus allowing modification of traffic flow [3].

The commonality in most road-crash analysis, represented by these studies, is examination of homogeneous or near homogeneous roadway with above-mentioned limitations. This study extends the capability of the second wave by demonstrating a data-driven method of analysis for the heterogeneous data representing the whole network, and allowing the effect of all participating variables to be expressed.

The debilitating problem in our scenario was missing skid resistance values. With the analysis focused on assessing the skid resistance of all sealed roads, a solution was required to accommodate the 60% of crash locations without skid resistance values.

Statistical imputation methods for replacing a small proportion of missing values with a value or set of values from a similar class have been in long usage. For a larger proportion of missing values, methods have applied data mining, and include deployments of association mining [10], decision tree [11], multilayer perceptron [12] and nearest neighbor [13].

Recently, Kwak's single replacement imputation method developed for a very large proportion of missing values, utilizing the multivariate DM environment [14], demonstrated that imputation of the proportion of missing values in our study was well within reach. The method, relying upon the intrinsic similarity between instances in like classes found in a small proportion of the population, selected the "optimal" value before data mining, and demonstrated the capability of managing between 50% and 90% of missing values in a given attribute.

Our non-replacement method, also conducted in the multivariate DM environment, relies on the similarity of relationships within classes in a small but representative population. However the method benefits from non-rejection of any imputations, with the benefits including the presence of a powerful context for understanding the behavior of the predictions across the value range of the attribute of interest, skid resistance, and identification of aberrant predictions.

Inspired by the *"what if"* DM deployment study [15], our testing of crash rate with the regression tree model across the full range of skid resistance solved both the missing value problem and provided a model deployment method to produce the skid resistance/crash rate curve.

3. UNDERSTANDING THE DATA

While road systems are evolutionary with a cyclic, unknown dynamic dependent on improvement, maintenance, degradation and weather, this study treats the four year period as a snapshot. The road data is evolutionary in the sense that crash locations were populated with the latest roadway data by crash date. However, because of natural variability and complexity, crash rates were averaged over the full four years and do not reflect the potential evolutionary change. Non-crash roads were excluded from the analysis because of the lack of detailed site data and lower significance. All available crashes were included because of their potential for traffic disruption [3], and the crash data set, while believed to be reasonably complete, suffers from being reported crashes only [4]. Extensive data imputation of the road data was required using comparative annual studies and transactional files, but with insufficient data, a small proportion of crashes was dropped.

3.1 Representing the Crash Risk Target

The fundamental premise accounting for roads with elevated crash rate is that crashes are caused by bad decisions made by drivers in an environment resulting from surrounding traffic conditions and the unforgiving geometric designs created by engineers (or wear/damage) [3]. Logic dictates that crash-prone roads [16] would maintain high crash-rate from year to year if road dynamics were a major contributor to crash. Thus the first role of the proposed *Strategy of Analysis* was to demonstrate that roads actually maintained their crash rate from year to year.

The commonly used absolute road segment crash rate (crashes/km /yr) was selected as the target. With awareness of the increasing randomness with reducing interval size, the industry 1km distance interval was selected [4] and a 4 year time span selected to moderate the fluctuations between individual years. The road segment crash counts were calculated by counting the total number of crash instances per 1 km road segment.

Crash counts ranged from 1 to 100 crashes for the four years, making the range of annual crash averages from 1 to 25 crashes, while annual totals ranged from 0 to 32 crashes. To investigate the annual variation in crashes, the average annual crash count of each road segment was plotted against the annual crash count. Examination of Figure 1 shows that roads at a given average crash count maintain a range distributed around the average value, and collectively yield a Poisson-like distribution.

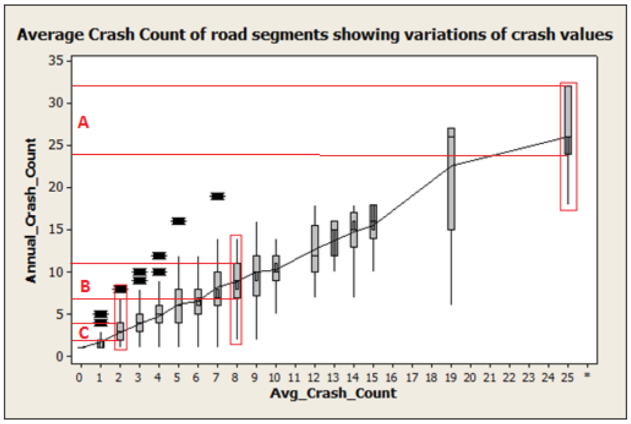


Figure 1 Poisson distribution of annual road segment crash counts for crash count averages

Thus, roads did maintain a characteristic crash count range from year to year. While the population count of road segments of each average dropped exponentially as the crash count increased, each average had sufficient data elements to maintain its "normal like" distribution through to the average count of 15. Above a crash rate of 15 crashes/km/4 years, the distributions became random, but members of the range remained capable of discrimination. A careful examination of the median confidence levels in the quartile box of each average shows only minor overlapping, ascertaining that the averages are generally significantly different. Labels A, B and C show the quartile projections on to the annual crash count axis, showing quite distinct differences.

