Journal of System Simulation

Volume 32 | Issue 11 Article 19

11-17-2020

Generalized Zero-inflated Binomial Distribution Model Aimed at Air Quality Data Analysis

Benyue Su
1.School of Mathematics and Computer,Tongling University,Tongling 244061,China; ;3.The University Key Laboratory of Intelligent Perception and Computing of Anhui Province,Anqing 246133,China; ;

Pengpeng Xu
2.School of Computer and Information,Anqing Normal University,Anqing 246133,China; ;3.The University Key Laboratory of Intelligent Perception and Computing of Anhui Province,Anqing 246133,China; ;

Sheng Min
3.The University Key Laboratory of Intelligent Perception and Computing of Anhui Province,Anqing 246133,China; ;4.School of Mathematics and Computational Science,Anqing Normal University,Anqing 246133,China;

Generalized Zero-inflated Binomial Distribution Model Aimed at Air Quality Data Analysis

Abstract

Abstract: *For the problem of the quality monitoring and counting of excessive gas emissions in chemical industry parks, a generalized zero-inflated binomial distribution model is constructed.* Statistics show that the times of number of excessive gas emissions has a typical zero-inflated feature. The traditional zero-inflated Poisson model and negative binomial regression model and so on will underestimate the probability of zero inflation. A generalized zero-inflated binomial distribution model is constructed by extending the traditional binomial regression model to a more general form. This model satisfies the characteristic that the expectation is less than the variance, and better solves the problems of both over-dispersed and zero-inflated in excessive gas emissions. Experiments show that *the generalized zero-flated binomial distribution model has a good fitting effect, strong adaptability and robustness.*

Keywords

count model, generalized binomial distribution, zero-inflated model, zero-inflated binomial regression model, air quality analysis

Recommended Citation

Su Benyue, Xu Pengpeng, Sheng Min. Generalized Zero-inflated Binomial Distribution Model Aimed at Air Quality Data Analysis[J]. Journal of System Simulation, 2020, 32(11): 2226-2234.

第 32 卷第 11 期 系统仿真学报© Vol. 32 No. 11
2020 年 11 月 Journal of System Simulation Nov., 2020

Generalized Zero-inflated Binomial Distribution Model Aimed at Air Quality Data Analysis

Su Benyue[1,3], *Xu Pengpeng*[2,3], *Sheng Min*[3,4]

(1.School of Mathematics and Computer, Tongling University, Tongling 244061, China;
(2.School of Computer and Information, Anqing Normal University, Anqing 246133, China;
3.The University Key Laboratory of Intelligent Perception and Computing of Anhui Province, Anqing 246133, China;
4.School of Mathematics and Computational Science, Anqing Normal University, Anqing 246133, China)

Abstract: *For the problem of the quality monitoring and counting of excessive gas emissions in chemical industry parks, a generalized zero-inflated binomial distribution model is constructed.* Statistics show that the times of number of excessive gas emissions has a typical zero-inflated feature. The traditional zero-inflated Poisson model and negative binomial regression model and so on will underestimate the probability of zero inflation. A generalized zero-inflated binomial distribution model is constructed by extending the traditional binomial regression model to a more general form. This model satisfies the characteristic that the expectation is less than the variance, and better solves the problems of both over-dispersed and zero-inflated in excessive gas emissions. Experiments show that *the generalized zero-flated binomial distribution model has a good fitting effect, strong adaptability and robustness.*

Keywords: count model; generalized binomial distribution; zero-inflated model; zero-inflated binomial regression model; air quality analysis

面向空气质量数据分析的广义零膨胀二项模型研究

苏本跃[1,3]，徐鹏鹏[2,3]，盛敏[3,4]

（1. 铜陵学院数学与计算机学院，安徽 铜陵 244061; 2. 安庆师范大学计算机与信息学院，安徽 安庆 246133;
3. 安徽省智能感知与计算重点实验室，安徽 安庆 246133; 4. 安庆师范大学数学与计算科学学院，安徽 安庆 246133）

摘要：*针对化工园区气体超标排放的质量监控和超排计数问题，构建了广义零膨胀二项分布模型。*在统计气体超标排放次数时，发现超标次数具有典型的零膨胀特征。传统的零膨胀泊松模型和负二项回归模型等会低估零膨胀概率，将传统的二项回归模型推广到更为一般的形式，构建了广义零膨胀二项分布模型。该模型满足了期望小于方差的特性，较好解决了超标排放中出现的既有过离散又有零膨胀的问题。实验表明，*广义零膨胀二项分布模型具有较好的拟合效果，适应性和鲁棒性均较强。*

