Principal Component Analysis in Used Oil Data for Condition Based Maintenance Modeling

Burairah Hussin and Abdul Samad Shibghatullah
Fakulti Teknologi Maklumat dan Komunikasi
Universiti Teknikal Malaysia Melaka
Kurung Berkunci 1200, Hang Tuah Jaya, Ayer Keroh, 75450, Melaka
{buraizah|samad}@utem.edu.my

Abstract—**This paper reports on a study using the available oil monitoring information that is obtained from the Spectrometric Oil Analysis Programme (SOAP), to predict the residual life of a set of ship engines. The analysis of oil samples taken from an engine gives an indication of the suitability of the oil for continued use and provides important information about the condition of the engine. This could allow the identification of wearing components before severe failure could occur without dismantling the engine. Given this condition-monitoring data, maintenance decisions may be taken as required and, most importantly, maintenance may be done in an effective and efficient way. This paper starts with some analysis, assumptions and techniques necessary to gain insight into SOAP data that will be useful to our modelling development. Several issues regarding the consistency, incompleteness and dimensions of the data used are discussed, which include the implementation of principal component analysis technique. This research proposed an approach called the 'total metal concentrations' calculation, which is used to explain the relationship between the residual life and the total wear concentrations which are available from SOAP data. Once the 'clean' data is attained, the next modelling steps can be established to recommend the optimal maintenance actions in terms of cost, availability or any criterion of interest.**

Index Terms—**Principal Component Analysis, Condition Based Maintenance, Data Analysis, Data Mining.**

I. INTRODUCTION

The Spectrometric Oil Analysis Programme (SOAP) is a technique for identifying the elemental composition of particles up to approximately 10 microns (Edwards et al., 1998) entrained in machinery oil samples. The rationale behind this technique is that as mechanical components wear, they shed small metallic particles that become entrained in the oil. Furthermore, particles over 10 microns are likely to exit the oil circulation via some filtration and this could lead the small particles which are less prone to the filter to remain suspended within the engine. If this measure is obtainable it could provide an indication of machine condition (Edwards et al., 1998). Wear metals such as iron (Fe), aluminium (Al), chromium (Cr), copper (Cu), tin (Sn), lead (Pb), silver (Ag), titanium (Ti) and nickel (Ni) are measurable, as well as lubricant additives such as calcium (Ca), barium (Ba), zinc (Zn), phosphorus (P), magnesium (Mg), boron (Be) and molybdenum (Mo). Other contaminants such as silicon (Si), sodium (Na) and potassium (K) are also detectable. By running periodic sampling and testing, SOAP enables the observation of trends in the metal concentrations of the engine oil.

ISBN 978-967-5048-36-4

II. DATA COLLECTION

At present, we have a set of data from diesel engines used in ships. The dataset consists of the condition indicators obtained from observed SOAP data reveal that there are 28 element indicators which can be broken down into three categories: lubricant condition, contamination and metal concentrations. Lubricant condition assesses whether the oil itself is fit for further service or is ready for a change. Assessment of contaminants measures the dirt, water, etc., which could degrade the oil. Metal concentrations measures several wear particles that become entrained in the oil due to component wear. At every check, the oil sample is analysed and all the elements quantified as parts per million (ppm). Generally, if the quantity of any element is higher than the tolerable level, maintenance actions may be performed, such as repair or replacement of a component, or topping-up or changing the oil. The data for the SOAP analysis, from a third party company, was unsorted, contained missing values, was incomplete and inconsistent, and needed further explanations, as is common in practice (Ascher et al., 1995; Mathur et al., 2001). However, much effort has been put into understanding and sorting the data, to give us a 'clean' dataset that is appropriate for our model. Hence, the discussion in this paper is related on how we attained the required 'clean' dataset. The data that we received was stored in a row-wise format, which needs to be transformed into a column-wise format for easy manipulation. The difference between these formats is illustrated in Fig. 1 and Fig.2 below.

