Preprints, 10th IFAC International Symposium on
Advanced Control of Chemical Processes
Shenyang, Liaoning, China, July 25-27, 2018

Hurdle Modeling for Defect Data with Excess Zeros in Steel Manufacturing Process

Xinmin Zhang[a], **Manabu Kano**[a*] **Masahiro Tani**[b]
Junichi Mori[c] **Junji Ise**[c] **Kohhei Harada**[c]

[a] *Department of Systems Science, Kyoto University, Kyoto 606-8501, Japan*
[b] *Nippon Steel & Sumitomo Metal Corp., Kimitsu Works, Quality Management Div.,1 Kimitsu, Kimitsu, Chiba 299-1141, Japan*
[c] *Nippon Steel & Sumitomo Metal Corp., Kimitsu Works, Production & Technical Control Div.,1 Kimitsu, Kimitsu, Chiba 299-1141, Japan*
(e-mail: manabu@human.sys.i.kyoto-u.ac.jp).

Abstract: The modern steel industry aims to produce high-quality products with higher product yield, lower costs, and lower energy consumption to meet market demands. To accomplish these goals, it is necessary to reduce or eliminate product defects. However, the relationship of operating conditions to the defect formation is not fully understood. There is increasing interest in developing models to monitor the quality and predict the number of defects in real time. Modeling and analyzing the defect count data is a very challenging problem because the defect count data exhibit the unique characteristics of non-negative integers, overdispersion, high skewed distribution, and excess zeros. To explicitly account for these unique characteristics, the present work develops an on-line quality monitoring and prediction system based on the hurdle regression model. The basic idea of the hurdle model is that a binomial model governs the binary outcome of the dependent variable being zero or positive. If the dependent variable takes a positive value, "hurdle is crossed", and the conditional distribution of the positives can be modeled by a zero-truncated Poisson or negative binomial (NB) model. Compared to Poisson and NB models, the hurdle model is not only suitable for modeling discrete and non-negative integer data, but also sufficient for handling both overdispersion and excess zeros data. The effectiveness of the hurdle model was verified through its application to the real defect data of a steelmaking plant. The results have demonstrated that the hurdle NB model is superior to the Poisson, NB, hurdle Poisson, and PLS models in the prediction performance.

Keywords: Defect, casting-rolling process, count data, quality improvement, Hurdle model

1. INTRODUCTION

To meet the high competitive market demands, the steel industry aims to improve product quality and productivity with low production costs and sustainable production environment. However, products and production processes are always subject to variations. Equipment malfunctions, process perturbations, and inappropriate operation will lead to a variety of defective products. Defects not only increase the production cost because of reworking or reproducing, but also waste materials, energy consumption, and lead time. Therefore, it is important to develop a model to monitor the quality and predict the number of defects in real time.

Virtual sensing or soft-sensor is a key technology for predicting product quality or other key variables in real time, and has been successfully applied to many industrial processes (Kano and Nakagawa (2008); Kadlec et al. (2009); Wang et al. (2010); Zhang et al. (2015, 2017)). The basic idea of soft-sensor is to construct an inference model that relates product quality (response) to process operating conditions (predictors). Multiple linear regression (MLR) and partial least squares (PLS) regression are the most popular approaches (Geladi and Kowalski (1986); Kano and Ogawa (2010); Yin et al. (2015)). However, they cannot work well in modeling defect data due to their basic assumptions of normality and homoscedasticity, because defect count data often violate these assumptions and show overdispersion (or heteroskedasticity) and high skewed distribution. Furthermore, MLR or PLS models might result in the prediction of negative counts although the defect count data are characterized by non-negative integers.

