Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2013, Article ID 136241, 10 pages
http://dx.doi.org/10.1155/2013/136241

Research Article

An Improved Generalized-Trend-Diffusion-Based Data Imputation for Steel Industry

Ying Liu, Zheng Lv, and Wei Wang

School of Control Sciences and Engineering, Dalian University of Technology, Dalian 116023, China

Correspondence should be addressed to Ying Liu; liu_ying@dlut.edu.cn

Received 5 January 2013; Accepted 20 February 2013

Academic Editor: Jun Zhao

Integrality and validity of industrial data are the fundamental factors in the domain of data-driven modeling. Aiming at the data missing problem of gas flow in steel industry, an improved Generalized-Trend-Diffusion (iGTD) algorithm is proposed in this study, where in particular it considers the sort of problem with data properties of consecutively missing and small samples. And, the imputation accuracy can be greatly increased by the proposed Gaussian membership-based GTD which expands the useful knowledge of data samples. In addition, the imputation order is further discussed to enhance the sequential forecasting accuracy of gas flow. To verify the effectiveness of the proposed method, a series of experiments that consists of three categories of data features in the gas system is presented, and the results indicate that this method is comprehensively better for the imputation of the periodical-like data and the time-series-like data.

1. Introduction

Data missing is one of the major obstacles to obtain valid data samples [1], which also might be common or even inevitable in some data-driven-based research fields, such as sample surveys, industrial productions, medical research, soft engineering, and wireless broadcast environment [2, 3]. The data missing problem might destroy the samples integrality since every cell in database may not be independent, and furthermore a single missing value might call for dropping the entire observed values or the useful information [4, 5]. As such, some useful information or knowledge could be lost from the data set. Moreover, the data missing will also lead to the nonresponse bias of samples which could be a serious concern for the data-driven-based studies [6–10]. In the literatures, most of the existing methods for such problem were mainly based on the statistical techniques. For instance, the multiple imputation (MI), a kind of popular technique, was used to resolve the missing data of gross domestic product (GDP) [11], cancer databases [12], and sample surveys [13]. And, a similar response pattern imputation (SRPI) was also implemented in [14]. In [15], the authors used the classic expectation-maximization (EM) algorithm, principal component analysis (PCA), and singular value decomposition, while [16, 17] utilized the maximum likelihood technique to carry out the missing data imputation. However, all the techniques mentioned previously might be hard to reflect the relationship among regression variables, since the imputed values were mere approximations of unknown values. Besides the statistical techniques, the machine learning was paying more and more attentions nowadays, as presented in [18, 19].

In industrial manufacturing process, the phenomenon of data missing often occurs due to the events such as data collector failures, transmission errors, or information storage errors, which directly result in some obstacles for establishing data-driven-models, such as scheduling models, data-driven based regression prediction models, and stochastic optimization models [20–22]. There were different types of approaches for industrial practitioners in the literatures to deal with these data missing problems. In [23, 24], the authors proposed a method called list-wise deletion that was easy to be implemented; however, it tended to reduce the sample data size. Considering that a lot of missing data in industry have the form of time series sampled in equal intervals in most cases, the integrality of sample data has to be broken

by such deleting the missing points. Besides wasting a lot of costly collected data, this method also led to invalid results if the excluded group was a selective subsample from the entire sample [25]. Mean imputation presented in [26] was another widely employed method. However, the mean values of the sample might eliminate the samples diversity in time series whose amplitude dramatically changes, and the distortion of samples was usually unacceptable for industrial practitioners. With respect to the other statistical or machine learning techniques, the maximum likelihood estimation and the linear interpolator were, respectively, proposed in [27, 28], where the effective experiments were used to validate the time series imputation. Yet, all of these experiments showed high demands of samples, and as a result, their applications in real industrial process were rather limited. As for all of the above mentioned methods, few of them can bring satisfying imputation accuracy, once the consecutive missing happens, or the missing rate is high, and the sample size is too small.

The Generalized-Trend-Diffusion (GTD) is a method of sample construction aiming at small data sets. As the virtual examples presented in [29] and the functional virtual population in [30], the so-called shadow data and membership functions were employed to increase the knowledge of small data sets; see more details in [31]. And, the expanded samples were provided for Back Propagate-based (BP) neural networks to carry out the forecasting, resulting in the prediction accuracy higher than that without expanding. Thus, the most significant advantage of GTD was that it could bring satisfactory forecasting accuracy with relatively small data sets. On the other hand, the original GTD described the membership degree to the mean value of observed sample via a triangular membership function. As such, each observed data point deviation from the mean value is proportional to the difference between the membership function values; that is, the observed data points linearly deviate from the mean point. However, such description of deviation cannot bring excellent accuracy in the imputation tasks for real industrial manufacturing process.

This paper aims at the missing data imputation of blast furnace gas flow in steel energy system. An improved GTD modeling algorithm based on Gaussian membership function is proposed considering the diversity of the gas flow data and the complex missing situations. The Gaussian membership shows that the observed data deviate from the mean value nonuniformly, and this deviation makes the close-to-mean values more likely to appear in the imputation. The samples, expanded by the membership function, make the predicted values by BP-based network lean to the mean. And, such predicted values do not make the samples single as those by the mean imputation do. In addition, the imputation order is essential to the accuracy of time series problem. A both-side-toward-middle (BSTM) order is proposed in this paper which is indicated to be more appropriate than the chronological order. And the tests are implemented to verify the effectiveness of the proposed method, in which the sample data comes from the practice of Shanghai Baosteel Co. Ltd. The results demonstrate that the improved GTD method is much better than the original version and other methods in several cases.

This study is organized as follows. In Section 2, the practical conditions of blast furnace gas in Baosteel is described. And then, the original GTD and its improved version are established in Section 3, where the details of how to use the improved GTD for the industrial missing data imputation are discussed. In Section 4, the validity of the improved GTD is verified by a series of comparative experiments. Finally, this study is summarized in Section 5.