Selecting Inputs

The road and crash attributes are listed in Table 1. Skid resistance values were averaged from 100m averages to match the 1 km road segment values. The survey results were relatively scarce; having been performed on only about 25% of the road segments either prior to or during the data sampling period. When allocated to the crash locations, only 40% had skid resistance values available, either before or after the crash. The presence or absence of skid resistance effectively partitioned the crash location dataset into two discrete sections and the dataset with skid resistance values became the training set. The relationship between these data sets is represented by:

DS* = *TR* U *NS where *DS* is the whole crash/road segment dataset, *NS* is the non-skid resistance data subset, and *TR* is the data subset with skid resistance values and is used as the *training set.*

The modeling objective was developed to analyze the behavior of the one km road segments using the enhanced data at the crash sites, i.e. *to predict the one kilometer road crash segment crash count at available crash sites using road characteristics at the crash sites and the general characteristics of the road segment itself.*

Table 1 Contributing Attributes

Id	Variable Class	Name	Type	Range/example
d	**Dependent**	Crash count (4 year)	interval	1-100
	Independent			
a1	Design	Roadway type	class	highway, main ..
a2		Crash speed limit	interval	10 to 110
a3		Lane count	interval	0 to 4.7
a4		Divided road	class	yes/no
a5		Has_intersection(s)	class	yes/no
a6		Carriageway type	class	single, dual
a7	Geometry	Horizontal alignment	class	straight, curve ..
a8		Vertical alignment	class	Level, grade, crest, dip
a9	Roadway surface	Avg Friction at 60 1km (skid resistance)	interval	0.19 to 0.65 ROAR Method
a10		Texture depth	interval	0.4 to 15.0
a11		Seal age	interval	0 to 20
a12		Seal type	interval	Spray seal, DGA,
a13	Wear /damage	Roughness average	interval	0 to 406
a14		Rutting average	interval	-2 to 29
a15	Roadway features	Roadway features	class	intersection, bridge, rabout ….
a16		Traffic Control	class	none, give way ..
a17	Demography & settlement	rural or urban	class	Urban/rural
a18	Traffic	Annualized average daily traffic(AADT)	interval	1 to 84,232
a19		Percent heavy vehicle	interval	0 -95
a20	Crash conditions	Wet road surface	class	yes/no
a21		Atmospheric	class	Clear, raining, foggy, smoke ..
a22	Crash details	Crash severity(max)	class	1-fatal-5 property
a23		Crash nature	class	Sideswipe, head on, read end etc.

4. THE PROPOSED METHODOLOGY

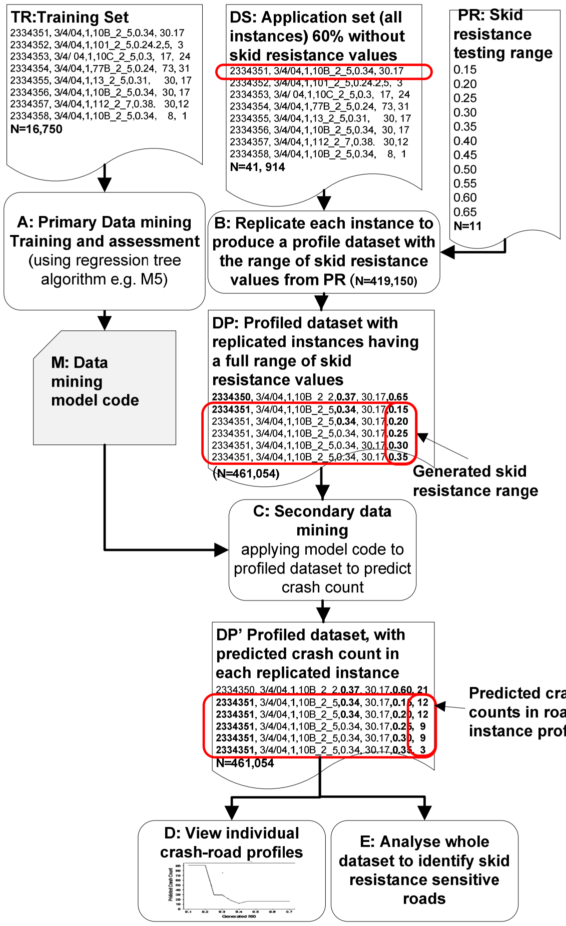
This section proposes the skid resistance/crash rate profiling method and outlines the method of evaluation.

4.1 The Risk Profiling Method

This method builds and applies the model in an extrapolated skid resistance/crash rate profile to produce a skid resistance/crash curve and queries the curve to identify road segments with elevated crash risk due to skid resistance deficit. The process is shown in Figure 2.

Process A performs training of the model (M) on the data set (TR) with allocated skid resistance values (40% of the total road/crash dataset), using the algorithm *bagged regression tree* method M5 [5]. Bayesian, decision trees, and random forests methods were inappropriate because of their inability to predict numerical values. The road crash regression tree model demonstrated that roadway features and traffic relationships can well describe crash count [17]; however a dynamic process was required to apply the model to identify the roads with skid resistance deficit. Haas, 2011 [15] proposes applying data mining models in dynamic "*what if*" extrapolation frameworks for unlocking the deployment potential of the model. We propose a *profiling* method that is a *simulated experiment* and has all of the hallmarks of the statistical hypothesis testing scenario, with rules that are *imbued with an understanding of the fundamental mechanisms that govern a system's behavio*r [15].

Our method distinguishes itself from the normal scenario of extrapolation, which calculates predictions using a singular expression. This is achieved by deploying the regression tree method as the predictive engine. The set of expressions delivered by the regression tree model is expected to be more capable of accurate prediction than a single expression applied over the whole range. As mentioned above, we coined the term *regression tree extrapolation* for this process.