The Poisson regression model assumes that the response variable or error structure follows a Poisson distribution, which is a discrete distribution expressing the probability of only nonnegative integers, and it is a basic model for modeling and analyzing non-negative integers (Cameron and Trivedi (2013); Fox (2015)). The Poisson regression belongs to the family of generalized linear regression, where the (canonical) link function is the natural log. The Poisson regression is suitable for modeling count data, but it assumes that the mean and variance are equal. This assumption may be restrictive for its application to the overdispersed data, such as the defect data investigated in this study. In the overdispersion circumstance, the Poisson regression model tends to underestimate the dispersion of the observed count data. A remedy to this overdispersion problem is the use of a negative binomial (NB) model, which is derived as a gamma mixture of Poisson random variables (Cameron and Trivedi (2013)). Although the NB model is more suitable for modelling overdispersed data, it is not appropriate for modeling count data when an excess of zeros exists. Since the observed defect data are characterized by a large number of zeros, the NB model will not work well.

To explicitly account for the unique characteristics of non-negative integers, overdispersion, high skewed distribution, and excess zeros in the observed defect count data, the present work develops an on-line quality monitoring system based on the hurdle regression model. The hurdle model is a two-component mixture model that combines a binomial model governing the binary outcome of the dependent variable being zero or positive and a zero-truncated Poisson or NB model for strictly positives. In contrast with the single Poisson or NB model, the hurdle model is not only suitable for modeling discrete and non-negative integer data, but also sufficient for handling both over-dispersion and excess zeros data. This research is motivated to verify the suitability of potential applications of hurdle modeling technique to the defect count data in a steel manufacturing process.

The rest of this paper is organized as follows. Section 2 gives a brief introduction of the Poisson regression and negative binomial regression. Then the hurdle modeling technique for defect data with excess zeros is presented in section 3. In section 4, the effectiveness of hurdle modeling is verified through its application to the real defect data of a steelmaking plant, and its application results are compared with those of Poisson, NB, and PLS models. The conclusions of this work are presented in section 5.

2. PRELIMINARIES

2.1 Poisson regression

Poisson regression is a basic tool for modeling and analyzing count data, which follows the Poisson probability distribution. Given the input-output data pairs $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$, where $\boldsymbol{x}_i^T \in \Re^K$ denotes the i-th observation of the predictor variables and y_i is the corresponding output. The Poisson probability mass function is expressed as

$$f(y_i|\boldsymbol{x}_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}, (y_i = 0, 1, 2, \cdots) \tag{1}$$

where $y_i!$ is the factorial of y_i, e is the base of the natural logarithm, and μ_i is the average number of events (also called rate parameter or shape parameter). The conditional mean and conditional variance are given by

$$\mathrm{E}(y_i|\boldsymbol{x}_i) = \mathrm{Var}(y_i|\boldsymbol{x}_i) = \mu_i. \tag{2}$$

In Poisson regression, the conditional mean of y_i is parameterized as an exponential function of the predictor variables $\boldsymbol{x}_i$:

$$\mu_i = \exp(\boldsymbol{x}_i^T\boldsymbol{\beta}) \tag{3}$$

where $\boldsymbol{\beta}$ is a vector of unknown parameters (regression coefficients). This exponential parameterization ensures the non-negativity of μ_i. Furthermore, equation (3) also implies that the Poisson model is a multiplicative regression model.

Traditionally, the maximum likelihood technique is used to estimate the parameters of Poisson regression (Cameron and Trivedi (2013); Fox (2015)). The log-likelihood function is given by

$$\begin{aligned}\ell(\boldsymbol{\beta}) &= \sum_{i=1}^{N}\left[y_i \ln(\mu_i) - \mu_i - \ln(y_i!)\right] \\ &= \sum_{i=1}^{n}\left[y_i\boldsymbol{x}_i^T\boldsymbol{\beta} - \exp(\boldsymbol{x}_i^T\boldsymbol{\beta}) - \ln(y_i!)\right].\end{aligned} \tag{4}$$

A restrictive property of the Poisson regression is the equality of the mean and variance. When overdispersion becomes an issue, the estimates of Poisson model may be inefficient. The negative binomial regression model offers a remedy to this problem.

