Zero Inflated Time Series Analysis of Terrorism in India

Subhrangshu Bit*

25 January, 2021

Abstract

With recent advent of statistical analytics in various fields zero inflation has become a common nuisance. Such an explosive number of zeros has become a necessary detail to be addressed while analyzing the number of terrorist attacks in India. This article tries to exploit an observation-driven model to handle this issue. Time-series modelling of count data is a challenging topic and since popular auto-regressive and moving average (ARMA) models are restricted to continuous state-space they have been included in modelling the distributional parameters. The counts given the past history of the process and other information from covariates is assumed to follow a mixture of a Poisson distribution and a distribution degenerate at zero, with a time dependent mixing parameter π_t. Since, count data usually suffers from overdispersion, a Gamma distribution is used to model the excess variation, resulting in a zero inflated Negative Binomial (NB) regression model with mean parameter λ_t. Linear predictors with auto regressive and moving average type terms and trend are fitted to λ_t and π_t through canonical link generalized linear models.

1 Introduction

Terrorism is one of the biggest challenges the world is facing today. Over the last decade, it has reached unprecedented heights with dramatic consequences. Not surprisingly, social science has produced a vast amount of literature that elaborates this issue. The fact that there is no general consensus on the definition of terrorism, reflects the complexity of the phenomenon. Moreover, scholars are divided when it comes to explaining its causes. Quantitative research has emerged in recent years to add to the debate.
Time series modelling of count data is a popular research topic but is often flooded by zeroes. In such a situation it is difficult to understand if the observed counts, represent true terrorist attack or spurious one (i.e., there was an attack but was not observed/reported). However ignoring such zero counts while time series modelling may lead to incorrect estimates and significant loss of information. This study involves an observation driven model i.e., the

*Department of Computer Science, Ramakrishna Mission Vivekananda Educational and Research Institute

autocorrelation is modelled as a function of past observations. The approach combines both autoregressive as well as moving average components in the zero inflated count time series model. The terrorist count data is initially assumed to follow a Poisson distribution. Since the assumption equality of mean and variance of Poisson distribution rarely holds true in practical is further assumed as Negative Binomial distribution. However these models do not take into account the inflation of zeroes and therefore can be further improved using a zero-inflated negative binomial distribution. In order to incorporate an ARMA model to fit the random component in the distributional parameters ARMA terms along with trend are fitted to the mean parameter of the NB, λ_t, and the mixing parameter, π_t, using log and logit links, respectively leading to a zero-inflated NB-ARMA model.

The rest of the article is organized as follows. The dataset under study is described in Section 2. In Section 3 we discuss the restrictions of ARMA on count data followed by autocorrelation structure between the observations. Section 4 includes various models and their details along with the estimated fits. We conclude with the scope of further improvement in Section 5 and references in Section 6.

2 Dataset

The Global Terrorism Database (GTD) is an open-source database including information on terrorist attacks around the world from 1970 through 2017. The GTD includes systematic data on domestic as well as international terrorist incidents that have occurred during this time period and now includes more than 180,000 attacks. The database is maintained by researchers at the National Consortium for the Study of Terrorism and Responses to Terrorism (START), headquartered at the University of Maryland.
Here we have tried to model a subset of the GTD which includes monthly data of number of terrorist attacks in India ranging between - February, 1972 to December, 2017. Since, the database has failed to restore the observations corresponding to the year 1993 here those missing observations has been imputed by zeroes.

- Charachteristics of the dataset:
  - The values that the variable - *Total Number of Attacks* can take every month is a count ranging between 0 to ∞ (theoretically).
  - This is an example of discrete state space discrete time stochastic process.
  - Moreover, a slight look into the data set reveals that there are 116 zeroes i.e. 21.1 % of the data is zero. Therefore there is a serious zero inflation in the data that requires to be addressed.

3 Discussion

3.1 Restrictions with ARMA/ARIMA

The observation at each time step being integer-valued clearly cannot be normally distributed. The traditional ARMA/ARIMA models work well in describing series with Gaussian marginal distributions [1]. However there is no known result characterizing autocovariance functions of stationary count series [1].
$\gamma_X()$ is a symmetric non-negative definite function on the integers, if and only if there exists a stationary Gaussian sequence $\{X_t\}$ with -

$$\gamma_X(h) = Cov(X_t, X_t + h)\forall h \tag{1}$$

Unfortunately, no analgous result exists for say, a stationary series with Poisson marginal distributions. Thus arises the necessity of modelling the count data in a different approach.

