Daniel Emerson of QUT
Realising community value through data mining

Co-author of ‘A data mining driven risk profiling method for road asset management’

ISBN: 978-1-4503-2174-7

(Emerson, Weligamage, Nayak, 2013)
Recognition

• Research Team
  – Daniel Emerson: MasterIT(Research)/PhD Student QUT;
  – A. Prof Richi Nayak: DM researcher/lecturer QUT;
  – Justin Z. Weligamage: Principal Engineer-Asset Management, Toowoomba Regional Council

• Programmer:
  – Dr Reza Hassanzadeh: QUT

(All content graphics, unless stated, copyright, Emerson, Nayak, & Weligamage, 2015)
Part 1 The problem of analysis in complex data.

• **Current solutions**: manually-managed homogeneous models:

• **The drawbacks**: fragmented, enterprise resource hungry, limited in direct decision-support.

• **Proposed solution**: enterprise-wide machine-learning, rapid, direct application in decision support (Emerson, Weligamage, & Nayak, 2013), (Emerson, 2013).

• **Future Directions**: refinement; wider deployment

Questions
Themes

• Descriptive vs. Prescriptive analytics;
• Modelling on homogeneous data vs. heterogeneous data;
• First principles modelling vs. empirical modelling;
• The knowledge environment, need for dynamic analysis,
• human (engineer) vs. machine learning;
• Lift from the road-crash data.
Section 1: The Problem
The Research Goal

Generally

• Making decisions about *allocating scarce resources* (money, equipment, people, time, etc) *usually in the face of uncertainty* (Haas, et al., 2011)

Specifically – in road asset management:

• Allocation of *scarce financial resources* across the *whole road network* to meet the stated goals of the network, particularly road safety (QDTMR, 2013)
The Analytical Landscape

**Descriptive analytics**: defining problems with analytics
- DBMS/ BI technologies (Haas, et al., 2011, Dayal, 2009)
  - OLAP,
  - machine learning tools
  - visualizations

**Prescriptive analytics research**: solving problems
- **Expert systems**: modelling of underlying processes based on expert understanding (Haas, et al., 2011);
- **Machine-learning deployment**: empirical modelling of complex data to derive the models (this study).
Descriptive analytics: defining the problem

Resolving road types and their crash type probabilities from clustering with expectation maximization (EM).

(Emerson, Nayak, & Weligamage, 2011)
Descriptive analytics: defining the problem

(Emerson, Nayak, & Weligamage, 2011)
Descriptive analytics: defining the problem

(Emerson, Nayak, & Weligamage, 2011)
Descriptive analytics: defining the problem

(Emerson, Nayak, & Weligamage, 2011)
Descriptive analytics: defining problems

(Emerson, Nayak, & Weligamage, 2011)
Crash type profile for main road type (Cluster 13 high % of serious)

Cluster 13

<table>
<thead>
<tr>
<th>No.</th>
<th>Label</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Angle</td>
<td>30</td>
</tr>
<tr>
<td>2</td>
<td>Sideswipe</td>
<td>19</td>
</tr>
<tr>
<td>3</td>
<td>Rear-end</td>
<td>60</td>
</tr>
<tr>
<td>4</td>
<td>Hit fixed obstruction or other</td>
<td>197</td>
</tr>
<tr>
<td>5</td>
<td>Overturned</td>
<td>130</td>
</tr>
<tr>
<td>6</td>
<td>Hit parked vehicle</td>
<td>4</td>
</tr>
<tr>
<td>7</td>
<td>Hit pedestrian</td>
<td>5</td>
</tr>
<tr>
<td>8</td>
<td>Hit animal incl. ridden h...</td>
<td>31</td>
</tr>
<tr>
<td>9</td>
<td>Head-on</td>
<td>39</td>
</tr>
</tbody>
</table>
Descriptive analytics: defining problems

(Emerson, Nayak, & Weligamage, 2011)
Descriptive analytics: defining problems

(Emerson, Nayak, & Weligamage, 2011)
Descriptive analytics: defining problems

(Emerson, Nayak, & Weligamage, 2011)
### Descriptive analytics: defining problems

(Emerson, Nayak, & Weligamage, 2011)

