Information Sciences 367–368 (2016) 311–323

Contents lists available at ScienceDirect

Information Sciences

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

# Data imputation for gas flow data in steel industry based on non-equal-length granules correlation coefficient

Zheng Lv, Jun Zhao*, Ying Liu, Wei Wang

*School of Control Science and Engineering, Dalian University of Technology, China*

ARTICLE INFO

*Article history:*
Received 1 July 2015
Revised 8 May 2016
Accepted 29 May 2016
Available online 2 June 2016

*Keywords:*
Byproduct gas of steel industry
Data imputation
Non-equal-length granules correlation coefficient
Estimation of distribution algorithm

ABSTRACT

In the field of data-driven based modeling and optimization, the completeness and the accuracy of data samples are the foundations for further research tasks. Since the byproduct gas system of steel industry is rather complicated and its data-acquisition process might be frequently affected by the unexpected operational factors, the data-missing phenomenon usually occurs, which might lead to the failure of model establishment or inaccurate information discovery. In this study, a data imputation method based on the manufacturing characteristics is proposed for resolving the data-missing problem in steel industry. A novel correlation analysis, named by non-equal-length granules correlation coefficient (NGCC), is reported, and the corresponding model based on Estimation of Distribution Algorithm (EDA) is established to study the correlation of the similar procedures. To verify the performance of the proposed method, this study considers three typical features of the gas flow data with different missing ratios. The experiment results indicate that it is greatly effective for the missing data imputation of byproduct gas, and exhibits better performance on the accuracy compared to the other methods.



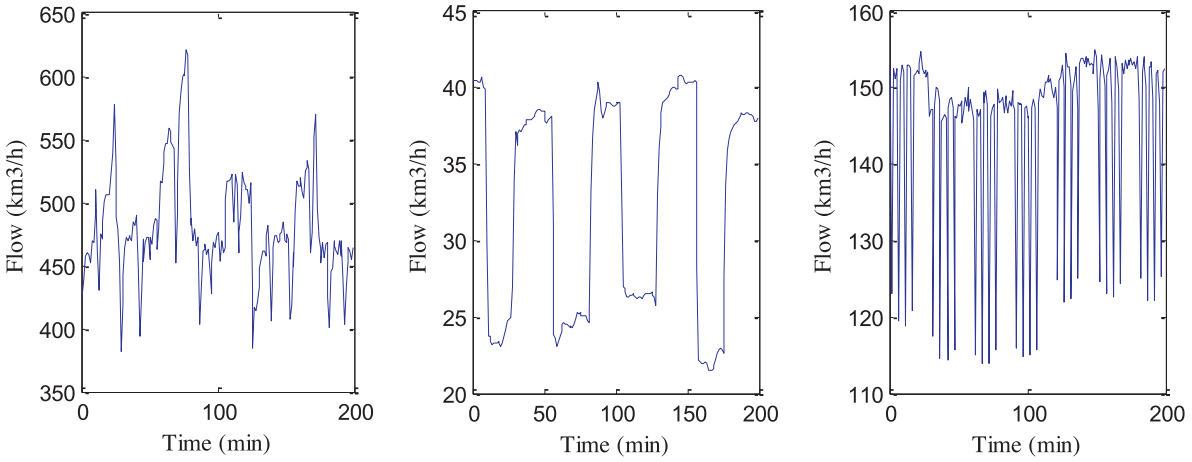
## 1. Introduction

With the rapid advancement of modern industrial information technology, a large number of real-time data related to manufacturing situations have been collected in steel industry. Due to the industrial environment problem, i.e., the failures of the collectors, the transmission deviation, the storage errors, etc., the acquired real-time data would encounter the missing problem, which belongs to the missing completely at random (MCAR) since there is no determinative missing reason [28]. If there is a large number of missing values in the archived data, then it is hard to evaluate the production situation. Furthermore, mining the information and knowledge behind the production data is becoming a hot research issue in the current field of steel industry, and the completeness and the accuracy of data are the significant prerequisites when adopting these approaches [17,19].

As for the data missing problem exhibited in industrial time series, an imputation method based on manufacturing procedure characteristics is designed in this paper, which takes into account the data correlation under same operating conditions in production practice. A correlation analysis method for non-equal-length granules is proposed in this study to construct the correlation of unfixed-cycle time series. For further searching the relationships between the sample sequence and the target one, an estimation of distribution algorithm (EDA)-based model is reported by transforming the calculation of the correlation into the evolution of probability matrix in the solution space. To exhibit the wide applicability of the proposed

* Corresponding author. Tel.: +86041184707582.
*E-mail address:* zhaoj@dlut.edu.cn (J. Zhao).

http://dx.doi.org/10.1016/j.ins.2016.05.046
0020-0255/

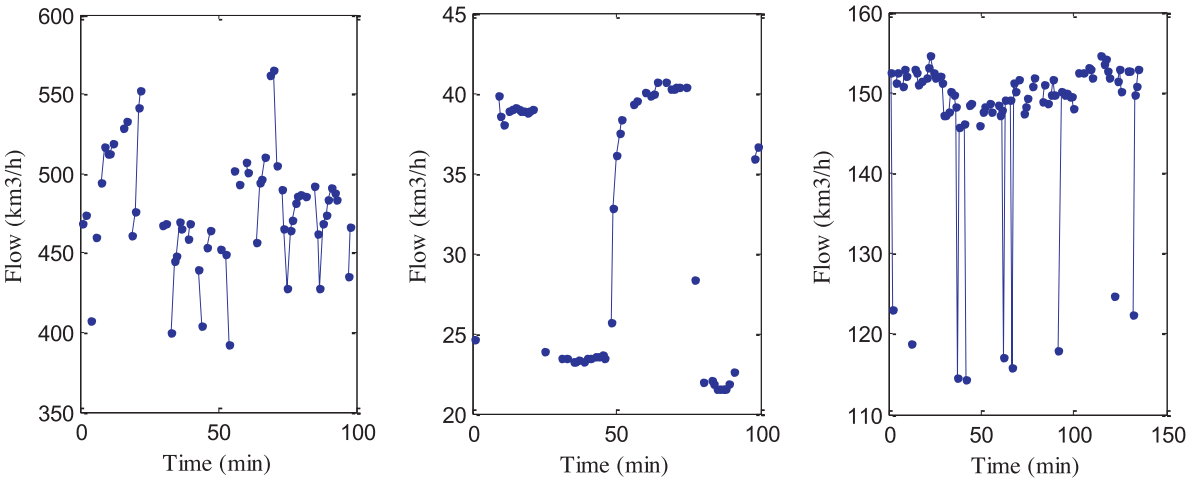



(a) BFG generation flow of blast furnace.

(b) LDG consumption flow of blast furnace.

(c) BFG consumption flow of coke oven.

**Fig. 1.** Gas flow of several typical production processes.




(a) BFG generation flow of blast furnace.

(b) LDG consumption flow of blast furnace.

(c) BFG consumption flow of coke oven.

**Fig. 2.** Gas flow data with missing points.

method, a series of missing ratios are respectively validated in the experiments by using the industrial data coming from the energy data center of a steel plant. The comparative results indicate that, the proposed method can greatly improve the imputation accuracy when facing with the industrial data missing problem.

## 2. Problem description

At present, the supervisory control and data acquisition (SCADA) systems have been widely established in large steel enterprises, and the byproduct gas flow data is of great importance for the energy scheduling work [39]. A series of typical flow data are shown in Fig. 1. The blast furnace gas (BFG) generation flow of blast furnace exhibits high level noises and large amplitude of fluctuations, see Fig. 1(a); the Linz Donawitz converter gas (LDG) consumption flow of a blast furnace with non-standard periodicity and variable amplitude, see Fig. 1(b); and the BFG consumption flow of coke oven exhibits large fluctuations with high frequency, see Fig. 1(c). Due to data collector failure, storage error, transmission deviation, etc., the data missing phenomenon might always occur, as shown in Fig. 2, which brings about difficulties and limitations for data-based system modeling and the energy scheduling tasks [18].

When the missing ratio becomes higher, the continuous missing points will increase. And when the missing situation becomes rather bad the data missing ratio might be close to 40% in current production practice. In such a condition, low imputation accuracy will fail to satisfy the practical demands for data-based analysis and modeling.

