Research on the detection method for abnormal data of wastewater treatment plants based on statistical theory

Liang Guo [1,2] , Ying Zhao [1,2 *] , Fuyi Cui [1,2*]

[1]School of municipal and environmental engineering, Harbin Institute of Technology, 150090, China

[2]State Key Laboratory of Urban Water Resource and Environment, Harbin Institute of Technology, 150090, China

[a] guoliang0617@hit.eud.cn, [b] zhaoying@hit.edu.cn, [c] cuifuyi@hit.edu.cn

Keywords: Statistical theory, wastewater treatment plant, abnormal data detection, t-test method, online monitoring, significant difference analysis

Abstract. In this study, statistical theory is employed to identify abnormal online monitoring data and substandard effluent water quality in a wastewater treatment plant (WTP) of northern China. The effluent water quality can be monitored and checked by the national standard (GB18918-2002). The detection of abnormal data is achieved for daily monitoring data by single-side t-test and two-side t-test methods. Meanwhile, significance difference analysis between manual data and online monitoring date is investigated to check whether there is abnormity in manual data. The established method is then applied to an actual case. The result indicates that the method can effectively identify the abnormal data in WTPs and provide powerful technical measures for regulatory authorities.

Introduction

Usually, some abnormal data are found from the equipment of process in a WTP resulting in substandard effluent water quality. This phenomenon is caused by the false data, low accuracy of instruments and malfunction of monitoring facilities. Therefore, certain method should be employed to detect the abnormal data, which are really beneficial for regulatory authorities to find the abnormal situation of WTPs and identify the reason of the fault in time.

Statistical method is first applied to the detection of abnormal data [1-2]. Jin and Xu [3] use three principles, Dixon method, Grubbs detection method and Boxplot method to detect respectively the abnormal data of lint yield in the cotton area of the Yangtze River Basin. Sun et al. [4] apply Boxplot method to the identification of abnormal animal health data, and accurately determine the imprecise data provided by statisticians. N.C. Schwertman [5] develops a modified Boxplot method for the identification of abnormal data of the gravity of wood substance and achieves excellent effects. An effective statistical method is used by Seok et al [6] to detect and correct the abnormal data among the historical data to achieve accurate and practical hourly water demand forecasting. In this study, the detection method for abnormal data in a WTP is investigated, which could provide powerful technical measures for regulatory authorities.

Methods

Study WTP and data sources

In this study, a northern WTP is selected as the research subject. According to the actual investigates, expert consultation and literature research, the daily management style and key technology of the WTP are acquainted. In addition, related parameters are collected and analyzed, including the water quality parameters (SS, TN, COD, NH3-N, TP) and operation parameters (influent flow Q, F/M, DO, ORP, SVI, C/N, et.).

T-test method

Significance analysis for singly normal population parameter Assuming $X_1, X_2 \mathbf{K}, X_n$ is a sample of the population $\mathrm{N}(m, s^2)$, $\bar{X}$ and S^2 are the sample mean value and sample variance, respectively. T-test method proceeds as follows: (1) Pose a statistical hypothesis and determine to use single-side test or two-side hypothesis test. For two-side hypothesis test, it is assumed that H_0 : $m = m_0$, H_1 : $m \neq m_0$; For single-side hypothesis test, it is assumed that H_0 : $m \geq m_0$ or $m \leq m_0$, and H_1 : $m < m_0$ or $m > m_0$; (2) Choose the sample statistic $t = \frac{\overline{X} - m_0}{S}\sqrt{n}$, and calculate its value; (3) Define the significance level a , which is always determined as a =0.05 or a =0.01; (4) For the given significance level a , obtain the marginal value of $t_{a/2}(n-1)$ (single-side) and $t_a(n-1)$ (two-single) according to the *t* distribution table, and determine the probability *p*; (5) Make a conclusion: H_0 will be rejected and H_1 will be accepted if $|t| \geq t_{a/2}(n-1)$ for two-side hypothesis test; H_0 will be rejected and H_1 will be accepted if $|t| \geq t_a(n-1)$ for single-side hypothesis test. The value of P can be used to determine whether to accept the original hypothesis. H_0 will be rejected and H_1 will be accepted if P < a.