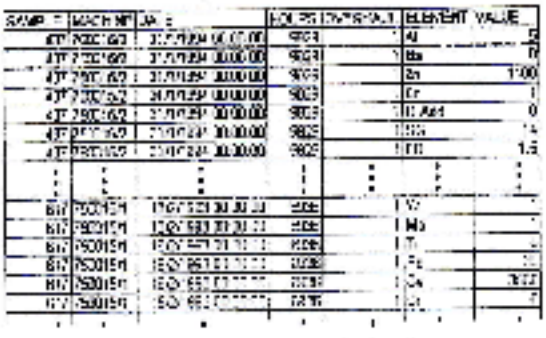

Fig.1. Row format for monitoring data

Fig.2. Monitoring data after column manipulation

III. MINING THE DATA

In order to correlate the data with the residual life, we had to model the relationship between them. This needs more explanation, as we did not use the information from the SOAP as it was collected, but transformed it into another measurement that we called 'total metal concentrations' that represents the cumulative metal concentrations since new. The rationale behind this transformation is that, to establish the relationship between the deterioration process from oil analysis and the residual life, we need some quantification for the deterioration process influenced by condition-monitoring variables. Here, the deterioration process is called wear, and we believe that wear is a non-decreasing process accumulated since new. If we want to use metal concentration to represent a function of wear, then the calculation of the cumulative metal concentration will provide an indirect measure of the cumulative wear since new with random noise. The other reason that we did not use the raw data is that oil top-ups or changes will distort the metal concentration content within the oil sample since a newly flushed engine will have very little metal concentration in the oil. The transformation was performed using the following formula.

Total Wear Metal,

$$y_t = y_{t-1} + [C*(e_t - e_{t-1})] + [\alpha*(e_t + e_{t-1})/2]$$

where

y_t = Total Wear Metal at time t, where t is the th checking time since new.

e_t = Element Concentration (ppm) in the th oil sample

C = Oil Capacity of component (litres)

α = Oil Added (litres)

ISBN 978-967-5848-36-1

In this study, we first used metal concentration measures only as monitoring indicators, because their characteristics directly provide important information on the wear condition of internal engine parts (Lukas et al., 1996; Barraclough et al., 2003). Noted however, we also used other indicators provide by SOAP.

Using equation above equation and considering only metal concentration measures, we plotted the value of the total concentration of each metal against operating hours, as shown in Figure 3. Only six metal concentrations from the metal group were available from the dataset that could be taken to characterize the residual life of the system.

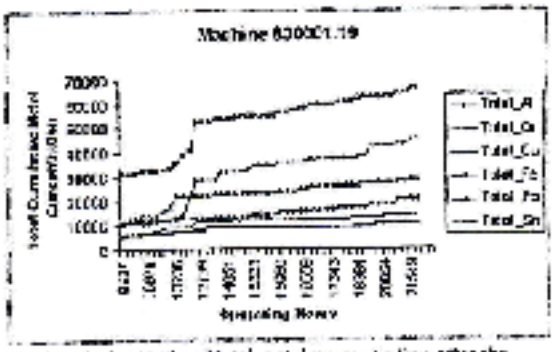

Fig.3. A sample of total metal concentration after the transformation

At this early stage, we have two major issues in analysing the dataset. The first is the incomplete nature of the data. This refers to the recorded monitoring indicators, which were not from new as we required. As an example, Fig.3 shows the observed data started around 10,000 hours before the engine was replaced to prevent failure, or at failure. To make a complete history of the SOAP for those engines with missing early data, we extrapolating these values to the starting point by using a simple linear regression.

As mentioned above, we have six element indicators of the total metal concentration that could be used into our next modelling procedure. We had difficulty in choosing which of these indicators are really useful and can be expected to produce the best results. Hence, we had two options regarding the dimensions of total metal concentration indicators to be used in our model. The dimension of a model is the number of independent or input indicators used by the model. The first option was to use all the total metal concentrations as given, but this would result in a complex model where all metal concentrations could be correlated with each other; we would have to use a joint probability density function for them, which is difficult and requires more parameters to be estimated. The second option was to reduce the correlation and the dimensions of the total metal concentrations, but at the same time we may lose some of our original information.

For model simplification, we chose the second option, and used the widely known data decomposition technique called 3 Principal Component Analysis (PCA) to simplify the data. Generally, PCA is a useful procedure dealing with dimension reduction techniques, especially when we have a set of sample measurements that are highly correlated (Jolliffe, 1986).

