Computers & Industrial Engineering 179 (2023) 109230

Contents lists available at ScienceDirect

Computers & Industrial Engineering

journal homepage: www.elsevier.com/locate/caie

A novel imputation based predictive algorithm for reducing common cause variation from small and mixed datasets with missing values

Raed S. Batbooti[1,2], Rajesh S. Ransing[*,2]

Zienkiewicz Institute for Modelling, Data and AI, Department of Mechanical Engineering, Faculty of Science and Engineering, Swansea University, Swansea, SA1 8EN, UK

ARTICLE INFO

Keywords:
Common cause variation
Missing data
Predictive analytics
Quality improvement
Tolerance limit optimization
7Epsilon

ABSTRACT

Most process control algorithms need a predetermined target value as an input for a process variable so that the deviation is observed and minimized. In this paper, a novel machine learning algorithm is proposed that has an ability to not only suggest new target values for both categorical and continuous variables to minimize process output variation but also predict the extent to which the variation can be minimized.

In foundry processes, an average rejection rate of 3%–5% within batches of castings produced is considered as acceptable and is considered as an effect of the common cause variation. As a result, the operating range for process input values is often not changed during the root cause analysis. The relevant available historical process data is normally limited with missing values and it combines both categorical and continuous variables (mixed dataset). However, technological advancements manufacturing processes provide opportunities to further refine process inputs in order to minimize undesired variation in process outputs.

A new linear regression based algorithm is proposed to achieve lower prediction error in comparison to the commonly used linear factor analysis for mixed data (FAMD) method. This algorithm is further coupled with a novel missing data algorithm to predict the process response values corresponding to a given set of values for process inputs. This enabled the novel imputation based predictive algorithm to quantify the effect of a confirmation trial based on the proposed changes in the operating ranges of one or more process inputs. A set of values for optimal process inputs is generated from operating ranges discovered by a recently proposed quality correlation algorithm (QCA) using a Bootstrap sampling method. The odds ratio, which represents a ratio between the probability of occurrence of desired and undesired process output values, is used to quantify the effect of a confirmation trial.

The limitations of the underlying PCA based linear model have been discussed and the future research areas have been identified.

1. Introduction

A manufacturing process produces products with consistent quality when it is capable of operating with acceptable variability around the desired process response or key product characteristic values (KPCs). The variation around a target process output is a natural process and is inherently present in any manufacturing processes. This natural variation is referred to as background noise, which results from unavoidable or unknown causes known as common causes (Montgomery, 2009). The other kind of variation that leads to undesired output is defined as special or assignable cause variation. The common cause variation is an allowable variation and hence the process response values remain within the process upper and lower specification limits (USL) and (LSL) respectively, whereas the assignable causes generate process response outside the specification limits. In other words, in the presence of assignable causes the process does not remain capable of operating within the desired specification limits. Both of these variation types are illustrated in Fig. 1. In the statistical process control (SPC) terminology (Montgomery, 2009), the process operating with assignable causes is referred to as out of control process. The process is in control when it is operating in the presence of common causes. The corrective actions are normally taken to remove special cause variation. The variation in the response values is usually associated with the variation of one or more process input or factor values. As a result, often a reduction in the variation of inputs contributes to the reduction in the variation of

* Corresponding author.
E-mail addresses: raed.hameed@stu.edu.iq (R.S. Batbooti), r.s.ransing@swansea.ac.uk (R.S. Ransing).
[1] Current Address: Basra Engineering Technical College (BETC), Southern Technical University, Basra, Iraq 61003.
[2] Both authors share equal credit and have contributed equally to the research.

https://doi.org/10.1016/j.cie.2023.109230
Received 23 July 2022; Received in revised form 27 March 2023; Accepted 2 April 2023
Available online 5 April 2023

Nomenclature

$x^{\#}$	missing part of x
x^{*}	observed part of x
Ω	Odds ratio of optimal response
Φ	Odds ratio of confirmation trial plan
π_v	Probability of avoid response values
π_p	Probability of optimal response values
$\tilde{t}_i$	Projection of score i on a variable
$\tilde{X}$	Data matrix after pre-treatment
B	Number of bootstraps
C_{miss}	Number of missing categorical values
D_e	Diagonal matrix containing the square roots of eigenvalues
D_s	Diagonal matrix containing the standard deviations of the columns of
E	Error matrix
$iqrj$	Interquartile of variable j
L	Loading matrix
Lim^j	Vector of values of variable j that lay in the optimal or avoid plain
LL^j	Lower (minimum) value of Lim_j
L_s	The standardized loading matrix
$L_{s,p}$	The standardized loading matrix for the first p principal components
M	Mean matrix
n_1	Number of quantitative variables
n_2	Number of original categorical variables
n_c	Number of correlated parameters resulted from applying CLI
n_x^j	Length of vector Lim
p	The first significant principal components
PO_j	Percentage of occurrences of dummy variable j
$\hat{\tilde{X}}$	Reconstruct matrix X from PCA model parameters
s_j	Standard deviation of factor j
Th_{max}	Maximum penalty matrix threshold
Th_{min}	Minimum penalty matrix threshold
Th_{op}	Optimal threshold
TL_i	Tolerance limits of factor i
T_p	Score matrix for the first p principal component
UL_j	Upper (maximum) value of Lim_j
V	Matrix of eigenvectors
X	Original data matrix

the response values (Steiner & MacKay, 2004). The statistical process control (SPC) methods (George et al., 2005; Montgomery, 2009) are normally employed to discover the existence of special cause variation. Fig. 2 illustrates a schematic relationship between process input variation A to process output variation B. The input variation A is referred to as special cause variation as it produces a much wider spectrum of variation B on the output. The corresponding variations C and D are within process specifications and represent common cause variation. With the technological enhancement of machines, advanced feedback controls and sensors, it is proposed that it may be possible to further reduce variation D by adjusting variation C in Fig. 2.

A foundry process is a complex process with many sub-processes such as pattern making, mould and core making, melting and pouring process. There are number of casting processes but an example of investment casting process has been discussed in this research. For an investment casting process, the mould (or shell) making process is further divided into sub-processes such as coating and drying processes. Investment casting foundries produce complex shaped and super-alloyed components such as turbine blades for aerospace and power industries and turbocharger wheels for automotive industries. Sometimes, it takes weeks to produce a turbine blade from the initial wax processing stage to the final casting. A typical continual process improvement study may have over hundred measurable process inputs that govern the quality of the final turbine blade.

On average precision foundries lose about 3%–5% of their revenue in rejected or reworked castings. In a foundry environment such a process is referred to as a stable and capable process and is approved by the customer during the product validation stage. Many foundries have higher internal rejection rate. The challenge for foundry process engineers is to be able to make changes to several process parameters (e.g. slight adjustments to the operating ranges of various parameters such as alloy compositions at various stages of melting and pouring process, pouring temperature, and moulding parameters etc.) Undertaking one change at a time is not sufficient. Even for experts, it is not easy to choose critical process variables that can be shown as being responsible for causing the 3%–5% rejection rate which is a representative of common cause variation. The aforementioned rejection rate is a cumulative rate which gives a general indication about the process, which is usually constant. In order to understand the variation of process defects detailed statistics on rejection rates is needed.

Recently, many methods have been developed to interpret prediction from available data of manufacturing process. A data based prediction model is presented for casting surface related defects (Chen & Kaufmann, 2022), where six regression methods are used. A Extremely Randomized Trees Regression model showed the best prediction performance in comparison to remaining five methods, whereas the maximum prediction error is obtained when the ridge regression method is used. The effect of data features (factors) on metal penetration of an iron casting is studied. Three factors from 282 factors showed a significant impact on the output (Uyan et al., 2022). A cloud-based process variable measurement system is developed to extract data. The work included the use of supervised machine learning model to predict the porosity defects in an aluminum low-pressure die casting process. The extreme boosted decision tree (XGBoost) model is used. The obtained results indicate that the model accuracy for predicting the good parts is 87 percent and is equal to 74 percent for defective parts (Sika & Ignaszak, 2020). A knowledge discovery based approach is introduced to reduce the defects of selected iron castings. The data acquisition and data mining methods are used to manage production parameters and to discover parameters that lead to increase or decrease the occurrence of defects. The objective was to use results to discover process knowledge. For surface monitoring and control applications, a novel 3D point cloud surface monitoring method is proposed. It uses an Earth Mover's distance (EMD) based control chart to measure the deviation of the cloud sample from the nominal sample. This helps to locate process shifts when data is collected with laser point cloud technique during an inspection process (Zhao, Lui, Du, Di, & Shao, 2023).

There are many applications of machine learning methods for process control applications. In these applications, a target value for process variables is generally known and the control method is used to bring the process variable value as close to the target value as possible. What if the target value aimed by the process control algorithms is sub-optimal? Can machine learning algorithms detect this situation by observing in-process data and suggests optimal target values and their tolerance limits for multiple process variables? This challenge is addressed in this paper. The major difference of the proposed algorithm is that it is designed to discover optimal target values with corresponding limits for multiple continuous and categorical variables using small observational data sets with missing values and it can predict the combined effect of optimal values on the process response.

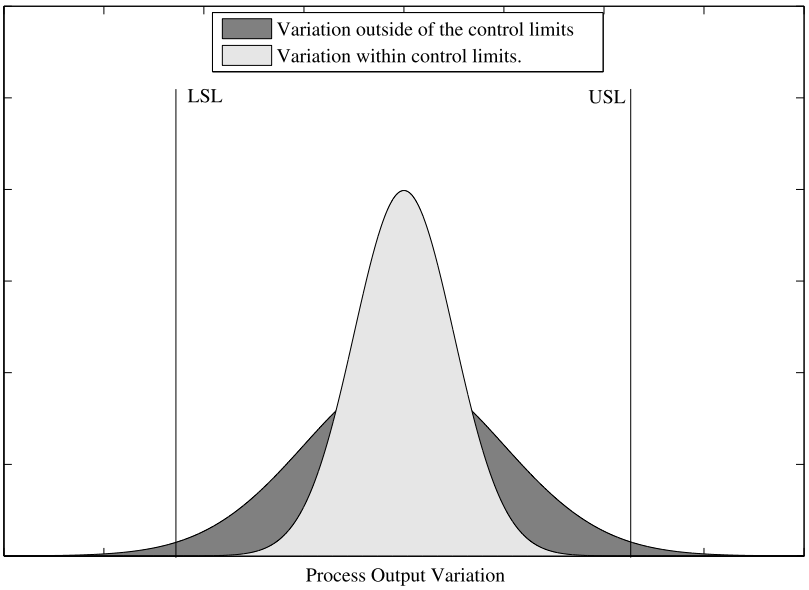

Fig. 1. Visualization of variation in factor values with reference to control limits in a Manufacturing process.

