DE GRUYTER OPEN

Folia Oeconomica Stetinensia
DOI: 10.1515/foli-2017-0001

WYDZIAŁ NAUK EKONOMICZNYCH I ZARZĄDZANIA
UNIWERSYTET SZCZECIŃSKI

FORECASTING RANDOMLY DISTRIBUTED ZERO-INFLATED TIME SERIES

Mariusz Doszyń, Ph.D., Associate Prof.

University of Szczecin
Faculty of Economics and Management
Institute of Econometrics and Statistics
Mickiewicza 64, 71-101 Szczecin, Poland
e-mail: mariusz.doszyn@usz.edu.pl

Received 25 November 2016, Accepted 13 April 2017

Abstract

The main aim of the article is to propose a forecasting procedure that could be useful in the case of randomly distributed zero–inflated time series. Many economic time series are randomly distributed, so it is not possible to estimate any kind of statistical or econometric models such as, for example, count data regression models. This is why in the article a new forecasting procedure based on the stochastic simulation is proposed. Before it is used, the randomness of the times series should be considered. The hypothesis stating the randomness of the times series with regard to both sales sequences or sales levels is verified. Moreover, in the article the *ex post* forecast error that could be computed also for a zero-inflated time series is proposed. All of the above mentioned parts were invented by the author. In the empirical example, the described procedure was applied to forecast the sales of products in a company located in the vicinity of Szczecin (Poland), so real data were analysed. The accuracy of the forecast was verified as well.

Keywords: forecasting zero-inflated time series, count data models, sales forecasting system, stochastic simulation, Bernoulli processes

JEL classification: C12, C15, C25, C53

Introduction

The main aim of the article is to propose a procedure that could be useful in forecasting a randomly distributed zero-inflated time series. The procedure is based on a stochastic simulation algorithm, in which the Bernoulli process is applied. In the literature it is sometimes stated that not all processes could be determined (Taleb, 2001). Sometimes data generating processes are just random. In such cases not econometric models but different kind of statistical methods might be applied (Asmussen, Glynn, 2007). Therefore, before using statistical models, it is a good idea to verify the hypothesis stating the randomness of the analysed process.

1. Literature review

In many economic contexts events appear rarely and are often very volatile. A dependent variable is then discrete (for count data) with a distribution that places probability mass only at nonnegative integer values. In such cases, special models for count data are recommended. These are nonlinear models with specific properties related to discreteness. Among them, the following regression models are most frequently used (Cameron, Trivedi, 1998; Cameron, Trivedi, 2005; Hilbe, 2011; Hilbe, 2014; Winkelmann, 2008; Yang, 2012; Biswas, Song, 2009):

1. Poisson (or negative binomial) models.
2. Zero-inflated Poisson (or negative binomial) models.
3. Hurdle Poisson (or negative binomial) models.
4. Zero-inflated Poisson (or negative binomial) time series models.

There are also other types of regression count data models, such as panel or multivariate models. Moreover, there are many kinds of special time series models for count data (Cameron, Trivedi, 1998; Biswas, Song, 2009; Yang, 2012).

Poisson regression models have many limitations. The probability of a zero count is usually underestimated. This is called the excess zero problem, as there are more zeros than the Poisson model predicts (Cameron, Trivedi, 2001). Contrary to the Poisson (or negative binomial) regression models, in the case of zero–inflated models there is a higher probability of a zero count.

Another obstacle is that the Poisson distribution implies that mean is equal to variance (equidispersion), but in many situations variance is much higher (overdispersion). The reason for overdispersion might be the unobserved heterogeneity. In such cases counts are treated as a realization of a Poisson process, where the rate parameter is treated itself as a random variable. This mixture approach leads to negative binomial models. The next reason for overdispersion

might be the fact that the processes determining events (zeros and non–zeros) might be different. The assumption that zeros and the positives come from the same data generating process is then relaxed. This leads to hurdle models (two-part models).