2. Problem Description

Blast furnace gas (BFG) is a kind of byproduct gas generated in the process of iron making [32]. As an important secondary energy for blast furnaces, coke ovens, power stations, and other units, its proper utility can not only reduce the energy consumption of steel enterprises but also improve their economic profits. Figure 1 shows the BFG system structure of a steel plant, where four blast furnaces supply the gas to consumers. However, BFG could be diffused if the flow prediction and the scheduling are inappropriately carried out, which will seriously pollute the environment. In this case, the supervision of BFG's generation and consumption becomes a crucial task for the steel enterprises.

Currently the on-site technicians perform the balance scheduling by estimating the BFG generation amount which comes from the observed data. However, the observed data often miss due to the collector failure, transmission errors, information storage errors, and so forth. Furthermore, the generating process of BFG is rather complex, and the output fluctuates irregularly, therefore the data missing makes the workers work hard to perceive the dynamics of gas flow via generic model. In practice, the gas engineers in Shanghai Baosteel employ the personal experience-based estimation as the current wide using method when encountering single point missing. However, there are more consecutive missing points in real manufacturing process, which make such method relatively weak. In addition, if the missing rate is high, the whole time series can be treated as a combination of several small size series. In this case, the existing methodologies like the recursive neural networks presented in [33, 34] cannot be utilized because they need a large amount of sample data to train the regression model.

Aiming at the various features of a large number of gas units of BFG system, we summarize the flow tendencies of the generation and consumption units as three categories, which involve (1) the periodicity-like flow data (the gas consumption amount of hot blast stove, see Figure 2), (2) the concussive flow data (the gas consumption of coke oven, see Figure 3), and (3) the ordinary time series flow (the generation amount of blast furnace, see Figure 4).

3. Improved Generalized-Trend-Diffusion

3.1. Generalized-Trend-Diffusion. The GTD is a method of sample construction aiming at small data sets which generates shadow data using the real data and the occurrence order of the observed data. The importance degree of those shadow data and observed data is quantified by the membership function values based on fuzzy theories. Both the

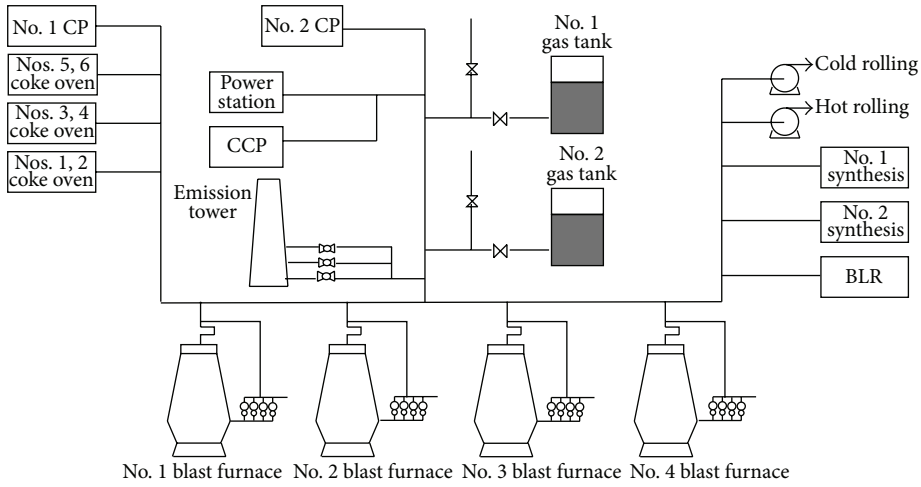

Figure 1: BFG network of Baosteel.

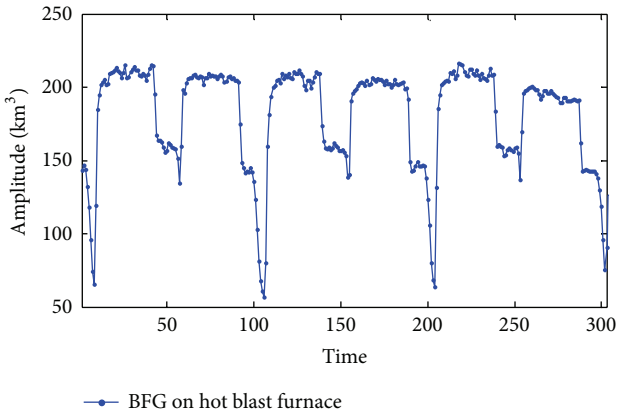

Figure 2: The consumption flow of hot blast stove.

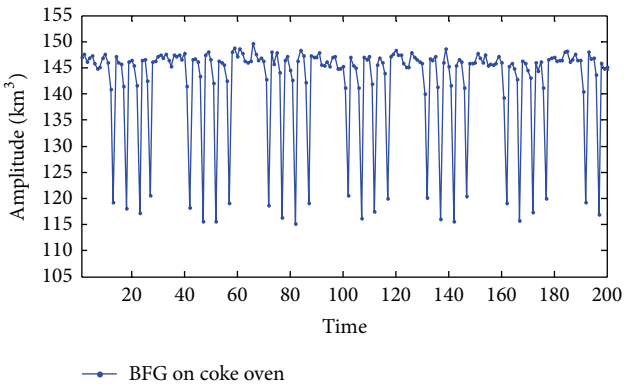

Figure 3: The consumption flow of coke oven.