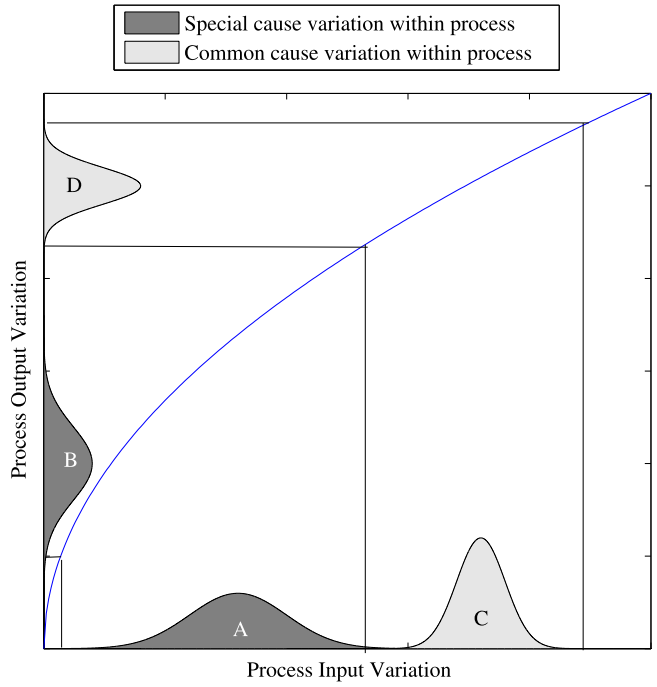

Fig. 2. The influence of input variation on the process output with reference to special (A&B) and common (C&D) cause variations.

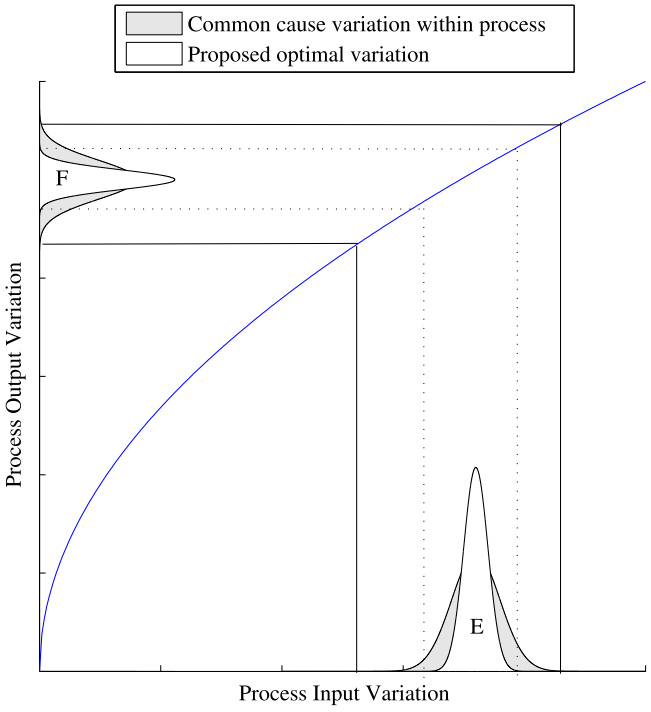

Fig. 3. Minimizing common cause variation to improve process outcome.

After discussions with foundry process engineers, it was discovered that they do try to reduce the common cause variation by further fine tuning the process manually using their experience and expertise. A quality correlation algorithm has been developed recently (Batbooti, Ransing, & Ransing, 2017; Ransing, Batbooti, Giannetti, & Ransing, 2016) to fine tune the process inputs to reduce a deviation from the desired process response (output) values. The developed algorithm is based using the co-linearity index (Giannetti et al., 2014; Ransing, Giannetti, Ransing, & James, 2013) as a measure to discover correlated variables. The principal component analysis (PCA) scores are projected on all variables and responses. The corresponding scores for correlated variable are collected based on direction of variable and response. These scores relate to either optimal or avoid settings with reference to the correlated variable. The observations corresponding to the collected scores generate a new operating range, which is considered as optimal or avoid based on the factor correlation direction. The obtained range is considered as an optimal range if the variable is correlated positively with low penalty values for the response vector. The range is considered as avoid if the variable is correlated positively with high penalty values for the response vector. The new operating ranges obtained by the QCA is equivalent to range E in Fig. 3. One of the objectives of this work is to develop a data based model to predict the corresponding response to the range E discovered by the QCA for all input factors. The schematic presentation of this problem is shown in Fig. 3.

In this work a typical example of an investment casting foundry manufacturing Nickel based superalloy castings is used. The variation

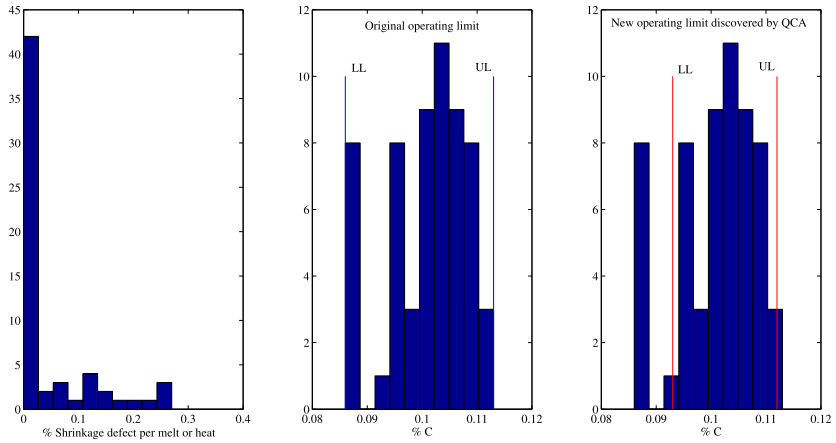

Fig. 4. The variation in rejection rate and the variation in values for an input factor %C for a Nickel based alloy for a dataset with 60 batches.

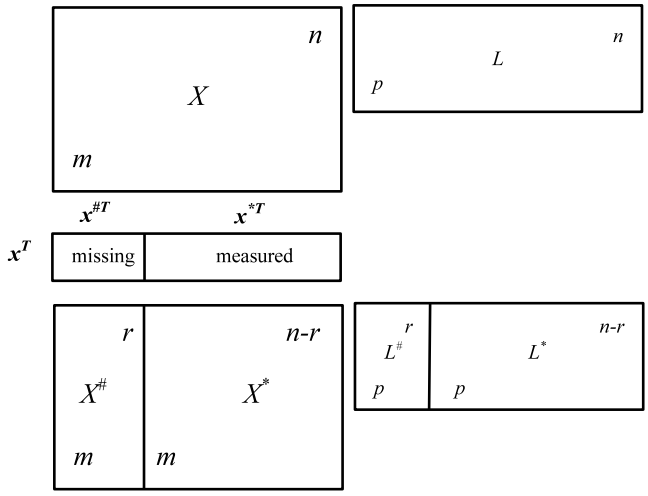

Fig. 5. Data set partition induced by observation x^T (Folch-fortuny, Arteaga, & Ferrer, 2015).

in number of castings rejected due to shrinkage related defects per melt is observed and noted as process response (i.e. castings produced per fixed amount of molten metal and the rejection rate and system parameters observed for each melt or batch).

The variation in the rejection rate with reference to the variation of one process input e.g. factor (%C) is shown in Fig. 4 with lower and upper operating range limits of (LL) and (UL). With reference to Fig. 2, the variation in the rejection rate corresponds to the variation D and the variation in %C in the middle corresponds to variation C. On the other hand, the upper and lower limit in the left corresponds to variation E in Fig. 3.

The overall aim of this work is to develop a data based predictive model to quantify the response corresponding to the operating ranges discovered by the QCA and estimate the QCA operating ranges in presence of missing data. This includes the following objectives:

1. Development of a missing data algorithm to impute missing values for mixed and small manufacturing dataset.
2. Prediction of the process response for any given choice of operating limits on selected input factors of mixed data types.

This paper is structured as follows. Section 2 reviews the missing data machine learning methods. Section 3 describes two PCA based methods, iterative based PCA algorithms and regression based PCA algorithms. This is followed by the proposed new algorithm for mixed datasets in Section 4. Section 5 discusses the use of the proposed missing data algorithm as a predictive tool to estimate the effect of operating limits on the response values on mixed datasets. The paper is concluded in Section 6.

2. Missing data imputation methods

The occurrence of missing data is a common problem in many industrial data sets. This may be due to many reasons, such as data collection errors, incorrect measurements and measuring instruments errors or any other reason can lead to miss the information. Several methods have been proposed in the literature to deal with the missing data. Mean imputation is a very common method based on replacing the missing entry by the attribute or variable mean. This method is very simple, but its limitation is that it underestimates the real variance of the variable (Little & Rubin, 2003). Laaksonen (2000) introduced an imputation method based nearest neighbour algorithm referred to as regression-based nearest neighbour hot decking. A measure of the distance among observations is used to group the data into clusters and the missing observation is replaced with the mean of the nearest neighbour cluster. A family of K-nearest neighbour algorithms has been developed based K-nearest neighbour imputation (Batista & Monard, 2003), weighted K-nearest neighbour imputation (Troyanskaya et al., 2001) and fuzzy K-means clustering imputation (Li, Deogun, Spaulding, & Shuart, 2004).

Schneider (2001) adapted the expectation and maximization (EM) algorithm (Dempster et al., 1977) to analyse an incomplete climate data. The missing values imputed from a conditional probability model. The mean and the covariance matrix (Expectation step) was determined followed by an estimate of the mean and the covariance matrix from observed and imputed observations (Maximization step). The Maximization step was taken into account the conditional estimate of the covariance matrix on imputation error. The iterative solution between the two steps continued until the convergence occurred. Nelwamondo, Mohamed, and Marwala (2007) compared the EM algorithm with an algorithm based on neural networks and genetic algorithms that has been developed by Mussa and Tshilidzi (2005). The difference between the target and actual output used as an objective function and the genetic algorithm used to estimate the missing values by minimizing the introduced objective function. The input in the objective function is represented by both the missing and observed values. The combined input derived from the imputed and observed values is supplied to an auto-encoder neural network. The genetic algorithm used in this work based on a population of string chromosomes, which corresponds to a point in the search space. The Multi-layer perceptron (MLP) network is used to construct an auto-encoder neural network and is trained with

