Loukopoulos P, Sampath S, Pilidis P, et al., Dealing with missing data for prognostic purposes, 2016 Prognostics and System Health Management Conference (PHM-Chengdu), 19/10/2016 - 21/10/2016. DOI: 10.1109/PHM.2016.7819934.

# Dealing with missing data for prognostics purposes

Panagiotis Loukopoulos, Suresh Sampath, Pericles Pilidis
School of Aerospace, Transport and Manufacturing
Cranfield University
Cranfield, UK

George Zolkiewski, Ian Bennett
Projects and Technology
Shell Global Solutions
Rijswijk, Netherlands

Fang Duan, David Mba
School of Engineering
London South Bank University
London, UK
mbad@lsbu.ac.uk

***Abstract*— Centrifugal compressors are considered one of the most critical components in oil industry, making the minimisation of their downtime and the maximisation of their availability a major target. Maintenance is thought to be a key aspect towards achieving this goal, leading to various maintenance schemes being proposed over the years. Condition based maintenance and prognostics and health management (CBM/PHM), which is relying on the concepts of diagnostics and prognostics, has been gaining ground over the last years due to its ability of being able to plan the maintenance schedule in advance. The successful application of this policy is heavily dependent on the quality of data used and a major issue affecting it, is that of missing data. Missing data's presence may compromise the information contained within a set, thus having a significant effect on the conclusions that can be drawn from the data, as there might be bias or misleading results. Consequently, it is important to address this matter. A number of methodologies to recover the data, called imputation techniques, have been proposed. This paper reviews the most widely used techniques and presents a case study with the use of actual industrial centrifugal compressor data, in order to identify the most suitable ones.**

***Keywords- missing data; imputation techniques; centrifugal compressor; prognostics***

## I. INTRODUCTION

Centrifugal compressors are one of the most critical components in oil industry, making the minimisation of their downtime and the maximisation of their availability a major target. Maintenance is considered a key aspect towards achieving this goal, leading to various maintenance schemes being proposed over the years [1], [2]. Condition based maintenance and prognostics and health management (CBM/PHM) [3], which is founded on the principles of diagnostics and prognostics, has been gaining popularity over the past years. It consists of three steps [3]–[6], and is presented in figure 1. Various types of measurements streaming from the machine, called condition monitoring data, are collected, pre-processed, analysed to extract the useful information and then fed to the CBM block. There, the status of the machine is determined and is presented as an input to the PHM block where, in the case of a fault, this information is used to estimate the time to failure, called remaining useful life (RUL), of the machine. As a result, it is possible to plan the maintenance schedule in advance which is why CBM/PHM is considered a step closer towards achieving minimisation of downtime.

The success of this methodology is heavily dependent on the quality of the data used due to its sequential structure. Missing data is one of the major issues that affect quality. It is a frequent phenomenon in industry that can manifest in various ways like sensor failure. According to [7], missingness is highly correlated with the amount of data gathered where higher amount of data recorded increases the probability of missing data. The presence of missing values may compromise the information contained within a set, introducing bias or misleading results [8], [9]. Consequently, as noted in [9], it is important to address this matter. To deal with this problem, various methodologies to recover the data, called imputation techniques, have been proposed. The purpose of this paper is to apply the most widely used imputation techniques in order to identify the most suitable ones regarding their accuracy, for centrifugal compressor data for prognostics purposes.

## II. LITERATURE REVIEW OF IMPUTATION TECHNIQUES

Although various imputation techniques have been developed [7], [9]–[27], to the authors' knowledge, none of them has been applied to centrifugal compressor data, though they have been applied successfully in other fields like biological studies. In [28], they compared Bayesian principal component analysis (BPCA) with singular value decomposition (SVD) imputation and k-nearest neighbours (KNN) imputation, applied to DNA microarray data where BPCA outperformed the other methods. In [10], they studied the imputation performance of various principal component analysis (PCA) methods, using artificial data was well as actual data from Netflix. For the artificial data, BPCA outperformed each method though it was the most time consuming. In the case of the high dimensional and sparse Netflix data, BPCA was not feasible being very computationally expensive on the authors' computer. The best method for these data was a variation of BPCA, called BPCAd, which was created in order to deal with cases of high dimensional sparse data. In [11], they presented a software package containing various PCA methods, applied to microarray data. BPCA was the best method while probabilistic principal component analysis (PPCA) was the fastest and is