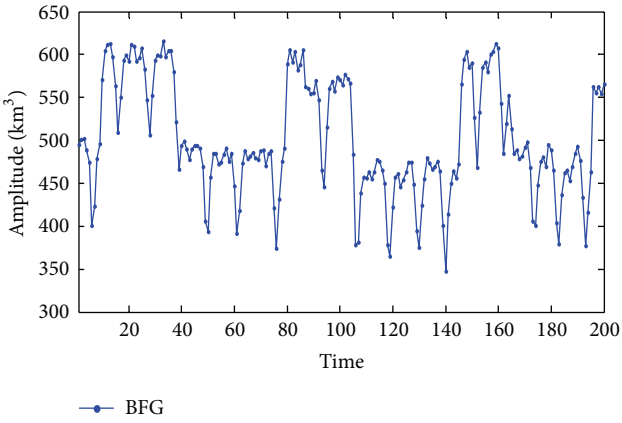

Figure 4: The generation flow of blast furnace.

membership values and the shadow data can be treated as the additional hidden data-related information, which helps to improve the imputation accuracy. All the previous features above make the GTD fit for the missing data imputation of time series because of their lack of more information except time [31]. One can start by considering that observations are collected with an empty set, where each point occurs with each observation (Figure 5). As the data increases, the central location, symbolized "C" in Figure 5, of the data for each observation moves from one location to another. If each point deviation from the central location can be obtained, then the detailed distribution of the whole sample is clear. As such, the GTD with membership function can be used to describe such deviation. Let the membership function value at "C" be 1, and let those of some missing data be MF_x. When these values get closer to 1, the missing data approach "C" and vice versa.

In the original GTD model, one can let MF_t be the membership function for the data collected at Step t. For example, MF_1 at Step 1 refers to Y_1 only, MF_2 at Step 2 refers to $\{Y_1, Y_2\}$, MF_3 at Step 3 refers to $\{Y_1, Y_2, Y_3\}$, and so forth. The data like Y_1 at Step 2 and $\{Y_1, Y_2\}$ at Step 3 are called the shadow data. They were called as such name because each of them was used repeatedly in each step when forming the corresponding membership functions, while it occurred actually once.

Then the imputation can be done by the shadow data. One can suppose that a sequence of data denoted as

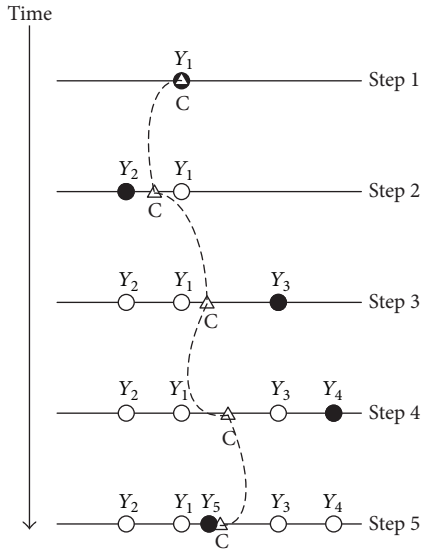

Figure 5: Time series and central location.

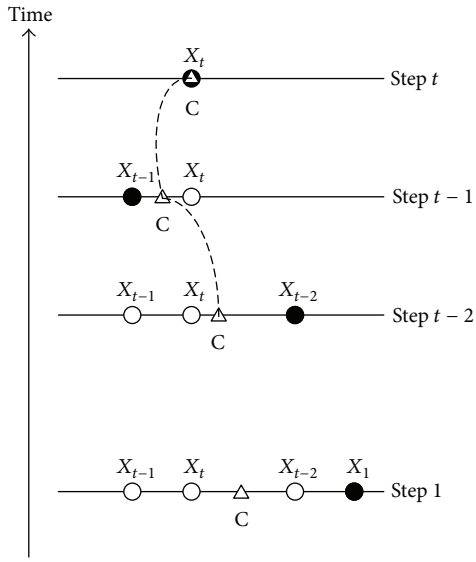

Figure 6: Backward tracking process.

$\{X_1, X_2, \ldots, X_{t-1}, X_t\}$ with X_{t+1} missing has been obtained. The shadow data can be built by unevenly repeating the more recent data which bring more important contemporary information of system variation than that provided by the previous data. As shown in Figure 6, the most recent point X_t is repeated t times, X_{t-1} is repeated $(t-1)$ times X_{t-2} is repeated $(t-2)$ times, and so forth. And, such repeating was called as the backward tracking progress, since it is done in the backward tracking progress. Then, these repeated data (shadow data) with their membership function values help to enlarge the sample knowledge.

3.2. Improved Generalized-Trend-Diffusion. A triangular membership function (Figure 7) was used to describe each point's deviation from the mean value in the original GTD.

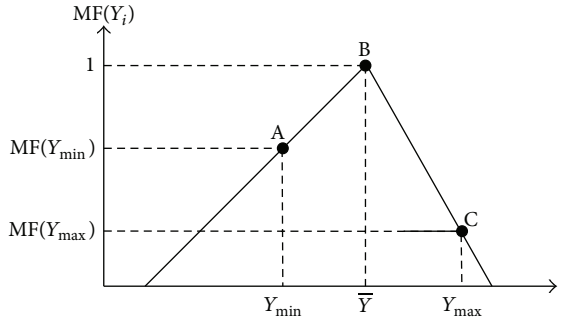

Figure 7: Triangular membership function.

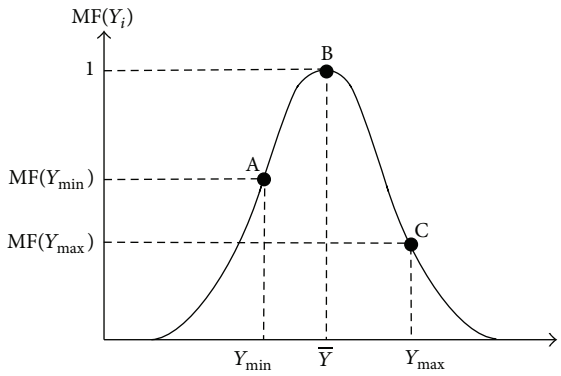

Figure 8: Gaussian membership function.

However, such description was somewhat unreasonable, since it restricted the deviation form as a linear one; that is, the deviation from the central location was proportional to the difference between the memberships. Under such condition, the possibility of the data value to appear in the imputation is equal. However, the mean-like data have actually a higher possibility of appearance in the industrial manufacturing process. In this study, we can call such data vividly as high frequency cloud. If a membership function can describe the high frequency cloud more like the mean value, then the data in the cloud (mean like) will reappear in the imputation with higher possibility. Considering such motivation, the Gaussian membership function (Figure 8) could be more competent to accomplish this job.