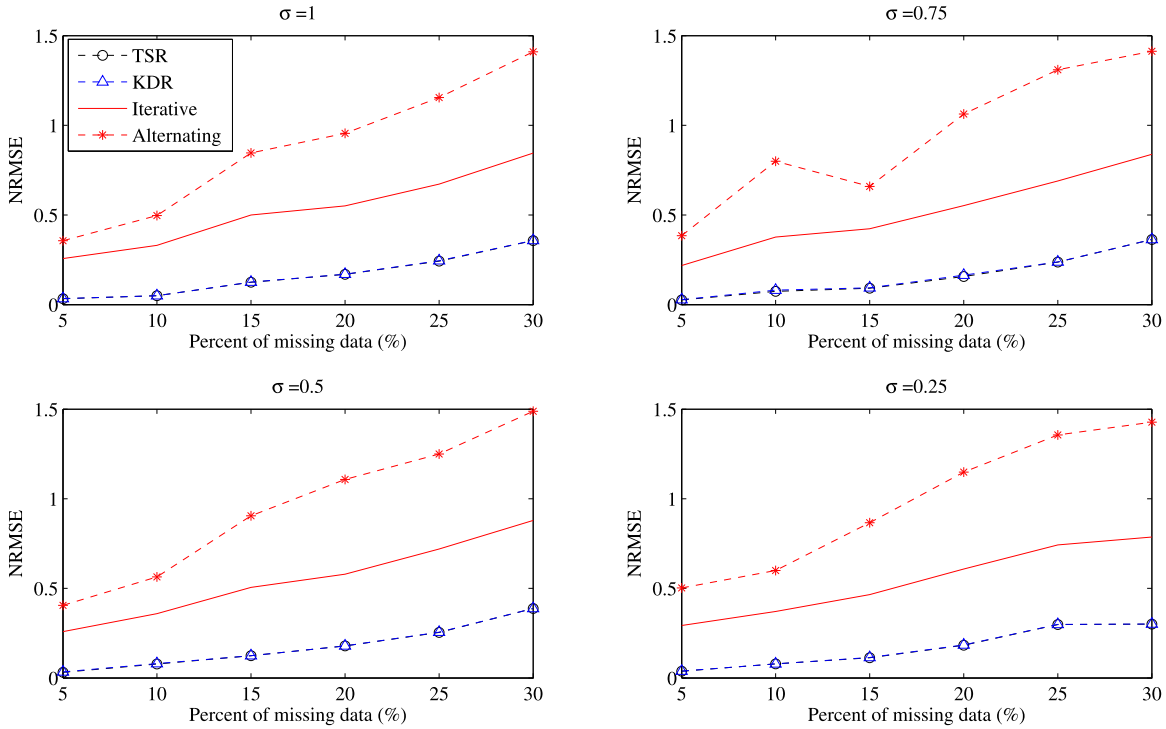

Fig. 6. Comparison of continuous data imputation algorithms.

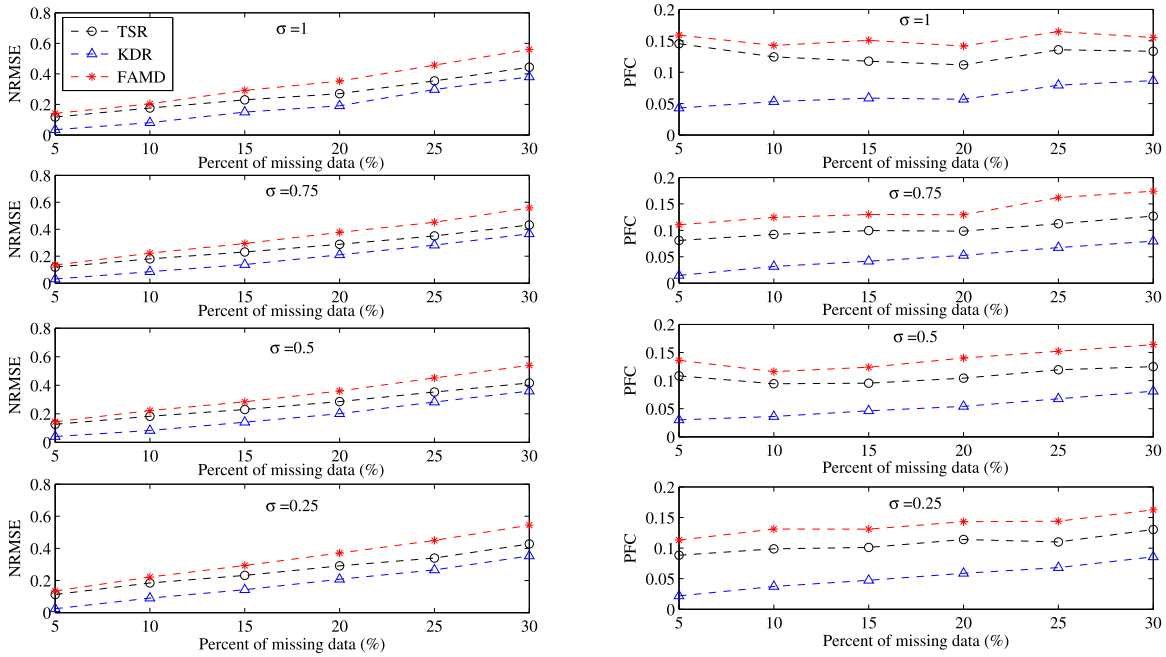

Fig. 7. Model based simulation errors: the plots on the left show quantitative variables error (NRMSE) and the PFC error with different values of σ is shown on the right.

back-propagation algorithm. The comparison of results from four historical data sets showed that the EM algorithm has better performance when there is little or no interdependency between the variables. The auto-associative neural network and genetic algorithm combination is used in cases where there is non-linear relationship between some of the given variables. However, genetic algorithms typically require large datasets.

Many missing value imputation methods in the literature are based on the Principal Component Analysis (PCA) for continuous data. The principal components on the complete dataset provide a new low rank subspace that achieves the maximization of variability of the projected data. The new projection aims to find two matrices $T_{m\times p}$ (the score matrix)and $L_{n\times p}$ (the loading matrix) such that the following reconstruction error is minimized (Diamantaras & Kung, 1996):

$$C = \|X - M - TL^T\|^2 = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(x_{ij} - m_j - \sum_{k=1}^{p} t_{ik}l_{jk}\right)^2 \quad (1)$$

Where:

$X_{m\times n}$: is the data matrix with rows represents the observations and the columns corresponding to the variables.

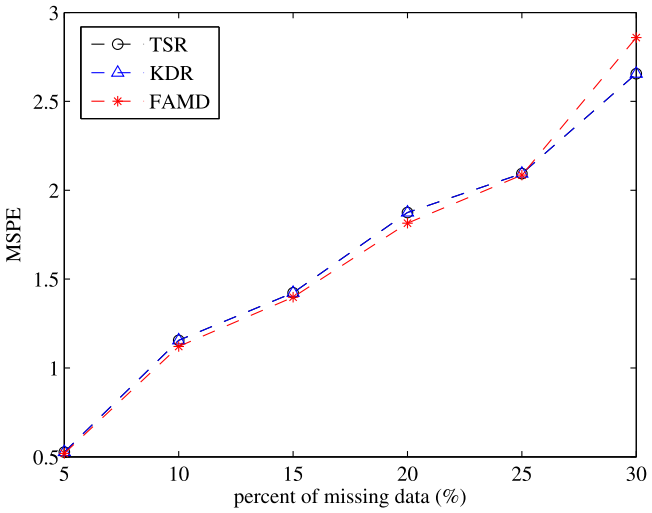

Fig. 8. Quantitative error for a non linear variable imputed as a continuous variable.

$M_{m\times n}$: the mean matrix that has the means of each columns of the data matrix $X_{m\times n}$ in each row.

In seventies, Christoffersson (1970) presented a procedure for missing data based on optimizing a least square problem for one component PCA. The loading matrix L was held constant while the score matrix T was optimized. Then the score matrix was fixed to optimize the loading matrix. The optimization target was to minimize a cost function as given in Eq. (1) for observed data. This resulted in an update rule for principal components (or loading matrix) L and mapping (or score) matrix T. This procedure was extended by Grung and Manne (1998) to include more than one principal component for missing data problems. The obtained results were further improved by Ilin and Raiko (2010) by updating the bias term in updating rules.

It should be noted that the alternating algorithm procedure is not efficient for a large number of principal components (Roweis, 1998) and is shown to have convergence properties only on a limited number of principal components (Ilin & Raiko, 2010). The computational cost of alternating algorithm can be improved by using gradient descent algorithm and Newton's method for optimization (Ilin & Raiko, 2010; Raiko, Ilin, & Karhunen, 2008).

A PCA imputation method that is widely used to impute missing values is referred to as an iterative PCA algorithm. It is based on minimization of the cost function. It introduces a weighted matrix with matrix element value of zero if there is missing value in the original dataset or one otherwise. The missing values are initialized with a mean, or any other value, followed by performing PCA on the complete data and then the missing values are reconstructed from the PCA projection space in an iterative procedure (Husson & Josse, 2013; Josse & Husson, 2012a). The iterative PCA is equivalent to an expectation maximization algorithm associated to the PCA model (Ilin & Raiko, 2010; Josse & Husson, 2012a), it has been called as EM-PCA algorithm (Josse & Husson, 2012a). Ilin and Raiko (2010) showed that the reconstruction step of imputation algorithm is corresponding to the E-step of the EM algorithm and the M-step of the EM algorithm is equivalent to performing the PCA on the complete dataset. It is also shown that the minimization of cost function with respect to the variation of the noise, as assumed in the probabilistic PCA (PPCA) model, has no effect on the imputation algorithm steps. The PPCA provides a Bayesian treatment of PCA that can be combined with EM algorithm to estimate the PCA model parameters iteratively. A Factorial Variational Approximations solution based PPCA, called as VBPCA, is introduced to deal with high-dimensional sparse data sets with high percent of missing values. The VBPCA showed better performance compared to the standard EM-PCA and iterative algorithm, but it computationally expensive. On the other hand, the iterative algorithm has ability to adapt with different missing data methods, such as regression based methods.

Regression based methods substitute the missing values by regressing the unknown data from observed data. Regression based methods have been developed, compared and studied in the prescience of missing data multivariate problems (Arteaga & Ferrer, 2002, 2005). The study included the standard imputation algorithm and other algorithms. Recently, Folch-fortuny et al. (2015) compared PCA regression based methods namely the trimmed score regression method (TSR) and the known data regression method (KDR) with other iterative algorithms (IA). The study built PCA models based on iterative algorithms and applied the developed algorithms for real case studies from literature. The regression based methods (TSR and KDR) showed fast and better performance in comparison to other methods such as standard iterative imputation algorithm.

Another point to note is that the PCA iterative algorithm has a mixed data version developed recently, which can be used to input missing values by using methods based on the factorial analysis for mixed data (FAMD) (Audigier, Husson, & Josse, 2016). FAMD is based on principal components method to describe and visualize multidimensional mixed data matrix by studying the similarities between each variables, the relationships between mixed variables and to study the contribution of each variable. Similar to PCA imputation methods, an iterative FAMD procedure developed by Josse and Husson (Audigier et al., 2016) imputes missing values for mixed data sets. The FAMD algorithm is similar to iterative algorithms for continuous data. For categorical variables, it has a scale step to convert categorical variables to continuous variables. This gives the algorithm an ability to impute mixed data. The proposed method has been compared to a random forest based method (Stekhoven & Bühlmann, 2012) and it showed an enhanced ability to impute mixed missing observations.

