文献标识码：B 文章编号：1003-0492（2021）12-064-04 中图分类号：TP309.3

基于相关因子的工业过程数据缺失值填补研究

Research on Missing Value Filling of Industrial Process Data Based on Correlation Factors

★ 徐瑞东（浙江正泰中自控制工程有限公司，浙江 杭州 310018）
★ 侯志鹏（杭州电子科技大学，浙江 杭州 310018）

摘要：文章分析了煤化工企业过程数据的特性，以及不同方法对过程数据缺失填补的优劣性，提出了一种基于相关性因子的工业缺失数据填补方法。通过引入相关因子，对传统的时间序列法进行改进，提高了对工业缺失数据估计的准确性。实验结果验证了论文方法的有效性。

关键词：相关因子；缺失数据；时间序列法

Abstract: This paper analyzes the characteristics of process data in coal chemical enterprises and the effect of different methods for filling in the lack of process data, and puts forward an industrial missing data filling method based on correlation factor. By introducing correlation factors, the traditional time series method is improved to improve the accuracy of industrial missing data estimation. The experimental results verify the effectiveness of the proposed method.

Key words: Correlation factor; Missing data; Time series method

1 引言

现今对于缺失数据主要是通过统计学知识和一些数据挖掘类的算法进行处理[1]。在早年对缺失数据处理时，主要是通过统计学的方法对数据进行简单的插补，后续的研究主要也是采用平滑指数或者插值法[2~5]。从预测角度出发，对于缺失数据的预测主要利用状态空间模型[6]。这类方法根据收集到的数据发现某种潜在的规律，从而预测缺失部分的数据[7]。文献[8]通过缺失点前的数据，以自回归滑动平均模型，对其后续的缺失进行预测，但是这样做并没有充分利用到缺失点后续的数据，从整体的利用程度来说，其对数据整体的利用程度较低。文献[9]主要通过SVM方法对缺失数据进行预测，但是SVM方法本身不能够对大量的数据缺失情况进行有效的填补。文献[10]主要是利用BP神经网络对缺失数据进行预测，由于神经网络的非线性拟合能力强，所以对于缺失值的填补有良好的效果，但是神经网络本身就是需要大量的数据进行训练才能达到预测的效果，而数据缺失的情况一般都是在数据一开始就会有缺失，很难给足够量的数据进行训练，所以神经网络法的处理方式和实际情况相差较远，并不能广泛适用。文献[11]利用统计方法进行缺失填补，通过数据统计特征得到缺失数据的概率分布情况，通过最大似然估计出缺失部分最适合的值进行填补。文献[12]发现主成分分析法对缺失值的填补也有很好的效果。文献[13]提出利用矩阵分解的方法可以对于高缺失数据有良好的填补效果。

煤化工业中由于工业生产过程中的记录情况和传感器传输过程以及传感器本身工作性能等多方面的因素造成数据缺失，由于工艺流程数据维度大，缺失时间点不一，所以存在很强烈的随机性。

在过程数据中瞬时缺失数据，即一组数据中缺失个数很少的情况下，由于过程数据本身具有平滑性的特点，用最简单的统计估计法其实能很有效地填补缺失数据。因为过程数据尤其是传感数据在采集时就会有系统噪声的存在，这是传感器本身的检测方式所决定的，而且工业流程数据采样周期短，且在工业流程运行中很难发生突变，所以前后数据变化幅度不大甚至没有变化，利用简单估计法既能节省时间，又能达到很好的填补效果。对过程数据的瞬时缺失研究并无特别大的意义，本

