A statistical approach for analyzing used oil data and enhancing maintenance decision making: Case study of a thermal power plant.

Wakiru, J.[1], Pintelon, L. [1], Muchiri, P.N. [2], Chemweno, P. [1]

[1]Centre for Industrial Management/Traffic and Infrastructure, KU Leuven, Celestijnenlaan 300A, Leuven 3001, Belgium.

[2]Dedan Kimathi University of Technology, Nyeri, Kenya.
Email: jamesmutuota.wakiru@student.kuleuven.be; liliane.pintelon@kuleuven.be; peter.muchiri@dkut.ac.ke; PeterKipruto.Chemweno@kuleuven.be;

Abstract A lubricant is an essential component for enhancing the equipment's functionality and durability. For this reason, used oil analysis (UOA) is becoming an integral part of the plant's lubrication program which is part of Condition Based Maintenance (CBM). By monitoring the lubricant's condition through the UOA, organizations can optimize the equipment availability by reducing failure incidents of rotating elements. This paper advances the use of a predictive model of used oil analysis data with a view of assisting maintenance decision making of critical power plant equipment. The steps of the proposed methodology include data pre-processing, principal component analysis (PCA) for dimension reduction, and logistic regression analysis to build the predictive model, where the lubricant's parameters are compared against set thresholds, or limit values from which, indications of significant lubricant deterioration may be derived. The framework is applied to a thermal power plant case study. The novelty of the framework is towards providing insights for maintenance decision making and moreover, highlighting critical used oil analysis parameters that are indicative of lubricant degradation. By addressing such critical parameters, organizations can better enhance the reliability of critical operable equipment.

Key words: Regression, Condition monitoring, PCA.

1.0 Introduction

Lubricant condition monitoring is a key monitoring technique applied under the condition based maintenance(CBM) strategy. The condition of the lubricant in use is monitored and changes would infer degradation of either the lubricant itself or the equipment being lubricated. The lubricants' performance level affects the performance and operability of the equipment. A lubricant plays major roles in the engine, for instance reducing friction between surfaces that are moving relative to each other, control and minimization of wear, heat, corrosion and contaminations[1]. The performance of a lubricant is mainly influenced by its deterioration level which further affects the working conditions of the engine or equipment.

In a lubricant condition monitoring or used oil analysis (UOA) program, four main areas are monitored and highlighted, that is, changes in the physical and chemical properties, contamination, component wear through ingression of wear particles and additive analysis which would indicate depletion or level of the main components in the additive. For each of the monitor-able areas, there are variables or parameters that are analyzed in the program, sample variables indicated in table 1.

Table 1 UOA Variables indicative sources

Variables	**Indicative sources**
Viscosity	An Increase is associated with oxidation or dilution with a denser product like heavy fuel oil from injector leaks, pump failures
Flash point	Associated with dilution with fuel or water as well as low lubricant quality
TBN	Depletion associated with the increase in acidity from oxidation (high temperatures) or fuel Sulphur.
Magnesium	Associated with leakages of hard water and detergents in the lubricant
Silicon	Associated with dust or anti-freeze ingression or anti-foam additives
Water	Indicates high condensation, water leaks or inefficient centrifuge operation
Iron	Associated with the wear of pistons, cylinder liners, oil pump, valves etc.
Chromium	Indicative of wear on piston rings, liners, valves
Aluminum	Indicative of wear on pistons, bushings, oil pumps and bearings
Nickel	Can indicate wear on an alloy with Iron and vanadium as well as HFO dilution

Once the oil has been analyzed, the analysts evaluate the different parameters against limits to confirm the condition of each parameter and hence the overall condition of the lubricant. The limits are often set by the original equipment manufacturers (OEM) or the lubricant supplier. In some cases, the maintenance team through experience and analysis can set the limits considering the operational environment and conditions like the sulphur level in the fuel, the age of equipment, humidity, temperature, dusty environment etc. Other bodies like the International Council on Combustion Engines (CIMAC) give typical limits for mandatory action for different types of equipment[2]. A single sample of oil may have over twenty parameters tested whilst the analyst should review each parameter, and possibly derive some trend graphs to confirm if the sample is okay. If the sample is okay/passed, the equipment can continue running using the lubricant, while if the sample has failed, then action need to be taken either to improve the condition or renew the state of the lubricant. The exercise takes considerable time and introduces errors due to the high dimensionality of the parameters which would require reduction without compromising importance.