In the literature there are many types of different time series models for count data (Yang, 2012; Cameron, Trivedi, 1998; Biswas, Song, 2009). Beyond the Poisson and negative binomial models, also DARMA models (Discrete ARMA Models), Markov models, serially correlated error models, state–space models, hidden Markov models, etc. could be applied.

It is worth emphasizing that generally, in the case of zero–inflated times series, two processes are modelled. At first, it is determined whether a given event (for example a sale) will occur. This could be treated as a realization of a binary Bernoulli process and such models as logit (or probit) might be useful. In the second stage, the conditional mean of a dependent variable is modelled, usually by means of the Poisson (or negative binomial) regression.

The models for count data are not always applicable. Before they are used, the hypothesis stating the randomness of a dependent variable should be verified. If we could not reject such a hypothesis, we should assume that the data generating process is random and the application of regression models for count data is pointless. In such cases, procedures based on stochastic simulation might be helpful. This procedure is a new approach that could be used to forecast a zero-inflated time series.

The proposed forecasting procedure was invented also because of the many difficulties with appropriately specifying the count data models (or zero-inflated models). The basic reason for this is that it is not easy to find good explanatory variables. In the case of time series models it is hard to find any plausible regularities, even for those series that have been classified as non-random (on the basis of the performed tests).

Generally, the estimated count data models turned out unsatisfactory. Additionally, estimation and forecasting by means of the count data models are very time consuming, because forecasts have to be computed and checked for about eighteen thousand products in each week. This makes the probability of making errors higher. Moreover, the time series for new offers are short, so it is often impossible to estimate a good model because of the low number of observations, which causes high estimation errors, etc.

It is also to be emphasized that in the zero-inflated models zeros have to be "structural". For example, in searching for fertility we could assume that infertile women have zero children. In the case of a sale this would mean that there is a zero sale because of the lack of a given product in the warehouse, which is hardly ever true. Because of all these problems, a new forecasting algorithm based on a stochastic simulation has been constructed.

In the article sales time series are analysed, but also different phenomena generated by similar data generating processes could be forecasted by means of the proposed procedure.

2. Theoretical model

2.1. Randomness of a time series

Two kinds of hypothesis to verify the randomness of a times series were performed:

1. At first, hypothesis stating the randomness of sales sequences was verified. If this hypothesis is rejected, there are not any regularities concerning the time intervals between sales (regardless to sale levels).
2. Secondly, the quartile test for the randomness of sales levels was performed.

Two kinds of hypotheses were applied because the data generating processes generally consist of two stages. In the first stage, it is settled whether the sale of the product in a given week will appear. In the second stage, the sales level is determined (if the result of the previous stage is positive).

The non-parametric tests for randomness, based on the number of series, were used (Domański, 1990). Many other non-parametric tests were considered, such as tests based on expected and observed lengths of series, different tests based on the lengths of series, tests based on the number of signs or Geary's test. The tests based on the number of series were used because of their high power (according to the second type error) (Domański, 1990).

In the performed tests, series were defined as a string of identical symbols. In the case of the hypothesis stating randomness of sales sequences (the first case), the series were constructed with regard to the occurrence of sales. So, the first symbol ("A") was assigned to weeks with positive sales and the next symbol ("B") to weeks with no sales. For all the products the data from the last 30 weeks were considered. The result of this test should give an insight into whether time intervals between the sales of products have been determined, at least to a certain extent.

In the quartile test the first symbol ("A") was assigned if the sales level was below the first quartile or above the third quartile. If the sales level was between these quartiles, the next symbol ("B") was introduced.[1] This test is helpful in determining tendencies in sales levels. For example, if there is a linear trend (both positive or negative), the number of series will be too low and a hypothesis stating randomness will be rejected. On the other hand, if there are

[1] The first and third quartiles included.

periodical fluctuations with proper amplitude, the number of obtained series could be too high and randomness will be rejected as well.