The information diffusion principle [35] is another reason for choosing the Gaussian membership function. Information diffusion has a function of filling in the blanks like the molecular diffusion, and its cause lies on that some data acquire little information from the sample knowledge, while molecular diffusion is caused by the heterogeneity in the space distribution. As for the molecular diffusion, it had been proved that current molecular density is proportional to the concentration gradient. If this principle is linked with the law of conservation of mass, the molecular diffusion can be described in the same form as the probability density of Gaussian distribution. As a kind of incremental learning method, the GTD is a representation of information diffusion. Since the causes of information diffusion and molecular diffusion are similar, we here get an inspiration to employ the Gaussian

membership function in the improved GTD. The form of Gaussian function is as follows:

$$f(x) = ae^{-(x-b)^2/c^2}, \tag{1}$$

where a, b, and c are real constants and $a > 0$. In order to make the function adaptive to the sample construction in this study, we use (2) as the general form of Gaussian function instead of (1) as follows:

$$f(x) = e^{-(x-\mu)^2/\sigma^2}, \tag{2}$$

where μ is the mean value of sample and σ is the standard deviation. Here, we make a as 1, since the membership value at the mean value should be 1. After its form confirmed, the Gaussian membership is capable to enlarge the sample knowledge instead of the triangular one.

3.3. Data Imputation. The BP algorithm is a supervised learning method in a network, which is effected by altering the weights to minimize the difference between the output value and the desired output value [36]. The enlarged knowledge then can be utilized by BP neural networks to finish the prediction.

Missing data points need to be imputed one by one, so that the order of imputation should be another concern in this study. If the imputation is real time, the chronological order has to be taken because one cannot currently acquire the future data points. However, the study in this paper is a data mining job which does not need real-time imputations. Furthermore, if the imputation is not real time, the BSTM order is superior to the chronological one. For instance, let there be five consecutive missing data, as Figure 9 shows. If the imputation order is chronological, the forecast error of point number 1 will be amplified so as to affect the forecast accuracy of point number 2, and the error of point number 2 will be again propagated to that of point number 3. In such way, the errors will be cumulative.

Besides, data points in time series are always fluctuating, as Figure 9 shows. If the missing happens on the hillside, the chronological imputation result is very likely to be the same as △, since it continues the peak. However, the result may be like □ if we use the BSTM order. That is, points number 1 and number 2 are on the peak, while points number 4 and number 5 are on the plain which both continue the trends. As for point number 3, we impute that it sings the mean of points number 2 and number 4. Obviously, △ deviates from the real values more than □ which shows that the BSTM order is superior to the chronological one. This summary is consistent if analyzing the missing points on the peak or on the plain.

Let $(X_1, X_2, \ldots, X_{t-1}, X_t)$ be a time series, the index of the first missing point denotes as n, the number of the consecutive missing points denotes as m, and the variable d represents the embedding dimension, then we have

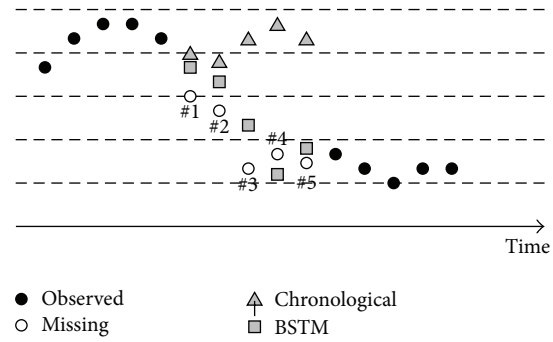

Figure 9: Imputation by BSTM and chronological manner.

$$\begin{aligned}
\widehat{X}_i^{\text{former}} &= \text{net}\left(\left(X_{i-d}, \text{MF}_{i-d}\right), \right. \\
&\qquad \left. \left(X_{i-d+1}, \text{MF}_{i-d+1}\right), \ldots, \left(\widehat{X}_{i-1}, \text{MF}_{i-1}\right)\right), \\
\widehat{X}_i^{\text{latter}} &= \text{net}\left(\left(\widehat{X}_{i+1}, \text{MF}_{i+1}\right), \right. \\
&\qquad \left. \left(\widehat{X}_{i+2}, \text{MF}_{i+2}\right), \ldots, \left(X_{i+d}, \text{MF}_{i+d}\right)\right),
\end{aligned} \tag{3}$$

where $\widehat{X}_i^{\text{former}}$ is the imputation of the former half, while $\widehat{X}_i^{\text{latter}}$ is in the latter half. Then all the imputations $\widehat{X}_i$ can be expressed as

$$\widehat{X}_i = \begin{cases} \widehat{X}_i^{\text{former}} & n \le i \le n + \dfrac{m}{2} - 1, \\ \widehat{X}_i^{\text{latter}} & n + \dfrac{m}{2} \le i \le n + m - 1, \end{cases} \quad \text{if } m = 2k,\ k \in N^*,$$

$$\widehat{X}_i = \begin{cases} \widehat{X}_i^{\text{former}} & n \le i \le n + \dfrac{m-1}{2} - 1, \\ \dfrac{1}{2}\left(\widehat{X}_i^{\text{former}} + \widehat{X}_i^{\text{latter}}\right) & i = n + \dfrac{m-1}{2}, \\ \widehat{X}_i^{\text{latter}} & n + \dfrac{m-1}{2} + 1 \le i \le n + m - 1, \end{cases} \quad \text{if } m = 2k - 1,\ k \in N^*. \tag{4}$$

4. Experimental Results and Analysis

The imputation tests of missing data in BFG flow are carried out with the proposed Gaussian membership function-based method, called iGTD here. First of all, the superiority of the BSTM order to the chronological one is tested and verified. A series of consecutive 800 data is picked from number 1 blast furnace in Baosteel dating from 14:34:00/13/8/2010 to 3:54:00/14/8/2010. Considering that it is difficult to guarantee quantities of consecutive valid data in real, industrial databases, and the small set of samples is our concerns in this study, the embedding dimension is empirically chosen

Table 1: Imputation accuracies with two kinds of orders.

