Empirical Evaluation of the Impact of Data Pre-Processing on the Performance of Predictive SHM of Jet Engines

Jean-Loup Loyer

7th European Workshop on Structural Health Monitoring
July 8-11, 2014. La Cité, Nantes, France

EMPIRICAL EVALUATION OF THE IMPACT OF DATA PRE-PROCESSING ON THE PERFORMANCE OF PREDICTIVE SHM OF JET ENGINES

Jean-Loup Loyer[1]

[1] *Instituto Superior Técnico, Universidade de Lisboa, Avenida Rovisco Pais 1, 1049-001 Lisbon*

jean-loup.loyer@tecnico.ulisboa.pt

ABSTRACT

We evaluate the impact of data pre-processing on the performance of predictive Structural Health Monitoring algorithm on a real case study involving dozens of jet engines. A simple robust four-step framework is designed to this effect, made of 1) outliers removal, 2) range scaling, 3) variable selection (either by "manually" evaluating variable correlations or by quantification of variable importance via random forests) and 4) evaluation of the predictive performance of a unique selected binary classifier (random forests). The results contrast with the intuition and the literature, since pre-processing raw data decreases predictive performance in half of the cases analyzed. The isolated influence of each of the pre-processing techniques rank in this order: important variables chosen through random forests has the highest positive impact, followed closely by variable scaling and outlier removal to a lower extent, while the "manual" variable selection via the correlation matrix exerts a slightly negative impact on predictive performance. The influence of combining pre-processing techniques is in line with the isolated influence of each technique. However, a detailed evaluation should be done for every application since these results might be due to the high data quality of aerospace engines or to the characteristics of random forests.

KEYWORDS : *Predictive SHM, outlier analysis, data cleaning, dimensionality reduction, gas turbine health monitoring.*

INTRODUCTION

Structural Health Monitoring (SHM) of gas turbines allows industrial companies to optimize operating performance, detect early potential part failures and increase economical returns through predictive maintenance [1]. To do so, dozens of sensors are typically installed at several stations in the rotating machinery in order to measure up to a few hundreds of parameters: absolute and marginal temperatures, pressure, shaft rotation speeds, vibration levels, fuel flow, oil characteristics... Such parameters are usually recorded continuously as unevenly-spaced multivariate time series over the operating life of the machine. However, the raw data from the sensors might exhibit unwanted features (outliers, offsets, trends related to instrument decalibration, interruptions in data acquisition, environmental disturbances…), which often require specific procedures to obtain data with a sufficient level of quality for the subsequent modeling stages. Moreover, even after the raw data had been properly transformed, the high number of predictors and their functional nature render the use of predictive algorithms particularly challenging. To remedy these two key issues, many data pre-processing techniques (DPPT) have been developed over the last three decades and applied to fields such as SHM of turbomachinery [2]. DPPT can be divided in two categories: 1) the transformation of the raw data into higher quality data, 2) the reduction of the dimensionality of the dataset by removing the less relevant covariates. Within the first category, techniques to transform the raw data include outlier analysis [3], curve smoothing (moving average, LOESS/LOWESS, splines), robust estimation or Kalman filtering [4]. Regarding the second category of DPPT, the dimensionality of the dataset can be reduced by the extraction of

summarizing features from the time series, variable subset selection according to measures of correlation or variable importance [5], clustering [6], Principal Component Analysis (PCA) [7] or Independent Component Analysis (ICA) [8] of time series [9].

The ultimate objective of our SHM algorithm is to build statistics-based models of the health of jet engine components. Its role is to predict the workscope at the next maintenance visit of a given turbomachine, as measured by the number of components to scrap or repair. In more statistical terms, such statistical predictive models belong to the family of binary classifiers: a component inspected during a maintenance visit is considered as either "failed" or "not failed". Thus, they can predict whether a given part in the engine is likely to be scrapped (output variable Y=1) or not (Y=0) at the next maintenance visit, given the past history of similar components in other engines. Data pre-processing contributes to a better prediction of maintenance needs by generating a dataset with suitable characteristics for statistical modelling.