<table>
<thead>
<tr>
<th>No.</th>
<th>Label</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Angle</td>
<td>127</td>
</tr>
<tr>
<td>2</td>
<td>Sideswipe</td>
<td>164</td>
</tr>
<tr>
<td>3</td>
<td>Rear-end</td>
<td>102</td>
</tr>
<tr>
<td>4</td>
<td>Hit fixed obstruction or barrier</td>
<td>31</td>
</tr>
<tr>
<td>5</td>
<td>Overturned</td>
<td>32</td>
</tr>
<tr>
<td>6</td>
<td>Hit parked vehicle</td>
<td>7</td>
</tr>
<tr>
<td>7</td>
<td>Hit pedestrian</td>
<td>6</td>
</tr>
<tr>
<td>8</td>
<td>Hit animal inc. hidden human</td>
<td>2</td>
</tr>
<tr>
<td>9</td>
<td>Head-on</td>
<td>2</td>
</tr>
</tbody>
</table>

---

### Descriptive analytics: defining problems

(Emerson, Nayak, & Weligamage, 2011)

<table>
<thead>
<tr>
<th>No.</th>
<th>Label</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Angle</td>
<td>46</td>
</tr>
<tr>
<td>2</td>
<td>Sideswipe</td>
<td>46</td>
</tr>
<tr>
<td>3</td>
<td>Rear-end</td>
<td>74</td>
</tr>
<tr>
<td>4</td>
<td>Hit fixed obstruction or barrier</td>
<td>96</td>
</tr>
<tr>
<td>5</td>
<td>Overturned</td>
<td>15</td>
</tr>
<tr>
<td>6</td>
<td>Hit parked vehicle</td>
<td>9</td>
</tr>
<tr>
<td>7</td>
<td>Hit pedestrian</td>
<td>2</td>
</tr>
</tbody>
</table>

---

### Descriptive analytics: defining problems

(Emerson, Nayak, & Weligamage, 2011)

<table>
<thead>
<tr>
<th>No.</th>
<th>Label</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Angle</td>
<td>89</td>
</tr>
<tr>
<td>2</td>
<td>Sideswipe</td>
<td>66</td>
</tr>
<tr>
<td>3</td>
<td>Rear-end</td>
<td>311</td>
</tr>
<tr>
<td>4</td>
<td>Hit fixed obstruction or barrier</td>
<td>134</td>
</tr>
<tr>
<td>5</td>
<td>Overturned</td>
<td>22</td>
</tr>
<tr>
<td>6</td>
<td>Hit parked vehicle</td>
<td>8</td>
</tr>
<tr>
<td>7</td>
<td>Hit pedestrian</td>
<td>17</td>
</tr>
</tbody>
</table>
Descriptive analytics: defining problems-future
Descriptive analytics: defining problems - future

The crash risks on this road are......

<table>
<thead>
<tr>
<th>No.</th>
<th>Label</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Angle</td>
<td>30</td>
</tr>
<tr>
<td>2</td>
<td>Sideswipe</td>
<td>19</td>
</tr>
<tr>
<td>3</td>
<td>Rear-end</td>
<td>60</td>
</tr>
<tr>
<td>4</td>
<td>Hit fixed obstruction</td>
<td>197</td>
</tr>
<tr>
<td>5</td>
<td>Overturned</td>
<td>130</td>
</tr>
<tr>
<td>6</td>
<td>Hit parked vehicle</td>
<td>4</td>
</tr>
<tr>
<td>7</td>
<td>Hit pedestrian</td>
<td>5</td>
</tr>
<tr>
<td>8</td>
<td>Hit animal incl. hidden</td>
<td>31</td>
</tr>
<tr>
<td>9</td>
<td>Head-on</td>
<td>39</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>No.</th>
<th>Label</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Angle</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td>Sideswipe</td>
<td>46</td>
</tr>
<tr>
<td>3</td>
<td>Rear-end</td>
<td>74</td>
</tr>
<tr>
<td>4</td>
<td>Hit fixed obstruction</td>
<td>96</td>
</tr>
<tr>
<td>5</td>
<td>Overturned</td>
<td>15</td>
</tr>
<tr>
<td>6</td>
<td>Hit parked vehicle</td>
<td>9</td>
</tr>
<tr>
<td>7</td>
<td>Hit pedestrian</td>
<td>2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>No.</th>
<th>Label</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Angle</td>
<td>89</td>
</tr>
<tr>
<td>2</td>
<td>Sideswipe</td>
<td>66</td>
</tr>
<tr>
<td>3</td>
<td>Rear-end</td>
<td>311</td>
</tr>
<tr>
<td>4</td>
<td>Hit fixed obstruction</td>
<td>134</td>
</tr>
<tr>
<td>5</td>
<td>Overturned</td>
<td>22</td>
</tr>
<tr>
<td>6</td>
<td>Hit parked vehicle</td>
<td>8</td>
</tr>
<tr>
<td>7</td>
<td>Hit pedestrian</td>
<td>17</td>
</tr>
</tbody>
</table>