In general all PCA iterative methods consist of an 'initiate step' followed by the scale step for FAMD, to perform PCA step and reconstruct step. The regression based methods have better and fast performance in comparison to standard iterative algorithms (IA), however, these methods are yet to be shown to be able to impute on mixed data examples. FAMD is the PCA iterative algorithm for mixed data, which has shown inferior performance against PCA regression based methods for continuous data. In order to introduce a new mixed data imputation algorithm with a better performance, a new procedures based on FAMD and regression based methods (TSR and KDR) is needed to impute the missing data in mixed matrices. A new procedure has been developed in this work without taking the effect of outliers on imputation methods.

3. PCA iterative and regression based methods

3.1. PCA iterative methods

In the PCA imputation algorithm, the minimization of least squares criteria (Eq. (2)) achieved by introducing a weighted matrix (the weighted matrix W, whose elements take either zero if the original dataset value is missing or one otherwise), resulted in the following cost function.

$$C = \sum_{i=1}^{m}\sum_{j=1}^{n} w_{ij}\left(x_{ij} - m_j - \sum_{k=1}^{p} t_{ik}l_{jk}\right)^2 \tag{2}$$

The iterative PCA algorithm procedure consists of following steps (Husson & Josse, 2013; Josse & Husson, 2012a): impute initial values for missing observations cells to complete the data and update the missing values from PCA reconstruction of resulted full data matrix. The iteration between the update and reconstruct step continues until the desired convergence is achieved.

For mixed data, Audigier et al. (2016) adapted alternative algorithm by converting each categorical variable into dummy variables by taking a unit value if the corresponding category was occurring and zero

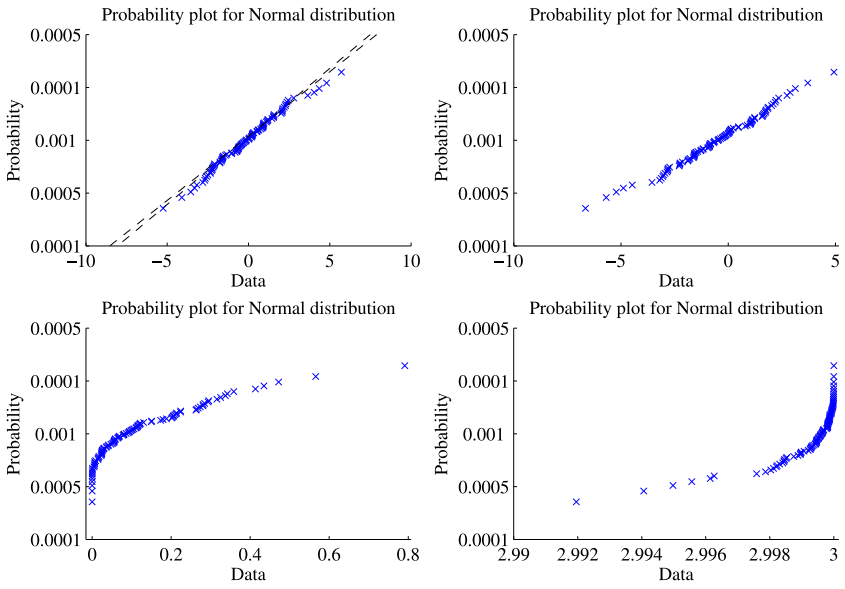

Fig. 9. Probability plot for normal distribution for four variables, two of them are linear(in the top of the figure) and the other two (the bottom two figures) are nonlinear.

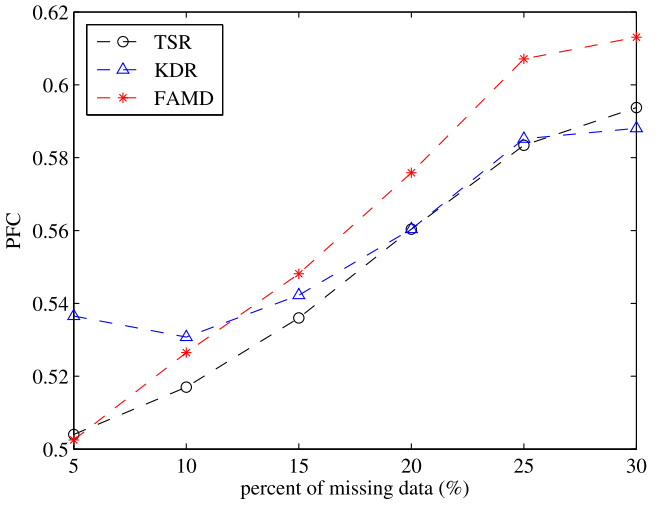

Fig. 10. Categorical error for non linear variables imputed by dividing each variable into three categories.

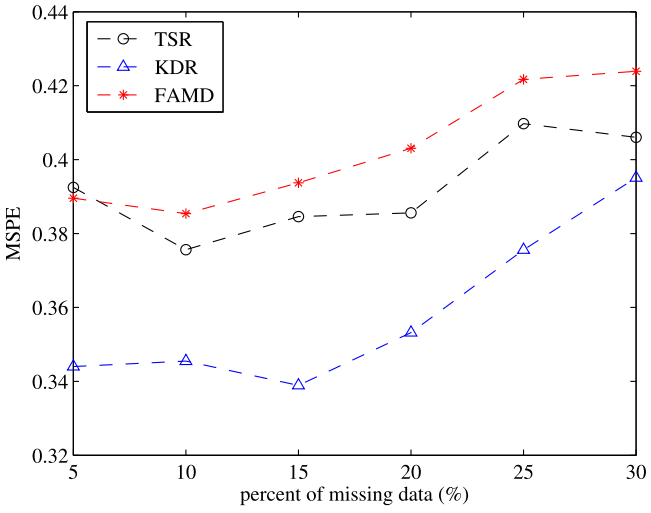

Fig. 11. PFC error for two categorical variables imputed with two quantitative linear variables.

otherwise. Each continuous variable was then standardized by dividing by its standard deviation and each dummy variable is divided by the root square of the proportion of the variable. This iterative FAMD can be implemented as follows (Audigier et al., 2016):

1. Step 0: impute an initial value for each missing values (mean for quantitative and the proportion of the category for each category). Calculate scale and mean matrices: D_{Σ}^{0} scale matrix and M^{0} mean matrix.
2. For step i:
   (a) Apply SVD on the global matrix $(X^{i-1} - M^{i-1})(D_{\Sigma}^{i-1})^{-1/2}$ to obtain the matrices T^{i} (left singular vector, score matrix), L^{i} (right singular vector, loading matrix) as well as $(\Lambda^{i})^{1/2}$.
   (b) Reconstruct X^{i} from the fitted model:

$$\hat{X}^{i} = (T^{i}(\Lambda^{i})^{1/2}(L^{i})^{T})(D_{\Sigma}^{i-1})^{1/2} + M^{i-1} \tag{3}$$

   the imputed data set become:

$$X^{i} = WX + (1 - W)\hat{X}^{i} \tag{4}$$

   (c) from resulted complete data set update, D_{Σ}^{i} and M^{i}.
3. Repeat steps 2a, 2b and 2c until convergence occurs between the imputed and original observed values.

Where:

D_{Σ} is a diagonal matrix that contains the square of standard deviation for each continuous variable and the proportion of the category for each category of categorical variable, $D_{\Sigma} = \text{diag}\,(s_1^2, \ldots s_{n_1}^2, p_{n_1+1}, \ldots \ldots p_n)$,

$M_{m\times n}$: the mean matrix that has the means of each columns of the data matrix $X_{m\times n}$ in each row,

n_1: number of quantitative variables,

n : total number of columns in matrix X, $n = n_1 + \sum_{i=1}^{n_2} g_i$, g_i is the number of categories in variable i, and n_2 is the number of categorical variables,

m : the number of observation, which represent the number of rows of the data matrix X,

$\hat{X}$: the reconstructed data matrix.

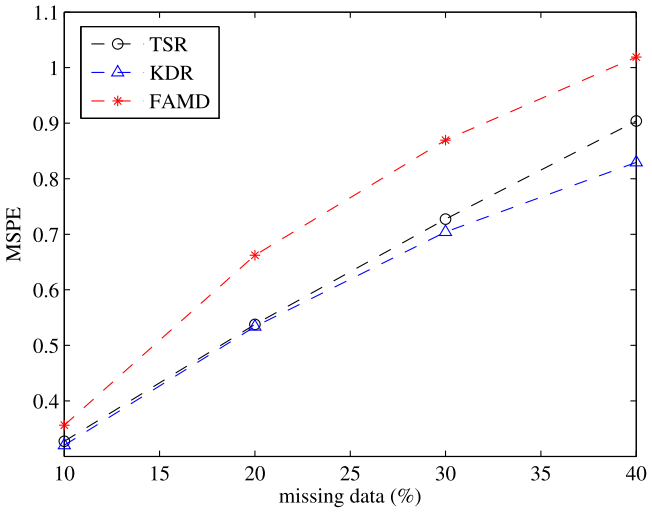

Fig. 12. Comparison of the Quantitative error for a Casting process dataset with 20733 observations.

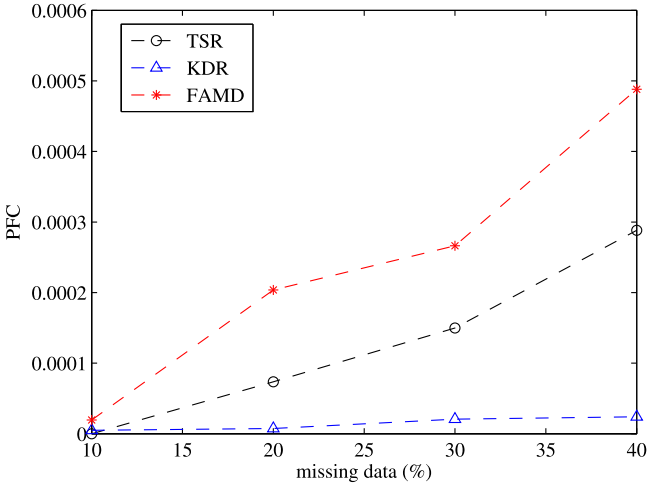

Fig. 13. Comparison of the Categorical error for Casting process dataset with 20733 observations.

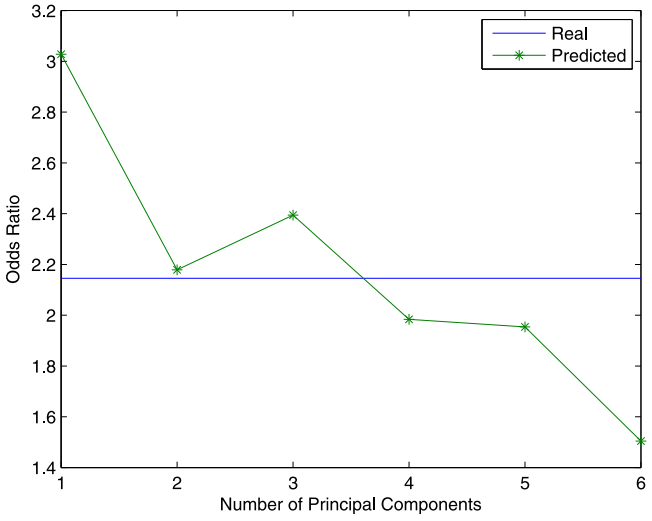

Fig. 14. Rule for selecting the number of PC's from odds ratio. The test is repeated 100 times and the most frequent number selected.