Paired t-test for two normal population It is assumed that the two samples $X_1, X_2, \mathbf{K}, X_{n_1}$ and $Y_1, Y_2, \mathbf{K}, Y_{n_2}$ obey the population $\mathrm{N}(m_1, s_1^2)$ and $\mathrm{N}(m_2, s_2^2)$ respectively, and they are independent of each other. According to the theorem:

$$\frac{\overline{X} - \overline{Y} - (m_1 - m_2)}{S_W\sqrt{\frac{1}{\mathrm{n}_1} + \frac{1}{\mathrm{n}_2}}} \square\ t(n_1 + n_2 - 2) \qquad (1)$$

where $S_W = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}$, S_1^2 and S_2^2 are the sample variances of two independent samples, respectively. When s_1^2 and s_2^2 are unknown, $n_1 = n_2 = n$, and t-test method is used to perform the significance test of the two normal population. Assuming H_0 : $m_1 = m_2$, and H_1 : $m_1 \neq m_2$. Define that $Z_i = X_i - Y_i \quad i = 1, 2, ..., n$, so $Z_1, Z_2, ..., Z_n$ is under independent and identically distribution. Meanwhile, $Z_i \sim \mathrm{N}(d, s^2)$, $d = m_1 - m_2$, and $s^2 = s_1^2 + s_2^2$. Here, whether $m_1 = m_2$ is tested, which is equivalent to H_0: d=0. The statistic *t* is determined in equtation (2), and $\overline{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_i$, $S = \sqrt{\frac{1}{n\text{-}1}\sum_{i=1}^{n}(Z_i - \overline{Z})^2}$. Then, the refused domain of H_0 can be determined as $|t| \geq t_{a/2}(n-1)$.

$$t = \frac{\overline{Z}}{S}\sqrt{n} \qquad (2)$$

Results and Discussion

Detection of abnormal values of online monitoring data

The monitoring data is collected every two hour, so there are 12 data in a day. In combination of historical data and expertise, the experience range of online monitoring data can be determined, including SS : 3 ~ 200mg/L ; COD : 10 ~ 300mg/L ; NH_3-N : 0 ~ 60mg/L ; TN : 4 ~ 70mg/L. If the data is out of the range, it can be determined that the abnormal data is caused by equipment failures. While, if the data is involved in the range, the first B category wastewater discharge standard (GB18918-2002) can be used to determine whether the real-time data has exceeded the standard.

The SS data presents abnormal value as 0.0 at 22:00, one day in 2014, which is not involved in the experience interval, and the values of NH_3-N, TN and COD are normal. According to the investigation, the abnormal value is caused by the malfunction of the suspended particle probe, therefore it should be eliminated and data revisions should be performed. Based on the analysis at lab, the abnormal value is replaced by 9.2mg/L, which is within the standard value (20mg/L).

Overproof detection of effluent water quality based on t-test of single sample

In this work, single-side t-test and two-side t-test are employed to assess whether the modified effluent data is within the standard value.

(1) T-test analysis result of standard data

Table 1 shows the probability values of Kolmogorov-Smirnov test and Shapiro-Wilk test of the standard data of effluent COD by the online monitoring equipment in a northern WTP. As shown in the table, all the probability values are greater than 0.05, which are in accordance with normal distribution, so two-side t-test is employed.