recommended when dealing with big data sets. In [29], they applied PPCA,BPCA, cubic spline interpolation and historical imputation to impute missing values on traffic data sets where BPCA outperformed the rest. In [14] they compared the internal imputation offered by the classifiers C4.5 and CN2 with that of KNN imputation and mean imputation, with KNN being the best. In [19], they applied mean imputation, normal ratio method, normal ratio with correlation method, multilayer perceptron network and multiple imputation to meteorological time series data sets. Multiple imputation, although being the most computationally demanding, was more robust and outperformed the rest. In [20], nearest/linear/cubic interpolation methods, regression based imputation, KNN, self-organising maps (SOM), multilayer perceptron neural network, hybrid methods and multiple imputation (MI), were compared to air quality data. MI despite being the slowest, offered the best results. In [30], they compared the self-organising maps (SOM) with linear regression and back propagation neural network when applied to water treatment time series data. The SOM outperformed the rest. In [31], They applied k-nearest neighbours, single value decomposition imputation and mean substitution on DNA microarray data, with KNN being the best.

## III. Application of Imputation Techniques to Compressor Data

### A. Ad Hoc method

This method replaces the missing values with a fixed one [12]. Common fixed values are the mean (method 2), the median (method 3) [12], [14], and the last measured value carried forward (method 1) [12]. The implementation was done in Matlab environment.

### B. Interpolation methods

These methods fit a curve along the missing data and try to estimate their values. There are various interpolation methods that can be used [20]: i. nearest neighbour (method 4), ii. linear interpolation (method 5), iii. cubic interpolation (method 6). Their implementation was done in Matlab [32].

### C. Time series methods

According to [33], the observed data are used to train a model to predict the missing values. This method (method 7) can be enhanced with the combination of forward and backward prediction [34], where data before and after the missing data are used to train two separate models and then predict the missing values by averaging the two predictions through an iterative procedure. The model used was an autoregressive model (AR), [33], and was selected for its simple structure. The analysis was done in Matlab [35].

### D. Self-organising maps (SOM)

Self-organizing map (SOM) (method 8) is used to project multidimensional data into a two-dimensional structure [20], [30], [36] in a way so that data with similar patterns are associated with the same neurons (best matching unit – BMU) or their neighbours [30]. The map is constructed with the available data. Then the data along with the missing values are given to the map to calculate their BMU. The missing values are estimated as their corresponding BMU values. This method was applied with the help of the SOM Toolbox for Matlab which can be found in (http://www.cis.hut.fi).

### E. K-nearest neighbours (KNN)

Assuming a variable within a set contains a missing value, KNN imputation (method 9) uses the K other variables that don't have a missing value at the same time stamp and are most similar to that variable, to estimate the missing sample [23],[31]. The number of neighbours (K) used affects strongly the performance of this method but there is not a global rule for selecting K, which has to be done intuitively. The analysis was done in Matlab [37].

### F. Bayesian principal components analysis (BPCA)

Principal components analysis (PCA) is the linear projection (scores) of the data where the retained variance is maximum (principal components) [38]. Probabilistic principal component analysis (PPCA) [39], is an extension where the principal components are assumed to have a prior distribution. Bayesian principal component analysis (BPCA) (method 10) [38], is a further enhancement where on top of that, the optimum number of principal components is selected automatically. In [28], BPCA is applied to the problem of imputing missing data where the missing values are estimated based on the observed ones, via an iterative procedure. It was implemented in Matlab using (http://ishiilab.jp/member/oba/tools/BPCAFill.html).

### G. Multipe imputation

Multiple imputation (MI) (method 11) was introduced, [13], [27], [40] to take into account the uncertainty that is caused by the existence of the missing data, by creating $m$ complete data sets [12], [15], [22], [27]. Usually, a small number of sets is adequate m=3-5. The backbone of the method is the data augmentation algorithm [27], a two-step iterative procedure where the missing data are simulated. The analysis was done in R [41] which is based on [42]. For this project, 10 sets were created and pooled together to form the final one.