TR:Training Set
2334351, 3/4/04,1,10B_2_5,0.34, 30.17
2334352, 3/4/04,1,101_2_5,0.24.2,5, 3
2334353, 3/4/ 04,1,10C_2_5,0.3, 17, 24
2334354, 3/4/04,1,77B_2_5,0.24, 73, 31
2334355, 3/4/04,1,13_2_5,0.31, 30, 17
2334356, 3/4/04,1,10B_2_5,0.34, 30, 17
2334357, 3/4/04,1,112_2_7,0.38. 30,12
2334358, 3/4/04,1,10B_2_5,0.34, 8, 1
N=16,750
DS: Application set (all instances) 60% without skid resistance values
2334351, 3/4/04,1,10B_2_5,0.34, 30.17
2334352, 3/4/04,1,101_2_5,0.24.2,5, 3
2334353, 3/4/ 04,1,10C_2_5,0.3, 17, 24
2334354, 3/4/04,1,77B_2_5,0.24, 73, 31
2334355, 3/4/04,1,13_2_5,0.31, 30, 17
2334356, 3/4/04,1,10B_2_5,0.34, 30, 17
2334357, 3/4/04,1,112_2_7,0.38. 30,12
2334358, 3/4/04,1,10B_2_5,0.34, 8, 1
N=41, 914
PR: Skid resistance testing range
0.15
0.20
0.25
0.30
0.35
0.40
0.45
0.50
0.55
0.60
0.65
N=11
A: Primary Data mining Training and assessment (using regression tree algorithm e.g. M5)
B: Replicate each instance to produce a profile dataset with the range of skid resistance values from PR (N=419,150)
M: Data mining model code
DP: Profiled dataset with replicated instances having a full range of skid resistance values
2334350, 3/4/04,1,10B_2_2,0.37, 30.17,0.65
2334351, 3/4/04,1,10B_2_5,0.34, 30.17,0.15
2334351, 3/4/04,1,10B_2_5,0.34, 30.17,0.20
2334351, 3/4/04,1,10B_2_5,0.34, 30.17,0.25
2334351, 3/4/04,1,10B_2_5,0.34, 30.17,0.30
2334351, 3/4/04,1,10B_2_5,0.34, 30.17,0.35
(N=461,054)
Generated skid resistance range
C: Secondary data mining applying model code to profiled dataset to predict crash count
DP' Profiled dataset, with predicted crash count in each replicated instance
2334350, 3/4/04,1,10B_2_2,0.37, 30.17,0.60, 21
2334351, 3/4/04,1,10B_2_5,0.34, 30.17,0.15, 12
2334351, 3/4/04,1,10B_2_5,0.34, 30.17,0.20, 12
2334351, 3/4/04,1,10B_2_5,0.34, 30.17,0.25, 9
2334351, 3/4/04,1,10B_2_5,0.34, 30.17,0.30, 9
2334351, 3/4/04,1,10B_2_5,0.34, 30.17,0.35, 3
N=461,054
Predicted crash counts in road instance profile
D: View individual crash-road profiles
Predicted Crash Count
Generated f60
E: Analyse whole dataset to identify skid resistance sensitive roads

Figure 2 Model Deployment

In *Process B*, the risk profile for each road segment was generated by populating the *x-axis* with a default range of skid resistance values (PR) between 0.15 and 0.65 with an increment of 0.05. Each instance present in the dataset DS was replicated by using the values of the PR range and stored in a data set (DP).

Let $DS_i = \{a_1, .., a_n\}$ where a_1 to a_{n-1} are the input attributes including the skid resistance attribute, and a_n is the original target value. Instances in *DS* are replicated by replacing the skid-resistance attribute value by constant values in the PR range:

$$DP_i = \{a_1, .., 0.15, .., \ a_n\}$$

………

$$DP_{i+10} = \{a_1, .., 0.65, .., \ a_n\}.$$

In *Process C*, the dependent *y-axis* value was predicted for each replicate through our process of extrapolation, where the method processes each skid resistance value along the *x-axis* using a series of rules, depending on the skid resistance value of the replicate. Thus, based on the feature combination in the replicate under processing, the model selects the appropriate formula used to predict the crash rate (*y-value*).

The predicted crash rate was added to the replicate records in the profile data set DP'. The prediction is represented by:

$$DP_i[a_n] = \ M(DP_i[a_1, .., PR, .., \ \ a_{n-1}]),$$

where *M* is the model, DP is profiled crash-road table, and each instance has *n* attributes, of which 1 to *n*-1 provided the input variables and n^{th} variable provided the target value. The model *M* was applied on each instance of DP with (*n-1*) attributes values to predict crash rate, stored in the n^{th} attribute, displacing the original target to the new position. This process creates the populated profile instance represented by:

$$DP_i' = \ DP_i[a_1, .., PR, .., a_{n-1}, a_n, a_{n+1}],$$

where attribute a_n denotes the predicted value by the model *M* and a_{n+1}denotes the original target value.

With completion of the predictions, *Process D* creates visualizations of the crash risk curve (crash risk vs. skid resistance) for a given road segment to show the progression of crash risk with the increase in skid resistance for the road segment of the crash location. *Process E* identifies skid-resistance sensitive roads with elevated crash rates, performing below their potential best, that are subsequently tagged as 'investigatory' for prioritization. The investigatory flag is set when the crash rate of the roadway segment was higher that the predicted optimal crash rate. The optimal skid resistance value is read from the curve, without knowing or referring to the existing skid resistance value.

4.2 Evaluating the method

Quality and configuration checks, planned in the *Strategy of Analysis,* were performed throughout. In the modelling stage, assessment included; (1) a comparison of competing predictive algorithms; (2) a configuration optimization of the selected algorithm, and (3) an investigation into the poor deployment performance and its consequence to the outcomes. Subsequently to model deployment, profiles were assessed for goodness of both individual profiles and the method. Firstly, profiles were compared to studies with measured changes [1] and probabilistic studies [18] for the expected form of the curve, i.e. *a drop in crash rate with an increase in skid resistance*. For roads with skid resistance values, the actual crash rate/skid resistance point was checked for its proximity to the curve. Processes were developed to examine the global properties of the profile dataset (DP'), i.e., examining the collective set of profiles for the proportion of erroneous profiles, the shift in distribution of skid resistance with optimization, and the shift in crash rates. Evaluations relied on finding the expected patterns as an assessment of goodness.

5. RESULTS AND DISCUSSION

5.1 Preliminary selection of algorithms selected on performance of test data

In the first stage, algorithms capable of making numeric predictions were compared to find the best performing group. The algorithms, in default configuration, were trained on the data set TR, and fitted with an improved version of the attribute list. The performance criteria were (1) high predictive capability when deployed on the whole network and (2) capability of prediction through the full range of crash values (1-100). The families tested included support vector machines, multilayer perceptron, nearest neighbor, regression and regression trees.

