MML Mixture Models of Heterogeneous Poisson Processes with Uniform Outliers for Bridge Deterioration

T. Maheswaran[1], J.G. Sanjayan[2], David L. Dowe[3] and Peter J. Tan[3]

[1] VicRoads, Metro South East Region, 12 Lakeside Drive, Burwood East, Vic 3151, (Previously Department of Civil Engineering, Monash University when the research presented in this paper was carried out), Mahes.Maheswaran@roads.vic.gov.au
[2] Department of Civil Engineering, Monash University, Clayton, Vic 3800, Australia, jay.sanjayan@eng.monash.edu.au
[3] School of Computer Science and Software Engineering, Clayton School of I.T., Monash University, Clayton, Vic 3800, Australia, dld@bruce.csse.monash.edu.au and Peter.Jing.Tan@infotech.monash.edu.au

Abstract. Effectiveness of maintenance programs of existing concrete bridges is highly dependent on the accuracy of the deterioration parameters utilised in the asset management models of the bridge assets. In this paper, bridge deterioration is modelled using non-homogenous Poisson processes, since deterioration of reinforced concrete bridges involves multiple processes. Minimum Message Length (MML) is used to infer the parameters for the model. MML is a statistically invariant Bayesian point estimation technique that is statistically consistent and efficient. In this paper, a method is demonstrated estimate the decay-rates in non-homogeneous Poisson processes using MML inference. The application of methodology is illustrated using bridge inspection data from road authorities. Bridge inspection data are well known for their high level of scatter. An effective and rational MML-based methodology to weed out the outliers is presented as part of the inference.

1 Introduction

Bridge asset management is an emerging concept in road authorities. Bridge management is a systematic process of maintaining, upgrading and operating bridge assets cost-effectively. It combines engineering principles with sound business practices and economic theory, and it provides tools to facilitate a more logical approach to decision-making. Thus, bridge asset management provides a framework for handling both short-and long-term planning. As defined by the American Public Works Association Asset Management Task Force, asset management is "… a methodology needed by those who are responsible for efficiently allocating generally insufficient funds amongst valid and competing needs."

Asset management of bridges has come of age because of (1) an increase in allowable truck loads of bridges, (2) changes in public expectations, and more importantly (3) extraordinary advances in information technology and data-mining. Currently, bridge investment and maintenance decisions are based on tradition, intuition, personal experience, resource availability, and political considerations, with systematic application of objective analytical techniques applied to a lesser degree. Many road authorities limit application of their management systems to monitoring conditions and then plan and program their projects on a "worst first" basis.

The deterioration of a reinforced concrete element is not a homogeneous process. It involves chloride ingress, corrosion initiation, crack initiation and crack propagation stages. Therefore, a multi-stage, non-homogeneous Poisson process with multiple deterioration rates to capture the entire phenomenon of concrete bridge deterioration is adopted. The application of the methodology is illustrated using real-life bridge inspection data from VicRoads. VicRoads is a state government owned agency responsible for the maintenance and management of bridges on state highways and main roads in Victoria, Australia. In Victoria alone, more than $50 million per year is spent on the maintenance and upgrade of bridges valued at more than $6 billion.

Bridge inspection data used in this paper for modelling deterioration is based on Level 2 inspections according to VicRoads [9]. Level 2 inspections are managed on a state-wide basis to assess the condition state of each structure and its main components. The frequency of inspection varies between 2 and 5 years depending on bridge rating. The bridge element condition state is described on a scale of 1 to 4, where 1 stands for "excellent condition" and 4 stands for "serious deterioration". The inspector records the condition states of the bridge element and the percentage of that element in a particular condition state.