### H. Types of missing data

According to [13]–[16], [21], [22], [25]–[27], [29], there are three mechanisms of missing data: i. missing completely at random (MCR), ii. missing at random (MR), iii. missing not at random (MNR). Most methods used in the paper, can perform only under the assumption of the first two types, which if broken, the analysis results can be of low quality, biased or misleading [13], [16], [21], [25]. In this paper, the missing at random (MR) type is assumed.

## IV. Missing Data Analysis

### A. Performance analysis

The data employed for this study were taken from an operational industrial compressor. After the preliminary analysis, it was observed that in 92% of the sets, the missing data had the form which can be seen in figure 2. For a given set, there is a single group of missing values for a specific variable within the set, with observed values before and after it. This was

decided to be called as continuous missingness and was the focus of the project.

For the analysis, a complete set of with 474 samples containing 25 variables was selected. Five percentages of missing data were simulated: 1, 5, 15, 25 and 50%. At any given time, only one variable within the set contains missing values, since as stated above, at any time only one variable presents missing data. For each percentage and the chosen variable, a sliding window with span the percentage of missingness translated into samples, starting from the beginning of the set, passes across the signal and removes the respective samples, creating new sets. This way, the effect of the position of the missing data on the quality of the imputation is also considered. Consequently, for each variable the new sets with missing data that are created are: 470 for 1%, 451 for 5%, 404 for 15%, 356 for 25% and 238 for 50%. In total, the number of sets created and analysed was (470+451+404+356+238)*25=47975. The metric used for the benchmarking was the normalised mean square error (NMSE) as given in [35].

For space reasons, only the results regarding the NMSE for 50% of missingness are presented in figure 3. In the x-axis, ranging from 1-25, are the variables within the set while in the y-axis, ranging from 1-11, are the methods used for the imputation. The graph is separated in a number of boxes, which are the combination of each variable and each method. Within each box there are colour-coded lines, regarding their NMSE value, representing the location of the missing data. Although NMSE ranges from $-\infty$ (no fit) to 1 (perfect fit), for scaling reasons the results range from -1 to 1. For example, box [1,1] corresponds to the results of applying method 1 to variable 1. For 50%, as mentioned previously, it contains 238 lines with the first line (top of the box) corresponding to missing data begin at the beginning of the signal and the last line (bottom of the box) to those at the end of the signal. While descending (moving from top to bottom), the location of the data shifts from the beginning towards the end of the set. Furthermore, it can be seen that this method performs poorly since the colour of the lines ranges from light to deep blue indicating a negative NMSE value meaning there is a bad fit.

Going through figure 3, regardless the percentage of missingness or the position of the missing data, multiple imputation was superior since most of its boxes ranged from light to dark red indicating high NMSE values. This method is followed by self-organising maps and k-nearest neighbours. Furthermore, the performance of these methods was highly related to the variable they were applied to. Despite having a robust performance for most variables, it can be seen that for some others (9, 17, 20 and 21) their boxes indicate negative NMSE values. Regarding the rest of the methods, they are considerably affected either by the position of the missing data or the percentage of missingness, thus being mostly blue. Also, it should be noted that for the time series method, a more powerful model like neural networks could give better results.

### B. *Computational time analysis*

Another aspect regarding the performance of the imputation method that was examined was time. The average time regarding each method for each percentage of missing data can be seen in table 1. It should be noted that the analysis was conducted in computer with dual core i7 and 6 gb ram. It can be seen that the most time consuming method is multiple imputation, with time increasing significantly as the amount of missing data grows, while the least is the last measurement carried forward.