3.2 Time Dependencies

The sample autocorrelation measures the linear relationship between the response variable's current value with it's past values at different lags. In order to get an idea of how the count of terrorist attacks over time are correlated we obtain the sample autocorrelation function (ACF) plot-

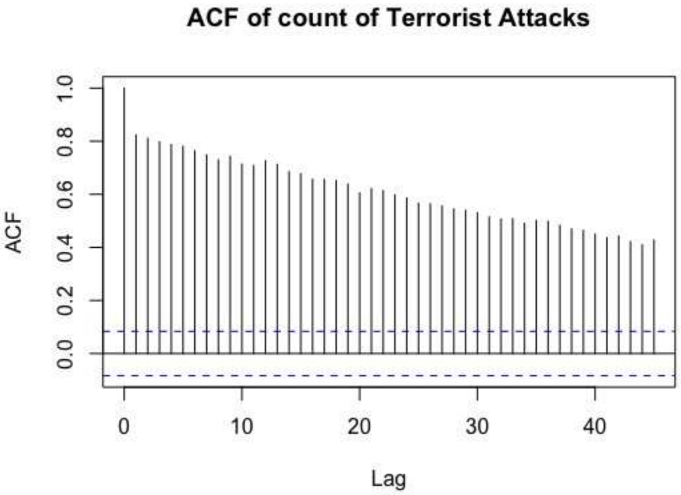

FIGURE 1: ACF Plot

We observe that there is a high autocorrelation in the observations. The autocorrelation value decreases over time very slowly. This indicates the presence of strong linear relationship among the observations.

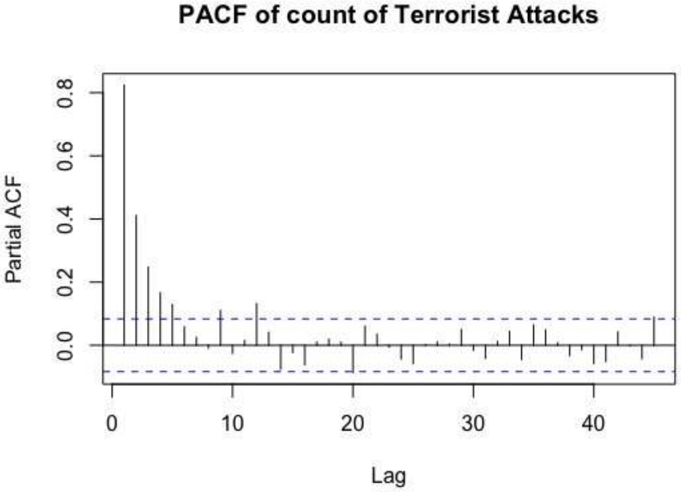

FIGURE 2: PACF Plot

The partial autocorrelation measures the correlation between X_{n+h} and X_n with the linear dependence of $\{X_{n+1}, ..., X_{n+h1}\}$ on each removed. Here, the observations have high partial autocorrelation uptil lag 5 and a few more at higher lags. The observations are correlated through the intermediate variables at higher lags.

4 Methodology

4.1 Zero Inflation

The data is said to be zero-inflated when the number of zeros is so large that the data do not readily fit standard distributions. However the source of the zeroes matters: Non-detection (false zeros) or true zeroes? Thus it requires to be taken into account.

There are primarily two methods of modelling zero-inflated data.

- Two-parts Model:
  - Sources of zeroes are NOT considered.
  - First step: Binomial model used to model probability of zeroes.
  - Second step: zero-truncated model used to model the count data.
- Mixture Model:
  - Zeroes are modeled explicitly as coming from multiple processes: true zeroes and false zeroes are modeled as being generated from different processes.

While our data seems to be zero-inflated, this doesn't necessarily mean we need to use a zero-inflated model. In many cases, the covariates may predict the zeros under a Poisson or Negative Binomial model. So let's start with the simplest model, a Poisson GLM.
Note: Here we have considered time as the only covariate.

4.2 Non-Autoregressive Model

4.2.1 Poisson Model

$$Attacks_t \sim Poisson(\lambda_t)$$

$$E(Attacks_t) = \lambda_t$$

$$Var(Attacks_t) = \lambda_t$$

$$log(\lambda_t) = \alpha + \beta * t$$

The mean is a function of time. We obtain $log(\lambda_t) = 2.56 + 0.007 * t$. Clearly the coefficient of time is not very significant. As a measure of comparison of different models we observe the Arkaike Information Criterion (AIC) value to be 8151.4. The standardized residuals are expected to have variance close to one. But it turns out to be 11.22. This suggests that our model produces overdispersion. One of the shortcoming of this model is that Poisson distribution enforces the *variance to be same as the mean*.