2 Non-Homogeneous Poisson Process

The non-homogeneous or non-stationary Poisson process is a process where the arrival rate, $r(t)$ at time t, is a function of t. The counting process $\{N(t), t \geq 0\}$ is said to be a non-homogeneous Poisson process with intensity function $r(t)$, $t \geq 0$, if

(i) $N(0) = 0$; (ii) The process has independent increments; (iii) $P\{N(t+h) - N(t) \geq 2\} = o(h)$; (iv) $P\{N(t+h) - N(t) = 1\} = r(t)h + o(h)$

Let $m(t) = \int_0^t r(s)\,ds$. Then it can be shown that the probability of n parts moving from condition state i to state $i+1$ can be expressed by Equation (1).

$$P\{N(t+s) - N(s) = n\} = e^{-[m(t+s)-m(t)]} \frac{[m(t+s) - m(t)]^n}{n!} \quad n \geq 0 \qquad \textbf{(1)}$$

3 Minimum Message Length

The Minimum Message Length (MML) principle [12][16][14][3][2][11] is widely used for model selection in various machine learning, statistical and econometric problems [12][16][14][13][15][10][1][2][5][8][11][17] and references therein. The principle is that the best theory for a body of data is the one that minimises the size of the theory plus the amount of information necessary to specify the exceptions relative to the theory.

A Bayesian interpretation of the MML principle is that it variously states that the best conclusion to draw from the data is the theory with the highest posterior probability or, equivalently, that theory which maximises the product of the prior probability of the theory with the probability of the data occurring in light of that theory. For a hypothesis (or theory), H, with prior probability $\Pr(H)$ and data, D, the relationship between the probabilities can be written [4][15] as shown in the Equation (2) by application of Bayes's Theorem.

$$\Pr(H \,\&\, D) = \Pr(H) \cdot \Pr(D \mid H) = \Pr(D) \cdot \Pr(H \mid D) \ . \qquad (\mathbf{2})$$

Equation (3) can be derived by re-arranging Equation (2).

$$\Pr(H \mid D) = \Pr(H) \cdot \Pr(D \mid H) / \Pr(D) \ . \qquad (\mathbf{3})$$

Since D and $\Pr(D)$ are given and H needs to be inferred, the problem of maximising the posterior probability, $\Pr(H \mid D)$, can be regarded as the one of choosing H so as to maximise $\Pr(H) \cdot \Pr(D \mid H)$. Elementary coding theory tells us that an event of probability, p, can be coded by a message length $l = -\log(p)$. So, the length of a two-part message $(MessLen)$ conveying the parameter estimates based on some prior and the data encoded based on these estimates can be given as in Equation (4). In this paper, natural logarithms are used and the message lengths are in nits.

$$MessLen(H \,\&\, D) = -\log(\Pr(H)) - \log(\Pr(D \mid H)) \ . \qquad (\mathbf{4})$$

Since $-\log(\Pr(H) \cdot \Pr(D \mid H)) = -\log(\Pr(H)) - \log(\Pr(D \mid H))$, maximising the posterior probability, $\Pr(H \mid D)$, is equivalent to minimising $MessLen(H \,\&\, D)$ given in Equation (4). The receiver of such a hypothetical message must be able to decode the data without using any other knowledge. The model with the shortest two-part message length is considered to give the best explanation of the data. For a discussion of the relationship of the works of Solomonoff, Kolmogorov and Chaitin with MML and the subsequent Minimum Description Length (MDL) principle [7], see [14] and [2].

The Poisson distribution is used in this paper for modelling bridge element deterioration. Let *r* be the rate at which the number of parts in a bridge element moving from condition state i to $i+1$, t_i be the length of the time interval and c_i be the number of parts moved in that time interval. In order to infer the rate of the process [13][15], first a Bayesian prior density on r is required. Let this prior be $h(r) = (e^{-r/\alpha}) / \alpha$ for some α. The message length can be expressed as in Equation (5), where L is the (negative) log-likelihood and F is the Fisher information [15][16][14][13].

$$\begin{aligned} MessLen &= -\log(h) + L + \frac{1}{2}\log(F) + \frac{1}{2}\big(1-\log(12)\big) \\ &= \log(\alpha) + \frac{r}{\alpha} + r\sum_{i=1}^{N} t_i - \log(r)\sum_{i=1}^{N} c_i - \sum_{i=1}^{N} c_i \log(t_i) + \sum_{i=1}^{N}\log(c_i!) \\ &\qquad -\frac{1}{2}\log(r) + \frac{1}{2}\log\left(\sum_{i=1}^{N} t_i\right) + \frac{1}{2}\big(1-\log(12)\big) \end{aligned} \qquad (5)$$