Principal component analysis (PCA) is a dimension reduction tool that can be used to reduce a large set of variables to a small set that still contains most of the information in the original large set. PCA has been used for dimension reduction in different areas, for instance, it was used to reduce the number of variables included in the CBM model[3], in UOA prediction models [4],[5],[6],[7],[8]. Other studies using PCA include [9][10][11], and [12]. Logistic regression is a predictive analysis used to describe data and explain the relationship between one binary variable (response variable which should be dichotomous or binary in nature for instance either 1 or 0 or either Pass or Fail) and one or more nominal, ordinal or ratio-level independent variables.

Logistic regression (LR) has also been used in different studies incorporating UOA, for instance, assessment of failure degradation and predicting propagation of failure from incipient to occurrence of actual failure [13],system condition estimation based on selected tribodiagnostic data [14], determination of maintenance inspection interval lengths on aircraft maintenance data [15], calculation of the probability of failure for given condition variables[16], and for machine health assessment[17].

2.0 Motivation of study

Manual evaluation and analysis of the used oil analysis results are prone to human errors and delay. The analysts have differences in exposure, experience and ideas, therefore no standard interpretation can be achieved. The use of the specific samples without reference to past data could imply the "noise" from the data is not considered. Errors in the sampling procedure can influence the results of the sample if no benchmarking is done, substantial information may go unnoticed or erroneously dismissed. Analysis of one sample can bring bias to the interpretation, which might not be a true representation. This can be accelerated by errors in sampling procedures if the sample is not representative of the oil in the system as is the requirement. The current procedure is cumbersome and time-consuming especially for management who require a snapshot with action items.

The factors motivate this study which seeks to develop a predictive model the maintenance team can use to confirm if a sample has passed, meaning all the tested parameters are within the acceptable level. On the other hand, the model can indicate whether the sample has failed, hence appropriate maintenance actions can be taken. The novel of this study is the quantitative approach of subjecting all the parameters or explanatory variables for criticality selection using PCA, afterward, use the selected variables to build the LR model and test the predictive accuracy of the model.

3.0 Methodology

The methodology as illustrated in figure 1 consist of a number steps. Step 1 deals with data collection, Step 2 involves data pre-processing in readiness for PCA and logistic regression, Step 3 incorporates the use of PCA to select critical variables, Step 4 building the LR model and Step 5 testing the developed model. We will use the statistical software R in both PCA and LR model.

3.1 Data collection

The data used in this study was from a thermal power plant that uses heavy fuel oil to drive the engines which eventually drive electric generators. The plant maintains data on used oil analysis for the engine oil which is sampled periodically and analysis done by an independent laboratory. The data used were collected from the years 2011 to 2015, and measuring twenty lubricant parameters as outlined in Table 1.

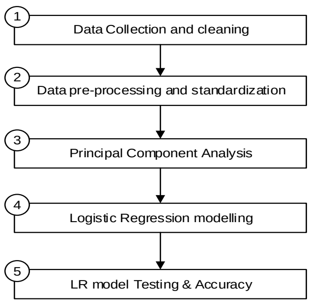

Fig 1 Schematic representation of methodology

3.2 Data pre-processing

The data was availed as individual reports for samples, while some had been organized in terms of specific variables. The pre-processing step adapts the data to the requirements of the data analysis hence enabling efficient analysis of the data which would be unfeasible otherwise. In this study, the data preparation or preprocessing includes a wide range of steps or phases, for instance, data transformation, integration and cleaning. The preprocessing stage also involved the maintenance team of the plant to verify the data and give some interpretation and linkages. Due to measurements scales for different parameters being different, the data was standardized.

3.3 Variables selection using PCA

PCA is used for identifying patterns in data and expressing it in a way to expose the similarities and differences. PCA is a powerful tool to draws out patterns in data of high dimension and reduces the number of dimensions without much loss of information, as a set of new orthogonal variables called principal components, reduces the size of data set whilst keeping important information and analyzes structure and variables. Factor loadings of the factor or component coefficients are used in selecting the important variables. This is correlation coefficients between the original variables and the principal components or factors and give an indication to which extent the original variables are important in creating new variables.
It is important while selecting variables to be used in a model, one can locate the important variables that influence the prediction. The data set had twenty variables which mean using all the variables in developing a model would become lengthy, moreover, some of the variables may not have key influence in the performance of the lubricant.