The following test statistic, based on the normal distribution, was used (Domański, 1990, p. 39):

$$U = \frac{K - \left(\frac{2n_1 n_2}{n} + 1 \right)}{\sqrt{\frac{2n_1 n_2 \left(2n_1 n_2 - n \right)}{n^2 \left(n - 1 \right)}}} \tag{1}$$

where:

K – number of series,

n – number of weeks,

n_1 – number of weeks marked with the first symbol (“A”),

n_2 – number of weeks marked with the second symbol (“B”).

The statistic can be used when n_1 and n_2 are not too high (Domański, 1990). It is also applicable if the probability of an event is different than 0.5, which was usually the case in the considered company.

2.2. Forecasting procedure

In the proposed procedure forecasts are generated in the following two stages.

In the first stage, based on the stochastic simulation, it is fixed whether the sale of the i-th product in the week T will appear. A random number w_{iT} from the uniform distribution $U\langle 0, 1\rangle$ is generated. The value w_{iT} is a realization of the random variable (W) that is uniformly distributed $W \sim U\langle 0, 1\rangle$. Then it is checked whether the relative sales frequency of the i-th product (c_i) is higher than the generated random number: $c_i > w_{iT}$. If it is true, it is assumed that the sale of the i-th product in the week T will appear. The idea is that the probability of sale is a realization of the binomial Bernoulli process $B(n, p)$ with the probability p equal to the relative sales frequency: $P(c_i > W) = \hat{p}_i = c_i$. The relative sales frequency (c_i) is then an estimator of the probability of sale $\left(\hat{p}_i \right)$ of a given product. For instance, if the relative sales frequency of the i-th product is $c_i = 0.2$, the probability of sale in a given week, for the uniformly distributed variable $W \sim U\langle 0, 1\rangle$, is equal: $P(0.2 > W) = 0.2$.

If in the previous stage it has been established that a sale will appear, the sales level is generated. Generally, forecasts are obtained consistently with the principle of unbiased

prediction, such as the expected value of sale, but only for weeks with a positive sale. Therefore, the expected value is calculated only be means of values for weeks in which the sale appeared.

The accuracy of forecasts calculated for many products should be analysed systematically, with respect to the distribution of a chosen *ex post* forecasts error. This error should be computable for all products and also should include all possible situations. Moreover, this has to be a relative measure if we want to have comparable results for all products. To compute typical relative *ex post* errors, such as *MPE*, *MAPE* or *Theil* coefficients, there should not be zeros in the denominator, which is not true in the case of the zero-inflated time series data. This is why a new *ex post* forecasts error (D), useful also in the case of the zero-inflated time series, is proposed. The proposed error is calculated for one week. The error could be calculated as well for more weeks and then averaged.

The proposed *ex post* forecast error is calculated for two cases:

$$D = \frac{y_{Tp} - y_T}{max\left\{y_{Tp},\ y_T\right\}} \qquad \text{if} \quad y_{Tp} \neq y_T \tag{2}$$

$$D = 0 \qquad \text{if} \quad y_{Tp} = y_T \tag{3}$$

where

y_{Tp} – *ex post* forecasts in week T,

y_T – sales in week T.

This error has the following properties:

1. If the sale is equal to the forecast ($y_{Tp} = y_T$), the error $D = 0$. This is the formula (3), that is also applicable when a sale and a forecast are equal to zero.
2. If the forecast and the sale are different ($y_{Tp} \neq y_T$), the error is calculated as a relation of the difference between the forecast and the sale (nominator) and the higher one of these two values (denominator) – formula (2). Taking a higher value in the denominator makes this error (D) always possible to calculate.
3. If $y_{Tp} < y_T$, the error is negative $D < 0$ and the forecast is underestimated. In such situations there is y_T in the denominator (the higher value).
4. If the forecast is equal to zero ($y_{Tp} = 0$) and the sale is positive ($y_T > 0$), then $D = -1$.
5. If $y_{Tp} > y_T$, the forecast is overestimated and $D > 0$.
6. If the sale is zero ($y_T = 0$) and the forecast is positive ($y_{Tp} > 0$), then $D = 1$.
7. The error is normalised: $D \in \langle -1, 1 \rangle$.