Phone image courtesy of ddpavumba at FreeDigitalPhotos.net
Prescriptive analytics: modelling of solutions-future

To lessen your elevated crash risk, reduce your...
Goal of prescriptive analytics research

Optimization methods to:

• Assess the consequences of alternative design, investment, or policy choices on the system of interest,
  – From the Institute for Operations Research and Management Science (INFORMS) (Haas, et al., 2011)

• Examples
  – ...diet in the context of residential segregation (Auchincloss, et al., 2012)
  – Policy on climate and food (Godfray, et al., 2011)
Roadway decision making
Current roadway decision making

A combination of
• **judgment** of the road engineer,
• following **best practice**, developed from experience and research,
• within the **constraints of road crash rate** and the specified **design parameters** (QDTMR, 2013).
## Decision-support heuristics: the skid resistance investigatory levels

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A - Motorway</td>
<td>0.35</td>
</tr>
<tr>
<td>B - Dual carriageway non-event</td>
<td>0.35 – 0.40</td>
</tr>
<tr>
<td>C - Single carriageway non-event</td>
<td>0.40 – 0.45</td>
</tr>
<tr>
<td>Q - Dual carriageway (all purpose) – minor junctions</td>
<td>0.45 – 0.55</td>
</tr>
<tr>
<td>Q - Single carriageway minor junctions &amp; approaches to and across major junctions (all limbs)</td>
<td>-</td>
</tr>
<tr>
<td>Q - Approach to roundabout</td>
<td>-</td>
</tr>
<tr>
<td>K - Approaches to pedestrian crossings and other high risk situations</td>
<td>0.50 – 0.55</td>
</tr>
</tbody>
</table>
Recognition for risk-based decision-support

• The **CIEAM** Skid Resistance Project (2007-2012) convened to develop a *risk-based method* to establish new skid resistance investigatory levels for minimization of crashes ([Bunker, 2007](#))

where *(we adopted the principal that)*

• the **risk-based method** will discover factors allowing optimizations of design and operational factors to reduce risk. ([Montewka et al., 2014](#))
Moving towards enterprise-wide models

complexity

Heuristics
Moving towards enterprise-wide models

- Complexity

Heuristics

Occam’s Razor (Keep it simple, stupid)
Moving towards enterprise-wide models

- **Deterministic**
- **Probabilistic**
- **Heuristics**

**Expansive complexity**

- Expert systems and law-based rules
  - (Golfarelli, et al., 2006, Saltelli, et al., 2006)
- Empirically-based rules and probabilistic solutions
  - (homogeneous data)
  - (most authors)

Jigsaw images courtesy of Jscreations at FreeDigitalPhotos.net
Moving towards enterprise-wide models

- Complexity
- Deterministic
- Probabilistic
- Heuristics
  - Expert systems and law-based rules
    (Golfarelli, et al., 2006, Saltelli, et al., 2006)
  - Empirically-based rules and probabilistic solutions
    (homogeneous data)
    (most authors)
  - Multi-class machine-learning
  - Bypassing missing values
  - Data-wide navigation index

Jigsaw images courtesy of jscreations at FreeDigitalPhotos.net
Moving towards enterprise-wide models

- Complexity
- Deterministic
- Probabilistic
- Heuristics

**Expert systems and law-based rules**
(Golfarelli, et al., 2006, Saltelli, et al., 2006)

**Empirically-based rules and probabilistic solutions**
(homogeneous data)
(most authors)

- Multi-class machine-learning
- Bypassing missing values
- Data-wide navigation index

**MRE**
Integrated, enterprise-wide modelling
(heterogeneous data)

Jigsaw images courtesy of jscreations, Salvatore Vuono at FreeDigitalPhotos.net
Moving towards enterprise-wide models

- Complexity
- Heuristics
  - Deterministic
  - Probabilistic

Expert systems and law-based rules
(Golfarelli, et al., 2006, Saltelli, et al., 2006)

Empirically-based rules and probabilistic solutions
(homogeneous data)
(most authors)

Multi-class machine-learning
Bypassing missing values
Data-wide navigation index

MRE
Integrated, enterprise-wide modelling
(heterogeneous data)

Prescriptive Analytics
(Haas, et al., 2012)

Enabling technology

Jigsaw images courtesy of jscreations, Salvatore Vuono at FreeDigitalPhotos.net
Roadway Research