IV. PRINCIPAL COMPONENT ANALYSIS

PCA encodes the most relevant information contained in a sample in a set of orthonormal vectors. This set defines a characteristic subspace that contains the main features of the sample, and the number of selected vectors defines the amount of variance that can be explained by the PCA model. To start with, we defined y_i where $i = 1,2,Km$ is a variable vector that represents sample data, and v_i, $i = 1,2,Km$ are the principal components; the linear relationship between y_i and v_i is given by Williams et al. (1995) as:

$$\begin{aligned} v_1 &= u_{11}y_1 - u_{12}y_2 + u_{13}y_3 + \ldots + u_{1m}y_m \\ v_2 &= u_{21}y_1 - u_{22}y_2 + u_{23}y_3 + \ldots + u_{2m}y_m \\ v_3 &= u_{31}y_1 - u_{32}y_2 + u_{33}y_3 + \ldots + u_{3m}y_m \\ &\ldots \\ v_n &= u_{n1}y_1 + u_{n2}y_2 + u_{n3}y_3 - \ldots + u_{nm}y_m \end{aligned} \quad (IV\text{-}1)$$

where $uk=[uk1,uk2,uk3,\ldots,ukm]$ is the kth eigenvector of the correlation or covariance matrix. The principal component v_i 's are uncorrelated and their variances are given by the corresponding eigenvalues. Basically, principal component analysis will generate the same dimension from the original data and rank them according to the value of the variance. The

ISBN 978-967-5048-36-4

question of how many variables should be retained needs to be answered. As a solution, a number of procedures have been suggested (Green et al., 1978). One flexible approach is to use the Kaiser criterion, which recommends that only principal components of the correlation or covariance matrix with eigenvalues greater than 1 need to be retained. Another technique, called the scree test, allows us to plot the order of eigenvalues and then look for elbows in the curves. A more convenient approach is to retain only those eigenvalues that account, on a cumulative basis, for some higher proportion of the total variance, such as 75 or 80%.

Using the latter techniques and the scree plot, the dimension of the variable was chosen to be 1, as most of the dataset produces similar results, as depicted in Fig. 4 and Fig. 5 below. In short, the mining procedure in this approach consists of the following steps:

1.Obtain a sample set of raw data of total metal concentrations. The sample data will form a matrix that consists of observed variables at every monitoring point.
2.Compute the correlation or covariance matrix.
3.Compute the eigenvalues and eigenvectors of the correlation matrix above.
4. Order the eigenvalues and eigenvectors from greater to smaller. Note that the number of eigenvectors is equal to the number of variable in the sample
5.Choose principal components and form a matrix of vectors.
6.Derive a new dataset. This can be done by taking the transpose of the vector and multiplying it with the original data set transposed.

After carrying out the analysis, we concluded that a single dimension of variables should simplify our model to represent overall total metal concentration. In subsequent analyses, we shall use the first principal component of total metal concentration as our monitoring information, y_1, unless otherwise specified.

The next problem that we encountered is irregular monitoring intervals. In fact, we could use the total metal concentration as it is, but this would increase modelling complexity as extra parameters may be needed to take the irregular interval into consideration and cause difficulty in parameter estimation. Since our interest is the total metal concentration, we need to re-organise the data and set up an imaginary regular monitoring interval for all data sets.

Eigenvalues 830001/19		Screen Plot	%Total Variance
1	2.83E−05	Scree Plot 830001/19	98.21%
2	2.68E−06	Eigenvalues 4.00E−05 2.00E−05 0.00E+00 1 2 3 4 5 Dimensions	0.93%
3	1.77E−06		0.61%
4	4.80E−05		0.17%
5	1.90E−05		0.07%
6	3.00E−04		0.01%

Fig. 4. Case 1 – Choosing the dimensions of principal component analysis for engine 830001/19

Eigenvalues 830001/19		Screen Plot	%Total Variance
1	2.83E+05	Scree Plot 830001/19	98.21%
2	2.68E+06	Eigenvalues 4.00E−05 2.00E−05 0.00E+00 1 2 3 4 5 Dimensions	0.93%
3	1.77E−06		0.61%
4	4.80E−05		0.17%
5	1.90E−05		0.07%
6	3.00E−04		0.01%