## V. CONCLUSIONS

Missing data is an important issue that needs to be resolved in order to apply prognostics successfully, with imputation being a common solution. This paper has reviewed and applied the most common imputation techniques to centrifugal compressor data for prognostics purposes. It has been shown that the best and most robust method is multiple imputation followed by self-organising maps and k-nearest neighbours, but are highly dependent on the variables they are applied to. Regarding the computational time, multiple imputation is the most time consuming, but gives the best results. If the desired outcome is a combination of good performance and speed, k-nearest neighbours and self-organising maps are a good choice. Consequently, when applying these methods one must be careful. Based on the work presented in this report, there are a number of aspects that need further exploration:

- The performance of the imputation techniques at the presence of missing data for more than one variable at the same time
- The effect of the imputation techniques regarding the outcome of prognostics.
- The further study of imputation techniques in other rotating equipment areas.

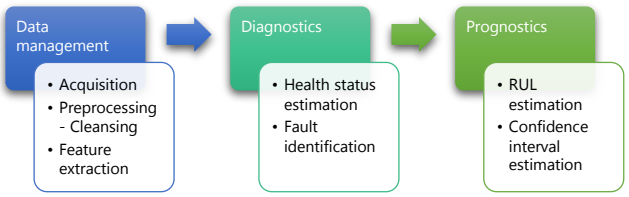


Figure 1 CBM/PHM process

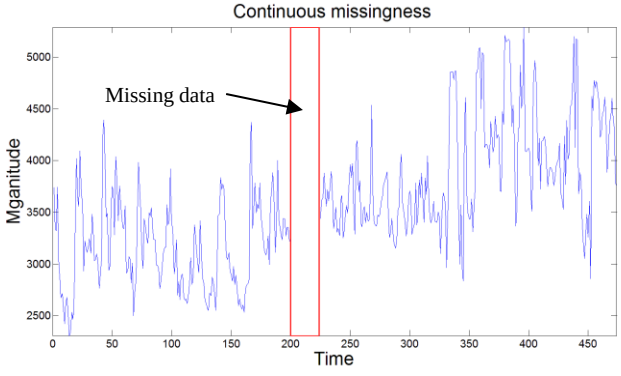

Figure 2 Continuous missingness

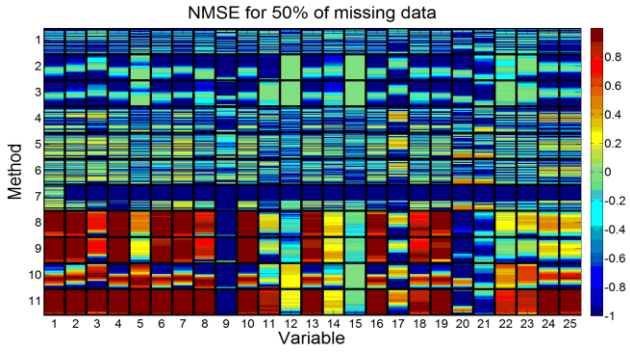


Figure 3 NMSE for 50% of missing data

Table 1 Average running time of imputation methods

| **Method** | **Percentage** | | | | |
|---|---|---|---|---|---|
| | **1%** | **5%** | **15%** | **25%** | **50%** |
| | **Time (s)** | | | | |
| Last measurement | 0.0007 | 0.0007 | 0.0007 | 0.0007 | 0.0007 |
| Mean | 0.0013 | 0.0013 | 0.0013 | 0.0013 | 0.0013 |
| Median | 0.0033 | 0.0033 | 0.0033 | 0.0033 | 0.0033 |
| Nearest interpolation | 0.0018 | 0.0019 | 0.0019 | 0.0019 | 0.0020 |
| Linear interpolation | 0.0018 | 0.0018 | 0.0019 | 0.0019 | 0.0020 |
| Cubic Interpolation | 0.0018 | 0.0019 | 0.0019 | 0.0020 | 0.0021 |
| Time series prediction | 0.5809 | 0.4638 | 0.4441 | 0.3407 | 0.2774 |
| SOM | 0.3707 | 0.3803 | 0.3520 | 0.3540 | 0.3692 |
| KNN | 0.0121 | 0.0125 | 0.0134 | 0.0158 | 0.0231 |
| BPCA | 0.2121 | 0.4167 | 0.9823 | 1.4936 | 2.3460 |
| MI | 2.6978 | 8.8361 | 24.1865 | 39.4370 | 77.9393 |

## REFERENCES


[1] J. Lee, F. Wu, W. Zhao, M. Ghaffari, L. Liao, and D. Siegel, "Prognostics and health management design for rotary machinery systems—Reviews, methodology and applications," *Mech. Syst. Signal Process.*, vol. 42, no. 1–2, pp. 314–334, Jan. 2014.