## 3. Literature review

### 3.1. Missing data imputation

The researches of data imputation has attracted more and more attentions of scholars recently. In the field of traditional imputation methods, mean substitution was an efficient and fast method, and did not affect the mean value estimation of the variable [32]. But this approach might produce a biased estimate and cause the variance and standard deviation to be smaller [27]. The hot-deck imputation searches for the record which was the most similar with the variables having missing values in database, and then performed the imputation using the value of the similar record [3]. Although that imputation method did not change the data distribution characteristics, it could be only used for the data with multiple attributes, and the accuracy became worse when the number of attributes with missing values increased. A regression imputation adopted in [11] estimated the missing data by the established regression equation, which means the conditional expectation value was used to replace the missing value. A multinomial logistic regression (MLR) method was employed in [30] to estimate the missing points, which could get unbiased estimation of the missing values. However, the random errors could be easily ignored, which became worse along with the increase of the missing information. Then, a multiple imputation for multivariate data-missing was proposed in [29]. To provide an imputation method which preserved interactions in the data automatically, three recursive partitioning techniques were implemented in the multiple imputation by chained equations framework in [7]. Besides, the discrete and consecutive missing values were filled by employing an improved multiple imputation in [5], and the results were compared with the classical ones in [16]. However, the forecasting accuracy must be guaranteed in that method, and the imputation procedure has to be repeated many times for each missing value, which produces huge computational cost. Subsequently, the Monte Carlo expectation-maximization (EM) algorithm was reported and tested by three data sets in [4], but the slow convergence rate failed to be well solved. In [15], the original EM algorithm was improved by using a vector algorithm, but the data source must be assumed as a Gaussian distribution model. Furthermore, the imputation accuracy of these methods described above could be greatly reduced when the data fluctuates largely or the consecutive missing values increase along with the missing ratio becomes higher.

In recent years, machine learning and statistical analysis are introduced to the data imputation field. In [21], a novel technique for missing data estimation using a combination of dynamic programming, neural networks and genetic algorithms (GA) on suitable subsets of the input data was presented, which is well suitable for decision making processes. The online and offline data imputation models were proposed based on the auto associative neural networks in [25]. In [2], a fuzzy c-means clustering hybrid approach that combined support vector regression and a genetic algorithm was presented. In [9], a k-nearest neighbors (k-NN) algorithm was improved by using an entropy weight matrix to identify the k-nearest relative data and performed the data imputation. In [36], the missing values in composition data were imputed by using a model-based iterative regression on the basis of the numerical characteristics of the neighboring data. In addition, an EM based method considering data characteristics were used in [38]. An autoregressive model considering the dynamics and local structure of the microarray temporal data was proposed in [6], which was applicable even the amount of missing data was large, or the missing points correspond to the same position. Some correlation-based data imputation models are also very popular in recent years. In [37], a hybrid missing data completion method named Multiple Imputation was addressed by using Gray-system-theory and Entropy based on Clustering (MIGEC). Firstly, the non-missing data instances are separated into several clusters. Then, the imputed value was obtained after multiple calculations by utilizing the information entropy of the proximal category for each incomplete instance in terms of the similarity metric based on Gray System Theory (GST). In [23], a novel method to impute missing data, named feature weighted grey KNN (FWGKNN) imputation algorithm was proposed, which employed mutual information (MI) to measure feature relevance. However, these methods focused on the data missing problem of multivariate data [14].

It can be noticeable that, all these mentioned methods exhibited a low level capability as for the industrial univariate time series data. In literature, there are still few studies for this category of data. In [8], four imputation methods for time series data, i.e., missing deletion, expectation substitution, average of adjacent values and maximum likelihood estimation, were reviewed. However, the standard normal distribution was assumed as the prerequisite of the mentioned imputation, which was rather hardly for the real-world industrial data. An autoregressive integrated moving average with exogenous inputs and hypothesis testing were combined in [1] to find and impute outliers in time series data sets, but this method focused on the imputation of outliers and is not suitable for continuous missing values. The regression analysis based on the principle of least square method was proposed in [20] to impute the missing values. However, there are many hypothesis conditions for this method, and the applicability for the real industrial data might be limitative. The mean imputation method [22] and k-nearest neighbor (k-NN) method [24] could also be used to deal with the univariate time series data. However, the industrial data such as the gas flow, are with high level noises and fluctuations along with the change of production situation, which greatly influence the imputation performance.

### 3.2. Data correlation analysis

This study reports a correlation-based data imputation method. Up to now, there are many methods can be used to quantify the similarity or correlation of two sequences. An adjusted Euclidean distance based method was proposed in [35], and a Mahalanobis distance measure termed as locally centred Mahalanobis distance was introduced in [38], which can be used as a similarity measure. However, the distance measure cannot accurately express the correlation. Pearson Correlation Coefficient is a statistical measurement of linear correlation between two variables, and is popular for the collaborative filtering recommender system [31]. The Spearman's rank correlation coefficient is shown to be a deterministic transformation of the empirical polychoric correlation coefficient, and the relation between these measures were analyzed in [13]. In contrast, the correlation coefficient is more simple and practical. In [33], a similarity measure based on the correlation coefficient was proposed for clustering time-course gene expression data, which could preserve the similarity information of trajectory patterns of the expression levels on the time interval and that of time points, where maximum and minimum expression levels were attained, respectively, between two profiles. Mutual information indicates the information contained in two random variables, which describes the relationship between two variables from the perspective of information theory. However, the probability distribution of two variables should be determined in advance [12]. The maximum information coefficient proposed in [26] captured a wide range of association both functional and not, which normalized the mutual information scores and compiled a matrix that stores the best grid at that resolution and its normalized score. However, these methods are sensitive to the deformation of timelines, which could obtain inaccurate results if the corresponding relation between the data is not synchronization. Moreover, these methods will be unavailable if the lengths of two sequences are not equal. In [10], a dynamic time wrapping method was proposed to deal with the non-equal-length sequences. However, this method analyzed the similarity by using the Euclidean distance, which cannot evaluate the correlation of the sequences.

## 4. Data imputation based on correlation of non-equal-length granules

The production equipments, e.g. the blast furnace, the coke oven and the converter, usually run with a rough rhythm. Thus, the corresponding gas generation and consumption data exhibit quasi-periodic features, and the data imputation model can be built on basis of the data correlation analysis for different production cycles in this study. Considering the lengths of the production cycles are different, a correlation criterion for non-equal-length information granules is designed here.

### 4.1. Correlation analysis for sequences with different lengths

In general, the correlation analyses are usually used to deal with the information with the same length and the pace of changes, in which the correlation coefficient is one of the approaches that reflects the relationships between two sequences directly [33]. However, if the information in two sequences is not strictly in the form of point-to-point, this method will be unavailable. In this study, a non-equal-length granules correlation coefficient (NGCC) is proposed to resolve such problems.

#### 4.1.1. Non-equal-length granules correlation coefficient

The NGCC is used to calculate the correlation value, which exhibits the maximum correlation degree. Taking the data sequence of the BFG generation flow as an example, as shown in Fig. 3, one can designate each manufacturing procedure (production rhythm) as an information block, see $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, and each block consists of a number of granules $g_i^j$, which represent the production features of a blast furnace.

It can be seen from Fig. 3 that, the granules $g_i^j$ generally have different lengths. Thus, in order to analyze the correlation of two manufacturing procedures, the matchups of them should be first obtained. Taking $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ as example, considering the features of each granule of them, the best matchups of these two sequences can be obtained, as the red lines in Fig. 4 show. Then, the correlation of these two sequences can be calculated.