This paper presents the evaluation of the impact of DPPT on the performance of predictive SHM of jet engines. The empirical evaluation is based on an industrial case study involving actual data acquired on Rolls-Royce's Trent 500 jet engines over the period 2002-2012. The dataset comprises a total of 12132 serviced components, corresponding to 337 maintenance visits performed on 176 different engines. The number of components serviced during the maintenance visits comes from the analysis of maintenance invoices while the dozens of predictors of the model corresponds to engine parameters (temperature, pressure, vibrations, rotations speeds…) extracted from the Engine Health Monitoring (EHM) database. The data pre-processing techniques covered in this paper are applied only to the predictors acquired by the EHM system.

After a brief presentation of the data pre-processing framework in Section 1, we cover in Section 2 the methodological aspects, including details of each steps of the framework. The Section 3 quantifies the impact of the data pre-processing techniques on the predictive performance of the SHM algorithms. Discussion of the results, opening to new problems and directions for future research closes the article.

1 FRAMEWORK FOR DATA PRE-PROCESSING

The pre-processing of data for Structural Health Monitoring of aerospace gas turbine proposed in this article is constituted of four steps (Figure 1). Raw data from the sensors are treated to handle the outliers, in our case by simply removing them. In the next step, variable without outliers are scaled before being selected by two concurrent techniques: assessment of correlation and quantification of variable importance given by an adequate Machine Learning algorithm (random forest). Finally, a test of the predictive performance evaluates the gain obtained by using a combination of the pre-processing techniques. The various steps are described in more details in the next section.

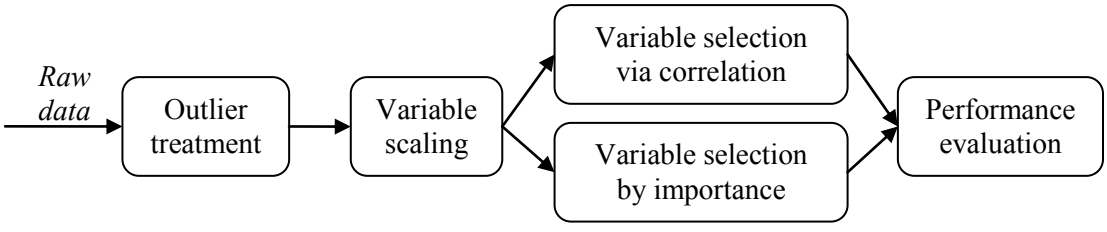

Figure 1: Framework for pre-processing of SHM data

For the sake of conciseness and respect of confidentiality, the framework presented in this article is simplified. Nonetheless, it offers a number of benefits: simplicity, robustness, rapidity and evaluability.

2 METHODOLOGICAL DETAILS ABOUT THE DATA PRE-PROCESSING TECHNIQUES

This section presents the details of the steps in the framework in Figure 1.

2.1 Outlier treatment

Raw data are acquired from dozens of sensors measuring heterogeneous physical parameters over several years in demanding conditions, which increases the likelihood of encountering outliers. Errors and outliers can be due to incorrect offsets, trends due to instrument decalibration, interruptions in acquisition, environmental disturbances… In this section, we are focusing on outliers regardless of their origin and can classify them into several categories (Figure 2).