3.3.1 Standardization of data

Since the method proposed is based on the principal components and eigenvalues of the covariance matrix on parameters measured in different scales, the data had to be standardized. This was to enable interpretation of the Principal components (PC) in terms of the original variables where each coefficient is divided by the standard deviation of the corresponding variable.

3.3.2 Selection of number of Principal components

There are three methods to expose the number of PCs to be considered. Firstly, the use the rule of thumb when using standardized data, retain those PCs with an eigenvalue larger or equal to 1, secondly examination of the scree plot (scree or elbow test) to find if there is an 'elbow' in the slope, keeping the components before the elbow. Lastly the percentage of total variance the PCs should account for. From the analysis, we had five PCs that had eigenvalues equal or larger than 1, while the scree plot was indicating the elbow at six components in figure 3, hence five PCs were selected which explain 70.34% of the total variation of the data.

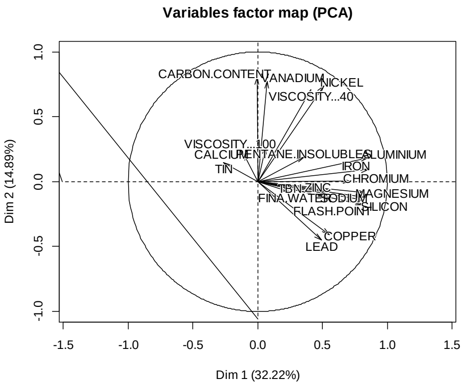

Figure 2 PCA factor map (PC1 and PC2)

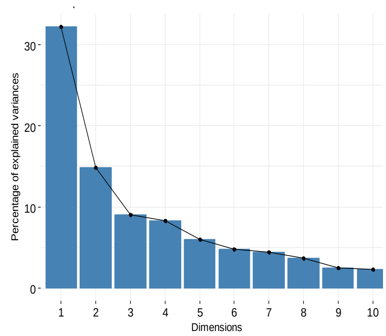

Figure 3 Scree plot

3.3.3 Variables selection

In selecting the variables, we disregard those loading below a certain threshold on each factor, for this study we use any below 0.5 as used traditionally[18].From the Principal component analysis in table 2, the following variables were extracted to be used as inputs for the logistic regression model: Viscosity at 40ºC, flash point, TBN, magnesium, calcium, zinc, silicon, sodium, water, carbon content, iron, chromium, copper, aluminum, nickel and vanadium.

The dimension reduction done using PCA exposed several correlations of the used oil parameters. From PC1 (in table 2), had silicon, iron, chromium magnesium and aluminum are wear metals, while flashpoint a physical property of the lubricant and water a contaminant. There is a possibility as indicated in table 1, the wear metal particles occurring in the lubricants due to wear of mostly the combustion parts, for instance the pistons, while the increase in water content increases flashpoint and catalyzes corrosive wear. PC2 was made up of viscosity characteristics with viscosity at 40ºC, Carbon, nickel and vanadium. For used oil analysis, the correlation of this three lubricant parameters comes into being when the viscosity increases due to ingression of heavy fuel oil which is used in such plants as in our case study[19]. The three parameters have a high effect of increasing the viscosity of a lubricant[20], carbon is an indication of soot presence which is a product of normal fuel combustion, moreover can be used to indicate the combustion efficiency of the engine. Nickel and vanadium are used in alloys with iron and aluminum and mostly appear in lubricants as a contaminant from heavy fuel oil used in medium speed engines (MSE)[2]. PC3 primarily had TBN characteristics, where calcium and magnesium form detergents/dispersants which are ingredients used for TBN boosters(additive)[19].TBN which is the alkaline reserve of a lubricant is critical in MSE due to its neutralization effect of acidity caused by either Sulphur from the fuel or oxidation of the lubricant.