[2] R. Kothamasu, S. H. Huang, and W. H. VerDuin, "System health monitoring and prognostics — a review of current paradigms and practices," *Int. J. Adv. Manuf. Technol.*, vol. 28, no. 9–10, pp. 1012–1024, Jul. 2006.

[3] G. Vachtsevanos, F. Lewis, M. Roemer, A. Hess, and B. Wu, *Intelligent Fault Diagnosis and Prognosis for Engineering Systems*. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2006.

[4] A. K. S. Jardine, D. Lin, and D. Banjevic, "A review on machinery diagnostics and prognostics implementing condition-based maintenance," *Mech. Syst. Signal Process.*, vol. 20, no. 7, pp. 1483–1510, Oct. 2006.

[5] J. Z. Sikorska, M. Hodkiewicz, and L. Ma, "Prognostic modelling options for remaining useful life estimation by industry," *Mech. Syst. Signal Process.*, vol. 25, no. 5, pp. 1803–1836, Jul. 2011.

[6] Y. Peng, M. Dong, and M. J. Zuo, "Current status of machine prognostics in condition-based maintenance: a review," *Int. J. Adv. Manuf. Technol.*, vol. 50, no. 1–4, pp. 297–313, Jan. 2010.

[7] M. L. Brown and J. F. Kros, "Data mining and the impact of missing data," *Ind. Manag. Data Syst.*, vol. 103, no. 8, pp. 611–621, 2003.

[8] A. Pantanowitz and T. Marwala, "Evaluating the Impact of Missing Data Imputation," *Adma Lnai*, vol. 5678, pp. 577–586, 2009.

[9] A. J. F. Patrick E. McKnight, Katherine M. McKnight, Souraya Sidani, *Missing Data: A Gentle Introduction*. The Guilford Press, 2007.

[10] [10] A. Ilin and T. Raiko, "Practical approaches to principal component analysis in the presence of missing values," *J. Mach. Learn. Res.*, vol. 11, pp. 1957–2000, 2010.

[11] W. Stacklies, H. Redestig, M. Scholz, D. Walther, and J. Selbig, "pcaMethods - A bioconductor package providing PCA methods for incomplete data," *Bioinformatics*, vol. 23, no. 9, pp. 1164–1167, 2007.

[12] N. J. Horton and K. P. Kleinman, "Much Ado About Nothing: A Comparison of Missing Data Methods and Software to Fit Incomplete Data Regression Models," *The American Statistician*, vol. 61, no. 1. pp. 79–90, 2007.

[13] C. K. Enders, "A Primer on Maximum Likelihood Algorithms Available for Use With Missing Data," *Struct. Equ. Model. A Multidiscip. J.*, vol. 8, no. 1, pp. 128–141, 2001.

[14] G. E. a. P. a. Batista and M. C. Monard, "An analysis of four missing data treatment methods for supervised learning," *Applied Artificial Intelligence*, vol. 17, no. 5–6. pp. 519–533, 2003.

[15] T. D. Pigott, "A Review of Methods for Missing Data," *Educ. Res. Eval.*, vol. 7, no. 4, pp. 353–383, 2001.

[16] A. N. Baraldi and C. K. Enders, "An introduction to modern missing data analyses," *J. Sch. Psychol.*, vol. 48, no. 1, pp. 5–37, 2010.

[17] I. Myrtveit, E. Stensrud, and U. H. Olsson, "Analyzing data sets with missing data: An empirical evaluation of imputation methods and likelihood-based methods," *IEEE Trans. Softw. Eng.*, vol. 27, no. 11, pp. 999–1013, 2001.

[18] C. K. Enders, *Applied Missing Data Analysis*. The Guilford Press, 2010.

[19] C. Yozgatligil, S. Aslan, C. Iyigun, and I. Batmaz, "Comparison of missing value imputation methods in time series: The case of Turkish meteorological data," *Theor. Appl. Climatol.*, vol. 112, no. 1–2, pp. 143–167, 2013.