Group	Order	RMSE	NRMSE	MAPE (%)
1	Chronological order	169.30	0.1486	45.65
	BSTM	110.54	0.0970	25.97
2	Chronological order	92.74	0.0669	18.80
	BSTM	82.40	0.0594	14.56
3	Chronological order	107.70	0.0842	23.95
	BSTM	57.32	0.0448	11.41
4	Chronological order	120.47	0.1022	29.75
	BSTM	95.64	0.0812	21.44

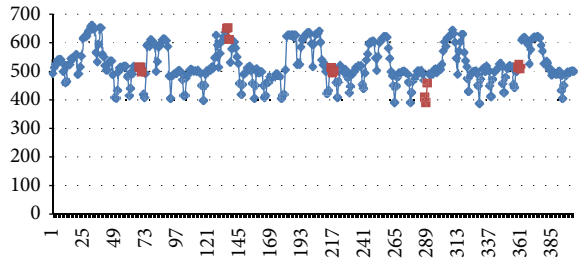

Figure 10: Overview of missing data.

as 15, and the hidden neuron size is chosen as 10 in the same manner. We divide the sample data into 4 groups. For each group, we randomly remove 3 consecutive points (A, B, and C) in 3 places. The tests are, respectively, implemented in the chronological order (A-B-C) and the BSTM order (A-C-B). Here, we use three indexes as the evaluation criterion of the imputation accuracy, which are root mean square error (RMSE), normalized root mean square error (NRMSE), and mean absolute percentage error (NRMSE) as follows:

$$\begin{aligned} \mathrm{RMSE} &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{Y}_i - Y_i\right)^2}, \\ \mathrm{NRMSE} &= \sqrt{\frac{1}{n\|Y\|^2}\sum_{i=1}^{n}\left(\widehat{Y}_i - Y_i\right)^2}, \\ \mathrm{MAPE} &= \frac{100}{n}\sum_{i=1}^{n}\frac{\left|\widehat{Y}_i - Y_i\right|}{Y_i}, \end{aligned} \tag{5}$$

where n is the total number of imputation, $\widehat{Y}$ is the imputation value, and Y is the real value. As for the separated 4 groups of data, the imputation accuracies for the different order are shown in Table 1. It is apparent that the effectiveness of BSTM is superior to that of the chronological order-based imputation method.

To further verify the effectiveness of the proposed Gaussian-based membership function, we comprehensively take the three categories of gas flow data mentioned in previous section, which include the periodicity-like flow data, shown like the BFG consumption amount by hot blast stove; the concussive flow data, shown like the gas consumption by coke oven; and the ordinary time series flow, shown like the generation amount by blast furnace. The comparative experiments are carried out by using the EM method, regression, spline, and the original GTD. EM algorithm is an iterative method for finding maximum likelihood or maximum a posteriori estimates of parameters in statistical models [37], which was widely used in dealing with missing data, since the maximum likelihood estimate of the unknown parameters can be determined by the incomplete data set. The regression method employed multiple linear regressions to estimate the missing values.

We still apply the real industrial data in Baosteel to complete the comparative experiment, where the collected data are divided into several groups, and some consecutive 3 points, 4 points, and 5 points are removed from the time series. In order to cover the all of the possible situations, the removed data involves the time series areas on peak, trough, and plain, as Figure 10 shows, in which the points in red are removed.

(1) BFG Consumption by Hot Blast Stove (Periodicity Like). The experimental data are from number 2 hot blast stove in Shanghai Baosteel randomly selected from 14:28:00/13/08/2010 to 21:55:00/14/08/2010. These data are divided into 3 groups, each of which is then divided into 3 subgroups, and each subgroup contains 200 points. The accuracies of the imputation result are presented in Table 2.

It can be found that the results by both the original GTD and the proposed iGTD are much more excellent in terms of the accuracy when the consecutive data missing occurs. Furthermore, the effectiveness of the iGTD is generally better than that of GTD. Then a conclusion can be drawn that the iGTD employed the Gaussian membership function can obtain the better data imputation results compared to the triangle-based membership of GTD.

(2) BFG Consumption by Coke Oven (Concussive). The experimental data are from number 1 coke oven in Shanghai Baosteel randomly selected from 07:14:00/14/08/2010 to 14:35:00/15/08/2010. The data-grouped measure is similar to that in the validation for periodicity-like data missing. And, the corresponding imputation accuracies are listed in Table 3. From the experiments results, all the five methods are almost same imputation accuracies, and in particular EM should be the best solution method of the five. However, it is

Table 2: Imputation accuracies with different methods (hot blast stove's BFG consumption).

Missing number	Methods	Group 1			Group 2			Group 3		
		RMSE	NRMSE	MAPE	RMSE	NRMSE	MAPE	RMSE	NRMSE	MAPE
3 points	Regression	57.16	0.0713	22.21	55.96	0.0777	28.56	35.56	0.0489	16.96
	EM	43.56	0.0543	17.88	45.82	0.0636	25.76	34.46	0.0474	17.58
	SPLINE	36.72	0.0458	13.80	60.62	0.0842	31.69	22.10	0.0304	10.68
	GTD	9.13	0.0114	2.88	14.90	0.0207	5.85	14.50	0.0199	3.42
	iGTD	7.19	0.0090	2.05	14.42	0.0200	5.71	6.23	0.0086	2.87
4 points	Regression	47.17	0.0578	23.07	29.10	0.0333	11.52	30.12	0.0356	11.56
	EM	26.15	0.0320	12.65	21.59	0.0247	10.96	26.44	0.0313	13.92
	SPLINE	25.18	0.0308	7.80	21.06	0.0241	6.51	14.99	0.0177	5.25
	GTD	15.73	0.0193	4.74	11.78	0.0135	3.82	17.42	0.0206	5.85
	iGTD	15.09	0.0185	4.25	9.80	0.0112	3.37	16.63	0.0197	5.13
5 points	Regression	28.84	0.0307	11.90	55.71	0.0612	30.84	39.92	0.0429	20.31
	EM	25.54	0.0272	13.35	39.99	0.0439	24.38	40.80	0.0439	25.78
	SPLINE	47.33	0.0505	21.45	15.21	0.0167	7.00	33.42	0.0359	19.18
	GTD	20.12	0.0215	6.70	7.89	0.0087	3.93	6.67	0.0072	2.70
	iGTD	15.03	0.0160	5.97	7.09	0.0078	3.55	5.15	0.0055	2.16

Table 3: Imputation accuracies with different methods (coke oven's BFG consumption).