We then estimate *r* by minimizing the message length [15] (Equation (6)).

$$\hat{r}_{MML} = \frac{C + \frac{1}{2}}{T + \frac{1}{\alpha}} \text{ where } C = \sum_{i=1}^{N} c_i \text{ and } T = \sum_{i=1}^{N} t_i \,. \qquad (6)$$

Cut-points can be found in many machine learning problems. They arise where there is a need to partition data into groups which are to be modelled distinctly [4][10]. A piece-wise function is used for partitioning of data and the rate of the Poisson process, r, is assumed to be constant in between the cut-points. The message length including the penalties for cutting the data into groups assuming a uniform prior can be roughly given as a first draft (see, e.g. [10]) by Equation (7):

$$MessLenCp = \log(n+1) + \log\left(\frac{n!}{(n-ncp)!\,ncp!}\right) + \sum_{i=1}^{ncp+1} MessLen(i) \qquad (7)$$

where $MessLenCp$ = Message length including penalties for cutting data; $MessLen(i)$ = Message length for data in between cut-point $(i-1)$ and cut-point (i) calculated using Equation (5); n = Maximum possible number of cut-points; and ncp = number of cut-points.

Note that it costs $\log_e(n+1)$ nits to specify ncp because $0 \le ncp \le n$. Letting cut-point $ncp+1$ refer to the end of the data, $MessLen(ncp+1)$ refers to the data after the ncp^{th} (i.e., last) cut.

A separate code-word of some length can be set aside for missing data. The transmission of the missing data will be of constant length regardless of the hypothesis classification, and as such will affect neither the minimisation of the message nor the (statistical) inference (Section 2.5 of [15] and Section 5, p.42 of [13]).

3.1 Models of Outliers and Multi-State Distribution

The bridge inspection data is from visual inspections, and the inspectors employed to gather data varied in their experience. Factors such as visibility at the time of inspection and resources available for the inspectors to carry out the bridge inspections, etc. may also affect the reliability of the data. A comprehensive study [6] on visual bridge inspection data concluded that significant measurement errors exist. Further, the measurement errors may show a seeming improvement in bridge element conditions (which is a physical impossibility) or give rather unusual data. Therefore, the data has to be screened or have an explicit model of outliers in order to take account of those errors before modelling the data.

If there are $n(t)$ members of class t $(t = 1,2,...,T)$ then the label used in the description of a thing to say that it belongs to class t will occur $n(t)$ times in the total message. If the relative frequency of class t is estimated as $p(t)$ (where *p(1)*+*p(2)*+…+*p(T)* = 1) then the information needed to quote the class membership of all things is given in Equation (8) [12]:

$$-\sum_{t=1}^{T}\left(n(t)+\frac{1}{2}\right)\log p(t) \qquad \textbf{(8)}$$

It was decided to have two classes in our bridge deterioration modelling problem, known as outliers and non-outliers. An outlier is a datum which is considered to be erroneous and therefore not used in the estimation of Poisson rate. However, there is a cost to classifying a datum as an outlier. The message length for the outlier data points is given – ignoring partial assignment (sec. 3.2 of [13] and sec. 4.2 of [15]) - by

$$MessLenOl = N_{ol}\left(-\log(q)+\log(H+1)\right) \qquad \textbf{(9)}$$

where $MessLenOl$ = Message length from (uniform) outlier data; N_{ol} = Number of data in outlier class; q = Frequency of outlier class = $\dfrac{N_{ol}+\frac{1}{2}}{N+1}$ (see Equation (12));
$H+1$ = Maximum possible values a data point can take (100+1= 101);
N = Total number of data. These two terms in Equation (9) are so because each outlier must be encoded as an outlier and then its value is encoded as being uniformly equally likely from (H+1) different values.