2.2 Negative binomial regression

The negative-binomial (NB) model can be viewed as a generalization of Poisson model that allows for overdispersion (Hilbe (2011); Cameron and Trivedi (2013)). In NB regression, the distribution of y_i given $\boldsymbol{x}_i$ is derived as a gamma mixture of Poisson distributions in terms of its mean μ_i:

$$f(y_i|\boldsymbol{x}_i) = \frac{\Gamma(y_i+\theta)}{y_i!\Gamma(\theta)}\left(\frac{\theta}{\theta+\mu_i}\right)^{\theta}\left(\frac{\mu_i}{\theta+\mu_i}\right)^{y_i} \tag{5}$$

where $\theta > 0$ is the shape parameter and $\Gamma(\cdot)$ is the gamma function. With $\alpha = 1/\theta$ ($\alpha > 0$), the NB distribution can then be rewritten as

$$f(y_i|\boldsymbol{x}_i) = \frac{\Gamma(y_i+1/\alpha)}{y_i!\Gamma(1/\alpha)}\left(\frac{1}{1+\alpha\mu_i}\right)^{1/\alpha}\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)^{y_i} \tag{6}$$

where α is referred to as the dispersion parameter in generalized linear models (Fox (2015)). The conditional mean and conditional variance of the NB distribution are

$$\mathrm{E}(y_i|\boldsymbol{x}_i) = \mu_i = \exp(\boldsymbol{x}_i^T\boldsymbol{\beta}) \tag{7}$$

$$\mathrm{Var}(y_i|\boldsymbol{x}_i) = \mu_i + \alpha\mu_i^2 = \mu_i(1+\alpha\mu_i). \tag{8}$$

Note that the conditional variance of the NB distribution is always larger than the conditional mean. Thus, the NB distribution is suitable for capturing the overdispersion in data. Furthermore, as $\alpha \to 0$, $\mathrm{Var}(y_i|\boldsymbol{x}_i) \to \mu_i$ and the NB distribution converges to the Poisson distribution. For NB regression, the parameters of α and $\boldsymbol{\beta}$ are calculated using maximum likelihood estimation (Cameron and Trivedi (2013); Fox (2015)). The log-likelihood function is given by

$$\begin{aligned}\ell(\alpha,\beta) = \sum_{i=1}^{N}\Bigg(& y_i \ln\alpha + y_i\boldsymbol{x}_i^T\boldsymbol{\beta} \\ &- \Big(y_i + \frac{1}{\alpha}\Big)\ln\left(1+\alpha\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})\right) + \ln\Gamma\Big(y_i + \frac{1}{\alpha}\Big) \\ &- \ln(y_i!) - \ln\Gamma\Big(\frac{1}{\alpha}\Big)\Bigg)\end{aligned} \tag{9}$$

3. HURDLE MODELING FOR DEFECT DATA WITH EXCESS ZEROS

As discussed above, the Poisson regression is designed under the equidispersion assumption. In practice, the observed defect data exhibit overdispersion. Thus, the estimates may be incorrect when the Poisson model is used. NB regression relaxes this constraint of Poisson model, and is more suitable for overdispersed data. However, both Poisson and NB models cannot fully account for the excess zeros, which are observed in the investigated defect data. In comparison, hurdle model can deal with both overdispersion and excess zeros problems effectively.

The basic idea of the hurdle model is that a binomial model governs the binary outcome of the dependent variable being zero or positive (Hu et al. (2011); Cameron and Trivedi (2013)). If the dependent variable takes a positive value, "hurdle is crossed", and the conditional distribution of the positives can be modelled by a zero-truncated Poisson or NB model. Let $f_1(\cdot)$ denote the probability distribution function for the hurdle part and $f_2(\cdot)$ be the probability distribution function for the positives part. The probability distribution of the hurdle model is given by

$$f(y_i|\boldsymbol{x}_i, \boldsymbol{z}_i) = \begin{cases} f_1(0|\boldsymbol{z}_i, \boldsymbol{\beta}_1) & \text{if } y_i = 0, \\ \Phi f_2(y_i|\boldsymbol{x}_i, \boldsymbol{\beta}_2) & \text{if } y_i > 0 \end{cases} \tag{10}$$