3. Empirical results

The sales of products in the last 30 weeks were considered (n = 30). Hypotheses were verified for as many as 16,309 products. For the remaining 1,690 products, the number of observations was not big enough (n < 30) to use the test statistic (1). The main tests results are presented in Table 1.

Table 1. General results of tests to verify the hypothesis stating the randomness of sales sequences and randomness of sales levels (quartile test) – significance level α = 0.01

Specifications	Sales sequence randomness	Quartile test
H_0 (randomness)	487	409
Rejection of H_0	15,822	15,900
Share of non-random series	0.030	0.025

Source: own work.

The main conclusion is that the very majority of the sales time series were randomly distributed. In the test in which the randomness of sales sequences was verified, the fraction of non-random series was equal to 3%. Only in the case of 487 products did the sales sequences appear to be non-random. In the quartile test the share of the non-random series was even lower and equalled 2.5%. As many as 409 products were classified as non-random. These shares are quite close to the chosen significance level (α = 0.01). In that kind of situation, the methods based on stochastic simulation could be useful in the process of forecasting.

The samples of a non-random times series, according to the two performed kinds of tests are shown below (Figures 1–2).

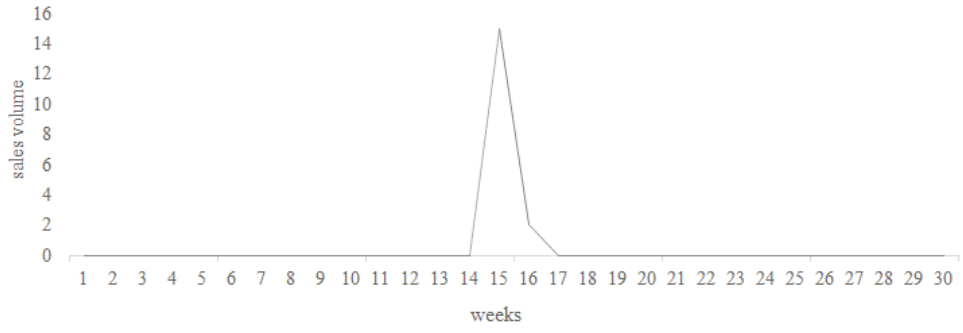

Figure 1. Non-random time series (according to sales sequences)

Source: own work.

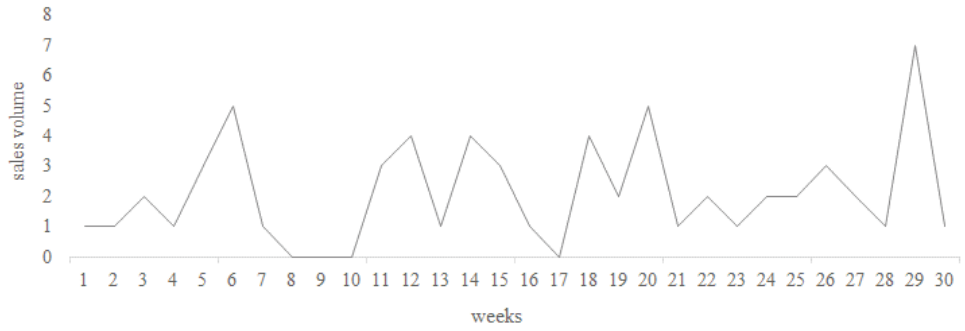

Figure 2. Non-random time series (according to sales sequences)

Source: own work.

In many cases the time series which were recognized as non-random are also difficult to model. The primary reason is that the number of series is often very low so it is hard to find any sensible regularity. Like in other cases, the estimation of zero-inflated time series models might be difficult.