The message length for non-outlier data points is similarly given by

$$MessLenNol = N_{Nol}\left(-\log(1-q)\right)+MessLenCp \qquad \textbf{(10)}$$

where $MessLenNol$ = Message length from non-outlier data; N_{Nol} = Number of data in non-outlier class = $N - N_{ol}$; and $MessLenCp$ is from Equation (7). The message length for all data points is then given – slightly inefficiently (sec. 4.1 of [15]) – by

$$MessLenMo = MessLenNol + MessLenOl \qquad \textbf{(11)}$$

Multi-State Distribution

For a multi-state distribution with M states, a uniform prior, $h(p)=(M-1)!$ is assumed over the $(M-1)$-dimensional region of hyper-volume $1/(M-1)!$ given by

$p_1 + p_2 + ... + p_M = 1; \quad p_i \geq 0$. Letting n_m be the number of things in state m and $N = n_1 + n_2 + ... + n_M$, minimising the message length Equation (5) gives that the MML estimate $\hat{p}_m$ of p_m is given by [13][15][12][16]:

$$\hat{p}_m = (n_m + 1/2)/(N + M/2) \qquad \textbf{(12)}$$

Substituting Equation (12) into the message length Equation (5) gives rise to a (minimum) message length shown in the Equation (13) for both stating the parameter estimates and then encoding the things in light of these parameter estimates [12],[15].

$$MessLenMs = \frac{M-1}{2}\left(\log\left(\frac{N}{12}\right)+1\right) - \log(M-1)! - \sum_{i=1}^{M}\left(n_i + \frac{1}{2}\right)\log\left(p_i\right) \qquad \textbf{(13)}$$

In this case, the distribution is binomial with M=2 in Equation (13). So, approximating partial assignment with total assignment [13][15], the total message length - including the cost of cut-points, modelling outliers and multi-state variables - is given by Equation (14).

$$MessLenT = MessLenMo + MessLenMs \qquad \textbf{(14)}$$

4 Application of Methodology

4.1 Bridge Inspection Data

There are four condition states defined in the bridge inspection data. The deterioration process of a bridge element can be treated as three separate Poisson processes as defined below (there can be deterioration observed, but we assume not improvement; and we also assume Cs13=Cs14=Cs24=0):

- Process Cs12: parts of element deteriorate from condition state 1 to state 2
- Process Cs23: parts of element deteriorate from condition state 2 to state 3
- Process Cs34: parts of element deteriorate from condition state 3 to state 4

It is assumed that the bridge element was new when constructed, and the initial condition state is assumed to be $pc1i$ = 100, $pc2i$ = 0, $pc3i$ = 0, and $pc4i$ = 0 at time 0. Since the bridge inspections are recorded in percentages, it is assumed that there are 100 parts in an element. $Cs12$, $Cs23$ and $Cs34$ during the period between the ith and (i+1)th inspections can be calculated from Equations (15) to (17).

$$Cs12 = pc1\,i - pc1\,f \qquad \textbf{(15)}$$

$$Cs34 = pc4\,f - pc4\,i \qquad \textbf{(16)}$$

$$\begin{aligned} Cs23 &= (pc3\,f - pc3\,i) + (pc4\,f - pc4\,i) \\ &= (pc1\,i - pc1\,f) - (pc2\,f - pc2\,i) \end{aligned} \qquad \textbf{(17)}$$

where $pc1i$, $pc2i$, $pc3i$, $pc4i$ = number of parts in condition state 1, 2, 3 and 4 respectively at the i^{th} inspection; $pc1f$, $pc2f$, $pc3f$, $pc4f$ = number of parts in condition state 1, 2, 3 and 4 respectively at the $(i+1)^{th}$ inspection.

The bridge element considered in this study is precast concrete deck/slab (element number: 8P) in the most aggressive environment [9]. This element includes all precast concrete deck slabs and superstructure units forming the span and the deck of a bridge. Bridge inspection records (from 1996 to 2001) of 22 bridges were selected from the VicRoads database for the analysis. Fig. 1 shows the condition states of these selected bridges versus the ages (ranging from 0 to 39) of the bridges in years.