文主要是研究中高型数据缺失量的恢复方法。

对于中大型的数据缺失情况，现今的研究主要是利用逻辑回归的插补法、时间序列法、稀疏贝叶斯恢复法[14]等。其中稀疏贝叶斯恢复法属于多维估计法，是采用基于压缩感知的信号恢复算法，这种算法主要利用时间序列时域平滑特性设计稀疏表示基，将一个缺失数据填补问题转化为一个稀疏向量的恢复问题。通过针对工业流程数据的特点，设计特定的稀疏表示基和相应观测矩阵，利用稀疏贝叶斯方法还原缺失数据。但在实际实验过程中发现多维估计法针对于工业缺失数据而言并未有良好的恢复能力，比较时间序列预测法和插补法发现时间序列预测法的准确性更高。但是时间序列预测法只是考虑了时序性，并未利用过程数据的相关性特征，为此本文提出了一种基于相关因子的时间序列预测方法（CR时间序列预测法）对煤化工业缺失值进行填补。CR时间序列预测法主要是利用工业传感数据中的强相关性，找到和缺失数据变化幅度相关的另一组过程数据，用同一时刻的采样数据通过相关因子分配相应权重对时间序列预测法的填补值进行补偿修正，从而使得结果更加准确。

2 一种基于相关因子的工业缺失数据恢复方法

本文所提基于相关因子的时间序列预测方法的具体步骤如图1所示。

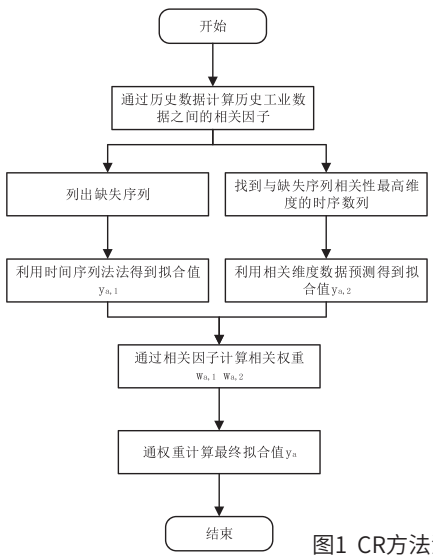

图1 CR方法流程图

Step1：通过时间序列预测算法（本文实验中利用的是自回归移动平均模型）对该缺失数据进行自相关的预测填补得到拟合值$y_{a,1}$。

Step2：通过历史流程过程数据计算出不同维度的时间序列之间的皮尔逊相关系数。找到与需要补缺的时间序列s_ϑ相关性最大，即相关因子为$r_{\theta,\rho}$的一组过程数据s_ρ，利用二者的强相关性，用数据s_ρ拟合出缺失估计值$y_{a,2}$，如式（1）所示。

$$y_{a,2}=\frac{(\max(s_\vartheta)-\min(s_\vartheta))*(s_{i\rho}-\min(s_\rho))}{\max(s_\rho)-\min(s_\rho)}+\min(s_\vartheta) \quad (1)$$

其中$\max(s_\vartheta)$和$\max(s_\rho)$分别是s_ϑ、s_ρ在历史数据中的最大最小值，$s_{i\rho}$为和缺失数据在同一时刻s_ρ中的采样值。

皮尔逊相关系数计算如式（2）所示：

$$r_{\theta,\rho}=\frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{P_i-Mean(s_\vartheta)}{\sigma_P}\right)\left(\frac{Q_i-Mean(s_\rho)}{\sigma_Q}\right) \quad (2)$$

其中P_i和Q_i表示时间序列s_ϑ和s_ρ的第i个值，$Mean(s_\vartheta)$和$Mean(s_\rho)$表示时间序列s_ϑ和s_ρ的平均值，σ_P、σ_Q是其标准差，n为进行相关性分析的数据个数。

Step3：最终拟合值如式（3）所示：

$$y_a=w_{a,2}y_{a,2}+w_{a,1}y_{a,1} \quad (3)$$

其中$w_{a,2}=\frac{r_{\theta,\rho}}{2}$，$w_{a,1}=\frac{2-r_{\theta,\rho}}{2}$

3 实验仿真

实验选取的数据为采样周期中的传感器数据，每一个采样周期中的采样次数均为721，即每一组时间序列的维度为721，人为缺失数量为300，即每一组数据的缺失数量为300个，且缺失位置随机分布。

本节实验选取了5组缺失变量，分别使用稀疏贝叶斯方法、逻辑回归插补法、时间序列法和CR时间序列法进行填补。5组变量及其相关变量和相关因子如表1所示。

表1 变量相关因子表

序号	缺失变量	相关变量	相关因子
1	气化炉煤浆进料流量	高压煤浆泵出口流量	0.84
2	洗涤塔出口流量	P1051去洗涤塔顶流量	0.98
3	气化炉液位	煤浆贮槽液位	0.97
4	气化炉出口流量	气化炉去V1051/V1054流量	0.205
5	气液分离器入口温度	煤浆贮槽液位	0.9