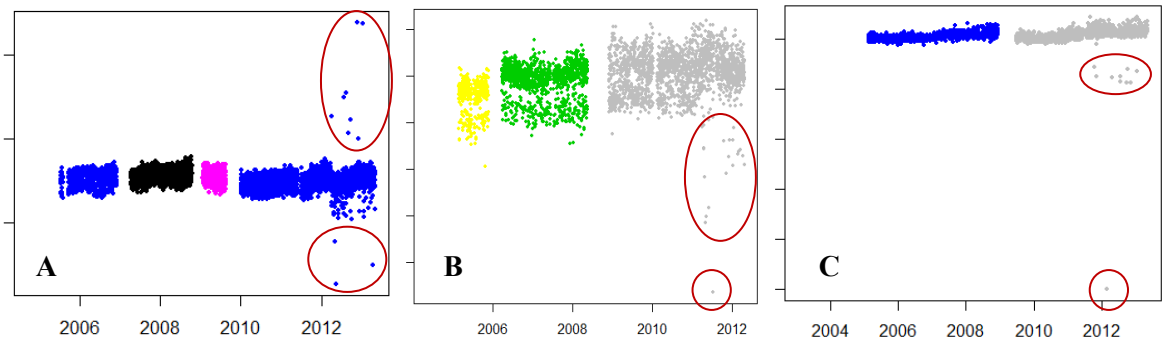

Figure 2: Examples of light (A), strong (B) and extreme (C) outliers for various engines and parameters

We defined outliers in our dataset according to a rather conservative way, as values departing from the mean of the variable by plus or minus four standard deviations of the probability distribution. After identifying the outliers variable by variable, we can apply several imputation techniques to handle them: simply remove them, replace them by the mean or median of the variable or apply (multiple) regression techniques with or without grouping by other categorical variables. Given the hundreds of data points and the relatively low proportion of outliers, we opted for the simplest solution, namely removal of the outliers. This decision has been confirmed by verifying that the moments and medians of the variables were not significantly affected (typically by less than 1%).

2.2 Variable scaling

Machine Learning techniques can be sensitive to input variables with heterogeneous scales [10]. SHM sensors measuring a large variety of physical parameters in jet engines, our dataset contains explanatory variables with scale ratios varying from 1 to 100. Scaling the variables enables us to attribute similar « weights » or importance to all the explanatory variables when estimating the parameters of the Machine Learning algorithms. To simplify the data pre-processing framework and ensure comparability between variables, the scaling function is identical for all the predictors:

$$Y_{scaled} = \frac{Y - \overline{Y}}{\max(Y) - \min(Y)} \quad (1)$$

Despite the standard deviation being the most popular choice of denominator in the scaling function, we selected the range R=max(Y)-min(Y) for its robustness and increased reduction of the scale ratios between the variables, as confirmed by comparing probability distributions before and after scaling.

2.3 Two concurrent methods of variable selection

After removing the outliers and scaling the variables, the next step in the framework consists in selecting the variables with the highest influence on the output variable. Indeed, out of the hundreds of variables present in our dataset, some exhibit a weak dependence with the explained variable while others are strongly correlated: selecting a subset of variables often leads to reduce the variance of the input variables and ultimately to increase the predictive performance of the models.

Variable selection via the quantification of correlation

We selected Pearson's correlation coefficient to measure the dependence between pairs of variables in the dataset. It is indeed the most popular correlation coefficient and is easy to interpret. Moreover, Pearson's coefficient is fast to compute on large datasets (millions of observations and hundreds of variables), contrary to Kendall, Spearman or energy-based correlation coefficients. However, it only quantifies the linear correlation between two variables and thus preliminary checks should be done to ensure that no nonlinear dependences between variables are present in the dataset. Such preliminary checks are done by examining bivariate scatterplots for every pair of variable and for several categorical variables in order to identify potential underlying clusters (Figure 3). Since our dataset contains 50 parameters in time series and 3 categorical variables are used to define clusters, we examined 3*50*49/2=3675 scatterplots. Although visual checking is relatively time-consuming, it has to be done only once as the structure of the dataset doesn't evolve significantly over time as new observations are recorded.