The main aim of the article is to present a procedure that could be useful in forecasting randomly distributed zero-inflated time series. The procedure is based on a stochastic simulation and is applied in the real sales forecasting system that encompasses about eighteen thousand products, in an enterprise located in the vicinity of Szczecin (Poland). The company, in which the mentioned sales forecasting system is functioning, is a distribution centre with a medium-sized warehouse. Each week sales forecasts for the next five weeks for all products are computed. It is worth adding that the described above forecasting system is extensive and has many other features related to the control of forecasts. There are many restrictions connected with the forecast level, volatility, consistency, etc. which are not described here. All of the data presented in the article are real. To show some specifics of the analysed data, a random sample of the two time series is presented in Figures 3–4.[2]

According to the presented data, we can see that there are many outliers that increase volatility. It is also hard to find any tendencies (regularities). The weekly sales frequency is usually very low (Shukur, Doszyń, Dmytrów, 2017). By the sales frequency (for a given product) the share of weeks with positive sales in all of the considered weeks is understood. The distribution of products according to sales frequency in the last 30 weeks is presented in

[2] As it was mentioned, there are about eighteen thousand products. This is just a small (random) sample of them. There are weeks on the horizontal axis and sales volume on the vertical axis.

Figure 5.[3] The distribution is strongly skewed with a high, positive asymmetry. As many as 67% of the products were sold at most in one per every ten weeks. Only 5% of the products were sold in four in every five weeks (or more often).

Figure 3. Sales of a randomly chosen product (weekly data)

Source: data from the analysed company.

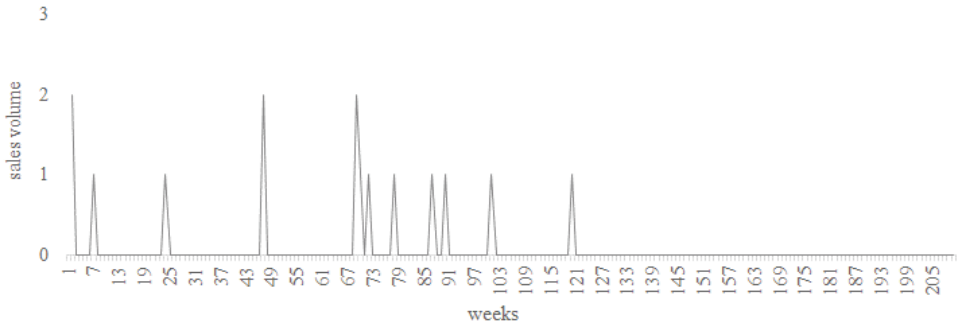

Figure 4. Sales of a randomly chosen product (weekly data)

Source: data from the analysed company.

A question arises of whether the series which have been classified as non–random should be forecasted by means of the count data models? In the case of the analysed forecasting system it seems to be difficult. Beyond randomness, there are other problems with forecasting sales by means of the count data models in a given company.

The presented forecasting procedure was applied to predict the sales of 17,678 products for five weeks ahead. The forecasts for the sample product are shown in Figure 6.

[3] For all products, if possible, the data from the last 30 weeks were taken to compute the sales frequency and the forecasts. In some cases the number of observations was lower, because new products had been introduced.