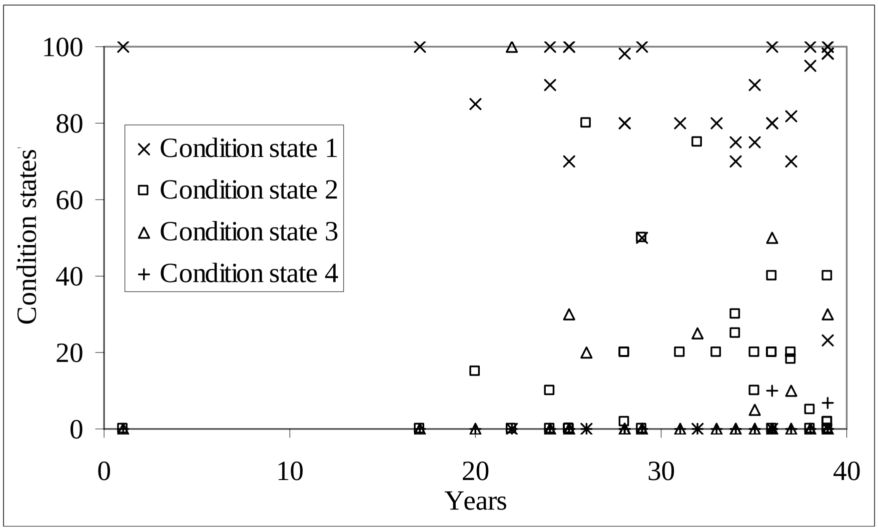

Fig. 1. Bridge element 8P in aggressive environment – condition states

4.2 Number and Location of Cut-Points and Estimation of Parameters

The minimum size of the time interval for the estimation of cut-points is assumed to be less than or equal to two years because the minimum time period for Level 2 inspections is in two-year cycles [9] and the deterioration expected for a bridge element within this time is relatively small.

The cut-points and the rates of the Poisson processes are estimated for each process by minimising the Message Length for each process separately. The multi-state (binary) distribution is used with both the Outlier model and the non-homogenous Poisson process model.

Fig. 2 shows the message lengths for various cut-points for Poisson process Cs12. The minimum message length of 165.24 nits was found to be with two cut-points (ncp=2) - hence three Poisson rates are estimated. Table 1 gives the cut-points followed by the Poisson rates of the processes. A closer examination of Poisson rates

r_1, r_2 and r_3 reveals that r_3 =0.041 could not occur unless there is an improvement in the condition states of the bridge element. An improvement in condition states of bridge element can occur by carrying out repair works or by measurement errors. A closer look at the bridge inspection data revealed that this is most probably the improvement works carried out to bridge elements (after 37 years of service) rather than measurement errors. But, this conclusion cannot be confirmed, since the data available is for 39 years only. We therefore decided to exclude the Poisson rate r_3 from further calculations.

Table 1. Poisson rates and cut-points for Cs12

<table>
<tr><td>Cut number (i)</td><td>Start</td><td colspan="2">1</td><td colspan="2">2</td><td>End</td></tr>
<tr><td>Cut location cp_i (years)</td><td>0</td><td colspan="2">29</td><td colspan="2">37</td><td>39</td></tr>
<tr><td>Range between cut-points</td><td colspan="2">&lt;1</td><td colspan="2">1 - 2</td><td colspan="2">&gt;2</td></tr>
<tr><td>Rates r_i</td><td colspan="2">0.017</td><td colspan="2">0.664</td><td colspan="2">0.041</td></tr>
<tr><td>No. of Data in Outlier</td><td colspan="2">8</td><td colspan="2">4</td><td colspan="2">1</td></tr>
</table>

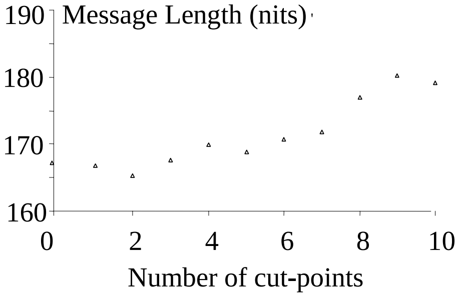

Fig. 2. Variation of message length for Poisson process Cs12

The message lengths for the Poisson process Cs23 for various cut-points were estimated in similar manner. The minimum message length of 85.66 nits was estimated for this Poisson process with no cut-point and a (very small) rate of 0.0006.