Results in Table 2 show the correlation coefficients (r) for both training and deployment over the whole network, and the capability of predicting the full ranges of crashes between 1 and 100 crashes in training in columns Min P.V. (*minimum predicted value*) and Max P.V. (*maximum predicted value*).

Table 2 Comparative model testing

Algorithm	Training (on TR)	Deployment (on DS)	Min P.V.	Max P.V.
LeastMedSq	0.5972	0.612	-1.2	21.85
Lazy.Kstar	0.9774	0.67	1	100
Decision Table	0.9598	0.674	-4.26	98.7
Linear Regression	0.7332	0.6851	-20.5	53.9
Lazy.Ibk	0.8991	0.6859	1	100
Support Vector Machine	0.68	0.7	-8.4	42.9
M5Rules	0.9226	0.7023	-15.82	97.67
M5P	0.9556	0.7055	-0.46	86.9
MultiLayer Perceptron	0.9221	0.712	-53.9	104.8
Dagged M5Rules	0.8849	0.7705	-0.46	87
Dagged REPTree	0.8553	0.775	2.2	88.6
REPTree	0.9238	0.7848	1.01	100
Bagged M5Rules	**0.9569**	**0.8043**	**-4.26**	**98.7**
Bagged REPTtree	0.962	0.8136	1.27	100

Some instances of Min P.V. show negitave values, indicating an erroneous outlier value below the lower range of 1. The deployment correlation coefficients (r) are ranged from 0.612 to 0.816 with the highest returns and the capability to predict through the full range posted by bagged M5Rules (a variant of M5) and REPTree. Multilayer perceptron was found to produce excessive outliers. Of the models, *Bagged M5Rules* was selected over *Bagged REPTree* because of the compact readable rules.

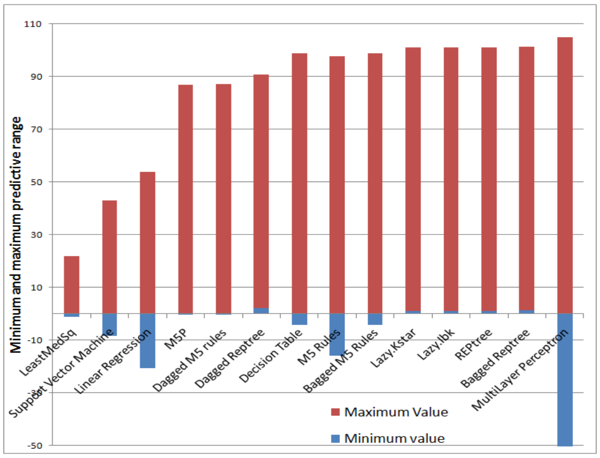


Figure 3 Crash range predicting ability of tested algorithms

Examination of Figure 3 shows that the poorest performers included Least Mean Squares and linear regression, attempting to perform predictions with a single rule. Success of the regression tree was attributed to firstly classifying the road segments, then producing a rule for each class, rather than describing all relationships in a single rule.

The results show that bagged M5Rules was among the best performers on training data, but all in deployment demonstrated a less than stellar correlation coefficient (r) of 0.8 (r-sq: 0.64). This degradation was thought to be caused by the high proportion of instances with null skid resistance values.

5.2 Investigating the effect of null skid resistance values on the deployment accuracy

To ascertain the effect of the missing skid resistance values on deployment classification accuracy, a model (M1) was trained with the same data (DS), except for the removal of the skid resistance attributes, making the data set (D1). In training, model M1 returned a correlation coefficient of 0.92, marginally below the model with skid resistance (0.95). But in deployment the order was reversed, with the model with non-skid resistance returning 0.805 (r-sq 0.648), slightly above the skid resistance model deployed in the dataset degraded by the high proportion of skid resistance null values. Thus we conclude that the model (M) with skid resistance has the potential to perform much better than the model without skid resistance, should the skid resistance values be known. However, the important question is how the change in skid resistance affects crash rate and how the nulls impact on behaviors of predictions in the profiles data set (DP') on the instance replicates, with each set having a range of F60 values.

5.3 Further assessing the impact of the change in skid resistance on crash prediction

To ascertain the effect of the changes in skid resistance on crash rate, a semi-random data subset was chosen from the training dataset (TR), and sorted on crash count during the selection process to ensure that a representation of all crash rates was present. The chosen instances were replicated making three sets of each, and the skid resistance values 0.2, 0.34 and 0.5 were inserted, one in each replicate. Modeling with the new data presented a shift into new territory, because the standard model statistics no longer applied since the value of a significant variable had been changed.

With model (M) deployed, the results in Figure 4 show significant changes in crash count with an increase in the values of skid resistance.

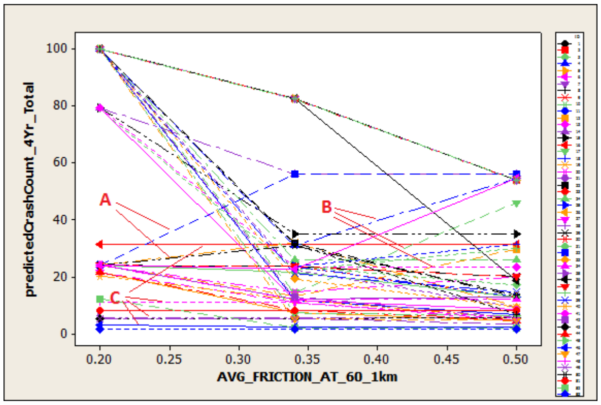


Figure 4 Change in predicted crash count with increasing values of skid resistance, values 0.2, 0.34 and 0.5

Examination of the chart shows most instances experience a drop in crash rate with an increase in skid resistance, with the exception of some that increase (label A & B), suspected of producing the erroneous skid resistance/crash profiles, possibly due to a missing attribute. Further investigation is required. A third set show no crash rate response to skid resistance at all (label C). These results provide an excellent guide to the likely behaviors of predictions in the profiling method, and show that crash rate is very mobile, exhibiting dramatic changes in some cases, with an increase in skid resistance. These results provide evidence for the strong inverse relationship between skid resistance and crash rate, with the magnitude depending on the class of roadway.