with

$$\Phi = \frac{1 - f_1(0|\boldsymbol{z}_i, \boldsymbol{\beta}_1)}{1 - f_2(0|\boldsymbol{x}_i, \boldsymbol{\beta}_2)} \tag{11}$$

where the numerator of Φ represents the probability of crossing the hurdle, and the denominator is the summation of $f_2(\cdot)$ on the support of the conditional density (i.e., the truncation normalization). The notations of $\boldsymbol{z}_i$ and $\boldsymbol{x}_i$ indicates that the regressors for the zero hurdle part and the positive count part could be different. The conditional mean and conditional variance are given by

$$E(y_i|\boldsymbol{x}_i) = \sum_{i\in\Omega_1} y_i f_2(y_i|\boldsymbol{x}_i)\Phi \tag{12}$$

$$Var(y_i|\boldsymbol{x}_i) = \sum_{i\in\Omega_1} y_i^2 f_2(y_i|\boldsymbol{x}_i)\Phi - \left[\Phi \sum_{i\in\Omega_1} y_i f_2(y_i|\boldsymbol{x}_i)\right]^2 \tag{13}$$

where $\Omega_1 = \{i|y_i \neq 0\}$, and its complementary set is $\Omega_0 = \{i|y_i = 0\}$, and $\Omega_0 \cup \Omega_1 = \{1, 2, \cdots, N\}$. The variance-mean ratio (VMR) is

$$VMR = \frac{Var(y_i|\boldsymbol{x}_i)}{E(y_i|\boldsymbol{x}_i)} \tag{14}$$

For example, if $f_2(\cdot)$ is a Poisson distribution and $\Phi = 1$, the VMR is equal to 1; this is the case of equidispersion of the Poisson distribution. In contrast, if $\Phi \neq 1$, (10) is the hurdle Poisson model, in which $0 < \Phi < 1$ corresponds to the overdispersion case, and $\Phi > 1$ corresponds to the underdispersion case. The hurdle model provides a more flexible framework for modeling and analyzing count data.

Estimation of the hurdle model is realized by using the log-likelihood parameterization of the hurdle probability distribution, with the aim of finding parameter values that make the data most likely (Hu et al. (2011); Cameron and Trivedi (2013)). The likelihood function is given by

$$L = \prod_{i\in\Omega_0} \{f_1(0|\boldsymbol{z}_i, \boldsymbol{\beta}_1)\} \prod_{i\in\Omega_1} \frac{1 - f_1(0|\boldsymbol{z}_i, \boldsymbol{\beta}_1)}{1 - f_2(0|\boldsymbol{x}_i, \boldsymbol{\beta}_2)} f_2(y_i|\boldsymbol{x}_i, \boldsymbol{\beta}_2) \tag{15}$$

Taking the natural logarithm of both sides of (15), we can obtain the log likelihood function:

$$\begin{aligned} \ell(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2; \boldsymbol{z}_i, \boldsymbol{x}_i) &= \sum_{i\in\Omega_0} \ln\{f_1(0|\boldsymbol{z}_i, \boldsymbol{\beta}_1)\} \\ &+ \sum_{i\in\Omega_1} \ln\{1 - f_1(0|\boldsymbol{z}_i, \boldsymbol{\beta}_1)\} + \sum_{i\in\Omega_1} \Big[\ln\{f_2(y_i|\boldsymbol{x}_i, \boldsymbol{\beta}_2)\} \\ &- \ln\{1 - f_2(0|\boldsymbol{x}_i, \boldsymbol{\beta}_2)\}\Big] \\ &= \ell_1(\boldsymbol{\beta}_1) + \ell_2(\boldsymbol{\beta}_2). \end{aligned} \tag{16}$$