Missing number	Methods	Group 1			Group 2			Group 3		
		RMSE	NRMSE	MAPE	RMSE	NRMSE	MAPE	RMSE	NRMSE	MAPE
3 points	Regression	16.73	0.0300	7.90	10.33	0.0185	3.56	9.37	0.0167	3.24
	EM	9.95	0.0178	5.43	7.25	0.0130	4.05	3.72	0.0066	2.40
	SPLINE	18.77	0.0336	7.73	16.81	0.0301	7.79	22.64	0.0402	9.49
	GTD	15.83	0.0283	8.30	15.82	0.0283	7.47	15.67	0.0278	7.49
	iGTD	13.94	0.0250	6.76	13.76	0.0246	6.55	13.74	0.0244	7.30
4 points	Regression	16.66	0.0260	7.81	11.10	0.0171	4.18	10.82	0.0168	4.49
	EM	10.32	0.0161	5.47	6.72	0.0104	3.80	6.71	0.0104	3.16
	SPLINE	15.77	0.0246	6.50	38.26	0.0591	9.33	14.97	0.0231	6.37
	GTD	15.93	0.0249	7.44	19.75	0.0305	10.55	29.51	0.0459	13.58
	iGTD	13.81	0.0216	6.07	19.36	0.0299	9.39	13.02	0.0203	6.51
5 points	Regression	13.24	0.0186	5.73	13.24	0.0184	5.55	15.98	0.0227	7.78
	EM	10.34	0.0145	5.31	7.52	0.0104	4.20	10.78	0.0153	5.27
	SPLINE	16.90	0.0237	9.06	21.57	0.0300	8.51	13.85	0.0196	6.20
	GTD	15.49	0.0217	8.33	22.83	0.0317	11.00	20.23	0.0287	11.48
	iGTD	11.81	0.0166	5.45	19.84	0.0276	9.02	19.27	0.0273	9.45

mentionable that the effectiveness of the proposed iGTD still does better than GTD in this test.

(3) BFG Generation Amount (Normal Time Series). The experimental data are from number 1 blast furnace in Shanghai Baosteel randomly selected from 02:28:00/27/03/2010 to 18:33:00/01/04/2010. And, the comparative accuracies are listed in Table 4.

From Table 3, we can discover that the regression method presents the worst performance, while iGTD obtains the best one. For the data with normal property of time series, iGTD is better than GTD, while GTD wins all the other three methods.

A conclusion can be drawn from Tables 2–4 that the proposed iGTD and the GTD are superior to regression, EM, and spline for the periodicity-like data and the normal time-series-like data. As for the data with concussive amplitude, both iGTD and GTD do not have an advantage, and yet iGTD still beats GTD which means the proposed Gaussian membership function is superior to the triangular one in the real industrial manufacturing process. And, for the visual imputation results of the BFG generation and consumption, the comparative imputation curves are randomly chosen as Figures 11, 12, and 13 show, where the advantage of the method proposed in this study can be easily presented.

5. Conclusion

This study aims at the imputation of missing data of gas flow in steel industry. In order to improve the imputation accuracy,

Table 4: Imputation accuracies with different methods (BFG output).

Missing number	Methods	Group 1			Group 2			Group 3		
		RMSE	NRMSE	MAPE	RMSE	NRMSE	MAPE	RMSE	NRMSE	MAPE
3 points	Regression	108.17	0.0598	21.94	97.04	0.0476	17.11	109.20	0.0537	16.60
	EM	87.43	0.0483	18.04	83.43	0.0409	15.38	71.65	0.0352	10.72
	SPLINE	92.99	0.0514	17.55	86.74	0.0437	15.41	129.67	0.0638	17.59
	GTD	75.21	0.0416	16.06	70.73	0.0347	12.50	56.69	0.0279	9.18
	iGTD	69.70	0.0385	14.74	68.64	0.0336	12.26	55.48	0.0273	9.32
4 points	Regression	98.25	0.0457	17.99	93.09	0.0394	15.45	95.38	0.0408	15.49
	EM	77.94	0.0363	14.70	75.53	0.0319	13.09	73.74	0.0315	11.57
	SPLINE	128.22	0.0597	20.51	85.83	0.0363	14.29	66.95	0.0286	10.86
	GTD	75.70	0.0352	14.21	63.36	0.0268	10.47	62.31	0.0266	9.44
	iGTD	70.91	0.0330	13.15	63.25	0.0267	10.32	41.51	0.0178	6.74
5 points	Regression	88.76	0.0362	17.09	79.13	0.0296	11.22	76.24	0.0289	11.47
	EM	72.61	0.0296	12.98	71.74	0.0268	11.98	71.39	0.0270	10.97
	SPLINE	97.83	0.0399	16.72	105.18	0.0393	16.78	102.59	0.0389	15.60
	GTD	67.44	0.0275	11.90	61.00	0.0228	9.63	67.23	0.0255	10.47
	iGTD	67.19	0.0274	11.95	57.75	0.0216	9.29	62.95	0.0238	10.14

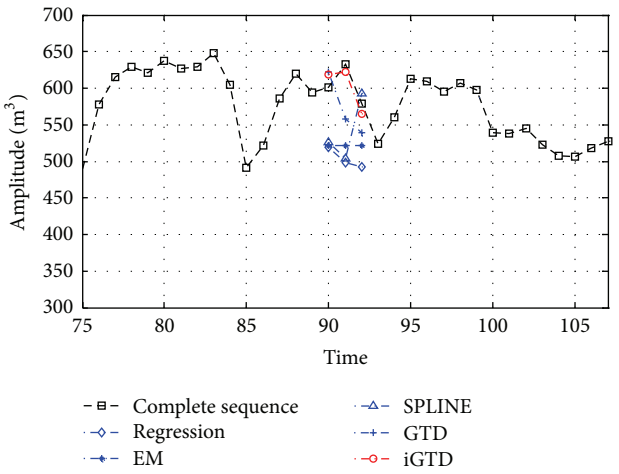

Figure 11: Comparison of different methods (3 points).