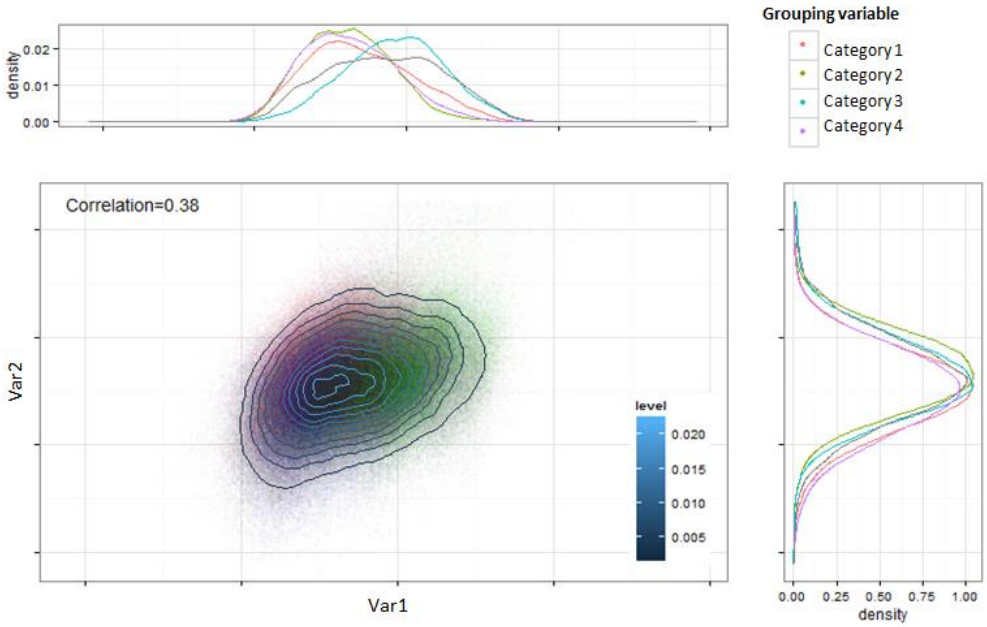

Figure 3: Example of a bivariate scatterplot used to check linear dependence between two variables

The identification of the correlated variables is the second step in the process of variable selection. Since the visual inspection of the scatterplots ensured that dependences between all pairs of variables are linear, Pearson's coefficient can be used to this effect. The correlation matrix is displayed for all the pairs of variables in the dataset (Figure 4): it assesses the strength of the dependence and clusters the variables according to the complete-linkage hierarchical clustering method. Linearly correlated variables essentially carry similar information and can be considered redundant: we finally kept 13 variables that are either the most independent (i.e. with the lower

number of high correlations) or "pivotal" (i.e. correlated with many variables as given by a high number of high correlation coefficients).

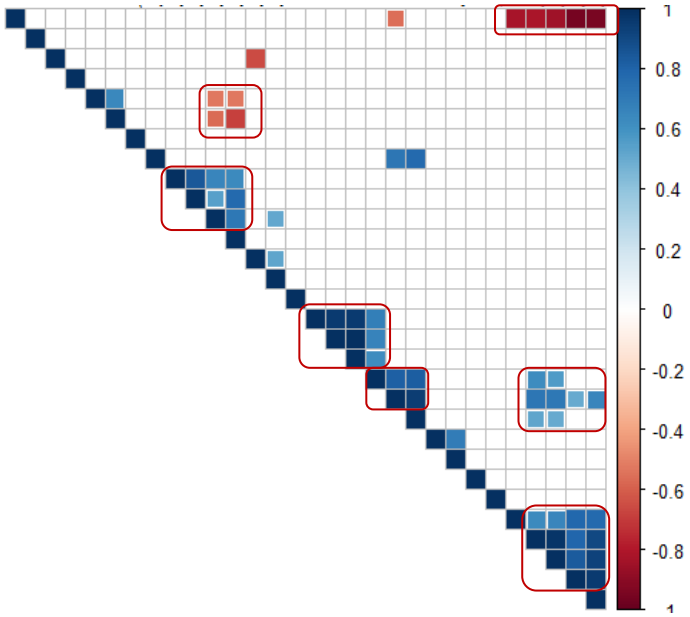

Figure 4: Correlation matrix with anonymized variables grouped by hierarchical clustering (red rectangles)