Table1 Normality test

	Normality test			
	Kolmogorov-Smirnov		Shapiro-Wilk	
	D statistic	Probability	W statistic	Probability
COD	0.153	0.200	0.908	0.202

Assuming H_0 ： $m = m_0 = 60$mg/L ， H_1 ： $m \neq m_0$. According to sample data, it could be founded that $N = 12$ ， $\overline{X} = 37.1392$ ， the standard deviation $S = 2.7015$ and the degree of freedom $df = n\text{-}1 = 11$. Thus, the calculation of the statistics is $t = \frac{\overline{X} - m_0}{S / \sqrt{n}} = \frac{37.1392 - 60}{2.7015 / \sqrt{11}} = -29.314$. According to t distribution table, when $a = 0.05$, $t_{0.05/2}(11) = 2.201$, it could be founded that $|t| > t_{0.05/2}(11)$. Therefore, H_0 is negated and H_1 is accepted, indicating that the effluent COD is not equal to 60mg/L. Further single-side test is performed. Assuming H_0 ： $m \geq m_0$ ， H_1 ： $m < m_0$. According to t distribution table, $t_{0.05}(11) = 1.796$, and $t = -29.314 < -1.796$. Therefore, H_0 is negated, indicating that the effluent COD is less than 60mg/L.

(2) T-test analysis result of substandard data

A set of substandard data is tested by two-side t-test. Assuming the effluent COD is 60 mg/L, which indicates that H_0 ： $m = m_0 = 60$mg/L ， H_1 ： $m \neq m_0$. Based on the sample data calculation, it is founded that $N = 12$ ， $\overline{X} = 62.6075$ ， the standard deviation $S = 4.1244$ and the degree of freedom $df = n - 1 = 11$. Thus, the calculation of the statistics is $t = \frac{\overline{X} - m_0}{S / \sqrt{n}} = \frac{62.6075 - 60}{4.1244 / \sqrt{11}} = 2.190$. According to t distribution table, when $a = 0.05$, $t < t_{0.05/2}(11) = 2.201$. Therefore, H_0 is accepted, which indicates that the effluent COD reaches 60mg/L. Further single-side test is performed. Assuming H_0 ： $m \leq m_0$ ， H_1 ： $m > m_0$. According to t distribution table, $t_{0.05}(11) = 1.796$, and therefore t=2.190> 1.796. Thus, H_0 is negated, indicating that the COD is greater than 60mg/L.

Significant difference analysis for data based on paired t-test

Paired t-test is used to judge whether significant difference exists between manual data and online monitoring date. The average value of the overall 12 data values is defined as the effective daily mean value. Effluent water quality data of May, 2014 are selected to research the defference between manual(lab) data and online monitoring data of a WTP in northern China.

(1) The comparison of effluent SS: t-test is used to test the following two sets of data. Assuming H_0 ： $m_1 = m_2$ ， H_1 ： $m_1 \neq m_2$. Define $Z_i = X_i - Y_i \quad i = 1,2,...,31$, where X_i represents the data tested in lab, Y_i is the online monitoring data. Thus, $\overline{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_i = 0.00161$, $S = \sqrt{\frac{1}{n\text{-}1}\sum_{i=1}^{n}(Z_i - \overline{Z})^2} = 0.32785$. It is founded that $t = \frac{\overline{Z}}{S}\sqrt{n} = \frac{0.00161}{0.32785}\sqrt{30} = 0.027$. When $a = 0.05$, $t' = t_{0.05/2}(30) = 2.042$, so $t = 0.027 <$ $t' = 2.042$. Therefore, H_0 is accepted, indicating that significant difference does not exist between manual data and online monitoring date (Fig. 1). (2) The comparison of effluent COD: It is founded that $\overline{Z} = 0.14258$ ， $S = 1.80444$, and $t = 0.440 < t' = 2.042$, therefore H_0 is accepted, indicating that significant difference does not exist between manual data and online monitoring date. (3) The

comparison of effluent NH_3-N: It is founded that $\bar{Z}$ =0.00710 ， S =0.19142, and t=0.206< t'=2.042, therefore H_0 is accepted, which indicates that significant difference does not exist between manual data and online monitoring date. (4) The comparison of the effluent TN: It is founded that $\bar{Z}$ =0.06613 ， S =0.67012, and t=0.549< t'=2.042, therefore H_0 is accepted, which indicates that significant difference does not exist between manual data and online monitoring date.