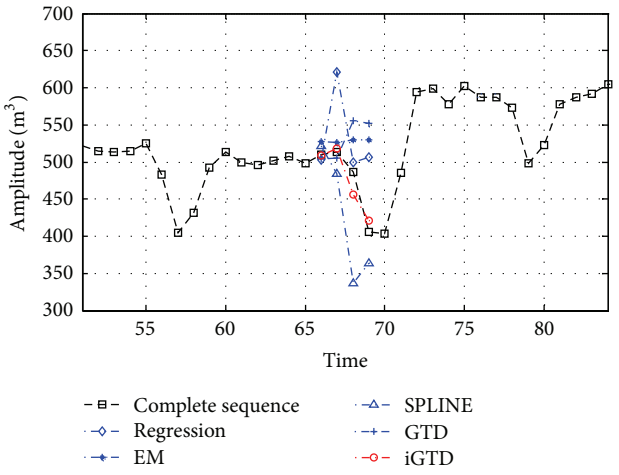

Figure 12: Comparison of different methods (4 points).

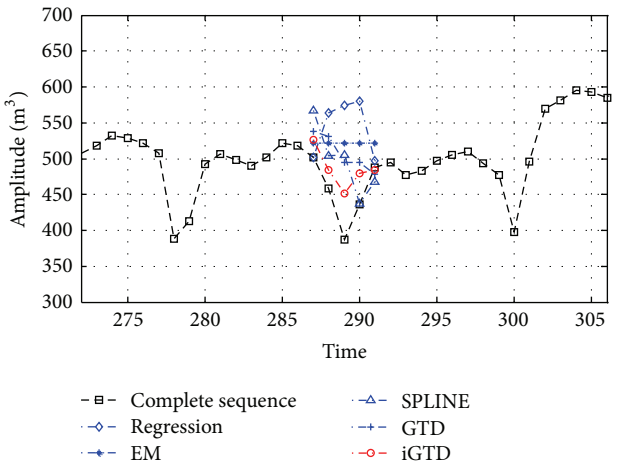

Figure 13: Comparison of different methods (5 points).

the proposed iGTD replaces the triangular membership function with the Gaussian one. Furthermore, the order of imputation is further discussed. The verification experiments show that the BSTM order brings less error than the chronological one does, since more observed data are utilized. As for the different data imputation method, compared to the original GTD, EM, regression, and spline, the proposed iGTD has some advantages in the problem with data properties of consecutively missing and small samples. And, the satisfying imputation accuracy provides the powerful support for the gas resources scheduling later. On the other hand, although the approach developed in this study can handle some types of missing in real industry, some theoretical analyses and the expanded application, for example, the type of concussive flow data, need to be given a further consideration in the future.

Acknowledgments

This work is supported by the National Natural Sciences Foundation of China (no. 61034003, no. 61104157, and no. 61273037) and the Fundamental Research Funds for the Central Universities of China (no. DUT11RC(3)07). The cooperation of energy center of Shanghai Baosteel Co. Ltd, China, in this work is greatly appreciated.

References

[1] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, Elsevier, 2006.

[2] P. C. Austin and M. D. Escobar, "Bayesian modeling of missing data in clinical research," *Computational Statistics and Data Analysis*, vol. 49, no. 3, pp. 821–836, 2005.

[3] S. Y. Yi, S. Jung, and J. Suh, "Better mobile client's cache reusability and data access time in a wireless broadcast environment," *Data and Knowledge Engineering*, vol. 63, no. 2, pp. 293–314, 2007.

[4] K. Lakshminarayan, S. A. Harp, and T. Samad, "Imputation of missing data in industrial databases," *Applied Intelligence*, vol. 11, no. 3, pp. 259–275, 1999.

[5] R. Lenz and M. Reichert, "IT support for healthcare processes—premises, challenges, perspectives," *Data and Knowledge Engineering*, vol. 61, no. 1, pp. 39–58, 2007.

[6] A. Mercatanti, "Analyzing a randomized experiment with imperfect compliance and ignorable conditions for missing data: theoretical and computational issues," *Computational Statistics and Data Analysis*, vol. 46, no. 3, pp. 493–509, 2004.

[7] T. Özacar, Ö. Öztürk, and M. O. Ünalir, "ANEMONE: An environment for modular ontology development," *Data and Knowledge Engineering*, vol. 70, no. 6, pp. 504–526, 2011.

[8] A. C. Favre, A. Matei, and Y. Tillé, "A variant of the Cox algorithm for the imputation of non-response of qualitative data," *Computational Statistics and Data Analysis*, vol. 45, no. 4, pp. 709–719, 2004.

[9] A. F. de Winter, A. J. Oldehinkel, R. Veenstra, J. A. Brunnekreef, F. C. Verhulst, and J. Ormel, "Evaluation of non-response bias in mental health determinants and outcomes in a large sample of pre-adolescents," *European Journal of Epidemiology*, vol. 20, no. 2, pp. 173–181, 2005.

[10] P. A. Patrician, "Multiple imputation for missing data," *Research in Nursing and Health*, vol. 25, no. 1, pp. 76–84, 2002.