Assuming these two information sequences $\boldsymbol{\alpha} = \{b_1^{\alpha}, b_2^{\alpha}, \ldots, b_m^{\alpha}\}$ and $\boldsymbol{\beta} = \{b_1^{\beta}, b_2^{\beta}, \ldots, b_n^{\beta}\}$, one can designate $s_k = (b_i^{\alpha}, b_j^{\beta})$ to describe the mapping relationship between the points in the two sequences, which means the $k$-th mapping that involves the points $b_i^{\alpha}$ and $b_j^{\beta}$. If $\{s_1, s_2, \ldots, s_t\}$ are the entire mapping relations of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, then this study converts $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ into two new blocks with the same length, which can be denoted by $\mathbf{b}^{\alpha} = \{s_1^{\alpha}, s_2^{\alpha}, \ldots, s_t^{\alpha}\}$ and $\mathbf{b}^{\beta} = \{s_1^{\beta}, s_2^{\beta}, \ldots, s_t^{\beta}\}$. And the NGCC of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ can be defined as follows,

$$NGCC(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \mathrm{MAX}\left\{ \frac{\sum_{i=1}^{t} \left(s_i^{\alpha} - \overline{\mathbf{b}^{\alpha}}\right)\left(s_i^{\beta} - \overline{\mathbf{b}^{\beta}}\right)}{\sqrt{\sum_{i=1}^{t} \left(s_i^{\alpha} - \overline{\mathbf{b}^{\alpha}}\right)^2 \cdot \sum_{i=1}^{t} \left(s_i^{\beta} - \overline{\mathbf{b}^{\beta}}\right)^2}} \right\} \tag{1}$$

$$s_1 = \left(b_1^{\alpha}, b_1^{\beta}\right), s_t = \left(b_m^{\alpha}, b_n^{\beta}\right) \tag{2}$$

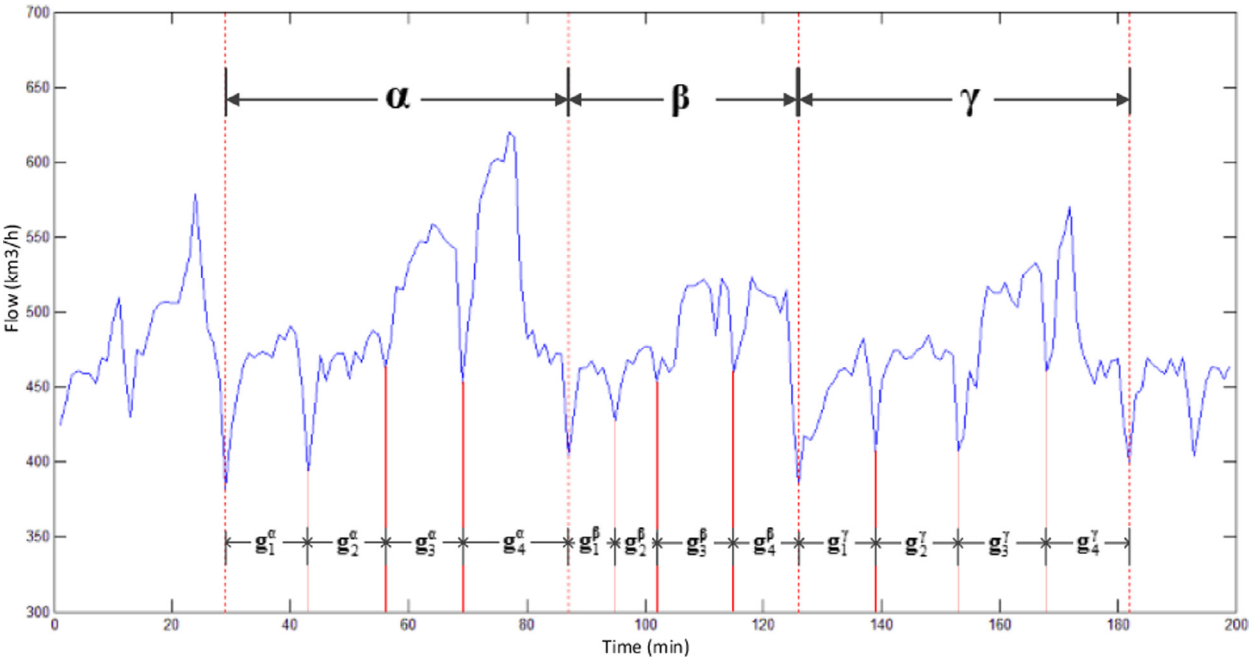



**Fig. 3.** Non-equal-length granules of BFG generation flow data.

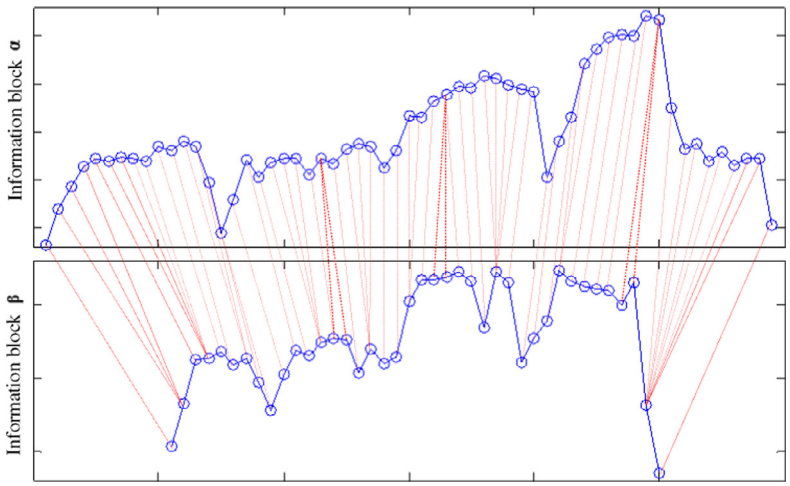



**Fig. 4.** The best matchups of two sequences with non-equal-length information granule.

$$0 \le i - i' \le 1, 0 \le j - j' \le 1 \quad \text{if } s_k = \left(b_i^{\alpha}, b_j^{\beta}\right) \text{ and } s_{k-1} = \left(b_{i'}^{\alpha}, b_{j'}^{\beta}\right) \tag{3}$$

$$\forall s_k = \left(b_i^{\alpha}, b_j^{\beta}\right), |i - j| \le w_{\text{Length}} \tag{4}$$

where, $\overline{\mathbf{b}^{\alpha}}$ and $\overline{\mathbf{b}^{\beta}}$ are the expectation of $\mathbf{b}^{\alpha}$and $\mathbf{b}^{\beta}$ respectively. The $NGCC(\boldsymbol{\alpha}, \boldsymbol{\beta})$ is the maximum correlation measurement among all the possible mapping relations of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. And, the constraints can be listed as the formulas (2)-(4). Since the blocks $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are time series data, all of the data points must be included in the mappings while keeping the time order in each block, see the constraint (3). Moreover, given the practical application, the constraint (4) depicts the time delay of the two points in a mapping must be within a reasonable range, and $w_{\text{Length}}$ is the largest time delay. As such, the NGCC can be used to quantify the correlation of non-equal-length granules and can be calculated by solving the discrete optimization problem formulated by (1)-(4).

*4.1.2. Calculation for NGCC*

As for solving the proposed NGCC, on one hand, it is impossible to calculate the value of NGCC if only knowing one mapping (a certain $s_i$) of two sequences; on the other hand, for two sequences, one cannot determine the number of the mappings in advance. The estimation of distribution algorithm (EDA) is an optimization method, which describes the evolutionary direction by building the probability distribution model of the solution group [34]. Enlightened by the concept of the EDA, this study constructs the probability model of the mapping relations and designs the calculation process for NGCC. Here, the probability model of the solution space is provided and represented by a matrix $\mathbf{P}$,

$$\mathbf{P}=\begin{bmatrix} \mathbf{p}^*_{(1,1)} & \mathbf{p}^*_{(1,2)} & \cdots & \mathbf{p}^*_{(1,n-1)} & \mathbf{p}^*_{(1,n)} \\ \mathbf{p}^*_{(2,1)} & & & \mathbf{p}^*_{(2,n-1)} & \mathbf{p}^*_{(2,n)} \\ \vdots & & \ddots & \vdots & \vdots \\ \mathbf{p}^*_{(m-1,1)} & \mathbf{p}^*_{(m-1,2)} & \cdots & \mathbf{p}^*_{(m-1,n-1)} & \mathbf{p}^*_{(m-1,n)} \\ \mathbf{p}^*_{(m,1)} & \mathbf{p}^*_{(m,2)} & \cdots & \mathbf{p}^*_{(m,n-1)} & — \end{bmatrix} \tag{5}$$

where $\mathbf{p}^*_{(i,j)}$ is a probability distribution, which determines the mapping relationship next to the one of $(b_i^{\alpha}, b_j^{\beta})$. Considering the constraints (3) and (4), each element of this matrix is given by a set of conditional probabilities, i.e.,

$$\mathbf{p}^*_{(i,j)}=\begin{cases} \begin{bmatrix} p_{(i,j)(i,j+1)} & p_{(i,j)(i+1,j)} & p_{(i,j)(i+1,j+1)} \end{bmatrix} & \text{if } i<m \text{ and } j<n \text{ and } |i-j|<w_{\text{Length}} \\ \begin{bmatrix} p_{(i,j)(i,j+1)} & p_{(i,j)(i+1,j+1)} \end{bmatrix} & \text{if } i<m \text{ and } j<n \text{ and } i-j=w_{\text{Length}} \\ \begin{bmatrix} p_{(i,j)(i+1,j)} & p_{(i,j)(i+1,j+1)} \end{bmatrix} & \text{if } i<m \text{ and } j<n \text{ and } j-i=w_{\text{Length}} \\ p_{(m,j)(m,j+1)} & \text{if } i=m \text{ and } j<n \\ p_{(i,n)(i+1,n)} & \text{if } i<m \text{ and } j=n \end{cases} \tag{6}$$

where one can take an example here, $p_{(a_1,a_2)(a_3,a_4)}$ can be calculated by (7), and the sum of the elements of $\mathbf{p}^*_{(i,j)}$ is 1.