Variable selection via importance quantification

Another variable subset selection method consists in ranking all the variables in the dataset by their influence on the output variable. Amongst all the variable selection methods available in the literature, we chose the quantification of variable importance returned by random forests, a recent efficient Machine Learning technique [11]. The importance v_k of the k^{th} predictor is measured by the summing and standardizing on m trees the difference $d_i = e_i - p_i$ of binary classification error rates on the i^{th} tree between the out-of-bag classification error rate before (e_i) and after (p_i) the permutation of the k^{th} predictor:

$$v_k = \frac{\overline{d}}{s_d} = \frac{\frac{1}{m}\sum_{i=1}^{m} d_i}{\sqrt{\frac{1}{m-1}\sum_{i=1}^{m}(d_i - \overline{d})^2}} \tag{2}$$

We ranked all the variables in the dataset by computing their individual importance v_k so as to keep only the 30 most influential ones for later predictive SHM algorithms.

2.4 Method for assessing the impact of the pre-processing techniques

We evaluated the data pre-processing techniques through their impact on the predictive accuracy of the SHM algorithms. Within the large family of existing Machine Learning algorithms, we selected random forests, a recent technique from the field of Machine Learning, as the model for predicting the failure of jet engine components for three main reasons: 1) coherence with the estimation of

variable importance performed in the previous step of the framework, 2) conciseness (it is not strictly necessary to compare several similar techniques) and 3) its renowned predictive performance.

The predictive performance of a binary classifier can be assessed by several criteria: the misclassification rate, the prediction accuracy, the area under the Receiver Operating Characteristics curve (AUROC), the area under the Precision-Recall curve (AUCPR)… We selected the AUROC as it is more adapted to our situation and is more robust than the simple misclassification rate or prediction accuracy. The AUROC indeed measures the "probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance" [12] and as such quantifies the quality of the classification of the binary classifier over the full range of a parameter or classification threshold.

3 IMPACT OF THE DATA PRE-PROCESSING ON PREDICTIVE ACCURACY

After selecting random forests as the binary classifier and the AUROC as the measure of predictive performance, we estimated via 100 simulations the gains or losses in performance - as compared to the reference situation (i.e. use of the initial raw dataset) - obtained for each of the 11 combinations of the 4 pre-processing techniques (Table 1). The reference absolute value is not provided to respect the agreement on industrial confidentiality.

Table 1: Impact of the pre-processing techniques on predictive power of SHM algorithms.

Pre-processing techniques	Relative impact on AUROC
None	Reference value
No outliers + 30 RF-selected variables	+1.14%
30 RF-selected variables	+.084%
Scaling	+0.82%
No outliers	+0.47%
No outliers + Scaling	+0.46%
No outliers + Scaling + 30 RF-selected variables	+0.26%
Manual variable selection	-0.09%
Scaling + Manual variable selection	-0.34%
No outliers + Scaling + Manual variable selection	-0.46%
Scaling + 30 RF-selected variables	-1.15%
No outliers + Manual variable selection	-1.21%

First of all, the reference situation occupies an intermediate position in terms of predictive performance, as 6 (resp. 5) combinations are more (resp. less) accurate. Regarding the isolated influence of individual pre-processing techniques, selecting the 30 most important variables through random forests ("30 RF-selected variables" item) has the highest positive impact, followed closely by variable scaling and outlier removal to a lower extent, while the "manual" variable selection via the correlation matrix exerts a slightly negative impact on predictive performance. Regarding the combined influence of pre-processing techniques, the selection the variables via random forests improves the performance in 75% of the combinations it appears in, against 66% for removing the outliers, 50% for scaling the predictors while the manual variable selection has always a negative impact. Finally, the highest increase in predictive performance is obtained be removing outliers and selecting variables via random forests.