This log likelihood function is separable with respect to the parameters $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$. In other words, the log likelihood function describes the sum of a log likelihood for the binary outcome model ($\ell_1(\boldsymbol{\beta}_1)$), and a log likelihood for a zero-truncated model ($\ell_2(\boldsymbol{\beta}_2)$). This separability indicates that the log likelihood can always be maximized under the condition of without loss of information through maximizing the two components separately (Hu et al. (2011); Fox (2015)). Clearly, the hurdle model is specified by the probability distributions $f_1(\cdot)$ and $f_2(\cdot)$. Generally, $f_1(\cdot)$ is specified by a binomial distribution with a logit link function. For the hurdle Poisson model, $f_2(\cdot)$ is specified by a zero-truncated Poisson distribution with a log link function. For the hurdle NB model, $f_2(\cdot)$ is specified by a zero-truncated NB distribution with a log link function.

4. CASE STUDY

In this section, the effectiveness of the hurdle modeling technique is validated through its application to the real defect data of a steelmaking plant. The application results are compared with those of Poisson, NB, and PLS models.

The defect dataset was recorded from a continuous cast-rolling process of the steelmaking plant. It consists of 3600 samples and 71 input variables. The objectives are to predict the number of defects, identify the factors that affect defect occurrence, and provide effective suggestions and countermeasures to reduce the number of defects. Thus, the output variable is the number of defects. To construct the predictive model, the entire dataset was partitioned into two subsets: a training set with 3000 samples and a testing set with 600 samples. Figure 1 shows the frequency distribution of the defects, which is characterized by a large number of zeros and high skewness. The percentage of no defect occurring is around 70%. Furthermore, its variance is much larger than its mean; that is, overdispersion exists in the data.

The widely used Akaike's information criterion (AIC) (Cameron and Trivedi (2013)) and Bayesian information criterion (BIC) (Cameron and Trivedi (2013)) are firstly used to evaluate the relative performances of Poisson, NB, hurdle Poisson, and hurdle NB models in modeling and analyzing defect count data. Both AIC and BIC are defined on the basis of the maximized log-likelihood function:

$$AIC = 2\log\ell + 2k \tag{17}$$

$$BIC = 2\log\ell + \log(n)k \tag{18}$$

where ℓ denotes the log-likelihood function, k denotes the number of model parameters, and n is the sample size. A smaller value of AIC or BIC means that the model is better. As shown in table 1, the hurdle NB model seems to be the optimal model with the smallest AIC and BIC values among the four models.

Although AIC and BIC are often used to evaluate the goodness of fit of count data regression models, they are based only on the maximized log-likelihood. In this work, we used the hanging rootogram to evaluate the goodness-of-fit of Poisson, NB, hurdle Poisson, and hurdle NB models for defect data. The rootogram is a visualization technique which evaluates the goodness-of-fit of the black-box model in a graphical way (Kleiber and Zeileis (2016)). Compared to AIC and BIC, the rootogram highlights the discrepancies between observed and expected frequencies. This property is particularly useful in diagnosing and exploring issues such as overdispersion and excess zeros when modeling count data. Figure 2 provides the hanging rootograms of four models for fitting the defect data. The red curved line represents the theoretical fit corresponding to each model. The hanging bar (or rootogram bar) on the red line represents the deviation between observed and expected counts on a square-root scale. The square-root transformation is employed to avoid smaller frequencies being obscured and overwhelmed by larger frequencies. The line at 0 is called the horizontal reference line, which allows us to easily visualize where the model is over- or underfitting. A bar hanging below or over 0 indicates underfitting or overfitting in each count, respectively. As shown in Fig. 2, the rootogram bars of the Poisson model form a 'wave-like' pattern around horizontal reference line, and the small counts $1, \cdots, 15$ are severely overfitted while both zeros and most counts from 16 to the end are underfitted. This indicates that a large amount of overdispersion in the data, but it is not captured by the fitted Poisson model. Furthermore, the clear lack of fit for 0 gives an additional indication of excess zeros. Compared to the Poisson model, the rootogram bars of the NB model looks better around the horizontal reference line, which indicates that the NB model coped with the overdispersion better than the Poisson model, as shown in Fig. 2. However, the NB model is still underfitting the number of zeros and overfitting the small counts. The rootogram of hurdle Poisson model shows a good fit for the number of zeros, as shown in Fig. 2. However, the wave-like pattern in the positive counts reflects that there is still massive overdispersion that is not captured by the hurdle Poisson model. The rootogram of hurdle NB model indicates that the hurdle NB model fits the defect data better than the hurdle Poisson model, as shown in Fig. 2. The deviations between observed and predicted frequencies are smaller for most of the number of defects, as compared to the other models. In summary, the Poisson model performed the worst, and the hurdle NB model performed the best.