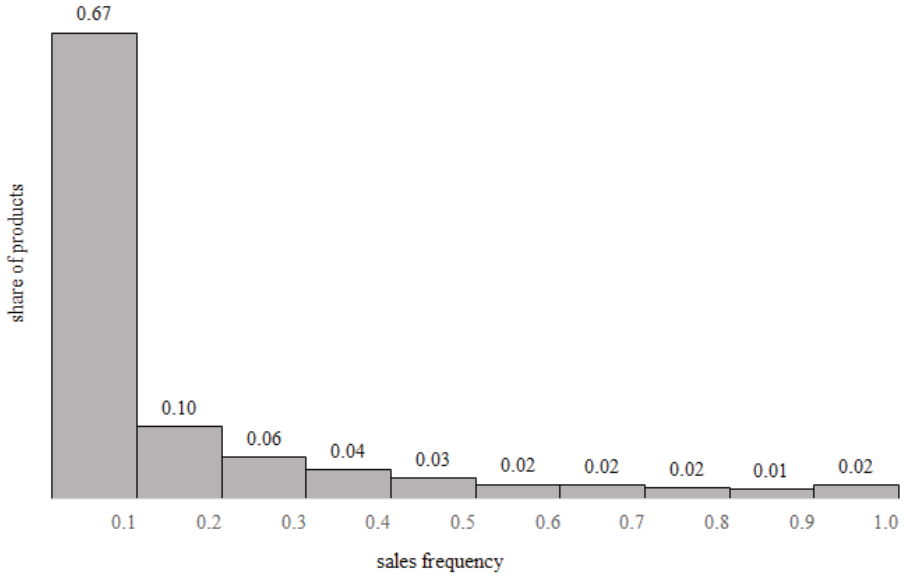

Figure 5. Distribution of products with respect to weekly sales frequency

Source: own work.

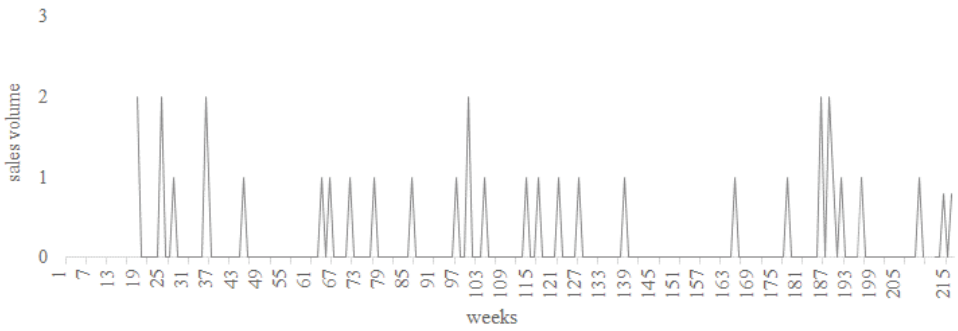

Figure 6. Sales forecasts of the exemplary (randomly chosen) product (weekly data)

Source: own work.

Figure 6 shows weekly data representing sales in the previous 30 weeks and the forecasts for the following 5 weeks. They have been computed by means of the proposed procedure based on the stochastic simulation.

It is not easy to analyse the accuracy of forecasts for so many products. One way to do this is to verify the distributions of chosen forecast errors. The problem with this method is that the most relative *ex post* forecast errors are not possible to compute in the case of a zero-inflated

time series (there are zeros in the denominator). The *ex post* forecast error that could be useful in such situations was proposed in the previous paragraph.

To evaluate the accuracy of the forecasts computed by means of the proposed stochastic simulation procedure, the proposed error (D) was computed for all 17,656 products (Figure 7). Forecasts computed for the last considered week were compared with real values.