• 1: Traditional aggregated data modelling from roadway segments, generalizing where crashes occurred: heuristics & goal setting

• 2: Real-time data collection to identify where a crash is likely to occur: e.g. used for moderating risk by speed reduction (Abdel-Aty & Pande, 2007)
Roadway Research

• 1: Traditional aggregated data modelling from roadway segments, **generalizing** where **crashes occurred**: heuristics & goal setting

• 2: **Real-time** data collection to identify where a crash is **likely to occur**: e.g. used for moderating risk by speed reduction (Abdel-Aty & Pande, 2007)

• **Our goal**: modifying the traditional approach for rapid **system-wide assessment** for design and operational decision-support.
Emerging roadway decision support

A risk-based system, the Australian National Risk Assessment Model:

• has 30+ generalized independent models (Austroads, 2014, Austroads, 2015)

• model deployment difficulty in matching roadway to the model, resulting in estimate inaccuracy.
A research question

• MRE: aims to provide an integrated, risk-based, enterprise-wide model for decision support in road asset management.

• ...where the research question is: given the inherent uncertainty of any model (Saltelli, 2007), can MRE’s accuracy outperform, and exceed the utility and economy of the emerging highly parameterized, multiple-model methodology.
Section 2: MRE

The Model Rule Extrapolation (MRE) research stream: realizing value through machine-learning with enterprise-wide modelling and deployment

Image courtesy of Salvatore Vuono at FreeDigitalPhotos.net
Inspiration

• Applying the model to make a prediction gave no value to risk abatement strategies,
• The effect of changing variables was sought.
• Thus, testing progressive change in the variable of interest on the outcome of the risk estimate was trialled, hence, an extrapolation-based scenario.

• Different to MARS;
  – allowing analysis of heterogeneous data;
  – producing multiple curves from the data.
MRE analytical objectives

To provide a decision-support tool that
• is trained on enterprise-wide data,
• produces an individual risk curve to automatically assess each instance,
• provide simulative what-if capability to optimize each instance.
Target areas of our analysis

• **Judge on latent risk**: rather than expressed risk,

• **Utilize observational data**: integrate the homogeneous pockets of roadway data, in single cohesive dataset, (Miaou & Lum, 1993, Abdel-Aty & Pande, 2007)

• **Overcome missing values**: overcome this significant barrier to enterprise-wide analysis,

• **Problem of the controlled experiment**: overcome the isolation of controlled variables, exacerbating the lack of understanding among variables, as described (Cairney, 2008).
Selecting an algorithm

- Numerically predictive capable of showing variable contributions
Algorithm requirements

Select an algorithm that:

• predicts numerical data;

• the output provides rules rather than individual point predictions.

• manage data that originates from diverse stochastically-generating processes and may not be well represented by a single rule or even a single tree
Numerically predictive algorithm for multi-class (heterogeneous) data

<table>
<thead>
<tr>
<th>Output</th>
<th>Single expression</th>
<th>Single structure Multiple expression</th>
<th>Multiple structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calculated value</td>
<td></td>
<td>Regression Tree: CART (Breiman et.al., 1984), REPTree (Pentaho, 2013), Neural Network (Minski, et.al. 1969, Matuško, et. al., 2008)</td>
<td>Random Forests (Breiman et.al., 1984)</td>
</tr>
</tbody>
</table>
Algorithm performance on the heterogeneous crash data

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Correlation coefficient (r)</th>
<th>Min Predicted value of 1</th>
<th>Max predicted value of 100</th>
<th>Rules/nodes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Support vector machines</td>
<td>0.67</td>
<td>-8.5</td>
<td>35.2</td>
<td>1</td>
</tr>
<tr>
<td>Linear Regression</td>
<td>0.71</td>
<td>-18</td>
<td>51.2</td>
<td>1</td>
</tr>
<tr>
<td>Neural Network</td>
<td>0.71</td>
<td>-44</td>
<td>131.8</td>
<td>46</td>
</tr>
<tr>
<td><strong>M5Rules</strong></td>
<td><strong>0.93</strong></td>
<td><strong>-0.86</strong></td>
<td><strong>98.5</strong></td>
<td><strong>138</strong></td>
</tr>
<tr>
<td>Decision Table</td>
<td>0.95</td>
<td>1</td>
<td>100</td>
<td>4416</td>
</tr>
<tr>
<td>M5P</td>
<td>0.95</td>
<td>-0.86</td>
<td>97.9</td>
<td>1187</td>
</tr>
<tr>
<td>Nearest Neighbor (Kstar)</td>
<td>0.98</td>
<td>1</td>
<td>100</td>
<td>-</td>
</tr>
</tbody>
</table>