Table 2 Factor loadings for the PCs

Variable	PC1	PC2	PC3	PC4	PC5	Variable	PC1	PC2	PC3	PC4	PC5
Viscosity @40°C	0.41	0.71	-0.05	- 0.29	0.16	Water	0.51	-0.11	-0.20	-0.25	-0.51
Viscosity @100°C	0.06	0.27	-0.10	- 0.05	0.02	Carbon/Soot	-	0.80	0.18	0.08	0.09
Flash point	0.81	-0.18	-0.02	- 0.36	- 0.11	Iron	0.85	0.09	-0.23	0.19	0.14
TBN	0.25	-0.07	0.86	0.10	0.05	Chromium	0.71	-	-0.26	0.33	0.20
Magnesium	0.82	-0.12	0.34	- 0.12	0.04	Lead	0.49	-0.45	-0.18	0.45	0.11
Calcium	-0.11	0.22	0.72	0.43	- 0.20	Copper	0.55	-0.41	-0.19	0.42	0.16
Zinc	0.56	-0.07	0.06	0.20	- 0.63	Tin	-0.26	0.14	0.02	0.08	0.47
Silicon	0.82	-0.14	0.24	0.07	0.24	Aluminium	0.85	0.19	0.10	0.05	0.04
Sodium	0.82	-0.08	0.10	- 0.34	0.11	Nickel	0.51	0.74	-0.18	0.14	-0.12
Pentane Insolubles	0.36	0.19	0.10	- 0.54	0.13	Vanadium	0.07	0.77	-0.20	0.38	-0.16

3.3 Logistic Regression Model training and testing

3.3.1 Data preparation

In this study, each sample was evaluated based on the twenty variables that had been tested. Each of the variables is individually evaluated against limits or thresholds to reveal the performance of the lubricant parameters, which eventually reveals the sample condition. Limits or alarm level for oil analysis results can be specified to monitor the machine's condition. Typically, the absolute thresholds are usually recommended by the equipment manufacturer and/or lubricant suppliers', which in practice, they are based on the average operational and performance situations that may not correspond to the actual application condition of the machine. The limits assist in evaluating the lubricant parameters that have deviated from the expected values indicating that either the lubricant or equipment or both conditions are not as expected. In this study, data was prepared such that the lubricant parameters were evaluated and a score for each sample updated to have either a PASS or FAIL. The data was split to training and testing data. For instance, in this study, values taken were 70:30 i.e. 70% of the data to use for training the model while 30% of the data will be used to test the model.

3.3.2 Fitting the model to the data

In this section, the model introduced in the earlier section was fitted to the training data, since the engines under study do not run to complete failure due to the preventive maintenance performed. However, the failure of the lubricant is not defined as an actual breakdown, but a deviation of one or more parameters that are important to its health and state. The variables exposed by the principal component analysis were taken as the input (explanatory variables) to the model, while the predictor variable is the binary state of 1 to represent sample PASS and 0 to represent sample FAIL.

3.3.3 Model Goodness of fit

Once a logistic regression model has been fit to a given set of data, the adequacy of the model is examined by overall goodness-of-fit tests and examination of influential observations. A goodness-of-fit test that is commonly used to assess the fit of logistic regression models is the Hosmer–Lemeshow test[21]. A hypothesis of H_o the LR model fits the data while the null hypothesis H_1, the LR model does not provide a good fit and is tested using the significance level of 5%. The model generated a p-value of 0.7948 hence we do not reject the null hypothesis based on a significance level of 5% and conclude that the logistic regression model fits the data.

3.3.4 Model evaluation

The built model using the training data was tested and evaluated using the testing data by evaluating its predictive power using several parameters.

(a) Classification table

The classification table, as seen in table 3, indicates the model prediction using the testing data. It can be used to calculate the sensitivity, which is the proportion of events (1 or PASS) predicted as events (1), specificity which indicates the proportion of non-events (0 or FAIL) predicted as non-events (0) and false positive which indicate the number of non-events (0) predicted as events (1). Classification of observations using the prediction where the test data was used, was done based on a cut-off value of 0.5 giving Sensitivity of 74.07% ({40/ (40+14)}), Specificity of 93.14% ({258/ (258+19)}) and a false positive of 6.83%, all computed from results in table 3.