Returning to the question unanswered in Section 5.2, this experiment shows that under the influence of the unknown but correct skid resistance value, crash rates are likely to move towards their "correct" value, and the deployment would be expected to be higher than 0.80 (R-sq 0.64), were the values known. Thus the model is better than indicated by the correlation coefficient. Further, the comparative model method in Section 5.2 shows only a small difference between models with & without skid resistance, indicating that this method is relatively ineffective in judging the significance of an attribute.

5.4 Configuring the candidate model

Section 5.1 described the selection process in choosing the bagged M5Rules model (M). However, in an effort to improve the performance of the model, configuration testing was performed, changing the rule splitting criterion of "the *minimum number of instances"* required for a split (*minNum*), to find the optimal level of generalization for the model. Results are shown in Table 3.

Testing was performed on the complete road/crash location dataset without skid resistance (NS), selected to remove the effect of the missing skid resistance values in deployment. The default settings were used for the M5Rules algorithm except the *minNum* parameter. The parameter *build Regression Tree* remained at *false* to build a model tree with a regression formula at each leaf node. The parameter *unpruned* was set at false to allow automatic tree branch pruning, and *use unsmoothed* set at false to deactivate the process to compensate for prior smoothing of data. The configuration variable *minNum* was the experimental variable.

As stated above, *minNum* configures the minimum number of instances required before branching is allowed on a condition, and thus determines the number of classes and rules in the model. The initial value set to 4 produced an over-fitted model with slight degradation in deployment. The split count value in Table 3 shows that the value of *minNum* was progressively doubled until a severely under-fitting model was produced at the value of 1024.

Table 3 Training and testing of the whole dataset

Model M1 Split Count (minNum)	Training Result onTR1 Correl. Coeff.(r)	Rule count	Testing result on all instances on D1, Correl. Coeff. (r)
4	0.9471	136	0.7371
8	0.9497	154	0.8079
16	0.9437	119	0.7371
32	**0.9337**	**111**	**0.8039**
64	0.9220	80	0.7474
128	0.9039	55	0.7710
256	0.8699	36	0.6937
512	0.8257	18	0.7183
1024	0.7415	8	0.6861

With a progressive increase in *minNum*, the results show a successive reduction in rules/class count. Results from deployment show that the most generalized model with near-best deployment performance has the *minNum* value of 32, providing a deployment correlation coefficient (r) of 0.8039 (r-sq 0.64) with the model having 111 rules. The production model (M) for this deployment in the method was built on the training data with skid resistance (TR) with default parameters, except for the *minimum Number of Instances (minNum) set* to this value 32. A plot of the crash rate vs. the predicted crash rate produced by the model (M) is shown in Figure 5.

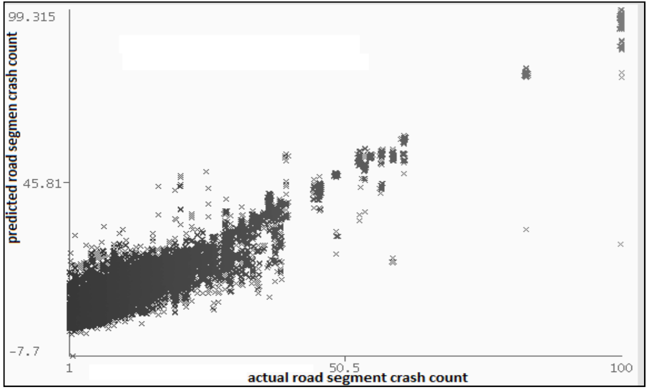


Figure 5 Predictive capability of bagged M5Rules

This plot shows a reasonably narrow variance, endowing the model with the ability to predict crashes within an acceptable tolerance, thus having the ability to discriminate. Examination of the vertical distribution of each prediction suggests them to be approximately Poisson distributed, with a region of highest density generally positioned centrally, with variation relative to the crash count. The variance is less of a problem than first appears, except in the lowest crash range, were below 10 crashes/km/4yr, the variation is greater than the crashes being predicted, and thus the model loses its ability to successfully discriminate.

5.5 Skid resistance/ crash profile results

The selected model (M), when deployed on the skid resistance/ crash profile dataset (DP), was used to predict the crash count values for each of the replicated instances.

Examination of the set of skid resistance/crash count values for a given crash location from the profile dataset (DP') shows that the profiling process does operate as stated. Table 4 shows that with an increase in skid resistance (F60), a substantial decrease in the four-year crash rate values (CC) is predicted. This profile is represented by curve A in Figure 7.

Table 4 Profile values of a sample crash site showing predicted 4 year crash count by skid resistance.

F60	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5	0.55	0.6	0.65
CC	101.3	100.1	35.1	33	31.5	19.9	19.2	14.7	12.6	10.4	8.2

A comparison of the measured skid resistance/crash point for this crash location at (1.8, 100) and the curve shows that the point is almost directly on the curve, thus providing an absolute link between the real world and the profile: an instance of assessment contributing to the external validation process.

As mentioned above, new metrics were needed to assess the result because the standard metrics were no longer relevant in a deployment where the data, skid resistance in each replicate, had been extensively modified. The first new method used a plot showing changes in predicted value (crash rate) in comparison to the target (original crash count). The plot shows the change in the predicted crash rate caused by the change in independent variable of the experiment (skid resistance). As noted in Section 5.3, road sections can experience a large predicted change in crash count with a change in skid resistance, and the mobility is shown in

Figure 6, with some road sections traversing the full range of crash rate with a corresponding change in skid resistance.