**M5Rules**: fewest rules, each highly information laden
M5Rules Process Flow

A-all instances as tempInstances
M5Rules Process Flow

A-all instances as tempInstances

B- loop & build tree
M5Rules Process Flow

A-all instances as tempInstances
B- loop & build tree
C- select best rule and save
M5Rules Process Flow

A - all instances as tempInstances
B - loop & build tree
C - select best rule and save
D - dump instances covered by the rule
M5Rules Process Flow

A- all instances as tempInstances
B- loop & build tree
C- select best rule and save
D- dump instances covered by the rule
E- non covered instances become tempInstances
M5Rules Process Flow
(from M5Rules code, (Pentaho, 2015))

A- all instances as tempInstances
B- loop & build tree
C- select best rule and save
D- dump instances covered by the rule
E- non covered instances become tempInstances
F Loop while instances remain
The M5Rules model

M5Rules model

R1 precedent consequent
R2 precedent consequent
...
Rn-1 precedent consequent
Rn constant

R2

IF

lane_count > 2.476
CRASH_SPEED_LIMIT <= 75
lane_count > 3.3
AADT > 19087
AVG_FRICTION_AT_60_1km <= 0.235

THEN

CrashCount_4yrTotal_1km =
- 8.9152 * seal_type=Cement
  Concrete, Open Graded Asphalt,
- 4.3939 * seal_age
+ 4.9779 * Text_Depth_SPTD_OWP
- 38.0987 * AVG_FRICTION_AT_60_1km
+129.8761 [104:24.083%]
M5Rules modelling results $r$-sq 0.9
Model rules extrapolation (MRE)

• The MRE method
Fundamentals of the method

- Produces a **model of linear rules** from the heterogeneous data;
- Generate a risk curve by model deployment across an **extrapolated range of the variable of interest** (i.e. skid resistance);
- Avoid the missing-values problem by **non-requirement** of observed value of **the variable of interest** during deployment;
- **Optimizing** the **variable of interest** on the risk curve, and **modelling changes to all risk factors** for risk reduction.
The MRE Method

With all sealed road-crash instances
The MRE Method

With all sealed road-crash instances

A1-Sample non-null instances for training
The MRE Method

With all sealed road crash instances
A1-Sample non-null instances for training
A2- Train the model
The MRE Method

With all sealed road-crash instances
A1-Sample non-null instances for training
A2- Train the model
B1-Create test range

\[ P: \text{Skid resistance test range (for } a_k) \]
\[ 0.15 \]
\[ 0.20 \]
\[ 0.25 \]
\[ 0.30 \]
\[ 0.35 \]
\[ 0.40 \]
\[ 0.45 \]
\[ 0.50 \]
\[ 0.55 \]
\[ 0.60 \]
\[ 0.65 \]
\( (p = 11) \)
The MRE Method

With all sealed road-crash instances
A1-Sample non-null instances for training
A2- Train the model
B1-Create test range
B2-Replicate each instance with the test range set of values
The MRE Method

With all sealed road-crash instances
A1-Sample non-null instances for training
A2-Train the model
B1-Create test range
B2-Replicate each instance with the test range set of values
C-deploy the model on the replicates

\[ P: \text{Skid resistance test range (for } a_k) \]

\[
\begin{align*}
0.15 & \\
0.20 & \\
0.25 & \\
0.30 & \\
0.35 & \\
0.40 & \\
0.45 & \\
0.50 & \\
0.55 & \\
0.60 & \\
0.65 & (p = 11)
\end{align*}
\]
The MRE Method

With all sealed road-crash instances
A1-Sample non-null instances for training
A2-Train the model
B1-Create test range
B2-Replicate each instance with the test range set of values
C-Deploy the model on the replicates
D&E optimize and simulate
Deployment: populating an instance profile

\[
\text{for } j: 1 \text{ to } p \quad DP'_{ij}[\hat{a}_n] \leftarrow M(DP_{ij}[a_1, \ldots, (a_k)^j, \ldots, a_{(n-1)}])
\]