DISCUSSION & CONCLUSION

The article presented a simple framework for data pre-processing applied to predictive performance of aerospace gas turbines, cumulating a number of advantages: understandable by non-experts, robust and easy to implement, computationally efficient to be deployed on production systems. The research showed that pre-processing raw data doesn't always have a positive impact on predictive performance of SHM binary classifiers such as random forests. This result is at odds with engineering intuition and a large part of the literature on signal processing and Machine Learning. We may explain it by two reasons: 1) recent techniques such as random forests are very robust to unprocessed datasets; 2) the raw data in our particular case study is of high quality. The second point can be explained by stringent requirements of jet engines monitoring system; yet, a secondary analysis of the dataset showed minor variations in the quality of predictors, probably due to differences between sensors or a higher variability in some physical phenomena or location in the engine. For example, parameters in the high pressure system or related to temperature have more outliers than the ones in the low pressure system or related to pressure and shaft rotation speeds. In any case and regardless of whether the impact is positive or negative, the influence of a given pre-processing technique on the quality of the prediction seems to go in the same direction, should they be used alone or combined with other

In terms of applicability to other case studies, the results highlight the importance of a careful analysis of cost-benefits. High quality datasets might not need computer-expensive and time-consuming data processing in order to achieve – at best - a marginal improvement of predictive performance. On the contrary, cases involving raw data with lower quality might require a rigorous quantitative assessment of combinations of pre-processing steps to ensure satisfactory results.

Future research could focus on evaluating the impact of other data pre-processing techniques (PCA, ICA…) on the predictive performance of the subsequent SHM models. However, early results, not presented in this paper, seem to demonstrate that the gain over the aforementioned simplified framework might not be important enough to justify the associated increase in computing resources, development time, maintenance cost and decrease in interpretability or usability of the models. Another research path might consist in comparing the results obtained on random forests with other binary classifiers (e.g. logistic regression, support vector machines, gradient boosted trees or neural networks) and measures of prediction performance (e.g. misclassification rates, prediction accuracy or AUCPR).

REFERENCES

[1] C.R. Farrar, K. Worden. An introduction to structural health monitoring. *Phil. Trans. R. Soc. A*, 365(1851) 303-315, 2007. doi: 10.1098/rsta.2006.1928

[2] H. Sohn, C.R. Farrar, N.F. Hunter, K. Worden. Structural Health Monitoring Using Statistical Pattern Recognition Techniques. *J. Dyn. Sys., Meas., Control*, 123(4): 706-711, 2001. doi:10.1115/1.1410933

[3] R. Baragona, F. Battaglia. Outliers detection in multivariate time series by independent component analysis. *Neural Comput.*, 2007, 19(7): 1962-84

[4] S. Borguet, O. Léonard. Coupling principal component analysis and Kalman filtering algorithms for on-line aircraft engine diagnostics. *Control Engineering Practice*, 17(4): 494–502, 2009. doi: 10.1016/j.conengprac.2008.09.008

[5] R. Genuer, J-M. Poggi, C. Tuleau-Malot. Variable selection using random forests. *Pattern Recognition Letters*, 31(14):2225–2236, October 2010

[6] S. da Silva, M. Dias Júnior, V. Lopes Junior, M.J. Brennan. Structural damage detection by fuzzy clustering. *Mechanical Systems and Signal Processing*, 22(7): 1636–1649, 2008. doi: 10.1016/j.ymssp.2008.01.004

[7] W. Sun, J. Chen, J. Li. Decision tree and PCA-based fault diagnosis of rotating machinery. *Mechanical Systems and Signal Processing*, 21(3): 1300–1317, 2007. doi: 10.1016/j.ymssp.2006.06.010

[8] M.J. Zuo, J. Lin, X. Fan. Feature separation using ICA for a one-dimensional time series and its application in fault detection. *J. of Sound and Vibration*, 287(3): 614–624, 2005. doi:10.1016/j.jsv.2005.02.005

[9] M. Gul, F. Necati Catbas. Statistical pattern recognition for Structural Health Monitoring using time series modeling: Theory and experimental verifications. *Mechanical Systems and Signal Processing*, 23(7): 2192–2204, 2009. doi: 10.1016/j.ymssp.2009.02.013

[10] T. Hastie, R. Tibshirani, J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Verlag, 2009

[11] R. Genuer, J-M. Poggi, C. Tuleau-Malot. Variable selection using random forests. Pattern Recognition Letters, 31(14):2225–2236, 15 October 2010

[12] T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861-874, 2006. doi:10.1016/j.patrec.2005.10.010