Similarly, the message lengths for Poisson process Cs34 for various cut-points resulted in the minimum message length of 36.42 nits for no cut-point and an even smaller Poisson process rate of 0.0005.

The estimated Cs12, Cs23 and Cs34 values are used to calculate the distribution of condition states of the bridge element. Fig. 3 shows the deterioration model for the bridge element estimated from the bridge inspection data shown in Figure 1. The number of parts moved from condition state 3 to state 4 (Cs34) and state 2 to state 3 (Cs23) were estimated to be zero and therefore there are no parts in condition states 3 and 4 in Fig. 3.

5 Conclusions

Concrete bridge elements in aggressive environments considered in this paper are normally expected to have an initiation period of about 30 years during which no deterioration occurs. The fact that this is accurately inferred by our (heterogeneous Poisson with outliers) model adds confidence to this modelling process adopted here.

Deterioration models for predicting the distribution of future condition states of bridge elements are an essential part of a bridge asset management system. It has been shown in this paper that concrete deterioration can be modelled using a non-homogeneous Poisson process together with an application of MML inference to estimate the Poisson rates. The estimated cut-points and Poisson rates in turn are used for predicting the distribution of the future condition states of bridge elements.

Bridge inspection data contain measurement errors or highly erroneous data due to a range of reasons including inspector subjectivity. Past attempts to model the data indicated that finding a structure in this type of data is very difficult. The methodology illustrated using bridge inspection data in this paper gives an objective and reasonably accurate way to identify and exclude measurement errors in the data.

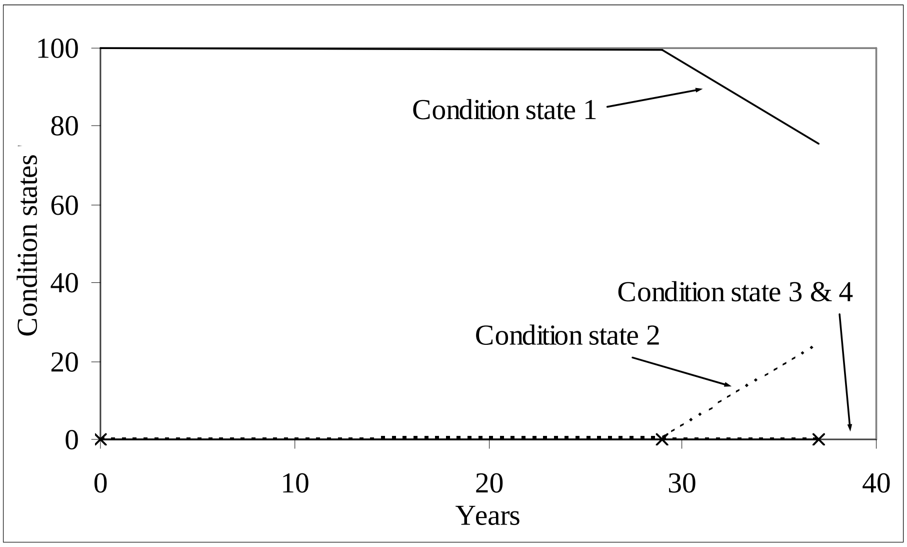

Fig. 3. Deterioration model for the bridge element 8P

Acknowledgement

The authors acknowledge Monash University's FACULTY OF ENGINEERING SMALL GRANTS SCHEME – 2003 for funding our project "Bridge Performance Modelling using Inspection Data and Data mining Techniques" for this research.