<table>
<thead>
<tr>
<th>j</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>((a_k)^j)</td>
<td>0.15</td>
<td>0.20</td>
<td>0.25</td>
<td>0.30</td>
<td>0.35</td>
<td>0.40</td>
<td>0.45</td>
<td>0.50</td>
<td>0.55</td>
<td>0.60</td>
<td>0.65</td>
</tr>
<tr>
<td>(\hat{a}_n)</td>
<td>45.8</td>
<td>44.2</td>
<td>42.6</td>
<td>40.4</td>
<td>37.2</td>
<td>27.8</td>
<td>22</td>
<td>20.5</td>
<td>18.9</td>
<td>17.3</td>
<td>15.7</td>
</tr>
</tbody>
</table>

Rule 1
Rule 2
Rule 3
Rule 4
the MRE profile risk profile curve

- The risk curve
the MRE profile risk profile curve

• Intersect the crash rate with the curve
the MRE profile risk profile curve

- Estimate an optimal crash rate and its skid resistance
Working with the MRE profile

• The risk curve supporting engineering-based decision-making.
Curve samples

(Emerson, Weligamage, Nayak, 2013)
Application Prototype

Application Functionality

A
1.1 Select model and dataset

B
2.1 Alter Attribute Values

C
3.1 Interactive crash count generation

D
3.2 Interactive road road segment profile generation

E
3.3 Selection of road segments for skid resistance treatment from whole dataset

Start

End

Navigation
Sample Predictions

Estimate is based on historic data: each of these roads has similar features to a set of roads with the average crash count of the predicted value.

2 lane highway, 100 km/h, F60 of 4.1: 3 → 2.8

4 lane highway, 110 km/h, F60 of 4.1: 12 → 8.6

4 lane highway, 60 km/h, F60 of 0.23: 100 → 91.1
Sample Curves

High risk mitigated with an increase in skid resistance

Approaching high risk

(note: these curves are produced using regression tree, not a model tree or rules learner, thus no intra-rule gradient)
Perform *what-if analysis*

- Change one or more risk factors
...changes the risk curve

• thus allowing the effect of decisions to be modelled
Testing multiple changes

Change: skid resistance from 0.35 to 0.4, Crash count estimate of: $15 \rightarrow 11.6$

Change: skid resistance from 0.35 to 0.4 carriageway type and single to dual: crash count estimate of: $15 \rightarrow 6$
Benefits

• **A single model** managed for the enterprise-wide data,

• **Rapid machine-learned rules** from complex heterogeneous data, **beyond human capability,**

• The profile provides the **environment of the controlled experiment,**

• **Prescriptive,** allowing modelling of risk factors.

• **Missing values managed** in the variable of interest no longer block enterprise-wide analysis.
Section 4: Future Work
Curve uncertainty analysis

Requires M5Rules reporting upgrade: for export of point predictions (rule_Id, instance_Id) for model uncertainty and sensitivity analysis (Saltelli, 2005)
System uncertainty vs. modifications

• Instance and system uncertainty will provide a metric to assess the success of optimization efforts:
- where a reduction of individual curve variance is sought.
Future work continued

• Model development & the mechanism of empty extrapolation;
• Assess and develop **domain utility** including enterprise-wide cost modelling;
• Test better **quality data** and subsequent **site testing**;
• Accommodate **streaming data** and **functionalize**;
• Trial **Domain case studies**: education, insurance, banking, medical.
Acknowledgement

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

- Modelling was performed with algorithms from Weka, SAS Enterprise Miner and Salford Systems.
Questions?

With Special thanks to:
Connect with other IAPA members

www.iapa.org.au

Stay current | Education

Industry recognised courses

www.iapa.org.au/courses

Stay involved | Events

Industry connected experts

www.iapa.org.au/events
Post-presentation slides

Question support and references
Section 5

The analytical framework
An industry DM framework

• The DM study of the CIEAM project (2009-2012) began with no glimpse of the final direction.

• Progress was based on an uncertain, step by step knowledge building process of data mining and statistical methods looking for pockets of concordance among results and literature:
  – accepting & associating pockets in confluence,
  – and on evidence, sidelining discordant knowledge.