3.2. PCA regression based methods

The PCA regression based methods for missing data partition the attributes with missing values into two parts: the missing part and the observed part. Suppose the observation x^T has some missing values, these will take as first r elements of the row vector, without loss of generality. This partitions the vector x^T into $x^T = [x^{\#T} x^{*T}]$. As a result, the data matrix becomes $X = [X^{\#}\ X^{*}]$, and the loading matrix L can be written as $\begin{bmatrix} L^{\#} \\ L^{*} \end{bmatrix}$

Where:

$x^{\#T}$: denotes the missing elements.

x^{*T}: the observed elements.

$X^{\#}$: is the submatrix containing the first r columns of X (corresponding to the missing variables in x^T).

X^{*}: contains the remaining columns corresponding to the observed values in x^T

$L^{\#}$: is the submatrix with r rows of L.

L^{*}: contains the rest of $L(n-r)$ rows.

The partition strategy is shown in Fig. 5.

Arteaga and Ferrer (2002, 2005), Folch-fortuny et al. (2015) presented two new regression-based methods for estimating incomplete observations.

In known data regression method (KDR), the missing parts for each row are estimated by the following regression model:

$$X^{\#} = X^{*}B + U \tag{5}$$

The least square estimation yielding

$$B = (X^{*T}X^{*})^{-1}X^{*T}X^{\#} \tag{6}$$

Now missing part is estimated from: $x^{\#} = X^{\#T}X^{*}(X^{*T}X^{*})^{-1}x^{*}$. This is written as,

$$x^{\#} = S^{\#*}(S^{**})^{-1}x^{*} \tag{7}$$

Where: $S^{**} = X^{*T}X^{*}/(m-1)$ and $S^{\#*} = X^{\#T}X^{*}/(m-1)$.

The second method is trimmed scores regression (TSR), which based on the following regression model:

$$X^{\#} = (X^{*}L^{*})B + U \tag{8}$$

Where $X^{*}L^{*}$ represent the scores matrix corresponding to observed values $T = X^{*}L^{*} + X^{\#}L^{\#}$, yielding

$$B = (L^{*T}X^{*T}X^{*}L^{*})^{-1}L^{*T}X^{*T}X^{\#} \tag{9}$$

Similar to KDR method, the missing part is estimated from: $x^{\#} = X^{\#T}X^{*}L^{*}(L^{*T}X^{*T}X^{*}L^{*})^{-1}L^{*T}x^{*}$ that is,

$$x^{\#} = S^{\#*}L^{*}(L^{*T}S^{**}L^{*})^{-1}L^{*T}x^{*} \tag{10}$$

Folch-fortuny et al. (2015) adapted iterative method for continuous data based on KDR and TSR as follows:

1. Assume initial values for missing elements such as mean value of each variable resulted in complete data matrix X^0.
2. Step i:
   (a) Perform PCA on the whole data matrix $\left(X^{i-1} - M^{i-1}\right)$ via SVD to estimate the matrices L^{i-1} (right singular vector or loading matrix) and S^{i-1} is the covariance matrix of X^{i-1}.
   (b) Update the whole data matrix by replacing missing values with the fitted values from a regression formula. For TSR use:

$$\hat{x}^{\#i} = S^{\#*(i-1)}L^{*(i-1)}(L^{*(i-1)T}S^{**(i-1)}L^{*(i-1)})^{-1}L^{*(i-1)T}x^{*(i-1)} \tag{11}$$

   and for KDR use:

$$\hat{x}^{\#i} = S^{\#*(i-1)}(S^{**(i-1)})^{-1}x^{*(i-1)} \tag{12}$$

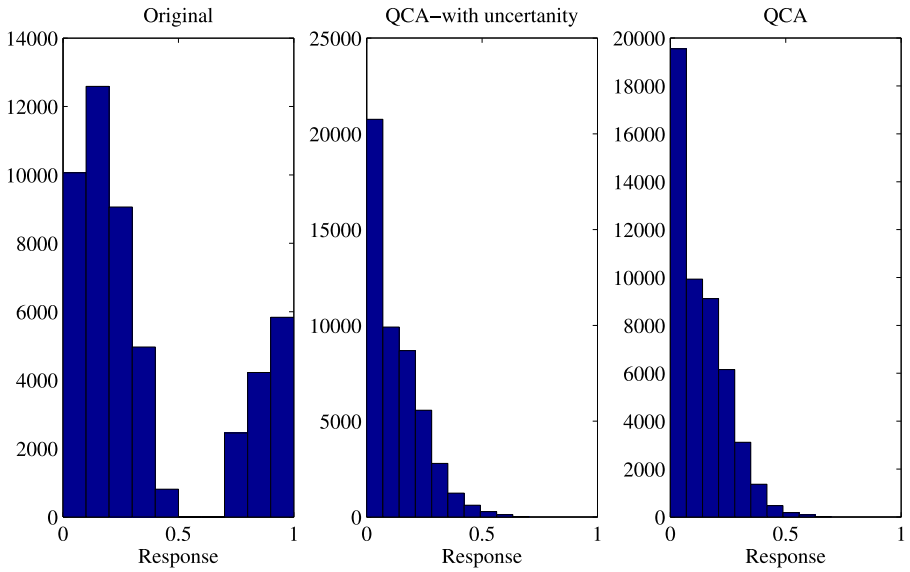

Fig. 15. Response histogram comparison before and after applying optimal operating limits with two types of limits: QCA and QCA with uncertainty estimation.

For both methods, observed values are not changed. The imputed data matrix is:

$$X^i = WX + (1 - W)\hat{X}^i \tag{13}$$

(c) Update mean from the imputed data set from step (2-b).

3. Repeat steps (2-a), (2-b) and (2-c) until convergence occurs.

Fig. 6 shows the comparison of regression based algorithms (TSR and KDR) with iterative and alternating algorithms for four standard deviation σ values (0.25, 5, 0.75 and 1). Regression based methods showed a better performance in comparison to other iterative algorithms such as PCA iterative algorithm and the alternating algorithm. A percentage of missing values were randomly generated from 5% to 30%. The methodology used for generating data used in this comparison is explained in Section 4.1.1.

4. A PCA regression based imputation algorithm for mixed data

All PCA based missing data algorithms discussed above, the iterative FAMD and regression based iterative algorithms TSR and KDR have two main steps: perform PCA step and update from reconstruction step. The reconstruction step implemented in different manner for each case. The FAMD algorithm has scale step, which gives the algorithm an ability to impute mixed data (quantitative and categorical data), whereas the regression based methods impute missing quantitative values without the scale step that used in FAMD to adapt the categorical values for imputation.

In the present work, the regression based methods TSR and KDR methods are adapted to impute mixed data by adding a scale step similar to the one used in FAMD. In other words, the reconstruction step in iterative FAMD algorithm is changed. The FAMD reconstruction step is step 2-b (Eq. (3)) as shown in the iterative FAMD algorithm above

$$\hat{X} = (T(\Lambda)^{1/2}(L)^T)(D_\Sigma)^{1/2} + M$$

The TSR and KDR reconstruction step is achieved by updating the missing values based on the following formula (Eqs. (7) and (10))

$$\hat{x}^{\#} = S^{\#*}L(L^T S^{**} L^*)^{-1} L^{*T} x^*$$

where L is the loading matrix for TSR and is equal to identity matrix in KDR. In order to use the TSR and KDR for mixed data, the reconstruction step (step 2-b in FAMD algorithm) is changed to coincide with the TSR and KDR requirements, so the missing part for the mixed data is updated from the following equation:

$$z^{\#} = S^{\#*}L^*(L^{*T} S^{**} L^*)^{-1} L^{*T} z^* \tag{14}$$

Where: $Z = (X - M)(D_\Sigma)^{-1/2}$, $z^{\#}$ and z^* are related to Z similar to the way $x^{\#}$ and x^* are related to X as shown in Fig. 5 and M is the mean matrix of X as defined in Eq. (3).

The steps for the proposed algorithm are described below:

1. Step 0: impute an initial value for each missing values (mean for quantitative and the proportion of the category for each category). Calculate matrices D^0_Σ, M^0, the mean of X^0, and calculate $Z^0 = (X^0 - M^0)(D^0_\Sigma)^{-1/2}$.
2. For step i:
   (a) S^{i-1} = covariance of matrix $(X^{i-1} - M^{i-1})$, apply SVD on the global matrix (Z^{i-1}), to find a loading vector, which represents the right singular vector (L^{i-1}).
   (b) For each row that has missing values, estimate the missing part from the following regression equations. For TSR method use:

$$\hat{z}^{\#i} = S^{\#*(i-1)} L^{*(i-1)} (L^{*T(i-1)} S^{**(i-1)} L^{*(i-1)})^{-1} L^{*T(i-1)} z^{*(i-1)} \tag{15}$$

   and for KDR method use:

$$\hat{z}^{\#i} = S^{\#*(i-1)} (S^{**(i-1)})^{-1} z^{*(i-1)} \tag{16}$$

   For both methods, observed values are not changed. The reconstructed $\hat{Z}^i$ matrix becomes:

$$Z^i = WZ + (1 - W)\hat{Z}^i \tag{17}$$

   The imputed data matrix is:

$$X^i = Z^i (D^{i-1}_\Sigma)^{1/2} + M^{i-1} \tag{18}$$

   (c) from resulted complete data set, update D^i_Σ and M^i.
3. Repeat steps 2a, 2b and 2c until convergence occurs.

For convergence test, the following two criteria are used: for continuous variables:

$$\frac{\sum_{j=1}^{n_1} \sum_{k=1}^{m} (x_{kj} - x^{old}_{kj})^2}{\sum_{j=1}^{n_1} \sum_{k=1}^{m} x^2_{kj}} \leq \epsilon_1 \tag{19}$$

for categorical variables:

$$\frac{\sum_{j=1}^{n_2}\sum_{k=1}^{m} x_{kj} \neq x_{kj}^{old}}{C_{miss}} \leq \epsilon_2 \tag{20}$$

with ϵ_1 equals to 10^{-6} and ϵ_2 equals to 10^{-10} for example. The algorithm with all details depicted in Table below is referred to as Algorithm 1.

Algorithm 1: A PCA regression based imputation algorithm.