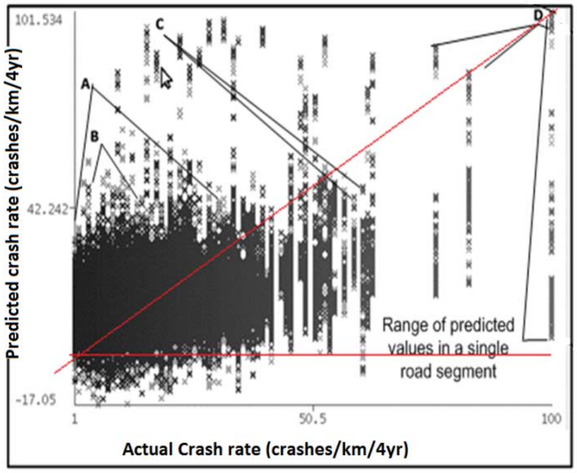


Figure 6 Actual crash vs. predicted crash from the populated profile (DP')

We can draw some generalizations from the plot, and foci of interest are highlighted by the labels A to D.

- A: a high proportion of crash rates do not have high crash counts, showing the existence of roads with lower skid resistance sensitivity;
- B: in the region of low crash rates, fewer road segments have extremely high predicted crash count rates compared to road segments with higher crash rates;
- C: even in the higher range of crash rates, not all crash ranges have extreme crash rates;
- D: instances with naturally high crash count can experience a significant drop in crash rate with increased skid resistance.

Representatives of the crash rate/skid resistance curves from the profiles for various crash sites are displayed in Figure 7.

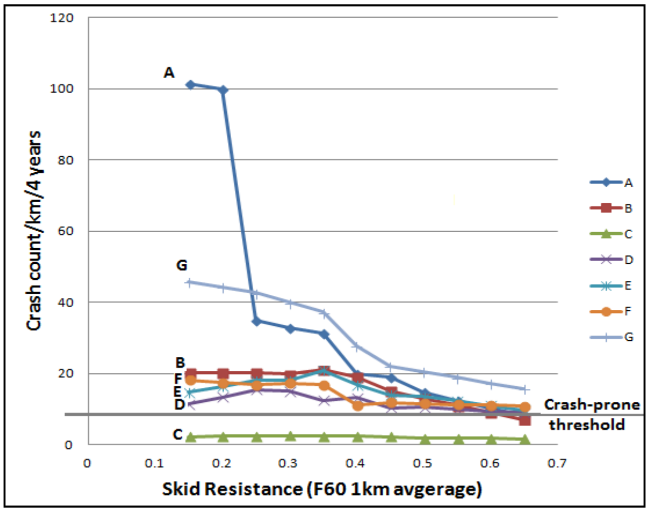


Figure 7 Sample profiles

Each of the locations shows the potential to substantially reduce its crash rate with an increase in skid resistance, with the exception of sample C which is a low crash-rate road. High crash-rate roads labeled A and G show the potential to substantially reduce their crash rate, with sample A experiencing a dramatic drop in crash rate with a skid resistance change from 0.18 to 0.25, but not falling below the crash-prone level until the F60 value of 0.6. Samples D & E show a suspected erroneous trend requiring investigation in the lower skid resistance ranges where crash-rates rise with a rise in skid resistance.

5.6 Assessing the profiling process

A prototype application was developed to demonstrate the value of the method to road asset managers. The screenshot shown in Figure 8 shows an example of the skid resistance/ crash rate curve drawn by querying the model. Note that the model was developed using M5Rules (not bagged) with a more comprehensive dataset.

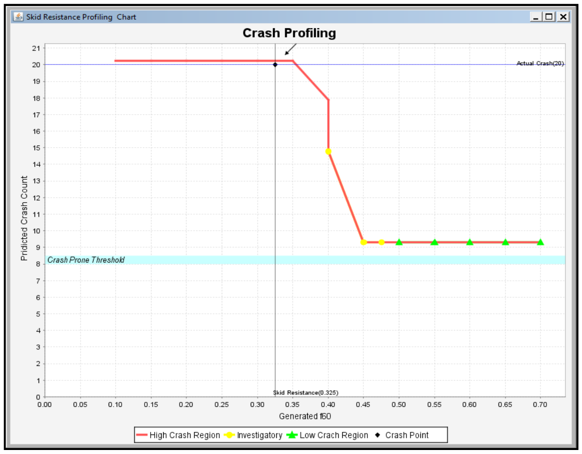


Figure 8 Predicted crash rate curve as crash location

This method utilized the model parameter of *building the Regression Tree*. With the parameter set to true, the model calculates an averaged results from all instances at the leaf nodes, rather than providing a rule. This configuration results in the flat plateau sections in the curve, each with a zero gradient for the F60 extent of a given rule.

The curve shows the expected pattern with a high average crash rate (20 crashes/km/4yrs) at lower skid resistance values, followed by a rapid drop in crash rate at the skid resistance (F60) threshold of 0.35, and subsequently reaching a low crash rate plateau of 9 crashes/km/4-yr beyond the skid resistance threshold of 0.45.

The change in skid resistance from 0.325 to 0.45 almost places the road into the target *non-crash prone zone* (Figure 8) [16]. The profile shows a level of consistency between the predicted curve and the real world measure with the actual skid resistance/crash rate point almost on the curve.

To assess the predictions collectively for the whole network, an algorithm was developed to scan a potential profile error state found in profiles inconsistent with the curve, i.e., of the crash minimum preceding the crash maximum. The results showed a high level of consistency, with only around 5% of profiles being highly inconsistent with the curve shown in Figure 7. Further analysis shows that predictions become progressively more erratic with low crash roads, as the crash count successively drops below 8 crashes/km/4yr, with this phenomenon thought to be a consequence of both the crash count dropping beyond the resolving power of the model and the natural unpredictability of low crash roads. However, examples of quality predictions are found in roads with uniformly low crash rates, as demonstrated by profile C in Figure 7.