Fig. 5. Case 2 – Choosing the dimensions of principal component analysis for engine 830001/26

ISBN 978-967-5048-35-4

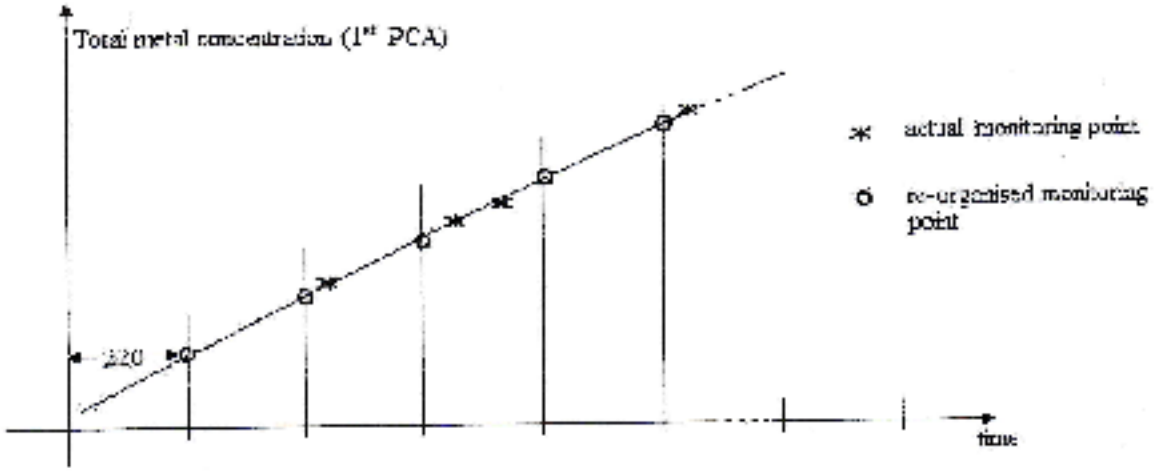

Fig. 6. Re-organizing condition-monitoring data from original reading

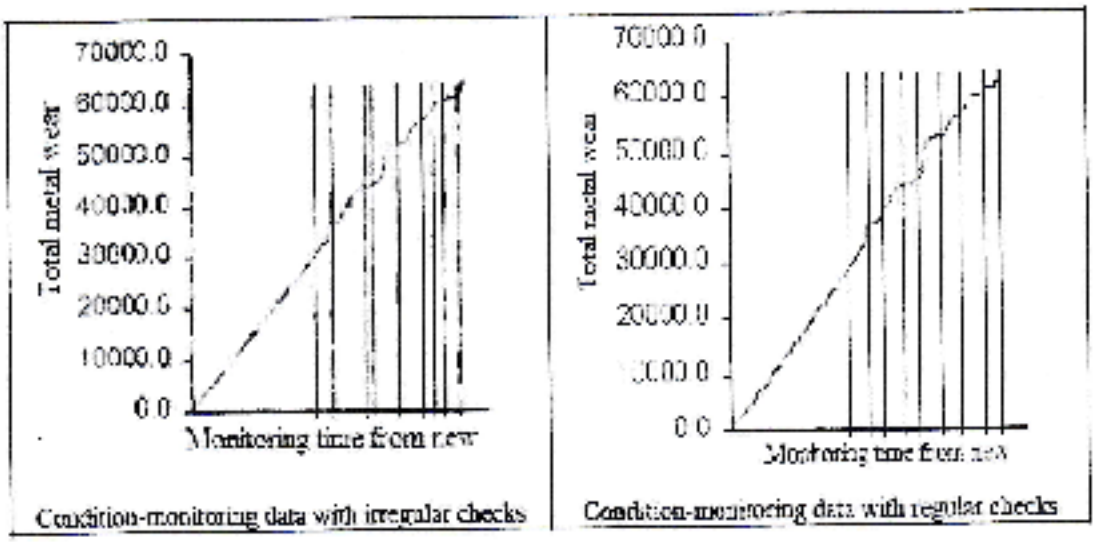

Fig. 7. The difference before and after re-organising

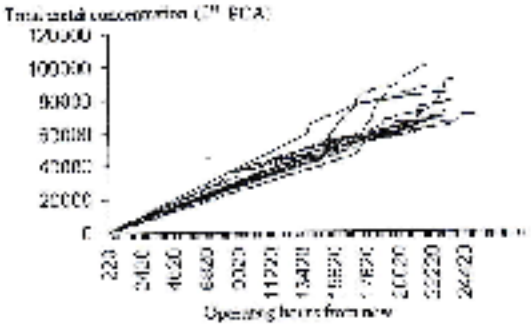

Fig 8. Example of regular total metal concentration (1st PCA) used in diesel engine since new.

In Fig. 6 the *, represents the original value of the total metal concentrations at irregular time checking points. To set up a regular interval, we find out the mean interval from the data as our approximation value, which is 220 hours. Using this as our regular interval for all engines over their lifetime, we have a representation shown in Fig. 6, where O denotes the imagined monitoring points. The difference before and after re-organising the dataset is shown in Fig. 7.

ISBN 978-967-5048-36-4

Having solved these problems, we now have a dataset, which contained information from new, and had equal monitoring intervals for ease of our modelling development.

An attempt to formulate the residual time prediction model using the cleaned data of the diesel engines was then carried out. A 1st PCA of the final 35 datasets for regular observed monitoring information from new, used in the subsequent analysis, is given in Fig. 8.

IV. CONCLUSION

This report has presented various mining technique based upon monitoring data for wear metal concentration obtained from SOAP. An analysis of such data was carried out to obtain the required format. Once the 'clean' data is attained, the next modelling steps can be established to recommend

REFERENCES

[1] H.T. Aschur, K.A.H. Kobbacy, and D.F. Percy, *Realistic Modelling of Preventive Maintenance*. University of Salford, 1995.

[2] T. Barraclough, M. Lucas, and D. Anderson, "Comparison of wear and contaminant particle analysis techniques in an engine test cell run to failure", 2003. Spectro-Inc Inter-net Homepage.