$i \leftarrow 0$;
if $x_{ij} = missing$ **then**
  if $j \leq n_1$ **then**
    $x_{ij}^0 \leftarrow mean(x_j)$;
  else
    $x_{ij}^0 \leftarrow 1/g_j$;
  end
end
while $\epsilon_1 \geq \epsilon_{1,th}$ & $\epsilon_2 \geq \epsilon_{2,th}$ **do**
  $i \leftarrow i + 1$;
  $X^{old} \leftarrow X^{i-1}$;
  $D_{\Sigma}^{i-1} \leftarrow diag(s_1^2, ..., s_{n_1}^2, p_{n_1+1}, ..., p_J)$;
  M^{i-1}, each row of $M^{i-1} \leftarrow mean(X^{i-1})$;
  $Z^{i-1} \leftarrow (X^{i-1} - M^{i-1})(D_{\Sigma}^{i-1})^{-1/2}$;
  $S^{i-1} \leftarrow cov(X^{i-1} - M^{i-1})$;
  $[T^{i-1}\ \Lambda^{i-1}\ L^{i-1}] \leftarrow SVD(Z^{i-1})$ (Perform PCA via SVD);
  Update Z ;
  $\hat{z}^{\#i} \leftarrow S^{\#*(i-1)} L^{*(i-1)} (L^{*T(i-1)} S^{**(i-1)} L^{*(i-1)})^{-1} L^{*T(i-1)} z^{*(i-1)}$(TSR);
  $\hat{z}^{\#i} \leftarrow S^{\#*(i-1)} (S^{**(i-1)})^{-1} z^{*(i-1)}$ (KDR) ;
  $z^{*i} \leftarrow z^{*i-1}$;
  $Z^i = WZ + (1 - W)\hat{Z}^i$;
  Update $X : X^i \leftarrow (Z^i (D_{\Sigma}^{i-1})^{1/2} + M^{i-1})$;
  Convergence Test;
  $\epsilon_1 = \frac{\sum_{j=1}^{n_1}\sum_{k=1}^{m}(x_{kj} - x_{kj}^{old})^2}{\sum_{j=1}^{n_1}\sum_{k=1}^{m} x_{kj}^2}$, $\epsilon_2 = \frac{\sum_{j=1}^{n_2}\sum_{k=1}^{m} x_{kj} \neq x_{kj}^{old}}{C_{miss}}$;
end

4.1. Missing data simulations

Two simulations are conducted to compare the proposed algorithms with FAMD algorithm. The strategy is to generate the missing data from a complete dataset by considering the following incremental levels in the first simulation (5%, 10%, 15%, 20%, 25%, 30%) and extend it to 40% with 10% incremental step for the second simulation. For each level, the missing data is generated randomly. The performance of the present work is assessed by calculating the normalized root mean squared error (NRMSE) for continuous variables and the proportion of falsely classified (PFC) for categorical variables. NRMSE values consider the variance of each variable. The imputed values should correlate with original values if the NRMSE value is equal to zero. They will correlate with the initial values (if the initial value is assumed as the mean) when the NRMSE value is equal to one.

$$NRMSE = \sqrt{\frac{\sum_{j=1}^{n_1}\sum_{i=1}^{m}(\frac{x_{ij} - \hat{x}_{ij}}{\hat{\sigma}_j})^2}{m \times n_1}} \tag{21}$$

$$PFC = \frac{\sum_{j=1}^{n_2}\sum_{i=1}^{m} x_{ij} \neq \hat{x}_{ij}}{C_{miss}} \tag{22}$$

where: C_{miss} is the number of missing categorical values.

4.1.1. Model based simulation

In this section, more than one data sets are generated according to the model based procedure proposed by Josse and Husson (2012b):

$$X_{ik} = M + T_{ip}(L_{kp})^T + \epsilon_{ik} \tag{23}$$

Where, the matrices T and L are generated from a standard normal distribution with zero mean and variance equal to the identity matrix.

Each column of the product matrix $T_{ip}(L_{kp})^T$ is divided by its standard deviation. The noise is added by drawing ϵ_{ik} from a normal distribution with mean equal to zero and variance equal to σ^2. The values in matrix M are assumed to be zero. Signal to noise ratio is defined as $1/\sigma$. In the current work, four data sets are generated to compare the performance of the proposed algorithm with the FAMD algorithm. The generated data sets consist of two quantitative variables and two categorical variables with four categories for each variable. The categories are generated from continuous data by dividing each variable into four segments. The number of observations is 100 and two principal components are selected to reconstruct the data. Four values of σ tested (0.25, 0.5, 0.75, 1) and the results displayed in Fig. 7. The obtained results showed a very good performance for KDR for the mixed data method for all categorical variables with small PFC error. Moreover, it showed a good imputation ability with quantitative variables as well. For all σ values, TSR method for mixed data and FAMD showed slightly different NRMSE and PFC errors, but TSR for mixed data gives better performance with NRMSE and PFC values smaller than that for the FAMD method.

In order to check the ability of algorithm to impute non-linear data, two non-linear variables were generated by adding non-linear functions x^2 and $3Cos(x)$ to the two variables generated from model in Eq. (23) with $\sigma = 1$. First, the variables were assumed as continuous. The imputation of missing data is depicted in Fig. 8, which shows a very high MSPE error value close to 1 for 10% or more missing data. To improve the performance of imputation, each variable is divided into categories based on the probability plot Fig. 9, which is usually used to check the linearity of distribution of the data. As it can seen from the plot that each variable can be divided into three categories. As a result, the imputation of the missing values performed for two categorical variables with three categories for each variable. The results of imputation are shown in Fig. 10, which gives a better prediction than continuous variable assumption.

Finally, another test is conducted to check the effect of imputation of non linear variables with linear ones. The same two non-linear simulated variables were merged with two linear variables to constitute a four variable data set. Fig. 11 shows the PFC error of the two non linear variables in the last simulation, which indicates a smaller imputation error compared to the imputation in Fig. 10 where only non-linear variables were used.

4.1.2. Manufacturing data set

A data set consists of 37 factors affecting the defect of a casting process, 21 of the factors are categorical factors such as percent of Mn with categories 0.002, less than 0.001 and between 0.002 and 0.001. Other factors are continuous factors like the percent of Co. The observed examples are 20733 observations, the results of the comparison of the proposed algorithm with FAMD algorithm is shown in Figs. 12 and 13

5. A novel imputation based predictive algorithm for mixed data

The main aim of prediction is to estimate the process response for a new batch for given values of input factors. In other words, the aim is to determine the ith response value R_i corresponding to factors vector F_{ij} (j = 1,.....,n). In the PCA context, this is similar to projection of a new observation with missing value (missing response) into a lower sub plane predefined by PCA loading matrix, which is known as new observations with missing data. This new observation with missing values can be obtained by finding its scores from the original loadings. It is obtained by an iterative procedure introduced by Arteaga and Ferrer (2002) for continuous data which alternates between estimating scores of missing values from Eq. (24) and reconstruct the missing values from Eq. (25), knowing that the first step includes initializing

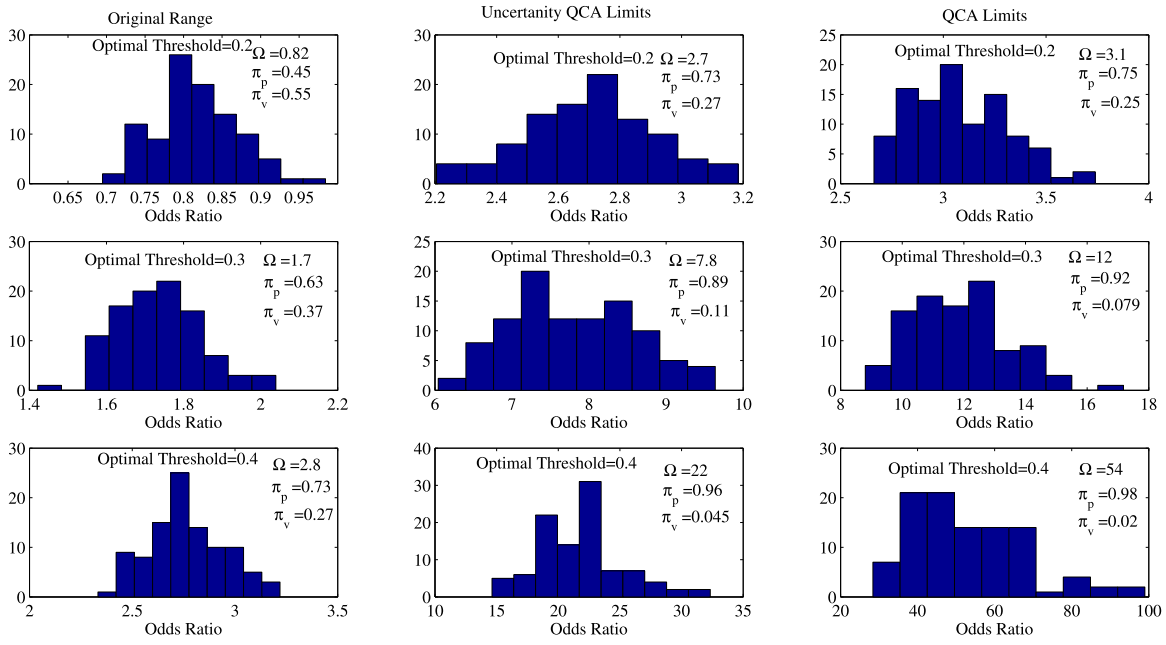

Fig. 16. Odds ratio for original ranges, QCA limits and QCA limits with uncertainty estimation.

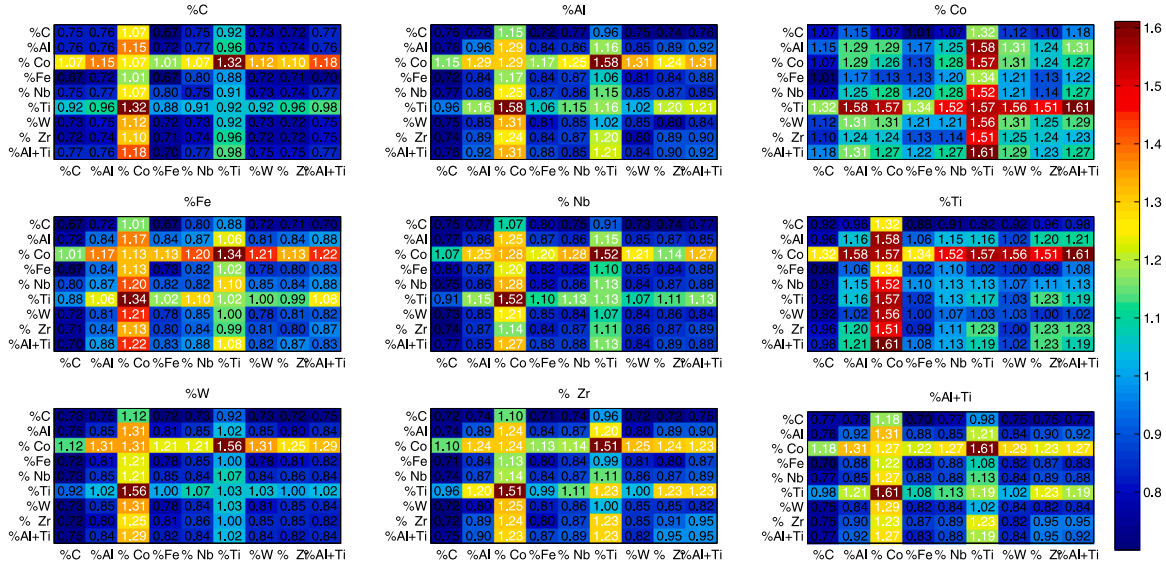

Fig. 17. Odds ratio for interaction of three factors with 0.2 penalty threshold for optimal response values, where the data is generated from QCA optimal limits.

a value for each missing value. These two steps are repeated until convergence occurs.

$$\hat{t} = L^T x = L^{\#T} x^{\#} + L^{*T} x^{*} \tag{24}$$

$$\hat{x}^{\#} = L^{\#T} \hat{t} \tag{25}$$

Arteaga and Ferrer (2002) showed that the obtained scores at convergence can be expressed as:

$$\hat{t} = (L^{*T} L^{*})^{-1} L^{*T} x^{*} \tag{26}$$

For the case of mixed dataset, the same concept is used with the following modification:

1- Add the scale step before perform PCA on the original data matrix.