References

1. Comley, J.W. and Dowe, D.L.(2003), General Bayesian Networks and Asymmetric Languages, Proc. 2nd Hawaii International Conference on Statistics and Related Fields, 5-8 June, 2003.

2. Comley, J. W. and Dowe, D. L. (2005), Minimum Message Length and Generalized Bayesian Nets with Asymmetric Languages, in Advances in Minimum Description Length Theory and Applications, edited by P D Grunwald, I J Myung and M A Pitt, Chapter 11 (pp. 265-294), April 2005 (MIT Press: London).
3. Dowe, D.L. and Wallace, C.S. (1998), Kolmogorov complexity, minimum message length and inverse learning, abstract, page 144, 14th Australian Statistical Conference (ASC-14), Gold Coast, Qld, 6 - 10 July 1998.
4. Fitzgibbon, L. J., Allison, L. and Dowe, D. L. (2000), Minimum Message Length Grouping of Ordered Data, in 11th International Workshop on Algorithmic Learning Theory, 2000, LNAI 1968, Springer, Sydney, Australia, pp. 56 - 70.
5. Fitzgibbon, L.J., D. L. Dowe and F. Vahid (2004). Minimum Message Length Autoregressive Model Order Selection. In M. Palanaswami et al. (eds.), International Conference on Intelligent Sensing and Information Processing (ICISIP), Chennai, India, 4-7 January 2004. ISBN: 0-7803-8243-9, IEEE, pp. 439-444.
6. Moore, M., Phares, B., Graybeal, B., Rolander, D. and Washer, G. (2001), Reliability of Visual Inspection for Highway Bridges, Volume: 1 and 2, Final Report, Report No: FHWA-RD-01-020, NDE Validation Center, Office of Infrastructure Research and Development, Federal Highway Administration, McLean, VA, USA.
7. Rissanen, J. J. (1978), Modeling by Shortest Data Description, Automatica, 14, pp. 465 - 471.
8. Tan, P.J., and Dowe, D.L. (2004). MML Inference of Oblique Decision Trees, Proc. 17th Australian Joint Conf. on Artificial Intelligence (AI'04), Cairns, Australia, Dec. 2004, Lecture Notes in Artificial Intelligence (LNAI) 3339, Springer, pp1082-1088.
9. VicRoads (1995), VicRoads Bridge Inspection Manual, Melbourne, Australia.
10. Viswanathan, M., Wallace, C. S., Dowe, D. L. and Korb, K. B. (1999), Finding Cutpoints in Noisy Binary Sequences - A Revised Empirical Evaluation, Proc. 12th Australian Joint Conference on Artificial Intelligence, Lecture Notes in Artificial Intelligence (LNAI) 1747, Springer, Sydney, Australia, pp. 405 - 416.
11. Wallace, C. S. (2005), Statistical and inductive inference by minimum message length, Springer, Berlin, New York, ISBN 0-387-23795-X, 2005.
12. Wallace, C. S. and Boulton, D. M. (1968), An Information Measure for Classification, Computer Journal, 11, pp. 185 - 194.
13. Wallace, C. S. and Dowe, D. L. (1994), Intrinsic Classification by MML - the Snob Program, Proc. 7th Australian Joint Conference on Artificial Intelligence, UNE, World Scientific, Armidale, Australia, pp. 37 - 44.
14. Wallace, C. S. and Dowe, D. L. (1999), Minimum Message Length and Kolmogorov Complexity, Computer Journal, 42(4), pp. 270 - 283.
15. Wallace, C. S. and Dowe, D. L. (2000), MML Clustering of Multi-State, Poisson, von Mises Circular and Gaussian Distributions, Statistics and Computing 10, Jan. 2000, pp. 73 - 83.
16. Wallace, C. S. and Freeman, P. R. (1987), Estimation and Inference by Compact Coding, J. Royal Statistical Society (Series B), 49, pp. 240 - 252.
17. Dowe, D.L., S. Gardner and G.R. Oppy (2007+), "Bayes Not Bust! Why Simplicity is no problem for Bayesians". forthcoming, British J. Philosophy of Science.