An industry-accepted framework was sought.
The industry DM process guide

• Cross Industry Standard Processes for Data Mining (CRISP-DM) (Chapman, 2000)
  – Business understanding,
  – Data understanding,
  – Data preparation,
  – Modelling,
  – Evaluation
  – Deployment

• The method provided the required process guide,

• but without information theoretic processes or reference to “knowledge discovery”
The Informational Paradigm  \cite{Coppi2002}(Coppi, 2002)

\[ I(P) \rightarrow I(P, B, S, L) \]

- **I**: the selected information function, **P**: prior knowledge,
- **B**: “an Empirical Basis, broadening the unusual notion of “Statistical Data”, including the more general types of complex information such as the databases (& data types) considered in DM”, managed a **Strategy of Analysis** and reliant on **external validation**,
- **S**: the Structure (i.e. model to be derived) from the data reflecting the underlying stochastic processes,
- **Links**: evaluating concordance/discordance between the components,
Capitalizing on the synergies
The Analytical Toolbox

• **Descriptive analytics**: summaries, averages, groups, correlation, cubes, visualization (Haas, 2011);

• **Statistical Tools** *(maximum likelihood)*
  – frequentist (Fisher, Neyman-Pearson & Wald)
  – Bayesian; (Coppi, 2002)

• **Data mining** with learning machines discovering the structure in existing data and new data types not meeting the classical assumptions:
  – Classification, Regression, density functions, neural networks, clustering, association mining, dimensional reduction etc (Coppi, 2002)
The *Structural Data Analysis* approach

Useful DM problem specification (Kruskal /Benzecri) to populate the IP (Coppi, 2002):

- the Data: $D$, the Structure: $S$ discovered from by an appropriate data analysis, representing the underlying stochastic processes.
- The operating function: $f$ acting on the data via the structural model to allow reconstruction via $f(S)$: $D \equiv f(S, \text{non random noise, random noise})$,
- Objective (loss) and optimization functions,
- Using probabilistic features and measures of confidence to qualify findings.
Questions of the *strategy of analysis*

• Do the crash-rate of roads generally maintain a consistency of value from year to year?
• Is there a relationship between skid resistance and crash?
• Does the data-fitted model have the ability to predict the crash rate target?
• Does the model apply well over the whole data?
• If yes to all: then proceed!
Section 6

Data understanding, and data transformation
Characterizing the 1 km stretch with data

Crash rate target
Surface properties (skid resistance)
Geometry
Roadway design
Features
Furniture
Wear & damage
Traffic
Weather
Adding variables showing increase in information lift

(Nayak, et al., 2010)

<table>
<thead>
<tr>
<th>Model</th>
<th>Variables</th>
<th>Splitting Criteria</th>
<th>Validation Method</th>
<th>Tree Dimension width depth</th>
<th>R Squared</th>
<th>Leaves</th>
<th>Dominant Road Attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>40_4_1a</td>
<td>AVG_F60</td>
<td>F.Test</td>
<td>Validation</td>
<td>5w*6d</td>
<td>0.201393</td>
<td>54</td>
<td>Avg_F60</td>
</tr>
<tr>
<td>40_4_3a</td>
<td>Add intersection</td>
<td>F.Test</td>
<td>Validation</td>
<td>5w*6d</td>
<td>0.207609</td>
<td>41</td>
<td>Avg_F60</td>
</tr>
<tr>
<td>40_4_4a</td>
<td>Add SEAL TYPE</td>
<td>F.Test</td>
<td>Validation</td>
<td>5w*6d</td>
<td>0.387185</td>
<td>53</td>
<td>SEAL TYPE Avg_F60</td>
</tr>
<tr>
<td>40_4_5a</td>
<td>Add SEAL AGE</td>
<td>F.Test</td>
<td>Validation</td>
<td>5w*6d</td>
<td>0.427332</td>
<td>89</td>
<td>SEAL TYPE Avg_F60 SEAL AGE</td>
</tr>
<tr>
<td>40_4_6a</td>
<td>Add TEXTURE DEPTH</td>
<td>F.Test</td>
<td>Validation</td>
<td>5w*6d</td>
<td>0.409653</td>
<td>93</td>
<td>1: SEAL TYPE Avg_F60 SEAL AGE</td>
</tr>
<tr>
<td>40_6_7a</td>
<td>Add ROAD_TYPE</td>
<td>F.Test</td>
<td>Validation</td>
<td>5w*6d</td>
<td>0.419679</td>
<td>105</td>
<td>SPEED_LIMIT SEAL TYPE</td>
</tr>
<tr>
<td>40_4_8a</td>
<td>Add SPEED LIMIT</td>
<td>F.Test</td>
<td>Validation</td>
<td>5w*6d</td>
<td>0.528407</td>
<td>97</td>
<td>SPEED_LIMIT SEAL TYPE</td>
</tr>
<tr>
<td>40_4_9a</td>
<td>Add LANE COUNT</td>
<td>F.Test</td>
<td>Validation</td>
<td>5w*6d</td>
<td>0.529152</td>
<td>92</td>
<td>SPEED_LIMIT SEAL TYPE</td>
</tr>
<tr>
<td>40_4_10a</td>
<td>Add PAVEMENT WIDTH</td>
<td>F.Test</td>
<td>Validation</td>
<td>5w*6d</td>
<td>0.576699</td>
<td>84</td>
<td>SPEED_LIMIT SEAL TYPE PAV_WIDTH</td>
</tr>
<tr>
<td>40_4_11a</td>
<td>Add ROUGHNESS AVERAGE</td>
<td>F.Test</td>
<td>Validation</td>
<td>5w*6d</td>
<td>0.576836</td>
<td>80</td>
<td>SPEED_LIMIT SEAL TYPE ROUGH_AVG PAV_WIDTH</td>
</tr>
<tr>
<td>40_4_12a</td>
<td>Add RUTTING AVERAGE</td>
<td>F.Test</td>
<td>Validation</td>
<td>5w*6d</td>
<td>0.577703</td>
<td>79</td>
<td>SPEED_LIMIT SEAL TYPE ROUGH_AVG PAV_WIDTH</td>
</tr>
<tr>
<td>40_4_13a</td>
<td>ADD CRACKING TYPES</td>
<td>F.Test</td>
<td>Validation</td>
<td>5w*6d</td>
<td>0.57984</td>
<td>82</td>
<td>SPEED_LIMIT SEAL TYPE ROUGH_AVG PAV_WIDTH</td>
</tr>
<tr>
<td>40_4_14a</td>
<td>Add AADT</td>
<td>F.Test</td>
<td>Validation</td>
<td>5w*6d</td>
<td>0.690838</td>
<td>101</td>
<td>AADT SPEED LIMIT</td>
</tr>
</tbody>
</table>