关键词：计数模型；改进二项分布；零膨胀；零膨胀二项回归模型；空气质量分析
中图分类号：TP391; O212; C81; 文献标识码：A 文章编号：1004-731X (2020) 11-2226-09
DOI: 10.16182/j.issn1004731x.joss.19-FZ0510E

Introduction

The pernicious gas emissions of chemical industry have a significant negative effect on air quality and human body. It is noteworthy that the "three wastes" national standards in industrial emission (GBJ4-73) have been developed. In the air, the exhaust gas mixed with each other can lead to a variety of complex chemical reactions. These chemical reactions usually subject to the

Received date: 2019-02-10 Revised date: 2019-09-08;
Foundation: National Nature Science Foundation of China(11475003), Science and Technology Major Project of Anhui Province (18030901021), Anhui Provincial Department of Education outstanding top-notch talent-funded projects (gxbjZD26);
Biographies: Su Benyue(1971-), Male, Wuhu, Ph.D., professor, research direction: statistical computing and statistical pattern recognition.

interferences of the different environments, human and other external factors, and unfortunately can cause unpredictable results. In many countries, the criterion for assessing air quality is the daily air quality index (AQI) that published by the national environmental protection department. In this paper we focus on the main emission gas components (SO_2, NO_2, CO, VOCS, airborne particulate matter and so on) of chemical enterprises that are collected by gas sensor in the chemical park.

From the initial classification analysis of the collected data, we found that the recorded gas emissions data are mostly safe, which means that no excessive emissions at zero. In this case the probability of obtaining a large number zeros values is incoming, which is called typical zero- inflated phenomenon [1]. All of Poisson model and the binomial model are common computational models on the count. In Poisson distribution, the mean and variance are equal. The Poisson model is suitable for "control node" data structure, such as queuing and car crossing. As for the problem of excessive gas emissions of statistical air is obviously not consistent with the characteristics of the Poisson distribution. In the statistical emissions when there are large numbers of zero values, these numbers of zero are called zero-inflated characteristic. The zero-inflated Poisson regression model (ZIP)[1] proposed by Lambert (1992) is very useful to model zero-inflated phenomenon with many zeros and was applied on the analysis of the welding defect data, which was focused on the phenomenon of excessive zero in the data. This model has later been improved by domestic and foreign scholars. Yip et al.[2] have achieved a good result by applying the zero-inflated regression model to non-life actuarial aspects. Hall[3] used a zero-inflated Poisson model in agriculture and conducted a related study. Agarwal et al.[4] have applied zero-inflated model to the field of environmental science. Denuit[5] and Winkelmann[6] have analyzed and discussed the zero-inflated phenomenon in their monograph. Guo[7] has made some improvements on the zero-inflated Poisson model by applying the zero-claim modeling. Meng and Yang[8] have addressed the problem of applying the random effect to the zero-inflated claims frequency. Yuan et al.[9] have used the negative binomial regression model and its promotion in the classification rate determination. Li et al.[10] analyzed the zero-inflated drink-driving problem under time-space effect.

Most of the models that mentioned above are based on zero-inflated Poisson model or its improvements such as zero-inflated negative binomial distribution model and so on. Based on intense analysis we can notice that: when we only discuss whether the air quality in the chemical industry park exceeds the standard per day, the air emissions will meet the characteristics of the binomial distribution. According to the statistical data, our proposed method has achieved the mean value 0.817 4 of dependent variable, which is less than the variance value 1.673 2 of the number of excessive emissions in one chemical park enterprises. From the view of data structure point, the general binomial distribution is expected to be greater than the variance (where, mean value is np, variance value is np $(1\text{-}p)$), which will lead to cause a larger fitting deviation and overestimate the significance level. So the above models such as ZIP, GZIP, ZINB et al. were not adapted to this kind of phenomenon due to that the data in this paper have obvious characteristics of binomial distribution. For this reason we have made some modifications on the

https://dc-china-simulation.researchcommons.org/journal/vol32/iss11/19
DOI: 10.16182/j.issn1004731x.joss.19-FZ0510E

第 32 卷第 11 期 系统仿真学报 Vol. 32 No. 11
2020 年 11 月 Journal of System Simulation Nov., 2020

binomial distribution model to satisfy the requirement of data structure, which we added the adjustment parameters α and adjusted the expectation to be less than the variance.