图2、图3分别给出了气化炉煤浆进料流量、洗涤塔出口流量的缺失值填补结果，表2为气化炉煤浆进料流量、洗涤塔出口流量、气化炉液位、气化炉出口流量、气液分离器入口温度的填补误差。

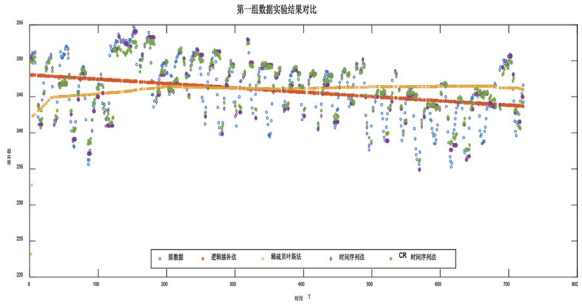

图2 汽化炉煤浆进料流量缺失值填补

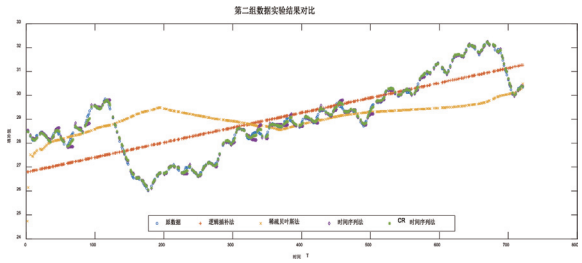

图3 洗涤塔出口流量缺失值填补

表2 实验误差表

	气化炉煤浆进料流量(m^3/h)	洗涤塔出口流量(m^3/h)	气化炉液位(m)	气化炉出口流量(m^3/h)	气液分离器入口温度(℃)
CR时间序列法	4.804	0.157	0.868	0.009	1.096
时间序列法	6.824	0.271	1.447	0.011	1.253
逻辑回归插补法	8.507	2.154	2.637	0.018	1.579
稀疏贝叶斯法	25.170	3.996	5.563	0.051	7.063

从实验结果来说用稀疏贝叶斯的方法对煤化工过程数据进行缺失值恢复的效果很差，这是因为多维估计法不能适应于过程数据的时序特征，并且在对数据的维度信息中提取能力不足。其次是逻辑回归插补法，由于插补法主要通过两点间的逻辑趋势进行填补，在随机缺失的情况下不能很好地利用一整列的时序特征，从而造成对时序信息抓捕不完善。对于时间序列预测法而言，CR时间序列预测的误差率均比一般的时间预测法小，填补效果更优。

但是上述实验中相关因子除了第四组数据外，相关因子均较大，不足以体现出CR时间序列法的普遍适用性。所以下面利用不同大小的相关因子进行对比实验，验证在不同相关因子下CR时间序列法仍然能对时间序列法做出有效的补偿修正，如表3所示。

表3 不同相关因子实验变量表

序号	1	2	3	4	5	6
缺失变量	洗涤塔循环液流量	水冷器液位	pH调节剂液位	脱氧水槽液位	气化炉液位	气化炉煤浆进料流量
相关变量	P1051去洗涤塔顶流量	水冷器出口压力	气化炉一氧化碳浓度	P1051去洗涤塔顶流量	高压煤浆泵出口流量	泵P1003出口流量
相关因子	0.990	0.880	0.780	0.730	0.3	0.180
CR时间序列法误差	0.155	1.055	0.758	1.156	1.389	6.502
时间序列法误差	0.255	1.313	0.959	1.185	1.447	6.824
逻辑回归插补法误差	2.114	1.895	1.337	1.635	2.637	8.507
稀疏贝叶斯法误差	4.021	4.322	7.417	7.400	5.563	25.170