Since the method is making prediction of improved crash rate by generally increasing the skid resistance, the range and distribution of skid resistance of optimized roads is expected to be above the existing range, and also be normally distributed. Data visualization shows this to be true. Examination of Figure 9 shows that the range of measured skid resistance ranges from 0.19 to 0.65 in a near normal distribution. In the predicted optimal low crash model, the skid resistance ranges from 0.45 to 0.6 in a skewed normal distribution. This pattern indicates that the method is operating in a reasonable and expected way.

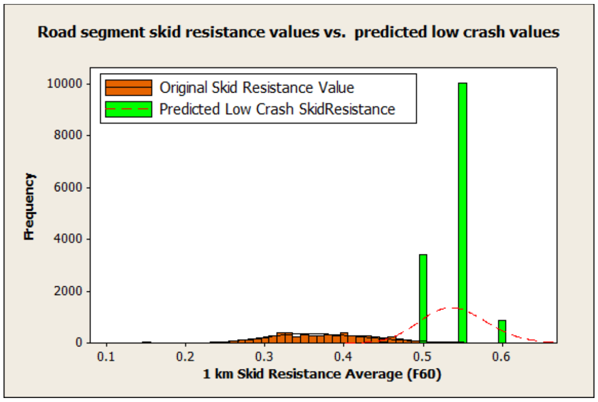


Figure 9 Comparison of the road segment skid resistance range vs. the predicted values for optimal crash rates

The goal of a prediction of an optimized skid resistance value in our context is the reduction in crash rate. The process is performed in stage F of the method (Figure 2) which predicts a consequential reduction in crash rate with the optimal skid resistance value. A comparison of the original crash count values and the potential improved crash count predicted by the process is shown in a distribution in Figure 10, with the predicted optimal crash rate generally below 20 crashes/km/4yr.

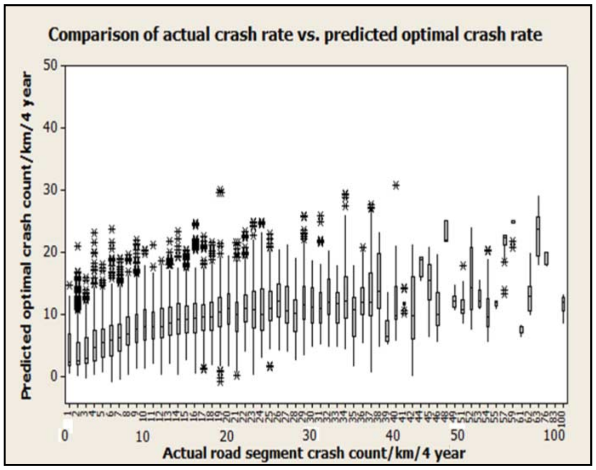


Figure 10 Comparison of recorded crash rates and predicted, optimized crash rates

The comparison between the distributions of recorded crash rates and predicted optimal rates shown in Figure 10 and Table 5, show a huge potential improvement throughout the network, with almost 75% of crash sites approaching or achieving crash rates below the crash-proneness level of 8 crashes/km/per 4 years.

Table 5 Potential improvement in crash rate

	min	Q1	mean	Q3	max
measured crash rate	1	3	13.0	19	100
predicted optimal crash rate	-1.3	3.3	7.4	10.3	30.8

In comparison, the probabilistically developed skid resistance profiles [18] show low-crash threshold ranges of skid resistance values lower than ours, with values between of 0.32 and 0.45, compared to the range from our model between 0.45 and 0.6. The difference is thought to be caused by modeling on low volume, wet roads where our method does not perform as well. In other comparisons, Cairney's before and after resealing studies show crash rate drops of 23% for dry and 68% for wet road crashes [1], in range with improvements in this study.

Since testing of the skid resistance/crash rate predictions is not yet an option, these assessments contribute to the requirement of Coppi's Informational Paradigm for external validation. Comfort is taken from the fact that the probabilistic profiles [18] show a curve similar to the inductively produced curve in Figure 8.

5.7 Discussion

The regression tree method is powerful algorithm that enables analysis of a heterogeneous dataset. With the ability to sidestep the necessity of having specific distributions patterns that limit some statistical methods, regression trees can be applied across the full domain. In addition, being capable of making numerical predictions allows regression trees to be applied in a *"what if"* data set in our novel extrapolation method to produce a risk profile. The attribute of interest provides the range of values for the *x-axis*, and the predicted *y-value* calculates the risk, producing a risk curve across the values of the attribute of interest. Since each variable is controlled, except for the independent variable, the profile approximates a controlled experiment.

In investigating the effect of a given attribute in a model, the method of modeling with and without the variable has been found to be ineffective, showing only a minor loss of predictive accuracy between the models. A more effective method creates replicates of instances and substitutes a range of values from the attribute of interest. When the model is applied, the corresponding change in value of the target indicates the magnitude of the significance of the variable of interest.

Data mining models capable of analyzing large and complex datasets have been available since the early 1990s, however road research has focused on problems on homogenous road sets, resulting in a lack of understanding about the interrelationships between road crash and the full set of roadway characteristics. The presence of missing values has further hampered progress. Dynamic methods such as extrapolation have solved both big and small problems. However the regression tree and its inherent powers has not been formally matched with extrapolation, and thus the problem of performing a road scan in the presence of the large proportion of missing skid resistance values has never been successfully attempted. This study was able to overcome these issues by using data mining in this new context.

6. CONCLUSIONS

This data mining method demonstrates a decision support methodology to assist road asset managers in the task of identifying both measured high crash roads and potentially dangerous road segments from the whole network, where skid resistance upgrade will reduce the crash risk.