2- Reconstruction step will be in terms of z instead of x :

$$\hat{z}^{\#} = L^{\#T} \hat{t} \tag{27}$$

3-Estimate the missing value from the below equation:

$$x^{\#} = \hat{z}^{\#} (D_{\Sigma})^{1/2} + \mathbf{m}^{\#} \tag{28}$$

Where $\mathbf{m}^{\#}$ is the mean vector of elements in $x^{\#}$.

The score of new observation is obtained from Eq. (26) instead of the iterative procedure. Full steps of this algorithm are depicted in the table below.

The above procedure can be used in the quality correlation algorithm (QCA) to check the behaviour of the discovered operating limits range by estimating the corresponding responses values. After estimating the response of operating limits, this tool can be used to compare the performance of the process before and after applying operating limits. In other words, it allows to estimate the probability of occurrence of the desired (optimal) and undesired (avoid) response values for a confirmation trial plan and the original plan.

The comparison of two proportions of occurring, such as success and failure, can be conducted by calculating odds of probabilities

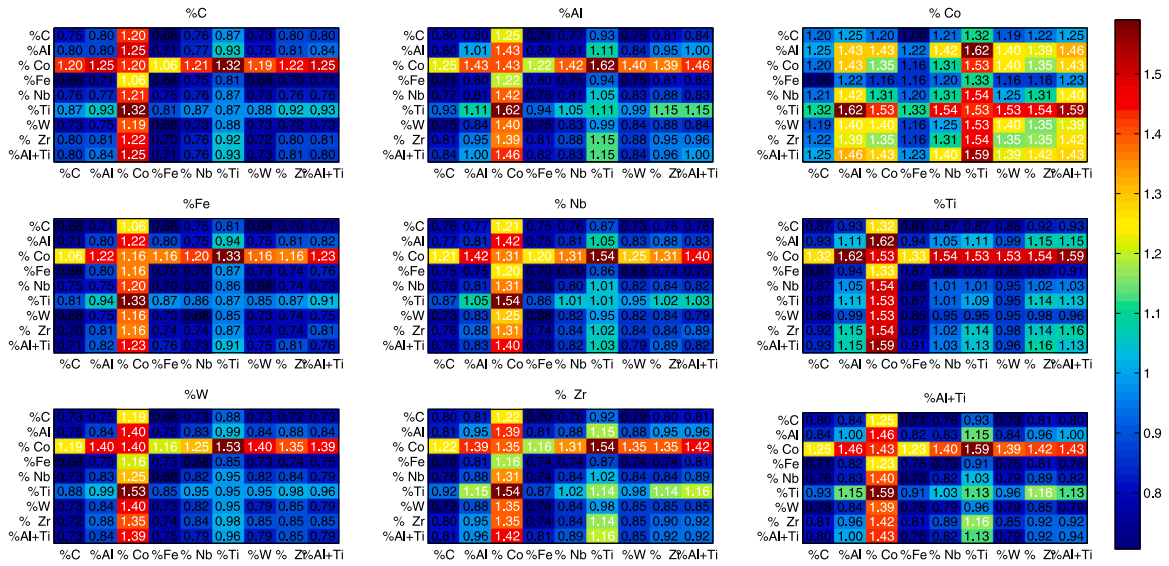

Fig. 18. Odds ratio for interaction of three factors with 0.2 threshold for optimal response values, where the data generated from QCA with uncertainty optimal limits.

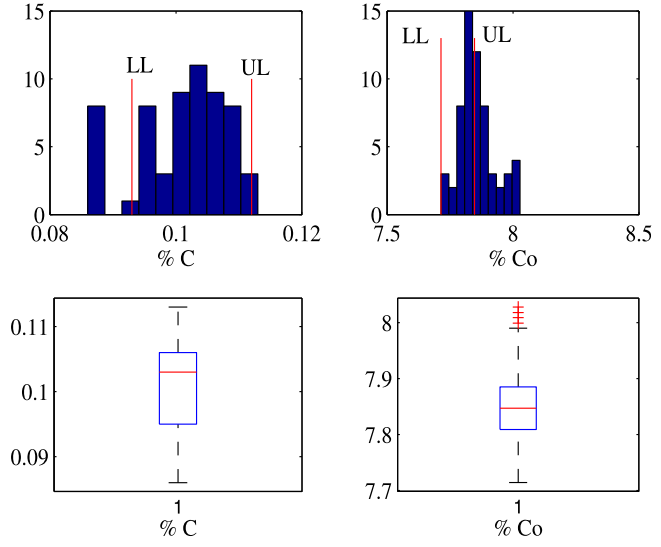

Fig. 19. The histogram and box plot of factors %C and %Co.

Algorithm 2: A novel imputation based predictive algorithm for mixed data

Full data Matrix ;
$\mathbf{m} \leftarrow mean(X_{full})$;
M, each row of $M \leftarrow mean(X_{full})$;
$D_\Lambda \leftarrow diag(s_1^2, ..., s_{n_1}^2, p_{n_1+1}, ..., p_J)$;
$Z \leftarrow (X - M)(D_\Sigma)^{-1/2}$;
$[T\ \Lambda\ L] \leftarrow SVD(Z)$ (Perform PCA via SVD);
New observation with missing values $\mathbf{x}^T = [\mathbf{x}^{\#T}\ \mathbf{x}^{*T}]$;
$\hat{t} = (L^{*T}L^*)^{-1}L^{*T}z^*$;
$\hat{z}^\# = L^{\#T}\hat{t}$;
$\hat{x}^\# = \hat{z}^\#(D_\Sigma)^{1/2} + \mathbf{m}^\#$;

of proportions; another test method used is the likelihood ratio test (see Giannetti & Ransing, 2016). The odds of success is defined as the ratio between probability of success divided by the probability of failure (Agresti, 2002; Liberman, 2005):

$$\Omega = \frac{\pi_s}{\pi_f} \tag{29}$$

Where:

Ω: the odds of success,

π_s: the probability of success(odds of success), and

π_f: the probability of failure(odds of failure).

In terms of manufacturing defects, the success represents the occurrence of desired response values such as lower percentage of defects in batches (optimal response values), while the higher percentage of defects, or the occurrence of undesired response values (avoid response values), refers to the failure. As a result, Ω will be the odds ratio respresenting odds of optimal response and probabilities of success (odds of success) and failure (odds of failure) will be replaced by the probability of optimal response values (π_p) and the probability of avoid response values (π_v) respectively. The Odds ratio equation above is rewritten as follows:

$$\Omega = \frac{\pi_p}{\pi_v} \tag{30}$$

Where:

$\pi_p = P(R \leq Th_{op})$

$\pi_v = P(R > Th_{op})$

Th_{op} = Optimal threshold.

Also, a relative odds ratio can also be defined between any two odds ratio, such as the odds ratio of confirmation trial plan and the original plan as:

$$\Phi = \frac{\Omega_{cp}}{\Omega_{op}} = \frac{(\frac{\pi_p}{\pi_v})_{cp}}{(\frac{\pi_p}{\pi_v})_{op}} \tag{31}$$

In general the proposed missing data algorithm can be applied as a prior to complete the data matrix followed by applying QCA to estimate the optimal operating limits. The next step creates a set of new examples from the obtained operating limits range to study the influence of the factors. The new example is generated by using the Bootstrap by replacement method from the optimal range for each factor. Finally, the odds ratio of the optimal range is compared with the odds ratio of the original range.

%Carbon				
Q1	Q2		Q3	Q4
Minimum		Median		Maximum
0.086	0.095	0.103	0.106	0.113
Q1: Avoid; Range: Bottom 25%, {>=0.086 & <=0.095}				
Penalty	Q1	Q2	Q3	Q4
0.8-1	13	3		2
0.6-0.8			1	3
0.4-0.6				
0.2-0.4				
0-0.2	3	13	9	13

% Cobalt				
Q1	Q2		Q3	Q4
Minimum		Median		Maximum
7.714	7.809	7.847	7.885	8.028
Q3 & Q4: Avoid; Range: Top 50%, {>7.847 & <=8.028};				
Penalty	Q1	Q2	Q3	Q4
0.8-1	3	3	3	9
0.6-0.8	1	3		
0.4-0.6				
0.2-0.4				
0-0.2	12	12	5	9

Fig. 20. Penalty matrices for %C and %Co (Ransing et al., 2013).