[20] H. Junninen, H. Niska, K. Tuppurainen, J. Ruuskanen, and M. Kolehmainen, "Methods for imputation of missing values in air quality data sets," *Atmos. Environ.*, vol. 38, no. 18, pp. 2895–2907, 2004.

[21] J. W. Graham, "Missing data analysis: making it work in the real world.," *Annu. Rev. Psychol.*, vol. 60, pp. 549–576, 2009.

[22] J. L. Schafer and J. W. Graham, "Missing data: our view of the state of the art.," *Psychol. Methods*, vol. 7, no. 2, pp. 147–177, 2002.

[23] L. Li, Y. Li, and Z. Li, "Missing traffic data: comparison of imputation methods," *IET Intell. Transp. Syst.*, vol. 8, no. 1, pp. 51–57, 2014.

[24] a. R. T. Donders, G. J. M. G. van der Heijden, T. Stijnen, and K. G. M. Moons, "Review: A gentle introduction to imputation of missing values," *J. Clin. Epidemiol.*, vol. 59, no. 10, pp. 1087–1091, 2006.

[25] D. Little, RJA and Rublin, *Statistical Analysis with Missing Data*, Second Edi. 2002.

[26] A. C. Acock, "Working with missing values," *J. Marriage Fam.*, vol. 67, no. 4, pp. 1012–1028, 2005.

[27] J. L. Schafer, *Analysis of Incomplete Multivariate Data*. CRC Press LLC, 2000.

[28] S. Oba, M. Sato, I. Takemasa, M. Monden, K. Matsubara, and S. Ishii, "A Bayesian missing value estimation method for gene expression profile data.," *Bioinformatics*, vol. 19, no. 16, pp. 2088–2096, 2003.

[29] L. Qu, L. Li, Y. Zhang, and J. Hu, "PPCA-based missing data imputation for traffic flow volume: A systematical approach," *IEEE Trans. Intell. Transp. Syst.*, vol. 10, no. 3, pp. 512–522, 2009.

[30] R. Rustum and A. J. Adeloye, "Replacing Outliers and Missing Values from Activated Sludge Data Using Kohonen Self-Organizing Map," *Journal of Environmental Engineering*, vol. 133, no. 9. pp. 909–916, 2007.

[31] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman, "Missing value estimation methods for DNA microarrays.," *Bioinformatics*, vol. 17, no. 6, pp. 520–525, 2001.

[32] *MATLAB® Mathematics*. The MathWorks, Inc., 2015.

[33] L. Ljung, *System Identification: Theory for the User*, Second Edi. Prentice Hall PTR, 1999.

[34] T. A. Moahmed, N. El Gayar, and A. F. Atiya, "Forward and backward forecasting ensembles for the estimation of time series missing data," in *Lecture Notes in Computer Science*, 2014, vol. 8774, pp. 93–104.

[35] L. Ljung, System Identification Toolbox ™ User ' s Guide. The MathWorks, Inc., 2015.

[36] L. Folguera, J. Zupan, D. Cicerone, and J. F. Magallanes, "Self-organizing maps for imputation of missing data in incomplete data matrices," *Chemom. Intell. Lab. Syst.*, vol. 143, pp. 146–151, 2015.

[37] Bioinformatics Toolbox ™ User ' s Guide. The MathWorks, Inc., 2015.

[38] C. M. Bishop, "Variational principal components," *9th Int. Conf. Artif. Neural Networks ICANN 99*, vol. 1999, no. 470, pp. 509–514, 1999.

[39] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis," *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 61, no. 3, pp. 611–622, 1999.

[40] O. Harel and X.-H. Zhou, "Multiple imputation: review of theory, implementation and software.," *Stat. Med.*, vol. 26, no. 16, pp. 3057–3077, 2007.

[41] A. A. Novo and J. L. Schafer, "Package ' norm .'" 2015.

[42] J. L. Schafer, "NORM users' guide (Version 2)." University Park: The Methodology Center, Penn State. Retrieved from http://methodology.psu.edu, 1999.