In our proposed model, we have improved the binomial distribution model and constructed the zero-inflated binomial model (ZIB) and combined the zero-inflated generalized binomial model (ZIGB) with the zero-inflated model. Furthermore, in this work, we try to: fit the number of excessive gas emissions to the chemical park, operate the multi-collinear screening of the sensor data, optimize the gas influencing factor, reduce the calculation and improve efficiency. Meanwhile, compare the zero-inflated generalized binomial model with the binomial distribution model and the zero-inflated binomial distribution model by using the AIC, BIC and likelihood ratio test statistic LR values [11].

The empirical analysis shows that the improved binomial regression model for zero-inflated can achieve better results according to the group data of actual gas emission.

1 Binomial distribution and generalized binomial distribution models

1.1 Binomial distribution model

Suppose the random variable Y_i is the number of excessive emissions of the *i*-th type of air data, where the event $\{Y_i=y_i\}$ obeys the binomial distribution with the probability p_i, then the probability function will be as follows:

$$P(Y_i = y_i) = \binom{n}{y_i} p_i^{y_i}(1-p_i)^{n-y_i}, i = 1,2,...,n \quad (1)$$

where, $y_i \in N$, $E(Y_i) = np_i$ is the expectation, $Var(Y_i) = np_i(1-p_i)$ is the variance of the binomial distribution, and $p_i \in [0,1]$, $\sum p_i = 1$. In the formula (1), the *i*-type air data refers to the collected air component corresponding to frequency *i*, and p_i is the probability of the *i*-th type of air data that exceed the standard.

In this paper we use the multivariate linear expression $z_i = x_i\beta + \beta_0$ to find the air quality data z_i (see Tab. 1) corresponding to the *i*-th air data, which $x_i = (x_{i1}, x_{i2}, \cdots x_{im})$ is the $1\times m$ dimension vector. Where β is $m\times 1$ dimensional parameter vector, that is $\beta = (\beta_1, \beta_2, \cdots, \beta_m)'$, and β_0 is constant. In case of $z_i > 100$, the gas emissions will exceed standards.

In this model, all of sulfide (SOx), nitrides (NOx), airborne particulate matter ($PM_{2.5}$, PM_{10}), CO, VOC_S, temperature and other indicators were selected as influencing factors.

In our experiments we recorded 7 days for one interval as the number of y_i which the gas emissions is exceeded the standards, and exceeded the number of y_i corresponding to the probability p_i which can be expressed as: $p_i = y_i / \sum y_i$.

Tab. 1 Technical Regulations of Environmental Air Quality Index (AQI) (Trial)

Air Quality Index (AQI)	Air Level
0~50	excellent
50~100	good
100~150	Mild pollution
150~200	Moderate pollution
200~300	Severe pollution
>300	Serious pollution

1.2 Generalized binomial distribution model

Let $E(Y_i) = np_i$ is the expectation and $Var(Y_i) = np_i(1-p_i)$ is the variance in the ordinary binomial distribution. From these two formulas it's quite obvious that $E(Y_i) \geq Var(Y_i)$. In order to make air emissions exceed the standards in the chemical park we assume $E(Y_i) < Var(Y_i)$.

Under normal circumstances, the frequency of excessive emissions is lesser than the number of normal emissions. For this reason, the ordinary binomial distribution model should be improved in this paper. Assume the random variable Y_i is times of the i-th type of air data, where Y_i is obeying the generalized binomial distribution. Then the probability density function can be expressed as:

$$P(Y_i = y_i) = \frac{np_i}{np_i + y_i\alpha}\binom{n+\frac{\alpha}{p_i}y_i}{y_i} p_i^{y_i}(1\text{-}p_i)^{n+\frac{\alpha}{p_i}y_i - y_i} \tag{2}$$

where, $i = 1,2,\cdots,n$, $y_i \in N$ and $E(Y_i) = np_i(1-\alpha)^{-1}$ is the mean, $Var(Y_i) = np_i(1-p_i)(1-\alpha)^{-3}$ is the variance and α is the global divergence parameter. In case of $\alpha = 0$, the model will be degenerated into an ordinary binomial distribution, which $E(Y_i) = np_i$ and the variance will be $Var(Y_i) = np_i(1-p_i)$. In this case the parameter α can be used to control the expectations and variance of the distribution as well as describe the data structure.