To test the prediction performances of Poisson, NB, hurdle Poisson, and hurdle NB models for the testing data, the root mean squared error (RMSE) and correlation coefficient criteria were adopted. In addition, the PLS model was also built as a reference model. The number of latent variables used in PLS was 40, which was determined by cross-validation. Table 2 summarizes the prediction performances of five models in terms of RMSEP (RMSE of Prediction) and R^2. The PLS model provided low prediction accuracy with large RMSEP and a small correlation coefficient. Compared to PLS, both the Poisson and NB models presented better performance. In contrast with the single Poisson or NB model, the hurdle Poisson and hurdle NB models performed significantly better with smaller RMSEP values and higher correlation coefficients. As a result, the hurdle NB model achieved the best prediction performance among the five methods. This is mainly because the hurdle NB model takes the advantage of modeling the zero counts and the overdispersed positive counts. Fig. 3 shows the detailed comparisons of the PLS model and the hurdle NB model, where the red line denotes the horizontal zero line. There are a large amount of values that below the red zero line, indicating that the PLS model provided massive negative predictions, although the number of defects should be nonnegative. Thus, the PLS model should not be used for modeling and analyzing the defect count data. In contrast, the Poisson, NB, hurdle Poisson, and hurdle NB models can ensure the predictions are nonnegative. The results have demonstrated that the hurdle NB model is preferred for modeling and analyzing the defect data.

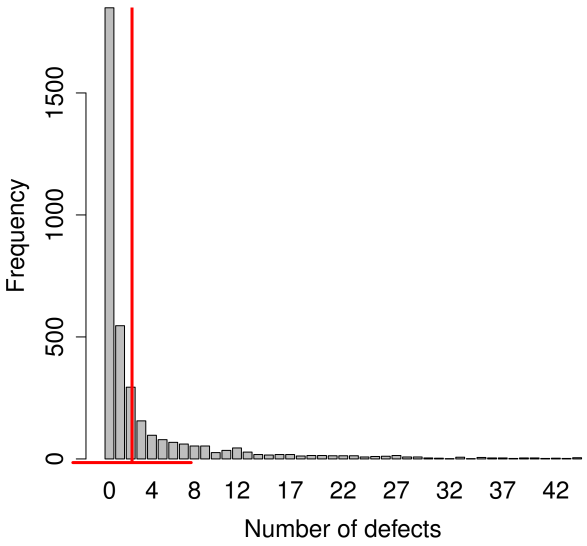

Fig. 1. Distribution of number of defects. The vertical red line represents the mean and horizontal line represents a range of one standard deviation (mean ±1 standard deviation).

Table 1. Goodness of fit tests by AIC and BIC criteria.