Table 3 Classification table -cut-off level of 0.5

	Actual 'FAIL'	Actual 'PASS'
Classified 'FAIL'	258	19
Classified 'PASS'	14	40

(b) AUC (Area Under the ROC curve)

The ROC curve, see figure 4, is a visual measure of the predictive ability or power of the logistic regression model. See figure 4. A model with a high AUC indicates to be one with a higher predictive power i.e. able to classify a 'FAIL' sample from the 'PASS' samples. The ROC curve developed is seen in figure 4, while the area under the ROC curve (AUC) for the LR model is 0.9473≃0.95.

(c) Proportion of deviance

This is also considered as a generalization of R^2

This can be interpreted as the proportion of deviance explained by the model.

$$Proportion\ of\ Deviance = \frac{Null\ Deviance - Residual\ Devince}{Null\ Deviance}$$

Where null deviance is the deviance for a model with only one constant term and residual deviance is the deviance of the fitted model. The model returned a proportional deviance of 75.47%.

The prediction of the model returns an overall error of 10% and an average class error of 16%.

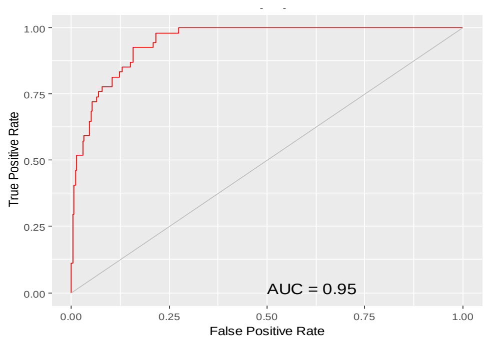

Figure 4 Receiver Operating Curve (ROC) curve

The LR model evidenced to classify approximately 90% of the oil samples correctly as seen in table 3. Logistic regression is a powerful tool for predicting class probabilities and for classification using predictor variables. Classification or scoring of data requires thresholding, which defines probability intervals for each class or score hence making it completely adaptable for the UOA sample classification. The use of different data set for model building(training) and testing here measures how the model will perform on previously unseen cases, moreover alludes to the importance of historical data to the future classification. This makes LR modelling a tool that is easily adoptable by the maintenance engineers to evaluate samples as the model is updated using newly generated data, moreover, this will reduce the time to make maintenance decisions.

4.0 Conclusion

The performance of the LR model has been evaluated for classification of oil sample data from a thermal power plant giving an accuracy of 90% in predicting. The results from PCA, are vital in root cause analysis (RCA) after a sample is classified to have failed. Though the classification would not be generalized for all kinds of data, it proves to be vital if accurate RCA should be carried out as it points at some links and patterns of the parametric performance. The methodology developed herein can be subtle to the maintenance decision support as it will shorten the time cycle of interpretation, hence ensure timely intervention before failure occurs. Moreover, the model developed can be used by anyone without much knowledge of statistics making it easy to adopt. Proposed future work would incorporate other predictive models in the study.

References

[1] "ICML Certification Series: Industrial Lubrication Fundamentals — Functions Of A Lubricant - Maintenance Technology." [Online]. Available: http://www.maintenancetechnology.com/2013/04.[Accessed:28-Apr-2017].

[2] CIMAC, "USED ENGINE OIL ANALYSIS -USER INTERPRETATION GUIDE on Combustion Engines," 2011.

[3] D. Lin, D. Banjevic, and a K. S. Jardine, "Using principal components in a proportional hazards model with applications in condition-based maintenance," *J. Oper. Res. Soc.*, vol. 57, no. 8, pp. 910–919, 2006.

[4] W. Wang and W. Zhang, "A model to predict the residual life of aircraft engines based upon oil analysis data," *Nav. Res. Logist.*, vol. 52, no. 3, pp. 276–284, 2005.

[5] K. Sepcic, M. Josowicz, and J. Janata, "Determination of the principal volatile compounds in degraded engine oil," *Tribol. Lubr. Technol.*, vol. 60, no. 3, pp. 40–47, 2004.

[6] H. B. Jun, F. Lo Conte, D. Kiritsis, and P. Xirouchakis, "A predictive algorithm for estimating the quality of vehicle engine oil," *Int. J. Ind. Eng. Theory Appl. Pract.*, vol. 15, no. 4, pp. 386–396, 2008.