While, if $\alpha < 1-(1-\max(p_i))^{1/2}$ or $\alpha > 1$, the expectation will be $E(Y_i) > Var(Y_i)$. And this phenomenon will call an under-dispersion phenomenon. And If $1-(1-\max(p_i))^{1/2} < \alpha < 1$, the expectation will be $E(Y_i) < Var(Y_i)$, and this case will call over-dispersion situation.

The model in formula (2) can deal with the over-dispersion problem, as well as solve the under-dispersion problem, this will lead to increase the scope and expand the model by adjusting the value.

For unknown data distribution structure, we can write the logarithmic likelihood expression of the generalized binomial distribution model as follows:

$$l(\alpha) = \prod_{i=1}^{n} P(y_i) =$$

$$\prod_{i=1}^{n}\left[\frac{np_i}{np_i+\alpha y_i}\binom{n+\frac{\alpha}{p_i}y_i}{y_i} p_i^{y_i}(1-p_i)^{n+\frac{\alpha}{p_i}y_i - y_i}\right]$$

$$\ln(l(\alpha)) =$$

$$\sum_{i=1}^{n}\ln\left\{\frac{np_i}{np_i+\alpha y_i}\binom{n+\frac{\alpha}{p_i}y_i}{y_i} p_i^{y_i}(1-p_i)^{n+\frac{\alpha}{p_i}y_i - y_i}\right\} =$$

$$\sum_{i=1}^{n}\left\{\ln np_i + y_i\ln p_i - \ln(np_i+\alpha y_i) + \ln\left(n+\frac{\alpha}{p_i}y_i\right)! - \right.$$

$$\left.\ln y_i ! - \ln\left[n+(\frac{\alpha}{p_i}-1)y_i\right]! + \left(n+\frac{\alpha}{p_i}y_i - y_i\right)\ln(1-p_i)\right\} \tag{3}$$

where, $i = 1,2,\cdots,n$.

In order to get the parameter α value for the maximum likelihood function, the likelihood equation can be rewritten as follows:

$$\frac{\mathrm{d}\ln l(\alpha)}{d\alpha} =$$

$$\sum_{i=1}^{n}\left[-\frac{y_i}{np_i+\alpha y_i} + \frac{y_i}{p_i}\ln(1-p_i) + \frac{y_i}{p_i}\sum_{j=0}^{y_i-1}\frac{1}{n+\frac{\alpha}{p_i}y_i - j}\right] = 0 \tag{4}$$

If $f(\alpha) =$

$$\sum_{i=1}^{n}\left[-\frac{y_i}{np_i+\alpha y_i} + \frac{y_i}{p_i}\ln(1-p_i) + \frac{y_i}{p_i}\sum_{j=0}^{y_i-1}\frac{1}{n+\frac{\alpha}{p_i}y_i - j}\right] \tag{5}$$

Then $f'(\alpha) =$

$$\sum_{i=1}^{n}\left[\frac{y_i^2}{(np_i+\alpha y_i)^2} - \frac{y_i^2}{p_i^2}\sum_{j=0}^{y_i-1}\frac{1}{\left(n+\frac{\alpha}{p_i}y_i - j\right)^2}\right] \tag{6}$$

By applying the Newton iterative method, the iterative formula will be as follows:

$$\alpha_{n+1} = \alpha_n + \frac{f(\alpha_n)}{f'(\alpha_n)}, n = 0,1,...$$

Using the iterative formula will allow us to calculate the estimated value $\hat{\alpha}$ of parameter α .

https://dc-china-simulation.researchcommons.org/journal/vol32/iss11/19
DOI: 10.16182/j.issn1004731x.joss.19-FZ0510E

第 32 卷第 11 期 系统仿真学报 Vol. 32 No. 11
2020 年 11 月 Journal of System Simulation Nov., 2020

2 Zero-inflated regression model

2.1 An overview of the zero-inflated regression model

Assume the random variable Y_i is frequency of the *i*-th type of air data that follows the distribution function of the zero-inflated structure, where ϕ is the ratio of the structure zero. Then the distribution function of the random variable Y_i can be can be expressed as[1]:

$$P(Y_i = y_i) = \begin{cases} \varphi + (1-\varphi) f(0), y_i = 0 \\ (1-\varphi) f(y_i), y_i > 0 \end{cases} \quad (7)$$

In formula (7) ϕ is the coefficient of inflation, which $\phi \in [0,1)$. While $f(y_i)$ is a distribution such as; Poisson distribution, generalized Poisson distribution, binomial distribution and so on. In this form, the zero value is usually composed of two different parts: the first part is that part which the probability of the structural-zero is ϕ, and the second part is called sampling-zero part, in this part, the sampling-zero is generated by the discrete distribution model (such as binomial distribution, et al.).