Methods	AIC	BIC
Poisson	19360	19642
NB	11238	11340
Hurdle Poisson	15521	15701
Hurdle NB	10959	11158

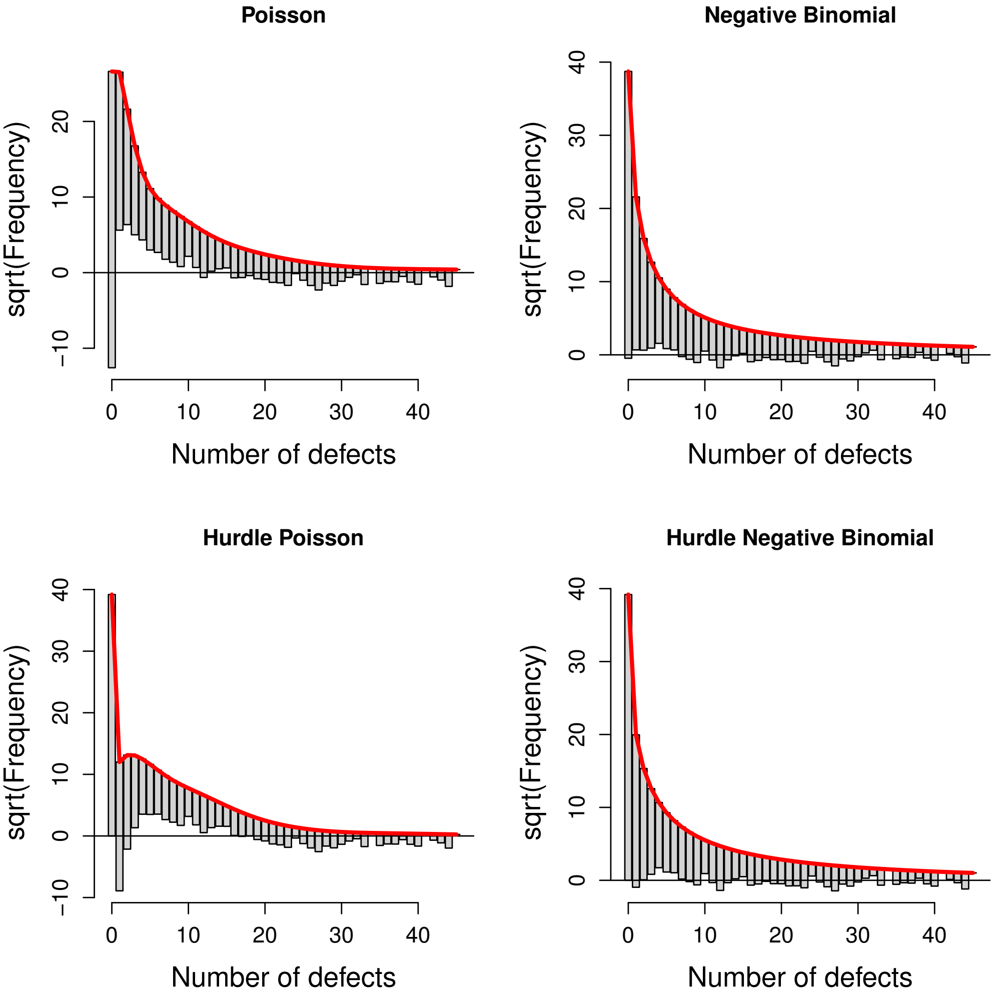

Fig. 2. Rootograms of four methods fit to the defect data.

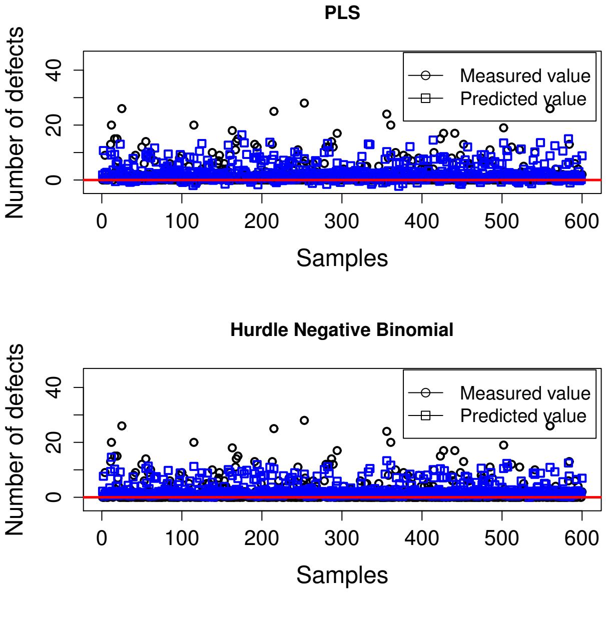

Fig. 3. Prediction results of PLS and hurdle NB model for the defect data.