$$p_{(a_1,a_2)(a_3,a_4)}=p\left(s_{k+1}=\left(b_{a_3}^{\alpha},b_{a_4}^{\beta}\right)\middle|s_k=\left(b_{a_1}^{\alpha},b_{a_2}^{\beta}\right)\right) \tag{7}$$

Let $\mathbf{S}'=\{\mathbf{s}_1,\mathbf{s}_2,\ldots,\mathbf{s}_l\}$ denote the optimal solutions, which has the largest value of formula (1). During the iterative process, $\mathbf{P}$ will be updated by formulas (8)-(11). Here, one can define a local optimal probability matrix $\mathbf{P}'$,

$$\mathbf{P}'=\begin{bmatrix} \mathbf{p}'^*_{(1,1)} & \mathbf{p}'^*_{(1,2)} & \cdots & \mathbf{p}'^*_{(1,n-1)} & \mathbf{p}'^*_{(1,n)} \\ \mathbf{p}'^*_{(2,1)} & & & \mathbf{p}'^*_{(2,n-1)} & \mathbf{p}'^*_{(2,n)} \\ \vdots & & \ddots & \vdots & \vdots \\ \mathbf{p}'^*_{(m-1,1)} & \mathbf{p}'^*_{(m-1,2)} & \cdots & \mathbf{p}'^*_{(m-1,n-1)} & \mathbf{p}'^*_{(m-1,n)} \\ \mathbf{p}'^*_{(m,1)} & \mathbf{p}'^*_{(m,2)} & \cdots & \mathbf{p}'^*_{(m,n-1)} & — \end{bmatrix} \tag{8}$$

where $\mathbf{p}'^*_{(i,j)}$ represents the probability distribution of the solutions set $\mathbf{S}'$. Let $s_k=(b_i^{\alpha},b_j^{\beta})$, if the mapping $s_k$ is not in $\mathbf{S}'$, then $\mathbf{p}'^*_{s_k}=\mathbf{p}^*_{s_k}$; otherwise, the probability will be calculated by (9), and smoothed by (10),

$$p''_{s_k,s}=\frac{\text{num}\left(\mathbf{s}_l|s_k\in\mathbf{S}'\&s\in\mathbf{S}'\right)}{\text{num}(\mathbf{s}_l|s_k\in\mathbf{S}')} \tag{9}$$

$$p'_{s_k,s_{k+1}}=\sum_s \exp\left(-\|s_{k+1},s\|\right)\cdot {p''}_{s_k,s} \tag{10}$$

where the operator $\text{num}(\cdot)$ aims at counting the amount of the individuals which meet the conditions, $p''_{s_k,s}$ represents the probability of the mapping $s$ in the solution set $\mathbf{S}'$, and the symbol $\|\|$ means calculating the Euclidean distance of two mapping relations. Then, the probability matrix $\mathbf{P}$ can be updated by (11),

$$\mathbf{P}=\eta\mathbf{P}'+(1-\eta)\mathbf{P} \tag{11}$$

where $\eta$ is a factor for the new result, which is less than 1. The probability matrix of the solutions is evolved by repeating the steps above, until the model converged or reaching the maximum number of looping. The steps of this algorithm are as follows.

Step 1: Establish the probability matrix $\mathbf{P}$ initialized using the average probability distribution.
Step 2: Generate a set of solutions according to the probability of $\mathbf{P}$.
Step 3: Calculate the objective function results of each solution, and get the set of better individuals $\mathbf{S}'$.
Step 5: Update the matrix $\mathbf{P}$ by using (8)-(11).
Step 6: Go back to step 2 until $\mathbf{P}$ is converged or reaching the maximum number of iteration. And the objective function value of the optimal solution is the NGCC.

### 4.2. Data imputation for byproduct gas flow data

In this study, a manufacturing procedure characteristics-based data imputation method is proposed for the byproduct gas flow in steel industry. The proposed NGCC is used to quantify the correlation of manufacturing procedures so as to impute the missing points by using the data correlation.

#### 4.2.1. NGCC-based correlation analysis

The searching process for NGCC can be listed as the following algorithm, where $\theta$ is the correlation evaluation parameter, and the outputs of the algorithm involve the correlation $NGCC(\mathbf{x}, \mathbf{y})$ and the corresponding mapping relations $\mathbf{S}'$.

**Algorithm** NGCC based correlation analysis

```
for i=1 to l − n + 1 do
    y_1 ← t_i
    Initialize P
    {NGCC(x, y), S'} ←Calculation for NGCC
    if NGCC(x, y) ≥ θ then
        break
    end if
end for
return {NGCC(x, y), S'}
```

Let $\mathbf{t} = t_1, t_2, t_3, \ldots, t_l$ represent the observed data, and the sequence to be imputed is denoted as $\mathbf{x} = x_1, x_2, \ldots, x_m$. Here, the proposed NGCC is used to find the sequence $\mathbf{y} = y_1, y_2, \ldots, y_n$, which has the highest correlation with $\mathbf{x}$. Since the length of $\mathbf{y}$ might not be equal to that of $\mathbf{x}$, the mapping relations between them also need to be determined. The probability model of the solutions space is determined by,

$$\mathbf{P} = \begin{bmatrix} \mathbf{p}^*_{(1,1)} & \mathbf{p}^*_{(1,2)} & \cdots & \mathbf{p}^*_{(1,m+w_{\text{Length}})} \\ \mathbf{p}^*_{(2,1)} & & & \mathbf{p}^*_{(2,m+w_{\text{Length}})} \\ \vdots & & \ddots & \vdots \\ \mathbf{p}^*_{(m-1,1)} & \mathbf{p}^*_{(m-1,2)} & \cdots & \mathbf{p}^*_{(m-1,m+w_{\text{Length}})} \end{bmatrix} \tag{12}$$

where $\mathbf{p}^*_{(i,j)}$ represents the probability distributions of the mapping $(x_i, y_j)$. And, we have

$$\mathbf{p}^*_{(i,j)} = \begin{cases} \begin{bmatrix} p_{(i,j)(i,j+1)} & p_{(i,j)(i+1,j)} & p_{(i,j)(i+1,j+1)} \end{bmatrix} & \text{if } |i-j| < w_{\text{Length}} \\ \begin{bmatrix} p_{(i,j)(i,j+1)} & p_{(i,j)(i+1,j+1)} \end{bmatrix} & \text{if } i - j = w_{\text{Length}} \\ \begin{bmatrix} p_{(i,j)(i+1,j)} & p_{(i,j)(i+1,j+1)} \end{bmatrix} & \text{if } j - i = w_{\text{Length}} \end{cases} \tag{13}$$

where $p_{(i,j)(i,j+1)}$ is the probability of mapping relation $(x_i, y_{j+1})$ when the mapping $(x_i, y_j)$ is determined.

#### 4.2.2. Correlation-based data imputation

In this study, the sample $\mathbf{y}$ could be converted to the same length as the imputed sequence $\mathbf{x}$, Here, one can define a mapping relation $x_i \to y_j$ to represent the $i$-th point of $\mathbf{x}$ corresponding to the $j$-th point of $\mathbf{y}$. As such, a new formed sequence can be formulated by $\mathbf{y}' = y'_1, y'_2, \ldots, y'_m$. Then, one have

(a) If $x_i \to y_j, y_{j+1}, \ldots, y_r$, then $y'_i = \text{Avg}(y_j, y_{j+1}, \ldots, y_r)$.
(b) If $x_i, x_{i+1}, \ldots, x_r \to y_j$, then $y'_i = y'_{i+1} = \ldots = y'_r = y_j$.