2.2 Zero-inflated binomial distribution model

By utilizing formula (7), where $f(y_i)$ is the binomial distribution, we can obtain the following zero-inflated binomial (ZIB) distribution model:

$$P(Y_i = y_i) = \begin{cases} \phi + (1-\phi)(1-p_0)^n, y_i = 0 \\ (1-\phi)\binom{n}{y_i} p_i^{y_i} (1-p_i)^{n-y_i}, y_i > 0 \end{cases} \quad (8)$$

where, $E(Y_i) = np_i(1-\phi)$ is the mean; $Var(Y_i) = (1-\phi)p_i[n-(n-\phi)p_i]$ is the variance of the model.

From formula (1) if p_0 is the probability when y_i is 0, then the likelihood function of the *i*-th observed value will be expressed as:

$$\ln l(\phi) = \begin{cases} \ln(\phi + (1-\phi)(1-p_0)^n), y_i = 0 \\ \ln\left((1-\phi)\binom{n}{y_i} p_i^{y_i}(1-p_i)^{n-y_i}\right), y_i > 0 \end{cases} \quad (9)$$

The logarithmic likelihood function expression can be written as follows:

$$\begin{aligned} &\ln l(\varphi) = \ln[\varphi + (1-\varphi)(1-p_o)^n] + \\ &\sum_{y_i>0} \ln\left((1-\varphi)\binom{n}{y_i} p_i^{y_i}(1-p_i)^{n-y_i}\right) = \\ &\ln[\varphi + (1-\varphi)(1-p_o)^n] + \\ &\sum_{y_i>0} \{\ln(1-\varphi) + \ln(n!) - \ln[(n-y_i)!] - \\ &\ln(y_i!) + y_i \ln p_i + (n-y_i)\ln(1-p_i)\} \end{aligned} \quad (10)$$

From (10) we calculation derivative of function $\ln l(\varphi)$ on variable φ, and let the derivative $\frac{d\ln l(\varphi)}{d\varphi}$ equals zero, then we can get:

$$\frac{d\ln l(\varphi)}{d\varphi} = \frac{1-(1-p_0)^n}{\varphi + (1-\varphi)(1-p_0)^n} - \sum_{y_i>0}\left(\frac{1}{1-\varphi}\right) = 0 \quad (11)$$

Similarly, we can use the Newton iterative method to solve the estimated value ϕ of the parameter ϕ.

2.3 Zero-inflated generalized binomial distribution model

In formula (7) if the distribution of $f(y_i)$ is the generalized binomial distribution, the zero- inflated generalized binomial (ZIGB) distribution model could be obtained, and this model can be expressed as follows:

$$P(Y_i = y_i) = \begin{cases} \phi + (1-\phi)(1-p_i)^n, y_i = 0 \\ (1-\phi)\frac{np_i}{np_i + \alpha y_i}\binom{n + \frac{\alpha}{p_i} y_i}{y_i} p_i^{y_i}(1-p_i)^{n+\frac{\alpha}{p_i}y_i - y_i}, y_i > 0 \end{cases} \quad (12)$$

In formula (12), because the mean is $E(Y_i) = n(1-\phi)p_i(1-\alpha)^{-1}$ and the variance is $Var(Y_i) =$

$np_i(1-\phi)(1-\alpha)^{-3}[1-p_i+\phi(1-\alpha)]$, so $E(Y_i)<Var(Y_i)$.

Therefore, the logarithmic likelihood function of the *i*-th observed value can be written as:

$$\ln l(\varphi)=\begin{cases}\ln(\varphi+(1-\varphi)(1-p_0)^n), y_i=0\\ \ln\left((1-\varphi)\dfrac{np_i}{np_i+\alpha y_i}\binom{n+\frac{\alpha}{p_i}y_i}{y_i}p_i^{y_i}(1-p_i)^{n+\frac{\alpha}{p_i}y_i-y_i}\right),\end{cases}$$
$$y_i>0 \tag{13}$$