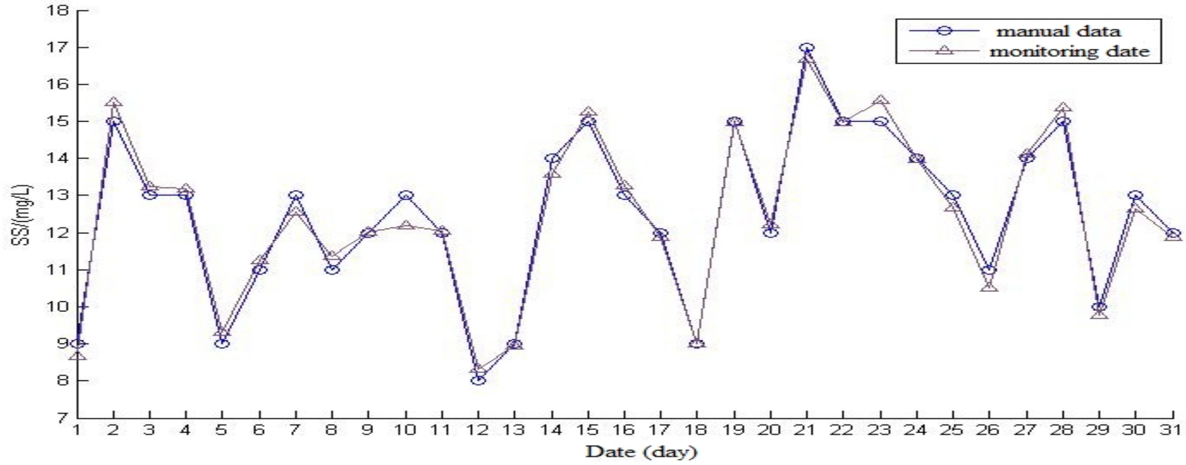

Fig. 1 The comparison of manual data and online monitoring data for effluent SS

Conclusions

In this study, several detection methods for different kinds of abnormal data are proposed based on statistical theory in a WTP of northern China. Firstly, the reasonable ranges are given for four effluent water quality parameters according to historical data and expertise. When the value of effluent water quality is out of the reasonable range, it indicates that there is fault in the online monitoring equipment. If the effluent water quality is in the reasonable range and there is no fault, the effluent water quality can be monitored and checked by the national standard. Secondly, the detection of abnormal data is achieved for daily monitoring data by single-side t-test and two-side t-test methods. Thirdly, the significant difference is calculated between manual data and online monitoring data by paired t-test method to check whether there was abnormity in manual data. The analysis result of the actual case indicates that this method is effective to detect abnormal data in WTPs and can provide technology support for fault diagnosis.

Acknowledgement

This paper is supported by Research and Innovation Fund project of Harbin Institute of Technology (HIT) in 2011.

References

1. L. Li, S, Das, R. J. Hansman, R.l Palacios, and A. N. Srivastava, Analysis of Flight Data Using Clustering Techniques for Detecting Abnormal Operations, Journal of Aerospace Information Systems, (2015). doi: 10.2514/1.I010329

2. L. Y. Zhao, L. Xu, Sewage Treatment Plant Detect Statistical Analysis of Research Data, Environmental science and management, 34 (12) (2009) 25-30.

3.S. Q. Jin, N. Y. Xu, The diagnostic method and application for abnormal data among test data in cotton area, JiangXi Cotton, 32 (6) (2010) 21-24.

4.X. D. Sun, Y. J. Liu, W. W. Chen, The application of Boxplot method in the identification of abnormal data in animal health area, China Animal Health Inspection, 27 (7) (2010) 66-67.

5. N. C. Schwertman, M. A. Owens, R. Adnan, A simple more general boxplot method for identifying outliers, Computational statistics & data analysis, 47 (1) (2004) 165-174.

6. J. H. Seok, J. J. Kim, J. Y. Lee, J. J. Lee, Abnormal data refinement and error percentage correction methods for effective short-term hourly water demand forecasting, International Journal of Control, Automation and Systems, 12 (6) (2014) 1245-1256.