The current method of roadway assessment relies on a set of heuristics based on the skid resistance investigatory levels for the various classes of roads, whereas our method, based on the historic data, potentially allows each road to be individually assessed and optimally configured.

The model, trained with available roadway data including skid resistance values, is subsequently deployed in a novel method combining extrapolation and the regression tree. A skid resistance/crash rate profile and crash rate curve is produced for each crash location, and interrogated to find the optimal skid resistance/crash rate for the road segment without knowing the skid resistance value of the road. Thus, the method provides a solution to the formerly-insurmountable problem of the high proportion of roads without skid resistance data. The internal measures of the method evaluate well, and the components and inductively deduced results generally agree well with external measures such as domain knowledge, observed phenomena and probabilistic studies.

Future work is required to consolidate this evidence. New metrics are required to assess the individual skid resistance/crash rate profiles and document the method's performance over the whole data set. The models and profiles describe many new relationships that could contribute to knowledge in the road crash domain, including covariate contributions to crash rate, sparking new areas of research. Work on data pre-processing and algorithm selection/configuration is required to improve the accuracy of outcomes, and the latest, improved data should be included. Issues associated with having road networks with predominantly high road surface friction for the various classes of roadway require investigation. The real test of success would be demonstrated value in the gradual integration of the method into road asset management practices to support decision making.

Potential exists for deployment in smart-vehicle technology for alerting the vehicle/driver to the crash type and risk level of the upcoming road segments. At a broader level, this risk profiling method may be explored to discover critical thresholds in other domains such as finance, insurance, engineering, and drug testing.

7. ACKNOWLEDGMENT

We acknowledge CIEAM for the financial support, Queensland Department of Transport Main Roads (QDTMR) for participation and datasets, and Dr Rune Rasmussen for advice. This research is developmental and does not reflect practices of QDTMR.

8. REFERENCES

[1] Cairney, P. 2008. Road surface characteristics and crash occurrence : a literature review. *Austroads Publication No. AP-T96/08"--T.p. versoBibliography: p. 40-4.* Accessed from http://nla.gov.au/nla.cat-vn4406507.

[2] Weligamage, J. 2006. *Skid Resistance Management Plan*, Department of Transport and Main Roads, Queensland, Road Asset Management Branch, Ed. Brisbane: State of Queensland, 2006.

[3] Abdel-Aty, M. and Pande, A. 2007. Crash data analysis: Collective vs. individual crash level approach, *Journal of Safety Research,* vol. 38, (2007), 581-587.

[4] Miaou, P.S. and Lum, H. 1993. Modelling vehicle accidents and highway geometric designs relatioships, *Accident Analysis & Prevention,* vol. 25(6), (1993), 686-709.

[5] Quinlan, J.R. 1992. Learning with continuous classes, *AI 92,* 1992. *AI'92*.

[6] Coppi, R. 2002. A theoretical framework for Data Mining: the Informational Paradigm, *Computational Statistics & Data Analysis,* vol. 38, (2002), 501-515.

[7] Das, A., Abdel-Aty, M. & Pande, A. 2009. Using conditional inference forests to identify the factors affecting crash severity on arterial corridors, *Journal of Safety Research,* vol. 40, (2009), 317-327.

[8] Yang, J.G., Xie, Y.L., Zhang, X. & Ma,W. 2009. Reliability analysis on pavement skid-resistant performance in expressway tunnel. In *Proceedings of ASCE Conference Proceedings*, (2009), 2679-2685.

[9] Fernandes, A. and Neves, J. 2013. An approach to accidents modeling based on compounds road environments, *Accident Analysis and Prevention,* vol. 53, (2013), 39-45, 2013.

[10] Wang, H. and Wang, S. 2009. Discovering patterns of missing data in survey databases: An application of rough sets. *Expert Systems with Applications*, 36, 3, Part 2, (2009), 6256-6260.

[11] Fortes, I., Mora-López, L., Morales, R. and Triguero, F. Inductive learning models with missing values. *Mathematical and Computer Modelling*, 44, 9–10, (2006), 790-806.

[12] Silva-Ramírez, E.-L., Pino-Mejías, R., López-Coello, M. and Cubiles-de-la-Vega, M.-D. 2011. Missing value imputation on missing completely at random data using multilayer perceptrons. *Neural Networks*, 24, 1 (2011), 121-129.

[13] Huang, X. and Zhu, Q. 2002. A pseudo-nearest-neighbor approach for missing data recovery on Gaussian random data sets. *Pattern Recognition Letters*, 23, 13 (2002), 1613-1622.

[14] Kwak, D. and Kim, K. 2012. A data mining approach considering missing values for the optimization for semiconductor-manufacturing process, *Expert Systems with Applicaions,* vol. 39, (2012), 2509-2596.

[15] P. J. Haas, Maglio,P.,Selinger, P. & Tan, W. 2011. Data is dead... without what-if models," in *The 37th International Conference on Very Large Data Bases 2011*, (Seattle, Washington, 2011).

[16] Nayak, R., Emerson, D., Weligamage, J. and Piyatrapoomi, N. 2011. Road Crash Proneness Prediction using Data Mining. In *Proceedings of the EDBT 2011* (Uppsala, Sweden 2011).

[17] Emerson, D., Nayak, R. and Weligamage, J. 2011. Using data mining to predict road crash count with a focus on skid resistance values. In *Proceedings of the 3rd International Road Surface Friction Conference,* (Gold Coast, Australia, May 17, 2011). Accessed from http://eprints.qut.edu.au/41458/.

[18] Piyatrapoomi, N. ,Weligamage, J. and Turner. L. 2008. Assessing wet crashes for the skid resistance management of road assets. In *Proceedings of the 3rd World Congress on Engineering Asset Management and Intelligent Maintenance Systems*, (Beijing, China, 2008), 1299-1307.