(Nayak, et al., 2010)
Crash prone-roads generally show more than 8 crashes per four years, whereas below that level the roads are more like non-crash roads (Nayak, et al., 2011)
Optimizing the crash rate target

Optimized intervals of the target (Fridstrøm, et al., 1995) maximize information available. In our data, domain intervals provide crash rate averages that are consistent, and statistically different, thus, suitable for analysis.
Skid resistance/crash relationship

(Emerson, 2013)
Assessing model goodness

- Assessing a Salford Systems CART model by comparing models with original data vs. confounded data.

B: Error rate during model build with randomized order of variable values

A: Error rate during model build with original data

(Emerson, 2013)
Section 7

Assessing the deployment
Applying the model to the replicates

\[ (a_k)^j \]

\[ \hat{a}_n \]

(Emerson, Weligamage, Nayak, 2013)
Quality measure 1:

External validation: expected similarity of MRE to manual method based on direct measurement.

(Piyatrapoomi, Weligamage, Turner, 2009)
Quality measure 2:

Expect sigmoid-like decreasing monotonic curves:

(Emerson, Weligamage, Nayak, 2013)
Quality measure 3:

Expect negative gradients reflecting skid resistance vs. crash rate relationship

(Emerson, Weligamage, Nayak, 2013)
Quality measure 4:

Expect proximity of the crash point to the curve

(Emerson, Weligamage, Nayak, 2013)
Quality measure 5:

Expectation of transformed skid resistance: ordered, normal and heightened.

(Emerson, Weligamage, Nayak, 2013)
Quality measure 6:

- expectation of transformed crash rate: ordered, normally distributed and depressed.

(Emerson, Weligamage, Nayak, 2013)
Summary of modeled improvement

<table>
<thead>
<tr>
<th></th>
<th>min</th>
<th>Q1</th>
<th>mean</th>
<th>Q3</th>
<th>max</th>
</tr>
</thead>
<tbody>
<tr>
<td>measured crash rate</td>
<td>1</td>
<td>3</td>
<td>13.0</td>
<td>19</td>
<td>100</td>
</tr>
<tr>
<td>predicted optimal crash</td>
<td>-1.3</td>
<td>3.3</td>
<td>7.4</td>
<td>10.3</td>
<td>30.8</td>
</tr>
<tr>
<td>rate</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Requiring assessment in the engineering, information and organizational contexts

(Emerson, Weligamage, Nayak, 2013)
Independent geneses of the method

• Dec 2010: Experimental results of extrapolated SAS regression tree model reported at the CIEAM Skid Resistance Meeting.
• Jun 2012: Final CIEAM report details the method.
• 2012: Rejection of initial KDD 12 submission.
• 2012: Dan Steinberg blogs of the possibility of applying Salford’s CART in an extrapolated framework (Steinberg, 2012).
References


References


References