So the logarithmic likelihood expression the zero-inflated generalized binomial distribution model will be as follows:

$$\ln(l(\alpha,\phi))=\ln(\phi+(1-\phi)(1-p_0)^n)+\sum_{y_i>0}\ln\left((1-\phi)\frac{np_i}{np_i+\alpha y_i}\binom{n+\frac{\alpha}{p_i}y_i}{y_i}p_i^{y_i}(1-p_i)^{n+\frac{\alpha}{p_i}y_i-y_i}\right)=$$
$$\ln(\phi+(1-\phi)(1-p_0)^n)+\sum_{y_i>0}\ln\left((1-\phi)\frac{np_i}{np_i+\alpha y_i}\binom{n+\frac{\alpha}{p_i}y_i}{y_i}p_i^{y_i}(1-p_i)^{n+\frac{\alpha}{p_i}y_i-y_i}\right)=$$
$$\ln\left(\phi+(1-\phi)(1-p_0)^n\right)+\sum_{y_i>0}\ln(1-\phi)+$$
$$\sum_{y_i>0}\ln\frac{np_i}{np_i+\alpha y_i}+\sum_{y_i>0}\ln\left(\frac{(n+\frac{\alpha}{p_i}y_i)!}{(n+\frac{\alpha}{p_i}y_i-y_i)!y_i!}\right)+$$
$$\sum_{y_i>0}y_i\ln p_i+\sum_{y_i>0}\left(n+\frac{\alpha}{p_i}y_i-y_i\right)\ln(1-p_i)_i \tag{14}$$

In (14), if the partial derivatives of the parameters α and ϕ are set to zero, the likelihood equations can be written as follows:

$$\begin{cases}\dfrac{\partial\ln l(\alpha,\phi)}{\partial\phi}=\dfrac{1-(1-p_0)^n}{\phi+(1-\phi)(1-p_0)^n}-\sum_{y_i>0}\dfrac{1}{1-\phi}=0\\ \dfrac{\partial l(\alpha,\phi)}{\partial\alpha}=\sum_{y_i>0}\dfrac{-y_i}{np_i+\alpha y_i}+\\ \sum_{y_i>0}\left[\dfrac{y_i}{p_i}\sum_{j=0}^{y_i-1}\left(\dfrac{1}{n+\frac{\alpha}{p_i}y_i-j}\right)\right]+\sum_{y_i>0}\left(\dfrac{y_i}{p_i}\ln(1-p_i)\right)=0\end{cases} \tag{15}$$

Which the estimated value ϕ and $\hat{\alpha}$ of the parameters ϕ and α can be solved by the Newton iterative method.

3 Case analysis

3.1 Data description

In this work we have the environmental data obtained from a chemical industrial park with real-time transmission by gas sensors. The environmental data mainly include sulfide (SOx), nitride (NOx), airborne particulate matter ($PM_{2.5}$, PM_{10}), CO, VOCS, temperature, air pressure, NH3 and so on. In our experiments we conducted a statistical survey of excessive emissions according to the Air Quality Index (AQI) that was published by the Ministry of Environmental Protection in two months. From Fig. 1 we can see that the frequency of excessive emissions in the statistical environmental data does not conform to the normal distribution. Meanwhile the degree of dispersion will be more serious.

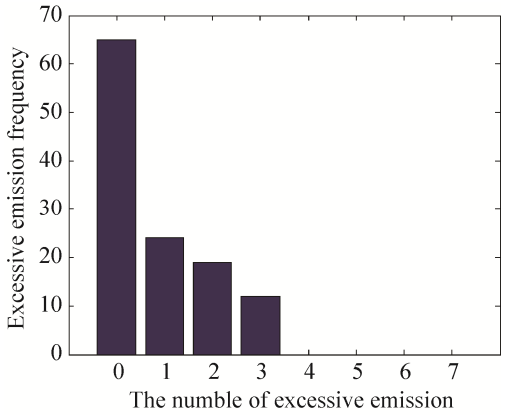

Fig. 1 The distribution of the frequency of emissions

3.2 Test multiple collinearity

In our experiments data have been collected from a variety of environmental indicators in the chemical park. In order to improve the accuracy of the number of simulated emissions, it is necessary to exclude the exact correlation between the indicators

https://dc-china-simulation.researchcommons.org/journal/vol32/iss11/19
DOI: 10.16182/j.issn1004731x.joss.19-FZ0510E

第 32 卷第 11 期 系统仿真学报 Vol. 32 No. 11
2020 年 11 月 Journal of System Simulation Nov., 2020

(or a high degree of correlation) and the irrelevant indicators removal. In this work, in order to maximize the retention of data collection indicators and improve reliability, the multiple collinearity of the collected environmental data index has been tested [8].