The imputation equations can be expressed as (14) for there are strong linear correlations in $\mathbf{x}$ and $\mathbf{y}'$.

$$\begin{bmatrix} \mathbf{x}_{obs} \\ \mathbf{x}_{abs} \end{bmatrix} = \begin{bmatrix} \mathbf{y}'_{obs} & 1 \\ \mathbf{y}'_{abs} & 1 \end{bmatrix} \cdot \begin{bmatrix} \alpha \\ \beta \end{bmatrix} \tag{14}$$

where $\mathbf{x}_{obs}$ and $\mathbf{x}_{abs}$ are the observed value and the absence value of $\mathbf{x}$, and the corresponding value of $\mathbf{y}'$ are $\mathbf{y}'_{obs}$ and $\mathbf{y}'_{abs}$, respectively. $\alpha$ and $\beta$ are the coefficient and the deviation of the linear relationship respectively. Thus, the missing value $\mathbf{x}_{abs}$ can be determined by (15).

$$\mathbf{x}_{abs} = \begin{bmatrix} \mathbf{y}'_{abs} & 1 \end{bmatrix} \cdot \begin{bmatrix} \mathbf{y}'_{obs} & 1 \end{bmatrix}^{-1} \cdot \mathbf{x}_{obs} \tag{15}$$

The calculating steps of the proposed imputation method are as follows:

Step 1: Select a long period of continuous historical data $\mathbf{t} = t_1, t_2, t_3, \ldots$ from the data base, and determine the length of the production cycle $m$ and the largest time delay $w_{\text{Length}}$ according to the production features and experience.

Step 2: Divide the original data associated with missing points into a number of data segments by the production cycle $m$, and select the segment with the least missing points as the imputed sequence $\mathbf{x}$.
Step 3: Search for the sample $\mathbf{y}$ and the corresponding mapping relations $\mathbf{S}'$ by using the NGCC-based correlation analysis method.
Step 4: Reconstruct $\mathbf{y}$ into $\mathbf{y}'$ based on the mapping relations $\mathbf{S}'$ and the converting rules, to make it equal length to $\mathbf{x}$.
Step 5: Impute the missing values by using (15) according to the correlation between $\mathbf{x}$ and $\mathbf{y}'$.
Step 6: If there are no missing points in the new sequence, stop the imputation. Otherwise, go back to step 2.

## 5. Experiments and analysis

In order to address the advantage of the proposed method for practical applications, the industrial data coming from the energy data center of a steel plant are employed, and the interval of the data is one minute, which can be viewed as the sampling frequency in application database on-site (Oracle® 11 g). In this study, the experimental data are obtained by randomly setting the missing points with various missing ratios.

### 5.1. Correlation analysis based on NGCC

In order to illustrate the performance of the proposed NGCC, the correlation analysis experiments by using the real data under different missing ratios are conducted. Taking the BFG generation flow data as an example, a common manufacturing procedure period of blast furnace can be empirically set to be 50 minutes. Thus, the length of the imputing sequence is set to be 50. Considering the data missing problems happened in the practical industrial process, the missing ratios of the imputing sequence are set from 10% to 40%. A series of 300-minute continuous data are selected as the history data, and the correlation analysis results under different missing ratios are shown in Fig. 6.

As shown in Fig. 6(a), the most relevant sequence $\mathbf{y}$, which has the maximum correlation to the sequence $\mathbf{x}$, can be searched from the history data by using the proposed NGCC. In order to obtain the sample sequence, which has one-to-one correspondence to $\mathbf{x}$. $\mathbf{y}$ is reconstructed by the length transmission process to form $\mathbf{y}'$, which can accurately describe the trend of $\mathbf{x}$. The proposed correlation analysis method is applicable even if there are missing values in $\mathbf{x}$, which can be seen from Fig. 6(b-e).

### 5.2. Industrial data imputation

In order to verify the effectiveness of the proposed imputation method, three kinds of data exhibited in the energy system that covers the typical features of the gas flow data in steel industry are studied, including 1) the large fluctuating data with high level noises; 2) the quasi-periodic data; 3) the data with high frequent vibrations. For each kind of data, the length of manufacturing procedure period, which equals to the length of each imputing sequence, is determined by production experiences. A series of continuous 1-day data are taken as the sequence to be processed, and the missing points are set at random. Considering that the data missing ratio might be close to 40% in practice when the situation is rather bad, it can be set from 10% to 40% in the following experiments.

For clarifying the quality of the proposed method, the neighboring mean method, the spline interpolation method and the k-nearest neighbor (k-NN) method are also employed for the comparative experiments. One can use the evaluation criterions of root mean square error (RMSE) and mean absolute percentage error (MAPE) as the indexes of imputing accuracy,

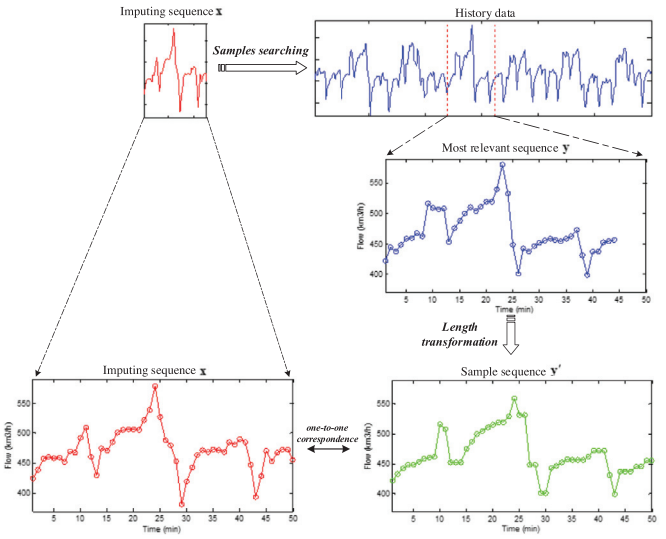
$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{16}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{\left|\hat{y}_i - y_i\right|}{y_i} \tag{17}$$

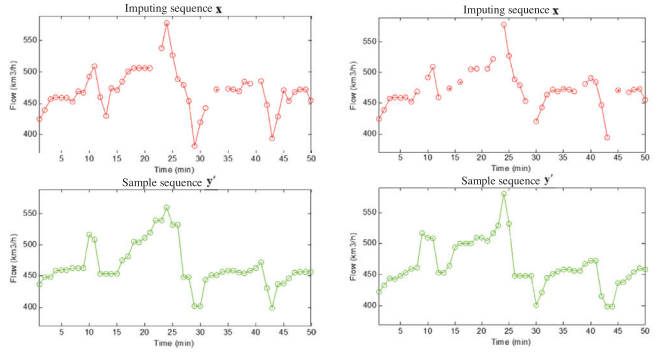
where $n$ is the amount of missing points, $\hat{y}_i$ is the imputed value and $y_i$ is the real data. In order to demonstrate the reliability of the imputation method, 50 experiments are conducted for each kind of data under each missing ratio, and the missing points are randomly set for each time. $MAPE_{\text{Media}}$ and $RMSE_{\text{Media}}$ denote the average of the MAPEs and RMSEs, which can be adopted to measure the imputing accuracy. The smaller $MAPE_{\text{Media}}$ and $RMSE_{\text{Media}}$ means the better performance. $MAPE_{\text{SD}}$ and $RMSE_{\text{SD}}$ are the standard deviation of the MAPEs and RMSEs, which denote the distribution of the around the average value and can be used to measure the variability.

#### 5.2.1. Data feature: large fluctuations and high level noises

In industrial practice, there are often many acquired original data with large fluctuations and high level noises, such as the BFG generation flow data. The length of each imputing sequence is set to be 50, considering the production rhythm of the blast furnace. In order to show the reliability and stability of the proposed method, the imputing results (errors) of different methods are listed in Table 1, where the proposed method exhibits the highest accuracy under different missing ratios validated by the original data.

