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.



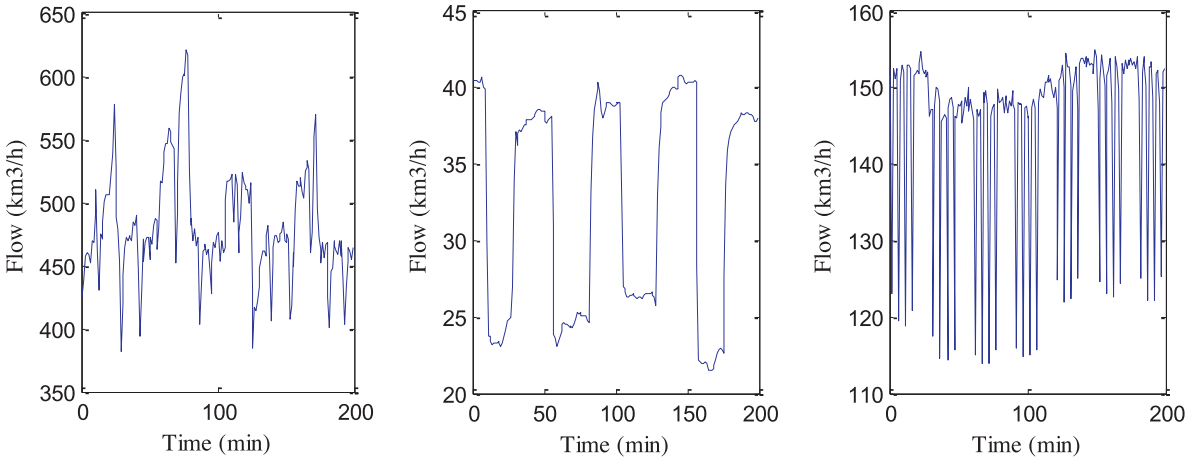
## 1. Introduction

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/

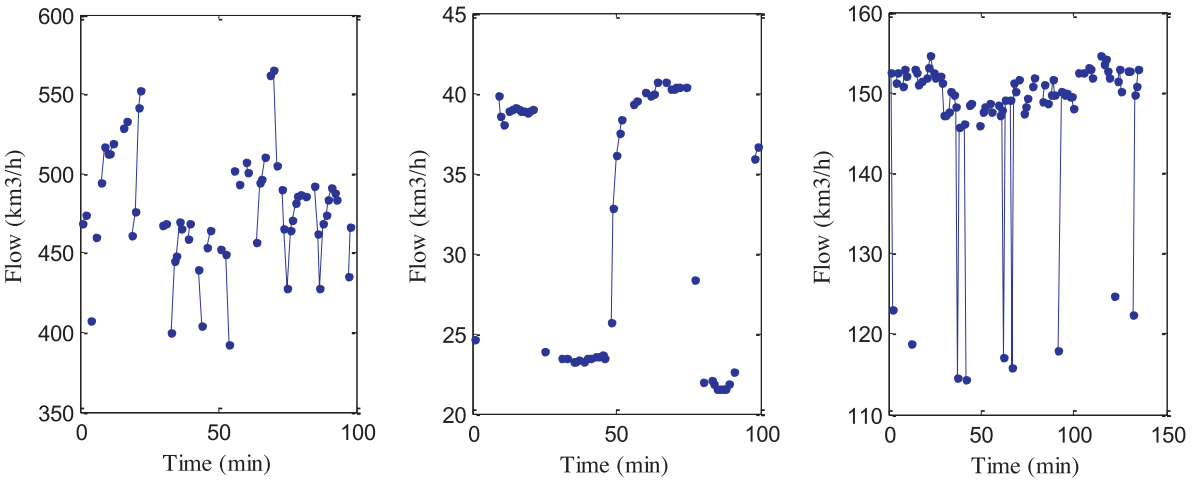



(a) BFG generation flow of blast furnace.

(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].