According to the variable variance inflation factor (VIF), the larger the VIF value, the stronger the collinearity between the variables. For the division of boundaries, we select VIF <10. From Tab. 2, it can be seen directly which of the variables are collinear, and whether it is suitable for the direct use to simulate the frequency of excessive emission.

Because the air quality data in the unit $\mu g/m^3$, mg/m^3, *ppm*, et al. are quite different in terms of size, so we have to start implementing preprocess the data before diagnosing the multicollinear of the indicators.

In order to normalize the dimension for the data, we will utilize the following equation:

$$y_i = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \tag{16}$$

where $(x_{\max} - x_{\min})$ is corresponding to the relevant national standard (GBJ4-73) of the indicators, and x_i is the actual value collected by the actual sensor.

According to the variable variance inflation factor (VIF), the larger the VIF value, the greater the collinearity between the variables. Therefore, in the practice we selected the set of variables with the smallest VIF value. In Tab. 2, the VIF values are listed after removing partial and different indicators.

In our experiments the indicators have been excluded out of $PM_{2.5}$ and PM_{10}, and the VIF values have exceeded 10. The VIF values are significantly changed when $PM_{2.5}$ or PM_{10} was removed, which means that there is collinearity between them.

When we exclude $PM_{2.5}$, the VIF value will be less than the one that resulted by omitting the PM_{10} value. However, in the actual situation the suspended particles of 2.5 mm diameter in the air must be more than 10 mm. For this reason we have chosen to keep $PM_{2.5}$. And the indexes of $PM_{2.5}$, O_3, CO, SO_2, NO_2, temperature, air pressure, VOCs and NH3 have been selected as influencing factors. In our experiments we have selected nine environmental indicators which indicated by m=9.

Tab. 2 Multiple collinearity diagnosis between the collecting data

$PM_{2.5}$	VOCs	O_3	CO	PM_{10}	SO_2	NO_2	T	AP	NH_3	VIF
1	1	1	1	1	1	1	1	1	1	12.728 9
0	1	1	1	1	1	1	1	1	1	5.686 3
1	0	1	1	1	1	1	1	1	1	12.719 6
1	1	0	1	1	1	1	1	1	1	12.244 6
1	1	1	0	1	1	1	1	1	1	10.598 9
1	1	1	1	0	1	1	1	1	1	6.624 5
1	1	1	1	1	0	1	1	1	1	12.699 8
1	1	1	1	1	1	0	1	1	1	12.210 2
1	1	1	1	1	1	1	0	1	1	12.200 1
1	1	1	1	1	1	1	1	0	1	12.721 6
1	1	1	1	1	1	1	1	1	0	12.586 3

Note: in Tab. 2, "1" indicates that the indicator is existed, while "0" indicates that the indicator is removed. T: temperature, AP: Atmospheric pressure

3.3 The goodness of fitting model

From Fig. 1, we can notice that there are large numbers of zero values when the emission count has exceeded 54%. This indicates that the zero-inflated phenomenon is existed, and the variance value of data is greater than mean value.

In Tab. 3, we compare between various models in terms of the statistics of goodness such as:

$$\begin{cases} AIC = -2L + 2k \\ BIC = -2L + k\ln(s) \end{cases} \tag{17}$$

AIC, BIC are the criteria of the information amount that will be modeled, where L is the value of the logarithmic likelihood function in formula (14), k, s are the number of parameters and the sample size respectively [11]. These criteria are considered as the measure of the superiority of statistical model fitting, which can balance between the complexity of the estimated model and the superiority of the model in terms of the fitting of the data. In general, we prioritize the model with the smaller AIC and BIC values.

Tab. 3 Comparison of the Goodness of Fitting Model

Model	-2L	AIC	BIC	Divergence α	φ	S
GB	333.54	335.54	338.54	0.52	-	135
ZIB	147.48	149.48	152.27	-	1/61	120
ZIGB	90.4	94.4	100.17	0.424	1.089/61	132

In Tab. 3, $S = \max\left(n + \frac{\alpha}{p_i} y_i\right)$, the data are represented the value that was obtained by subtracting 13 300 from each(-2L, AIC, BIC) data. In Tab. 3, in the generalized binomial distribution model (GB) and the zero-inflated binomial distribution model (ZIB), k is equal to one (k=1). While k=2 in the zero-inflated generalized binomial model (ZIGB). According to the past actual data of the park, the probability p_0=0.54 is indicated there is no excessive emissions. While p_1=0.2, p_2=0.16 and p_3=0.1 are indicated: once exceeded, twice exceeded and 3 times exceeded, respectively from Fig.1, and we use the probability values in formula (14) to compute the value of the logarithmic likelihood function L.