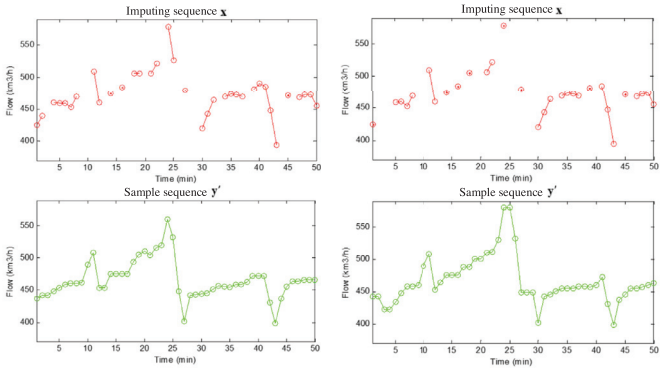



(a) The process of correlation analysis and the analysis results of no missing points.




(b) Analysis results of $\mathbf{x}$ with missing ratio 10%, where $NGCC(\mathbf{x},\mathbf{y}) = 0.9013$.

(c) Analysis results of $\mathbf{x}$ with missing ratio 20%, where $NGCC(\mathbf{x},\mathbf{y}) = 0.9390$.




(d) Analysis results of $\mathbf{x}$ with missing ratio 30%, where $NGCC(\mathbf{x},\mathbf{y}) = 0.8833$.

(e) Analysis results of $\mathbf{x}$ with missing ratio 40%, where $NGCC(\mathbf{x},\mathbf{y}) = 0.9428$.

**Fig. 6.** Correlation analysis results of different missing ratios.

To further illustrate the experimental results, a series of 100-minute data are selected from the 40% missing ratio experiment, as shown in Fig. 7. It can be apparently that, the mean imputation method exhibits good performance when there is few consecutive missing points. However, for the multiple consecutive missing values, the data feature could not be well reflected, see the results in the a-zone of Fig. 7. The spline interpolation uses the spline function to learn the characteristics of the data, which could reflect the dynamic features. But, due to the large fluctuations and high level noises, the results

**Table 1**
Imputing errors of different methods for the BFG generation flow data.

| | Missing ratio (%) | $MAPE_{Media}$ | $MAPE_{SD}$ | $RMSE_{Media}$ | $RMSE_{SD}$ |
|---|---|---|---|---|---|
| Mean method | 10 | 4.0737 | 4.3948 | 23.0606 | 20.4363 |
| | 20 | 2.1368 | 2.3234 | 16.3392 | 11.2443 |
| | 30 | 3.1056 | 2.5454 | 19.6332 | 13.4543 |
| | 40 | 3.7513 | 3.6934 | 24.6896 | 17.2843 |
| Interpolation method | 10 | 4.1307 | 4.3533 | 23.1844 | 19.3533 |
| | 20 | 2.9615 | 2.1353 | 21.2155 | 10.2433 |
| | 30 | 5.0913 | 3.2565 | 30.1697 | 19.7557 |
| | 40 | 5.4542 | 4.7868 | 30.8046 | 20.6546 |
| k-NN method | 10 | 5.6894 | 1.5685 | 32.5425 | 5.2353 |
| | 20 | 4.9752 | 1.1846 | 31.2106 | 6.6546 |
| | 30 | 4.7649 | 1.6496 | 29.6940 | 6.9443 |
| | 40 | 5.1302 | 2.4646 | 30.7083 | 8.3656 |
| Proposed method | 10 | 1.9450 | 1.1464 | 13.4572 | 5.3643 |
| | 20 | 1.4743 | 1.2759 | 11.3465 | 5.2464 |
| | 30 | 1.8353 | 1.5897 | 12.8745 | 6.3577 |
| | 40 | 2.3230 | 1.9058 | 14.8297 | 6.8395 |

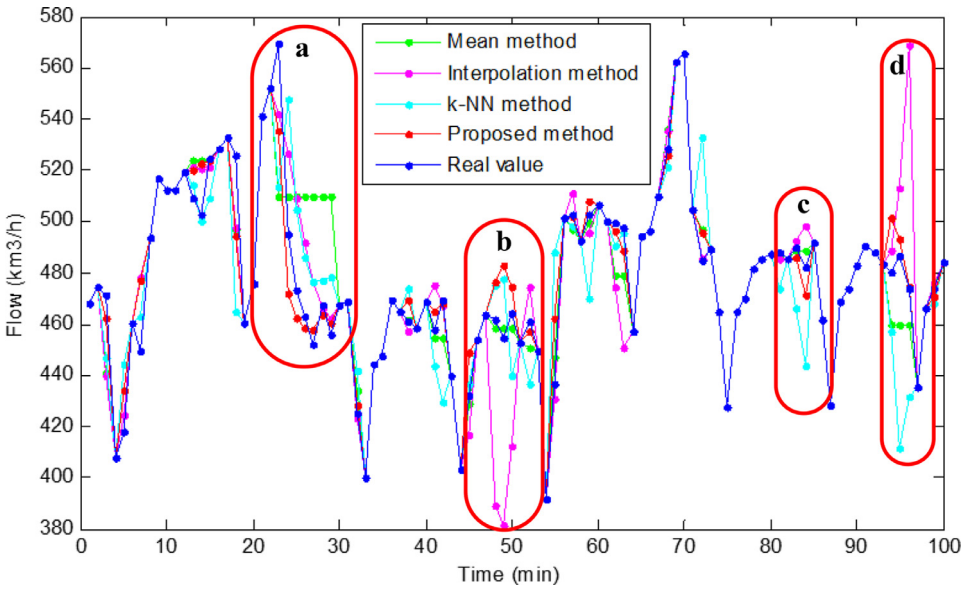



**Fig. 7.** Imputation results of the BFG generation flow data with missing ratio 40%.

might be large deviation due to the over-fitting problem of the spline function, see the results in the b-zone and d-zone of Fig. 7. The results of the k-NN method might bring out some deviations of data tendency because of the uncertain periodicity of a manufacturing procedure, see the results in the c-zone and d-zone. In contrast, the proposed method uses NIGCC to study the features of the industrial data and is capable of getting the more fitted samples, which greatly improved the imputation accuracy.

*5.2.2. Data feature: quasi-periodic fluctuations*

The LDG consumption flow of blast furnace is usually with quasi-periodic fluctuations. Similarly, considering the production rhythm of blast furnace, the imputing sequence can be set to be 50. The imputing errors of different methods are listed in Table 2. It is apparently that, the proposed method exhibits higher accuracy than the others. As for the k-NN method, due to the different length of the nearest neighbor sequence and the target sequence, the imputation accuracy was greatly influenced. However, the proposed NGCC could analyze the correlation of sequences with different length and find the optimal mapping relations of them, thus, could help to improve the accuracy of the results.

Fig. 8 also shows a series of typical 100-minute results for the LDG consumption flow with 40% missing points. The mean method and the interpolation method could only impute the missing points using the neighboring points, which can be regarded as the local information and cannot express the trend and characteristics of the original data. As shown in the a-zone, b-zone and c-zone of Fig. 8, the results of these two methods have large deviations. The proposed method and the k-NN method can both obtain the reasonable imputation results for this type of data, but the proposed method has higher accuracy and exhibits greater advantages even with unstable periods, see Table 2.

**Table 2**
Imputing errors of different methods for the LDG consumption flow data of the blast furnace.

| | Missing ratio (%) | $MAPE_{Media}$ | $MAPE_{SD}$ | $RMSE_{Media}$ | $RMSE_{SD}$ |
|---|---|---|---|---|---|
| Mean method | 10 | 7.1646 | 2.4567 | 1.0418 | 1.0486 |
| | 20 | 1.9429 | 1.5757 | 1.0519 | 1.2548 |
| | 30 | 3.3104 | 1.9486 | 2.0773 | 1.9847 |
| | 40 | 7.9504 | 2.8594 | 3.9306 | 2.7463 |
| Interpolation method | 10 | 6.5272 | 3.2574 | 2.4921 | 2.6949 |
| | 20 | 4.2425 | 3.7950 | 2.0694 | 2.8761 |
| | 30 | 3.6658 | 3.9747 | 1.9634 | 2.9764 |
| | 40 | 6.0669 | 4.6383 | 2.9214 | 3.5798 |
| k-NN method | 10 | 13.3994 | 2.9858 | 1.3622 | 1.0754 |
| | 20 | 2.4400 | 2.1242 | 1.0745 | 1.0837 |
| | 30 | 9.0096 | 2.4383 | 1.8726 | 1.7394 |
| | 40 | 3.2529 | 2.8764 | 1.7536 | 1.7635 |
| Proposed method | 10 | 3.3564 | 2.4644 | 0.9122 | 0.6434 |
| | 20 | 1.6332 | 1.1464 | 1.0321 | 0.9432 |
| | 30 | 1.9334 | 1.6744 | 1.0542 | 0.9534 |
| | 40 | 3.0109 | 1.7849 | 1.3759 | 1.1255 |