可以发现无论是高相关性下即第一组至第四组数据，还是低相关性下即第五组和第六组数据中，CR时间序列预测法均能有效地对时间序列方法做补偿修正，其填补效果优于另外3种方法。但也能看出随着相关变量的相关性越来越低，其所能补偿的幅度也会降低，效果渐渐趋近于时间序列法。

4 结语

论文主要介绍了煤化工过程数据的特性，当前工业缺失数据的相关方法，以及不同方法的优劣分析，总结了煤化工过程数据中处理缺失值的需求。针对工业缺失数据，利用煤化工过程数据的相关性特征，提出了基于相关因子的工业缺失数据恢复方法。通过实际对于所

选取出的煤化工业关键风险节点的传感器数据进行实际填补仿真，此方法可以提高煤化工工艺中的过程数据缺失值的预测准确性。AP

作者简介：

徐瑞东（1969-），男，江苏海安人，高级工程师，硕士，现就职于浙江正泰中自控制工程有限公司，研究方向为智慧化工及智慧水务领域工业自动化和信息化技术。

侯志鹏（1993-），男，江苏泰州人，硕士，现就职于杭州电子科技大学，研究方向为流程企业风控评估。

参考文献：

[1] Shin-Fu Wu, Chia-Yung Chang, Shie-Jue Lee. Time Series Forecasting with Missing Values[C]. 2015 1st International Conference on Industrial Networks and Intelligent Systems (INISCom), 2015, 1 (4) : 2 - 8.

[2] Balouji E, Salor O, Ermis M. Exponential smoothing of multiple reference frame components with GPUs for real-time detection of time-varying harmonics and interharmonics of EAF currents[J]. 2017 : 1 - 8.

[3] Kozera, R, Wilkoazka, et al. Natural Spline Interpolation and Exponential Parameterization for Length Estimation of Curves[J]. AIP Conference Proceedings, 2017, 1863 (1) : 1 - 4.

[4] Junninen H, Niska H, Tuppurainen K, et al. Methods for imputation of missing values in air quality data sets[J]. Atmospheric Environment, 2004, 38 (18) : 2895 - 2907.

[5] Hong S T, Chang J W. A New Data Filtering Scheme Based on Statistical Data Analysis for Monitoring Systems in Wireless Sensor Networks[C]. IEEE International Conference on High Performance Computing & Communications. IEEE, 2011 : 2 - 4.

[6] A Y K, B Y H, A M K. Traffic state estimation on a two-dimensional network by a state-space model[J]. Transportation Research Part C: Emerging Technologies, 2020, (113) : 176 - 192.

[7] Lao W, Wang Y, Peng C, et al. Time series forecasting via weighted combination of trend and seasonality respectively with linearly declining increments and multiple sine functions[C]. 2014 International Joint Conference on Neural Networks(IJCNN). IEEE, 2014 : 6 - 11.

[8] Lippi M, Bertini M, Frasconi P. Short-Term Traffic Flow Forecasting: An Experimental Comparison of Time-Series Analysis and Supervised Learning[J]. IEEE Transactions on Intelligent Transportation Systems, 2013, 14 (2) : 871 - 882.

[9] 朱梦成. 面向缺失数据处理的SVM算法研究[D]. 2019 : 2 - 35.

[10] Gao W, Niu K , Cui J, et al. A data prediction algorithm based on BP neural network in telecom industry[C]. International Conference on Computer Science & Service System. IEEE, 2011 : 4 - 11.

[11] Li L, Li Y, Li Z. Efficient missing data imputing for traffic flow by considering temporal and spatial dependence[J]. Transportation Research Part C Emerging Technologies, 2013, 34 (12) : 108 - 120.

[12] Qu L, Li L, Zhang Y, et al. PPCA-Based Missing Data Imputation for Traffic Flow Volume: A Systematical Approach[J]. IEEE Transactions on Intelligent Transportation Systems, 2009, 10 (3) : 512 - 522.

[13] Shi W, Zhu Y, Yu P, et al. Effective Prediction of Missing Data on Apache Spark over Multivariable Time Series[J]. IEEE Transactions on Big Data, 2017 : 1.

[14] Song X, Guo Y, Li N, et al. A novel approach for missing data prediction in coevolving time series[J]. Computing, 2019, 101 (11) : 1565 - 1584.