5. CONCLUSION

In this research, we focused on the modeling of defect data in the steel manufacturing process with the purpose of predicting the number of defects. However, modeling and analysis the defect data is a challenging problem because the defect data exhibit the unique characteristics of non-negative integers, overdispersion, high skewed distribution, and excess zeros. To explicitly account for these unique characteristics, a quality monitoring system based on hurdle modelling was proposed in this work. The hurdle model is a two-component mixture model that combines a binomial model governing the binary outcome of the dependent variable being zero or positive and a zero-truncated model for strictly positives. In contrast with the Poisson and negative binomial (NB) models, hurdle

Table 2. Prediction results of five models for the defect data.

Methods	RMSEP	R^2
PLS	4.2533	0.3848
Poisson	4.1723	0.3959
NB	4.1615	0.3775
Hurdle Poisson	3.8379	0.4766
Hurdle NB	3.7624	0.4964

model is not only suitable for modeling discrete and non-negative integer data, but also sufficient for handling both overdispersion and excess zeros data. The effectiveness of the hurdle model was verified through its application to the real defect data of a steelmaking plant. The results have demonstrated that the hurdle NB model is superior to the Poisson, NB, hurdle Poisson, and PLS models in the prediction performance.

REFERENCES

Cameron, A.C. and Trivedi, P.K. (2013). *Regression analysis of count data*, volume 53. Cambridge university press.

Fox, J. (2015). *Applied regression analysis and generalized linear models*. Sage Publications.

Geladi, P. and Kowalski, B.R. (1986). Partial least-squares regression: a tutorial. *Analytica chimica acta*, 185, 1–17.

Hilbe, J.M. (2011). *Negative binomial regression*. Cambridge University Press.

Hu, M.C., Pavlicova, M., and Nunes, E.V. (2011). Zero-inflated and hurdle models of count data with extra zeros: examples from an hiv-risk reduction intervention trial. *The American Journal of Drug and Alcohol Abuse*, 37(5), 367–375.

Kadlec, P., Gabrys, B., and Strandt, S. (2009). Data-driven soft sensors in the process industry. *Computers & Chemical Engineering*, 33(4), 795–814.

Kano, M. and Nakagawa, Y. (2008). Data-based process monitoring, process control, and quality improvement: recent developments and applications in steel industry. *Computers & Chemical Engineering*, 32(1), 12–24.

Kano, M. and Ogawa, M. (2010). The state of the art in chemical process control in japan: good practice and questionnaire survey. *Journal of Process Control*, 20(9), 969–982.

Kleiber, C. and Zeileis, A. (2016). Visualizing count data regressions using rootograms. *The American Statistician*, 70(3), 296–303.

Wang, D., Liu, J., and Srinivasan, R. (2010). Data-driven soft sensor approach for quality prediction in a refining process. *IEEE Transactions on Industrial Informatics*, 6(1), 11–17.

Yin, S., Li, X., Gao, H., and Kaynak, O. (2015). Data-based techniques focused on modern industry: an overview. *IEEE Transactions on Industrial Electronics*, 62(1), 657–667.

Zhang, X., Kano, M., and Li, Y. (2017). Locally weighted kernel partial least squares regression based on sparse nonlinear features for virtual sensing of nonlinear time-varying processes. *Computers & Chemical Engineering*, 104, 164–171.

Zhang, X., Li, Y., and Kano, M. (2015). Quality prediction in complex batch processes with just-in-time learning model based on non-gaussian dissimilarity measure. *Industrial & Engineering Chemistry Research*, 54(31), 7694–7705.