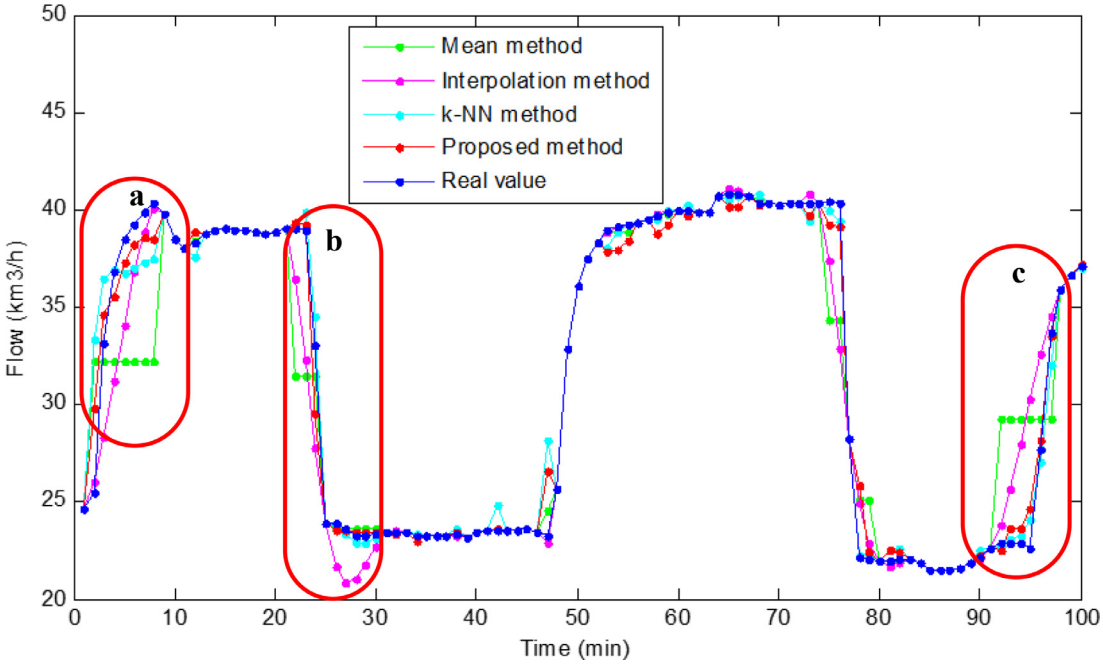



**Fig. 8.** Imputation results of the LDG consumption flow data of the blast furnace with missing ratio 40%.

*5.2.3. Data feature: high frequent vibrations*

In the process of coke oven production, the BFG consumption flow exhibits high frequent vibrations. Considering the production rhythm of coke oven, the information block of such data is set to be 30. The imputing errors of different imputing methods are listed in Table 3. The proposed method exhibits better performance than the mean method and the interpolation method. Due to the high frequent vibrations in the sequence, the k-NN method might have misjudgment for the imputation results. Thus, the proposed method exhibits higher quality than the k-NN method, especially for the data with higher missing ratios. Similarly, the comparative results can be illustrated in Fig. 9. The results in the a-zone of this figure indicate that the mean method and the interpolation are not applicable to the feature of high frequent vibrations. Furthermore, the interpolation might exhibit the over fitting problem, one can refer to the b-zone. In addition, the k-NN method, which cannot deal with the time stretched or compressed problems, could get wrong imputation values for the change points, presented the results in the c-zone. While the k-NN method uses the historical data as the imputation value, thus exhibits lower accuracy than the proposed method.

The experiment results above demonstrate that, the manufacturing procedure information could be studied effectively by using the proposed NGCC, and the imputation results with higher accuracy could show the true characteristics of the energy

**Table 3**
Imputing errors of different methods for the BFG consumption flow data of the coke oven.

| | Missing ratio (%) | $MAPE_{Media}$ | $MAPE_{SD}$ | $RMSE_{Media}$ | $RMSE_{SD}$ |
|---|---|---|---|---|---|
| Mean method | 10 | 6.2576 | 2.1535 | 13.7814 | 3.3643 |
| | 20 | 9.5184 | 3.5644 | 16.2527 | 4.4633 |
| | 30 | 6.4057 | 2.6436 | 13.8791 | 3.5765 |
| | 40 | 6.9154 | 2.6545 | 15.4932 | 3.4464 |
| Interpolation method | 10 | 11.9507 | 4.6437 | 23.2693 | 10.3643 |
| | 20 | 17.7832 | 5.3541 | 29.7351 | 12.5634 |
| | 30 | 8.9856 | 9.3456 | 16.3479 | 14.4634 |
| | 40 | 11.3473 | 10.3644 | 23.7314 | 17.3433 |
| k-NN method | 10 | 0.6380 | 0.3422 | 1.5392 | 0.2323 |
| | 20 | 0.9263 | 0.3634 | 1.6992 | 0.3322 |
| | 30 | 2.0111 | 0.4654 | 5.3583 | 0.3243 |
| | 40 | 1.6766 | 0.4233 | 5.8151 | 0.4453 |
| Proposed method | 10 | 0.5333 | 0.1425 | 1.0409 | 0.2534 |
| | 20 | 0.6198 | 0.1463 | 1.2932 | 0.3574 |
| | 30 | 0.6452 | 0.1444 | 1.5323 | 0.2454 |
| | 40 | 0.8473 | 0.1654 | 1.6062 | 0.4353 |

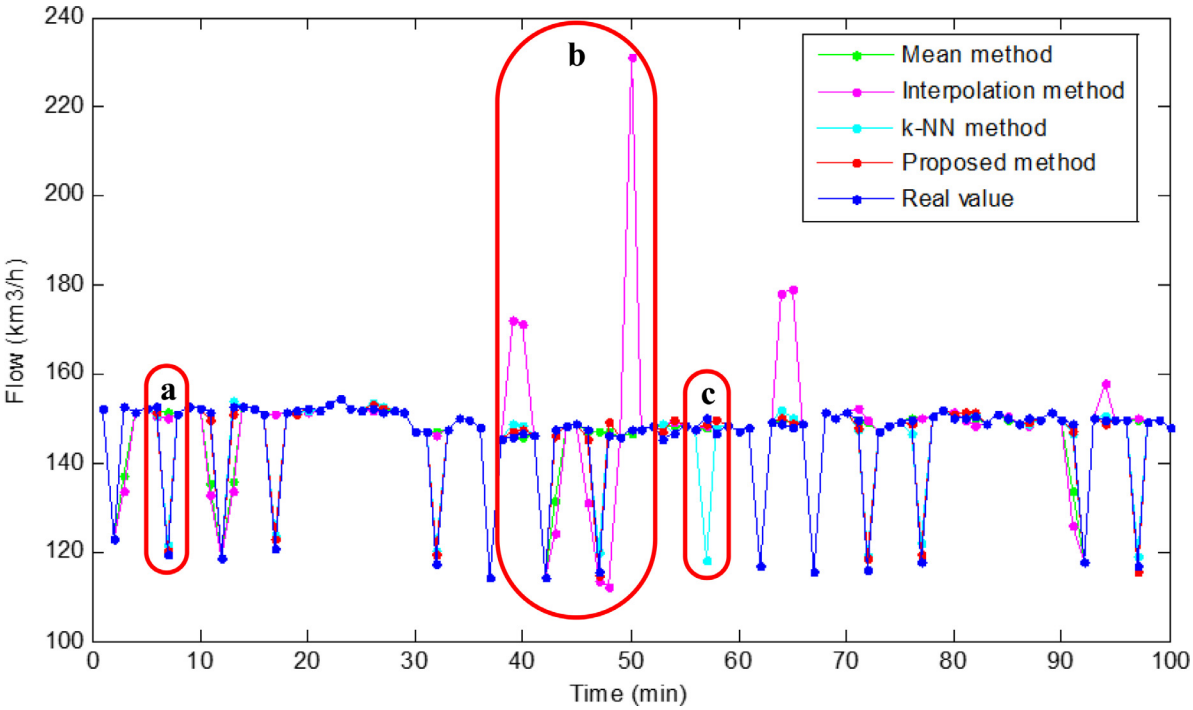



**Fig. 9.** Imputation results of the BFG consumption flow data of the coke oven with missing ratio 40%.

flow data. In the meantime, the proposed method exhibits better performance than the others with regard to different missing ratios, especially for the problem with higher missing ratio and continuous missing points.