[11] J. Honaker and G. King, "What to do about missing values in time-series cross-section data," *American Journal of Political Science*, vol. 54, no. 2, pp. 561–581, 2010.

[12] N. Sartori, A. Salvan, and K. Thomaseth, "Multiple imputation of missing values in a cancer mortality analysis with estimated exposure dose," *Computational Statistics and Data Analysis*, vol. 49, no. 3, pp. 937–953, 2005.

[13] J. M. Lepkowski, W. D. Mosher, K. E. Davis, R. M. Groves, and J. Van Hoewyk, "The 2006–2010 National Survey of Family Growth: sample design and analysis of a continuous survey," *Vital and Health Statistics*, no. 150, pp. 1–36, 2010.

[14] I. Myrtveit, E. Stensrud, and U. H. Olsson, "Analyzing data sets with missing data: an empirical evaluation of imputation methods and likelihood-based methods," *IEEE Transactions on Software Engineering*, vol. 27, no. 11, pp. 999–1013, 2001.

[15] A. L. Bello, "Imputation techniques in regression analysis: looking closely at their implementation," *Computational Statistics and Data Analysis*, vol. 20, no. 1, pp. 45–57, 1995.

[16] D. A. Newman, "Longitudinal modeling with randomly and systematically missing data: a simulation of Ad Hoc, maximum likelihood, and multiple imputation techniques," *Organizational Research Methods*, vol. 6, no. 3, pp. 328–362, 2003.

[17] R. H. Jones, "Maximum likelihood fitting of ARMA models to time series with missing observations," *Technometrics*, vol. 22, no. 3, pp. 389–395, 1980.

[18] G. Jagannathan and R. N. Wright, "Privacy-preserving imputation of missing data," *Data and Knowledge Engineering*, vol. 65, no. 1, pp. 40–56, 2008.

[19] J. Van Hulse and T. Khoshgoftaar, "Knowledge discovery from imbalanced and noisy data," *Data and Knowledge Engineering*, vol. 68, no. 12, pp. 1513–1542, 2009.

[20] A. Bagheri and M. Zandieh, "Bi-criteria flexible job-shop scheduling with sequence-dependent setup times—variable neighborhood search approach," *Journal of Manufacturing Systems*, vol. 30, no. 1, pp. 8–15, 2011.

[21] M. Zhang, J. Zhu, D. Djurdjanovic, and J. Ni, "A comparative study on the classification of engineering surfaces with dimension reduction and coefficient shrinkage methods," *Journal of Manufacturing Systems*, vol. 25, no. 3, pp. 209–220, 2006.

[22] M. S. Pishvaee, F. Jolai, and J. Razmi, "A stochastic optimization model for integrated forward/reverse logistics network design," *Journal of Manufacturing Systems*, vol. 28, no. 4, pp. 107–114, 2009.

[23] P. A. Ferrari, P. Annoni, A. Barbiero, and G. Manzi, "An imputation method for categorical variables with application to nonlinear principal component analysis," *Computational Statistics and Data Analysis*, vol. 55, no. 7, pp. 2410–2420, 2011.

[24] T. H. Lin, "A comparison of multiple imputation with EM algorithm and MCMC method for quality of life missing data," *Quality and Quantity*, vol. 44, no. 2, pp. 277–287, 2010.

[25] S. Van Buuren, H. C. Boshuizen, and D. L. Knook, "Multiple imputation of missing blood pressure covariates in survival analysis," *Statistics in Medicine*, vol. 18, pp. 681–694, 1999.

[26] K. Hron, M. Templ, and P. Filzmoser, "Imputation of missing values for compositional data using classical and robust methods," *Computational Statistics and Data Analysis*, vol. 54, no. 12, pp. 3095–3107, 2010.

[27] F. Dudbridge, "Likelihood-based association analysis for nuclear families and unrelated subjects with missing genotype data," *Human Heredity*, vol. 66, no. 2, pp. 87–98, 2008.

[28] Z. Lu and Y. V. Hui, "L_1 linear interpolator for missing values in time series," *Annals of the Institute of Statistical Mathematics*, vol. 55, no. 1, pp. 197–216, 2003.

[29] D. Decoste and B. Schölkopf, "Training invariant support vector machines," *Machine Learning*, vol. 46, no. 1–3, pp. 161–190, 2002.

[30] D. Li, L. Chen, and Y. Lin, "Using Functional Virtual Population as assistance to learn scheduling knowledge in dynamic manufacturing environments," *International Journal of Production Research*, vol. 41, no. 17, pp. 4011–4024, 2003.

[31] Y. S. Lin and D. C. Li, "The Generalized-Trend-Diffusion modeling algorithm for small data sets in the early stages of manufacturing systems," *European Journal of Operational Research*, vol. 207, no. 1, pp. 121–130, 2010.

[32] K. Goto, H. Okabe, F. A. Chowdhury, S. Shimizu, Y. Fujioka, and M. Onoda, "Development of novel absorbents for CO_2 capture from blast furnace gas," *International Journal of Greenhouse Gas Control*, vol. 5, no. 5, pp. 1214–1219, 2011.

[33] H. Jaeger, "A tutorial on training recurrent neural networks, covering BPTT, RTRL, EKF and "Echo State Network"

approach," GMD Report 159, German National Research Center for Information Technology, Berlin, German, 2002.

[34] H. Jaeger, "Adaptive nonlinear system identification with echo state networks," *Advances in Neural Information Processing Systems*, vol. 15, pp. 593–600, 2003.

[35] C. F. Huang, "Principle of information diffusion," *Fuzzy Sets and Systems*, vol. 91, no. 1, pp. 69–90, 1997.

[36] E. P. Zhou and D. K. Harrison, "Improving error compensation via a fuzzy-neural hybrid model," *Journal of Manufacturing Systems*, vol. 18, no. 5, pp. 335–344, 1999.

[37] G. Ward, T. Hastie, S. Barry, J. Elith, and J. R. Leathwick, "Presence-only data and the EM algorithm," *Biometrics*, vol. 65, no. 2, pp. 554–563, 2009.