From Tab. 3, we can find that $AIC_{GB} > AIC_{ZIB} > AIC_{ZIGB}$, $BIC_{GB} > BIC_{ZIB} > BIC_{ZIGB}$. Moreover, by calculating the Statistic value LR of the likelihood ratio, we have $LR(1)=2(L_{ZIB} - L_{GB})=186.06$, $LR(2)=2(L_{ZIGB} - L_{ZIB})=57.08$, and $LR(1) > R(2) > Z_{\alpha}^2/2 = Z_{0.025}^2 = 3.841$. We can observe that the fitting effect in the ZIGB model is better than those in ZIB and GB models. We can also observe that ZIGB model has the best fitting effect on the set of data, while the fitting effect in AIC and BIC statistics are smaller than other models.

4 Conclusions

In the actual analysis of pollution emissions process, we found that there are only two results about excessive (or non-excessive) and the two events mutually exclusive. Also we found that the data structure is unsuitable for the characteristics of binomial distribution in case of the depth study process. In order to make up for the deficiency of the traditional methods, we have tried to meet the requirement of the data distribution by improving the binomial distribution and the zero-inflated model. Experimental results show that the proposed model has achieved good results. It is worth mention that the excessive emissions have not only been affected by the enterprise emission behavior, but also by weather, wind speed, season, pressure, holidays and other factors. Since we have only considered the single factor of the enterprise gas emissions. So, the forecast results and accuracy will be negatively affected. In order to avoid these drawbacks we should

https://dc-china-simulation.researchcommons.org/journal/vol32/iss11/19
DOI: 10.16182/j.issn1004731x.joss.19-FZ0510E

第 32 卷第 11 期 系统仿真学报 Vol. 32 No. 11
2020 年 11 月 Journal of System Simulation Nov., 2020

consider the impact of space factors as well as establishing appropriate air diffusion model. So, our future work will focus on improving accuracy of the proposed model.

References:

[1] Lambert D. Zero-inflated Poisson Regression with an Application to Defects in Manufacturing[J]. Technometrics (S0040-1706), 1992, 34(1): 1-14.

[2] Yip K C H, Yau K K W. On Modeling Claim Frequency Data in General Insurance with Extra Zeros[J]. Insurance: Mathematics and Economics (S0167-6687), 2005, 36(2): 153-163.

[3] Hall D B. Zero-inflated Poisson and Binomial Regression with Random Effects: a Case Study[J]. Biometrics (S0006-341X), 2000, 56(4): 1030-1039.

[4] Agarwal D K, Gelfand A E, Citron-Pousty S. Zero-inflated Model with Application to Spatial Count Data[J]. Environmental and Ecological Statistics (S1352-8505), 2002, 9(4): 341-355.

[5] Denuit M, Marechal X, Pitrebois S, et al. Actuarial Model-ling of Claim Counts: Risk Classification, Credibility and Bonus-malus Systems[J]. Journal of the American Statistical Association (S0162-1459), 2007, (486): 856.

[6] Winkelmann R. Econometric Analysis of Count Data(5th)[M]. Berlin: Spring, 2008: 173-199.

[7] Guo N G. An Application of Adjusted Zero-inflated Poisson Model to Zero Claims[J]. Statistics & Information Forum (S1007-3116), 2010, 7: 22-25.

[8] Meng S W, Yang L. Random-effect Zero-inflated Claims Frequency Regression Models[J]. Statistical Research (S1002-4565), 2015, 32(11): 97-102.

[9] Xu X, Yuan W, Meng S W. Generalization of Negative Binomial Regression Model and Its Application to Classification Ratemaking[J]. Journal of Applied Statistics and Management (S1002-1566), 2010, 29(4): 656-661.

[10] Li Y C, Sze N N, Wong S C. Spatial-temporal Analysis of Drink-driving Patterns in Hong Kong[J]. Accident Analysis and Prevention (S0001-4575), 2013, 59: 415-424.

[11] Akaike H. A New Look at The Statistical Model Identification[J]. IEEE Transactions on Automatic Control (S0018-9286), 1974, 19(6): 716-723.