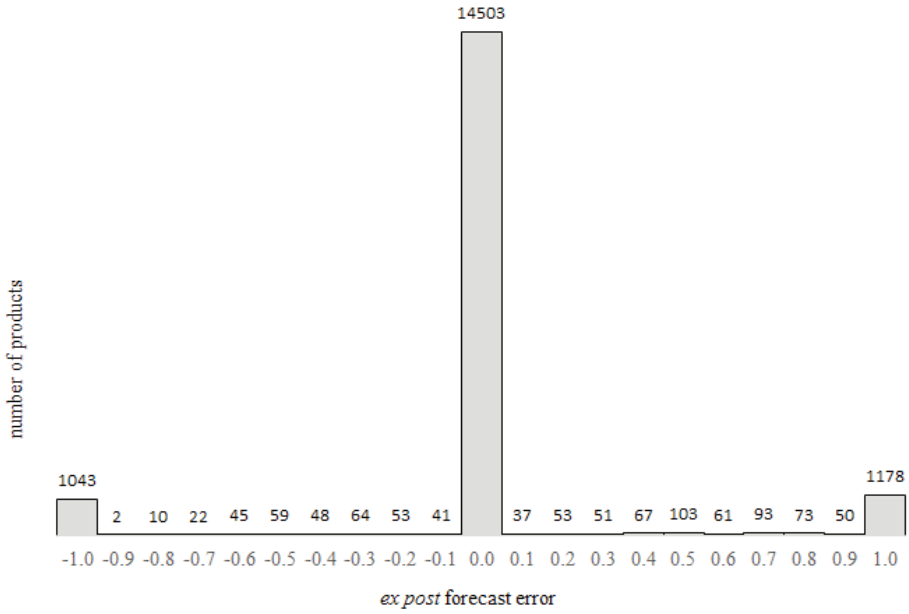

Figure 7. Distribution of *ex post* forecast error *D*

Source: own work.

For the vast majority of products (14,503 products) the *ex post* error was very small and, in most cases, equal to zero. In the analysed week, for 1,043 products the sales were positive but the forecasts were equal to zero ($D = -1$). In the case of 1,178 products the *ex post* forecasts were positive, but the sales was equal to zero ($D = 1$). In other intervals of the considered error, numbers were rather small (from 2 to 103). To sum up, the quality of forecasts due to the proposed *ex post* forecasts error seems to be very satisfactory.

Conclusions

The main aim of the article was to propose a forecasting procedure that could be useful in the case of a randomly distributed zero-inflated time series. According to the presented analysis, many economic time series are randomly distributed, so it is not possible to estimate any kind of statistical or econometric models. The considered procedure is based on a stochastic simulation. In the first step, a hypothesis stating the randomness of a times series was considered, with regard to both sales sequences or sales levels. If times series are randomly distributed it is not possible to apply models for count data and a stochastic simulation procedure might be helpful. The presented attitude is a novelty of the article. Moreover, the *ex post* forecast error that could be computed also for zero-inflated time series was proposed.

In the empirical example, the described procedure was applied to forecast the sales of products in a company located in the vicinity of Szczecin (Poland). The accuracy of the forecast was verified as well. Following the empirical verification it could be stated that the proposed forecasting system seems to work well.

By far, it was not possible to find in the literature other processes that could be forecasted also in the presented way. But it is probable that processes from different fields, which are generated by similar data generating processes, could be also forecasted in the presented way.

References

Asmussen, S., Glynn, P.W. (2007). *Stochastic Simulation: Algorithms and Analysis*. New York: Springer-Verlag.

Biswas, A., Song, P. (2009). Discrete-valued ARMA processes. *Statistics and Probability Letters 79*, 1841–1889.

Cameron, A.C., Trivedi, P.K. (2001). Essentials of Count Data Regression. In: *A companion to theoretical econometrics* (pp. 331–348). Oxford: Blackwell.

Cameron, A.C., Trivedi, P.K. (2005). *Microeconometrics. Methods and Applications*. Cambridge University Press.

Cameron, A.C., Trivedi. P.K. (1998). *Regression Analysis of Count Data*. Econometric Society Monograph No. 30. Cambridge University Press.

Domański, C. (1990). *Testy statystyczne*. Warszawa: PWE.

Hilbe, J.M. (2011). *Negative Binomial Regression*. Second Edition. Cambridge University Press.

Hilbe, J.M. (2014). *Modeling count data*. Cambridge University Press.

Taleb, N.N. (2001). *Fooled by Randomness. The Hidden Role of Chance in the Markets and in Life*. New York–London: Texere.

Shukur, G., Doszyń, M., Dmytrów, K. (2017). Comparison of the Effectiveness of Forecasts Obtained by Means of Selected Probability Functions with Respect to Forecast Error Distributions. *Communications in Statistics. Simulation and Computation*, *46* (5), 3667–3679. DOI: 10.1080/03610918.2015.1100734.

Winkelmann, R. (2008). *Econometric Analysis of Count Data*. Berlin, Heidelberg: Springer-Verlag.

Yang, M. (2012). *Statistical models for count time series with excess zeros*. University of Iowa.