5.1. Discussion of results

A Nickel based alloy data set used by Ransing et al. (2016) and Batbooti et al. (2017) to estimate the optimal limits is discussed here. In the current simulation, the QCA algorithm with six principal components is used. This resulted in nine correlated factors for which optimal operating limits were identified. The bootstrap method is used to generate 1000 examples from combination of optimal operating limits of factors and compared with the original range by estimating the odds ratios for the original nine factors and odds ratio based on operating limits. The prediction of odds ratio is dependent on the number of principal components chosen for the dataset defined by optimal factors (e.g. nine correlated factors in this case). The actual odds ratio, based on the original dataset, is 2.125 (Fig. 14). 100 bootstrapped examples were created to test the dependence of the predictive analytic ability of the algorithm on the number of principal components chosen. The real odds ratio of the data is compared with the predicted ones. Each bootstrap example gave slightly different Odds ratio value for the same chosen number of principal components. The most frequently occurring value is chosen and compared with the actual odds ratio value in Fig. 14. The number of principal components for which the most frequently occurring odds ratio value is closest to the actual one is chosen for the analysis. The response histogram and odds ratio values displayed for original range and bootstraped operating limits are shown in Figs. 15 and 16 respectively. The histogram of response to the left in Fig. 15 similar to the response histogram in Fig. 4, but the current one is based penalty value method used by the QCA and the Bootstrap sampling instead the original rejection rate in Fig. 4 and its original range. This approach can be extended to study the effect of interaction between factors, for example, by bootstrapping three factors with suggested optimal range and bootstrapped values for remaining factors taken from the original range. This procedure is repeated for all factors. The calculated odd ratio values are displayed in Figs. 17 and 18 for the QCA limits and QCA with uncertainty limits respectively. An optimal value threshold for penalty values chosen as 0.2 to classify an optimal process response is used. The high values of odds ratio resulting from each combination of factors indicate existence of an interaction among the factors. The value of odds ratio in the aforementioned two Figures is shown in a cell for factor names shown in the corresponding row and column for the given table associated with the corresponding factor name. For example, in the table for factor %C (top left table in Fig. 17), the odds ratio value of 1.32 represents the effect interaction among %C, %Ti and %Co. In other words, for bootstrapping, the values for these factors are chosen from the optimal limits where as the values for remaining factors are chosen from their original range. It can be also seen that if the values for factor %C are used from its optimal range, the resulting odds ratio is 0.75, whereas the %Co factor shows a much higher odds ratio value at 1.35. Both %C and %Co have similar strength co-linearity indices as determined by the QCA, however, the corresponding odds ratio values are significantly different. This is probably because of the linear assumption of prediction model. The predictive algorithm has underestimated odds ratio values for %C. The histogram for %C showed a skewness to the left where as the distribution of %Co showed behaviour close to the Gaussian distribution as shown in Fig. 19. The lower and upper optimal limits are shown as red lines. The measure of the skewness in the data is also observed in the box plot in the same Figure and the penalty matrices in Fig. 20. The first quartile of %C in the penalty matrix has 13 observations with high penalty values where as only three observations with lower penalty values. This range needs to be avoided. However, the performance of the process when %C is in quartile ranges 2, 3 and 4 remains similar. The skewness, or non-linearity, is defined by this step change in the process performance when %C is in quartile 1 as compared to quartiles 2, 3 and 4. Whereas, for %Co quartiles 1 and 2 are optimal (correlated with lower penalty values) with quartile 3 values associated with higher penalty values and quartile 4 with worst performance demonstrating strong correlation with higher penalty values. The variation in the association with low penalty values to high penalty values is linear as %Co range varies from the minimum to maximum value. It may also be possible that the low odds ratio of %C factor may come from the contribution of other factors.

6. Conclusions

A single imputation procedure to predict process response by selecting input factor values from any given range has been described. The procedure is designed to work for mixed datasets comprising quantitative and categorical variables with missing values. The proposed procedure is also required to work on mixed data sets where the number of observations are either smaller, or similar, than the number of input factors. It uses a dimensionality reduction method based on FAMD and investigates relationships between pairs of variables with an improved PCA regression based method. The proposed algorithm is used to impute real and model generated data. The generated data included linear and non-linear simulations. The imputation of non-linear data was improved by dividing the variable range into categories and convert quantitative (or continuous) variables as categorical variables. Also, it is shown that the imputation of non-linear variables with linear ones improves the performance of the algorithm. The obtained results showed a good performance, where the error of the proposed algorithm PCA regression based methods for mixed data (KDR and TSR for mixed data) was less than the error of FAMD based PCA imputation. The imputation of new observation missing values based on FAMD method conducted and used to estimate the response of in process data with known factors. The prediction simulation methodology is based on bootstrapping from original data to predict the behaviour of process when the operating limits discovered by QCA (or any other equivalent method). The odds ratio values are used as a reference to quantify the ratio of the desired to the undesired response values and to compare the behaviour of the process with the original range and the optimal range. The odds ratio values for a real Nickel Based alloy data set were estimated by bootstrapping from the original and optimal ranges respectively. The limitations of the linearity assumptions in potentially underestimating odds ratio values are discussed.

Further interdisciplinary research is needed to enhance and boost the ability of present algorithm for predict the behaviour of the process to reach the optimal process settings. The research efforts to extend the current research can include:

- Adaptation the PPCA to analyse mixed data and development of the quality correlation algorithm based PPCA.
- Development of a multiple imputation method for mixed missing data problem instead of the developed single imputation method in this work. A continuous data procedure based Markov Chain Monte Carlo algorithm used in Folch-fortuny et al. (2015) may be used as a starting point.
- Experimental study to test the optimal limits that obtained from QCA algorithm by building a hardware based system to embed the developed work.
- Incorporate the effect of non-linear factor–response relationships by either using non-linear PCA or Random Forest algorithms.

Data availability

The data used is shared as data in brief.

References

Agresti, A. (2002). *Categorical data analysis*. Hoboken, New Jersey, USA: John Wiley & Sons, Inc.

Arteaga, F., & Ferrer, A. (2002). Dealing with missing data in MSPC : several methods, different interpretations, some examples[2]. (pp. 408–418).

Arteaga, F., & Ferrer, A. (2005). Framework for regression-based missing data imputation methods in on-line MSPC. (December). (pp. 439–447).

Audigier, V., Husson, F., & Josse, J. (2016). A principal component method to impute missing values for mixed data. *Advances in Data Analysis and Classification*, *10*(1), 5–26.

Batbooti, R. S., Ransing, R. S., & Ransing, M. R. (2017). A bootstrap method for uncertainty estimation in quality correlation algorithm for risk based tolerance synthesis. *Computers & Industrial Engineering*, *112*, 654–662.

Batista, G. E. A., & Monard, M. C. (2003). An analysis of four missing data treatment methods for supervised learning. *Applied Artificial Intelligence*, *17*(5–6), 519–533.

Chen, Shikun, & Kaufmann, Tim (2022). Development of data-driven machine learning models for the prediction of casting surface defects. *Metals*, *12*(1).

Christoffersson, A. (1970). *The one component model with incomplete data* (Ph.D. thesis), Uppsala University.

Dempster, A. P., Laird, N. M., & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *39*(1), 1–38.

Diamantaras, K. I., & Kung, S. Y. (1996). *Principal component neural networks: Theory and applications*. New York, NY, USA: John Wiley & Sons, Inc.

Folch-fortuny, A., Arteaga, F., & Ferrer, A. (2015). PCA model building with missing data : New proposals and a comparative study. *Chemometrics and Intelligent Laboratory Systems*, *146*, 77–88.

George, M. L., Rowlands, D., Price, M., Maxey, J., Jaminet, P., Watson-Hemphill, Kimberly, et al. (2005). *The lean six sigma pocket toolbook*. Mc Graw-Hill.

Giannetti, C., & Ransing, R. S. (2016). Risk based uncertainty quantification to improve robustness of manufacturing operations. *Computers & Industrial Engineering*, *101*, 70–80.

Giannetti, C., Ransing, R. S., Ransing, M. R., Bould, D. C., Gethin, D. T., & Sienz, J. (2014). A novel variable selection approach based on co-linearity index to discover optimal process settings by analysing mixed data. *Computers & Industrial Engineering*, *72*, 217–229.

Grung, B., & Manne, R. (1998). Missing values in principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, *42*(1–2), 125–139.

Husson, F., & Josse, J. (2013). Handling missing values in multiple factor analysis. *Food Quality and Preference*, *30*(2), 77–85.

Ilin, A., & Raiko, T. (2010). Practical approaches to principal component analysis in the presence of missing values. *Journal of Machine Learning Research*, *11*(Jul), 1957–2000.

Josse, J., & Husson, F. (2012a). Handling missing values in exploratory multivariate data analysis methods. *Journal de la Société Française de Statistique*, *153*(2), 79–99.

Josse, J., & Husson, F. (2012b). Selecting the number of components in principal component analysis using cross-validation approximations. *Computational Statistics & Data Analysis*, *56*(6), 1869–1879.

Laaksonen, S. (2000). Regression-based nearest neighbour hot decking. *Computational Statistics*, *15*(1), 65–71.

Li, D., Deogun, J., Spaulding, W., & Shuart, B. (2004). *Lecture notes in computer science*: *vol. 3066*, *Towards missing data imputation: A study of fuzzy K-means clustering method* (pp. 573–579). Sweden: Springer-Verlag, Rough sets and current trends in computing (RSCTC).

Liberman, A. M. (2005). How much more likely? The implications of odds ratios for probabilities. *American Journal of Evaluation*, *26*(2), 253–266.

Little, R. J., & Rubin, D. B. (2003). *Statistical analysis with missing data*. New York: Wiley.

Montgomery, D. C. (2009). *Introduction to statistical quality control*. John Wiley & Sons Inc.

Mussa, A., & Tshilidzi, M. (2005). The use of genetic algorithm and neural networks to approximate missing data. *Computing and Informatics*, *24*, 577–589.

Nelwamondo, F. V., Mohamed, S., & Marwala, T. (2007). Missing data : A comparison of neural network and expectation maximisation techniques. *Current Science*, 1514–1521.

Raiko, T., Ilin, A., & Karhunen, J. (2008). Principal component analysis for sparse high-dimensional data. In Masumi Ishikawa, Kenji Doya, Hiroyuki Miyamoto, & Takeshi Yamakawa (Eds.), *Neural information processing: 14th international conference, ICONIP 2007, Kitakyushu, Japan, November 13–16, 2007, Revised selected papers, Part I* (pp. 566–575). Berlin, Heidelberg: Springer Berlin Heidelberg.

Ransing, R. S., Batbooti, R. S., Giannetti, C., & Ransing, M. R. (2016). A quality correlation algorithm for tolerance synthesis in manufacturing operations. *Computers & Industrial Engineering*, *93*, 1–11.

Ransing, R. S., Giannetti, C., Ransing, M. R., & James, M. W. (2013). A coupled penalty matrix approach and principal component based co-linearity index technique to discover product specific foundry process knowledge from in-process data in order to reduce defects. *Computers in Industry*, *64*(5), 514–523.

Roweis, S. (1998). EM algorithms for PCA and SPCA. In *Proceedings of the 1997 conference on advances in neural information processing systems 10* (pp. 626–632). Cambridge, MA, USA: MIT Press.

Schneider, T. (2001). Analysis of incomplete climate data: estimation of mean values and covariance matrices and imputation of missing values. *Journal of Climate*, *14*, 853–871.

Sika, R., & Ignaszak, Z. (2020). Cause-effect analysis using A&DM system for casting quality prediction. *Archives of Foundry Engineering*, *20*(2), 5–12.

Steiner, S. H., & MacKay, R. J. (2004). *Statistical engineering: An algorithm for reducing variation in manufacturing processes*. ASQ Quality Press.

Stekhoven, D. J., & Bühlmann, P. (2012). Missforest-Non-parametric missing value imputation for mixed-type data. *Bioinformatics*, *28*(1), 112–118.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., et al. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*, *17*(6), 520–525.

Uyan, T. Ç., Otto, K., Silva, M. S., et al. (2022). Industry 4.0 foundry data management and supervised machine learning in low-pressure die casting quality improvement. *International Journal of Metalcast*.

Zhao, C., Lui, F. C., Du, S., Di, W., & Shao, Y. (2023). An earth mover's distance based multivariate generalized likelihood ratio control chart for effective monitoring of 3D point cloud surface. *Computers & Industrial Engineering*, *175*, Article 108911.