[7] W. Wang, L. Sun, X. Lu, and H. Liu, "Application of DPCA based stochastic filtering model and comparison of optimal CBM policies," *QR2MSE 2013 - Proc. 2013 Int. Conf. Qual. Reliab. Risk, Maintenance, Saf. Eng.*, pp. 700–706, 2013.

[8] D. Vališ and L. Žák, "Assessment of Off-Line Diagnostic Oil Data with Using Selected Mathematical Tools," *Appl. Mech. Mater.*, vol. 772, pp. 141–146, 2015.

[9] A. M. Yan, G. Kerschen, P. De Boe, and J. C. Golinval, "Structural damage diagnosis under varying environmental conditions - Part II: Local PCA for non-linear cases," *Mech. Syst. Signal Process.*, vol. 19, no. 4, pp. 865–880, 2005.

[10] E. Gandino, "Diagnostics of machines and structures : dynamic identification and damage," Politecnico di Torino, 2013.

[11] F. Discenzo, D. Chung, M. Kendig, and K. Loparo, "Intelligent fluid sensor for machinery diagnostics, prognostics, and control," US7581434 B1, 2006.

[12] U. Neisser, "Pattern Recognition," *Cogn. Psychol. Key readings.*, no. February 2013, pp. 37–41, 2004.

[13] W. Caesarendra, A. Widodo, and B. S. Yang, "Application of relevance vector machine and logistic regression for machine degradation assessment," *Mech. Syst. Signal Process.*, vol. 24, no. 4, pp. 1161–1171, 2010.

[14] D. Vališ, L. Žák, and O. Pokora, "System Condition Estimation Based on Selected Tribodiagnostic Data," *Qual. Reliab. Eng. Int.*, vol. 32, no. 2, pp. 635–645, 2016.

[15] K. . Spezzaferro, “Applying logistic regression to maintenance data to establish inspection intervals,” in *Annual Reliability and Maintainability Symposium (RAMS)*, 1996, pp. 296–300.

[16] J. Yan, M. Koç, and J. Lee, “Production Planning & Control: The Management of Operations A prognostic algorithm for machine performance assessment and its application A prognostic algorithm for machine performance assessment and its application,” *Prod. Plan. Control*, vol. 15, no. 8, pp. 796–801, 2004.

[17] L. Liao, B. Huang, and J. Lee, “Logistic Regression-based Machine Health Assessment Method on Application of Smart Machine Tool,” *Proceedigns Int. Manuf. Leaders Forum*, no. May 2017, pp. 1–12, 2006.

[18] S. Sharma, *Applied multivariate techniques*. New York: John Wiley & Sons, Inc. New, 1996.

[19] R. Mortier, M. Fox, and S. Orszulik, *Chemistry and technology of lubricants*, vol. 53. 2010.

[20] M. R. Yakubov, D. V. Milordov, S. G. Yakubova, D. N. Borisov, V. T. Ivanov, and K. O. Sinyashin, “Concentrations of vanadium and nickel and their ratio in heavy oil asphaltenes,” *Pet. Chem.*, vol. 56, no. 1, pp. 16–20, 2016.

[21] K. J. Archer, S. Lemeshow, and D. W. Hosmer, “Goodness-of-fit tests for logistic regression models when data are collected using a complex sampling design,” *Comput. Stat. Data Anal.*, vol. 51, no. 9, pp. 4450–4464, 2007.

Authors’ Biography

	**James Wakiru** is a doctoral student at the Centre for Industrial Management, KU Leuven, Belgium.
	**Liliane Pintelon** is a professor at the KU Leuven (CIB), where she teaches logistics courses, including maintenance management. Her research interests are in asset management, both in industry and health care.

	**Peter Muchiri** is a professor at Dedan Kimathi University of Technology in Kenya, received both his masters and PhD in industrial management at KU Leuven(CIB). His research interests are in production management, performance management and maintenance decision making.
	**Peter Chemweno** completed his PhD in Mechanical Engineering in 2016. He is currently a postdoctoral researcher at the Centre for Industrial Management, KU Leuven, Belgium. His research interests are in risk management, maintenance performance management and maintenance decision making.