## 6. Conclusion

In this paper, the data missing problem of the byproduct gas flow in steel industry is studied, and the manufacturing procedure information is introduced to the data imputation research, which is useful for learning the characteristics of the industrial data. In order to describe the correlation of the data sequences with different length, a NGCC is proposed. The experimental results indicate that, the proposed method was suitable for the three different types of energy data, which can basically cover various industrial time series data features. The effective data imputation could provide solid support for the data-driven system modeling and the production performance analysis. On the other hand, due to the complexity of the industrial data, the efficiency of the computing process of NGCC could be relatively low and the EDA method tends to fall into the local optimum in some situations, which should be further theoretically considered in the future studies.

## Acknowledgements

This work is supported by the National Natural Sciences Foundation of China (No. 61273037, 61304213, 61473056, 61533005, 61522304, U1560102), National Sci-Tech Support Plan (No. 2015BAF22B01) and Fundamental Research Funds for the Central Universities (DUT13RC203, DUT15YQ113).

## References

[1] H.N Akouemo, R.J. Povinelli, Time series outlier detection and imputation[C], in: PES General Meeting | Conference & Exposition, 2014 IEEE, IEEE, 2014, pp. 1–5.
[2] I.B. Aydilek, A. Arslan, A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm, Inf. Sci. 233 (2013) 25–35.
[3] R.R. Andridge, J.A. Roderick, A review of hot deck imputation for survey non-response, Int. Stat. Rev. 78 (1) (2010) 40–64.
[4] J.G. Booth, J.P. Hobert, Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm, J. Royal Stat. Soc. Series B 61 (1) (1999) 265–285.
[5] S.V. Buuren, Multiple imputation of discrete and continuous data by fully conditional specification, Stat. Methods Med. Res. 16 (2007) 219–242.
[6] M.K. Choong, M. Charbit, H. Yan, Autoregressive-model-based missing value estimation for DNA microarray time series data, IEEE Trans. Inf. Technol. Biomed. 13 (1) (2009) 131–137.
[7] L.L. Doovea, S. Van Buurenc, E. Dusseldorp, Recursive partitioning for missing data imputation in the presence of interaction effects, Comput. Stat. Data Anal. 72 (2014) 92–104.
[8] T.C. Fu., A review on time series data mining, Eng. Appl. Artif. Intell. 24 (1) (2011) 164–181.
[9] P.J. García-Laencina, J.L. Sancho-Gómez, A.R. Figueiras-Vidal, M. Verleysen, K nearest neighbours with mutual information for simultaneous classification and missing data imputation, Neurocomputing 72 (2009) 1483–1493.
[10] A. Gedela1, G.S.N. Rao, Dynamic time warping an effective distance measure for time series-data mining, National Conference on Advanced Functional Materials and Computer Applications in Materials Technology, 2014.
[11] A. Gelman, J. Hill, Data Analysis Using Regression and Multilevel/Hierarchical Models, Cambridge University press, London, 2007.
[12] D. Gencaga, N.K. Malakar, D.J. Lary, Survey on the estimation of mutual information methods as a measure of dependency versus correlation analysis [C], in: BAYESIAN INFERENCE AND MAXIMUM ENTROPY METHODS, 2014, pp. 80–87.
[13] E. Joakim, On the Relation Between the Polychoric Correlation Coefficient and Spearman's Rank Correlation Coefficient, Department of Statistics, Ucla, 2013.
[14] W.L. Junger, A.P.D. Leon, Imputation of missing data in time series for air pollutants[J], Atmos. Environ. 102 (2015) 96–104.
[15] M. Kuroda, M. Sakakihara, Accelerating the convergence of the EM algorithm using the vector algorithm, Comutational Stat. Data Anal. 51 (2006) 1549–1561.
[16] R.J.A. Little, D.B. Rubin, Statistical Analysis with Missing Data, Second edition, Wiley, 2002.
[17] Y. Liu, Q.L. Liu, W. Wang, J. Zhao, H. Leung, Data-driven based model for flow prediction of steam system in steel industry, Inf. Sci. 193 (2012) 104–114.
[18] Y. Liu, J. Zhao, Z. Lv, W. Wang, Study on balance scheduling based on the gasholder level prediction for blast furnace gas system in steel industry, Appl. Mech. Mater. 128 (2012) 1464–1467.
[19] Y. Liu, J. Zhao, W. Wang, A time series based prediction method for a coke oven gas system in steel industry, ICIC Express Lett. 4 (4) (2010) 1373–1378.
[20] Y.F. Luo, Z.W. Ye, X.N. Guo, et al., Data missing mechanism and missing data real-time processing methods in the construction monitoring of steel structures[J], Adv. Struct. Eng. 18 (2015) 585–602.
[21] F.V. Nelwamondo, D. Golding, T. Marwala, A dynamic programming approach tomissing data estimation using neural networks, Inf. Sci. 237 (2013) 49–58.
[22] M.N. Noor, A.S. Yahaya, N.A. Ramli, A.M.M. Al Bakri. Mean Imputation Techniques for Filling the Missing Observations in Air Pollution Dataset 2015.
[23] R.L. Pan, T.S. Yang, J.H. Cao, K. Lu, Z. Zhang, Missing data imputation by K nearest neighbours based on grey relational structure and mutual information, Appl. Intell. 43 (2015) 614–632.
[24] S.A. Rahman, Y. Huang, J. Claassen, et al., Combining fourier and lagged k-nearest neighbor imputation for biomedical time series data.[J], J. Biomed. Inform. 58 (2015) 198–207.
[25] V. Ravi, M. Krishna, A new online data imputation method based on general regression auto associative neural network, Neurocomputing 138 (2014) 106–113.
[26] D.N. Reshef, et al., Detecting novel associations in large data sets, Science 334 (2011) 1518–1524.
[27] J.A. Roderick, Regression with missing x's: a review, J. Amer. Statist. Assoc. 12 (1992) 1227–1237.
[28] D.B. Rubin, Inference and missing data, Biometrika 63 (3) (1976) 581–592.
[29] J.L. Schafer, K.O. Maren, Multiple imputation for multivariate missing-data problems: a data analyst's perspective, Department of Statistics, Pennsylvania State University, 1998.
[30] P. Sentas, L. Angelis, Categorical missing data imputation for software cost estimation by multinomial logistic regression, J. Syst. Softw. 79 (2006) 404–414.
[31] L. Sheugh, S.H. Alizadeh, A note on pearson correlation coefficient as a metric of similarity in recommender system[C], AI & Robotics (IRANOPEN), 2015, IEEE, 2015.
[32] R.S. Somasundaram, R. Nedunchezhian, Missing value imputation using refined mean substitution, Int. J. Comput. Sci. Issues 9 (4) (2012).
[33] Y.S. Son, J. Baek, A modified correlation coefficient based similarity measure for clustering time-course gene expression data, Pattern Recognit. Lett. 29 (3) (2008) 232–242.
[34] J.Y. Sum, Q.F. Zhang, E.P.K. Tsang, DE/EDA: a new evolutionary algorithm for global optimization, Inf. Sci. 169 (2005) 249–262.
[35] H.F. Sun, Y. Peng, J.L. Chen, A new similarity measure based on adjusted Euclidean distance for memory-based collaborative filtering, J. Softw. 6 (6) (2011) 993–1000.
[36] H.M. Templ, P. Filzmoser, Imputation of missing values for compositional data using classical and robust methods, Comput. Stat. Data Anal. 54 (2010) 3095–3107.
[37] J. Tian, B. Yu, D. Yu, S.L. Ma., Missing data analyses: a hybrid multiple imputation algorithm using Gray System Theory and entropy based on clustering, Appl. Intell. 40 (2014) 376–388.
[38] R. Todeschini, D. Ballabio, V. Consonni, et al., Locally centred Mahalanobis distance: a new distance measure with salient features towards outlier detection, Anal. Chim. Acta. 787 (13) (2013) 1–9.
[39] J. Zhao, Y. Liu, X.P. Zhang, W. Wang, A MKL based on-line prediction for gasholder level in steel industry, Control Eng. Pract. 20 (6) (2012) 629–641.