[3] T.J. Edwards, G.D. Holt, and F.C. Harris, "Predictive maintenance techniques and their relevance to construction plant," *Journal of Quality in Maintenance Engineering*, vol 4, pp 35-37,1998.

[4] P.E. Green, and J.D. Carrol, *Analysing Multivariate Data*. Hinsdale Illinois: Dryden Press, 1978.

[5] I.T. Jolliffe, *Principal Component Analysis*. New York: Springer-Verlag, 1986.

[6] M. Lucas, and D.P. Anderson, "Lubricant analysis for gas turbine condition monitoring," *American Society of Mechanical Engineer*, pp. 1-12, 1998.

[7] A. Mathur, K.F. Cavanaugh, K.R. Pattipati, P.K. Willet, and T.R. Galie, "Reasoning and Modelling Systems in Diagnosis and Prognosis. Component and Systems Diagnostics." *SPIE The International Society for Optical Engineering*, pp 194-pp.203, 2001.

[8] J.H. Williams, A. Davies, and P.R. Drake, *Condition-based Maintenance and Machine Diagnostic*. Chapman & Hall, 1995.

BIOGRAPHIES

DR. BURAIRAH HUSSIN is a senior lecturer at Universiti Teknikal Malaysia Melaka from June 2001 until present. He performed research and teaching in the Faculty of Information and Communication Technology. In addition, he taught classes on various computer science subjects for undergraduate level and supervised postgraduate students in the pursuit of their Master and PhD. Research focuses include operational problem within industry, maintenance management, statistical analysis, creative media and software development.

DR. ABDUL SAMAD SHIBGHATULLAH is a lecturer at Universiti Teknikal Malaysia Melaka. He performed research and teaching in the Faculty of Information and Communication Technology. He received a B. Accounting from Universiti Kebangsaan Malaysia of Bangi. He received his M.Sc. in Computer Sciences from Universiti Teknologi Malaysia, Skudai, and received his Doctor of Philosophy at Brunel University. His research interests include scheduling, data mining, transportation and optimization.

ISBN 978-967-5048-36-4