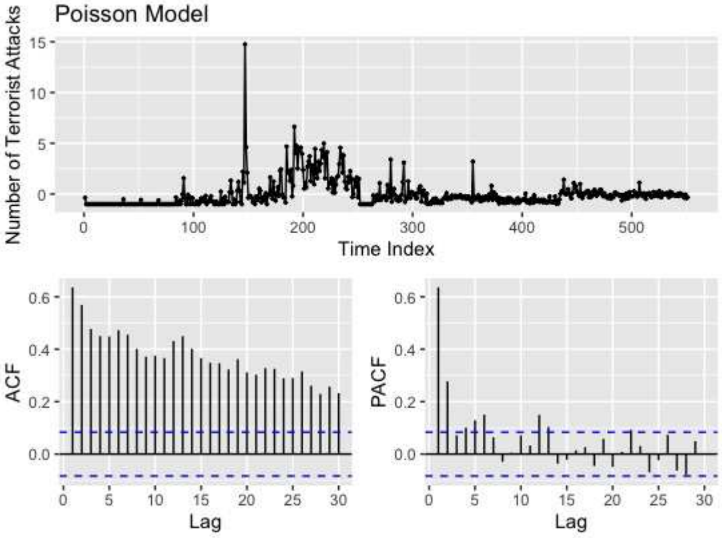

FIGURE 3: Visualization of residuals of Poisson Model

4.2.2 Negative Binomial Model

$$Attacks_t \sim NB(\mu_t, \theta)$$

$$E(Attacks_t) = \mu_t$$

$$Var(Attacks_t) = (\mu_i + \mu_i^2)/\theta$$

$$log(\mu_t) = \alpha + \beta * t$$

This model overcomes the shortcoming of same mean and varaince of Poisson. We obtain - $log(\mu_t) = 2.53 + 0.008 * t$. Still the coefficient of time is not very significant. We observe the AIC to be 3962. It has reduced significantly. Also the variance of the standardized residuals is 1.19. The amount of overdispersion has reduced compared to that of Poisson but this is further improved.

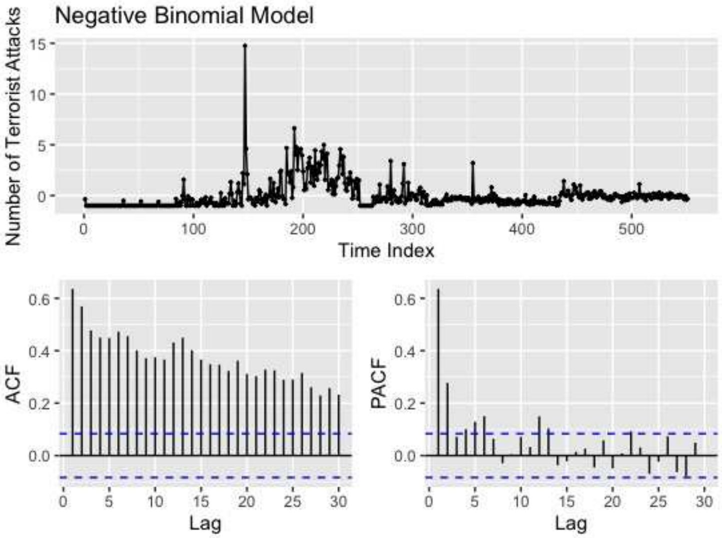

FIGURE 4: Visualization of residuals of Negative Binomial Model

4.2.3 Fitted Plot

We plot the estimates of the poisson and NB parameters over the count values as obtained from the pervious fits.

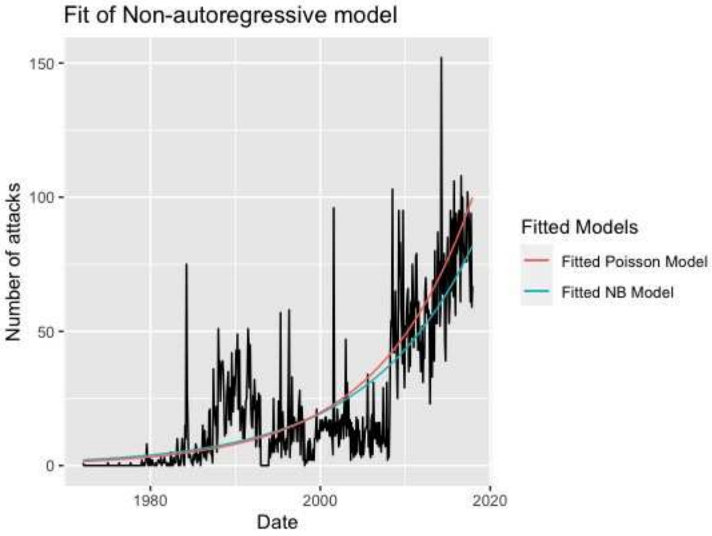

FIGURE 5: Fitted Models (Non-Autoregressive)

4.3 Zero-Inflated Model

4.3.1 Zero-Inflated Negative Binomial

We now require to address two issues:

- Overdispersion: We try to address this by including the zero inflation parameter or the mixing parameter(π_t).

- Autocorrelation: This is addressed by incorporating ARMA type structure in estimating the mean of the distribution.

We model the conditional distribution of the counts at the time t, Y_t, given H_t in two parts -

Part 1: Bernoulli

$$Y_t|H_t \sim \{^{0 \quad with\ probability\ \pi_t}_{>0 \quad with\ probability\ (1-\pi_t)} \tag{2}$$

Here H_t is the information available on responses till time (t-1) and on covariates till time t.

- Part 2: Truncated Negative Binomial

$$Y_t|H_t, Y_t > 0 \sim NB(\mu_t, k) \tag{3}$$

  or

- Part 2: Negative Binomial(Allowing zeroes)

$$Y_t|H_t \sim NB(\mu_t, k) \tag{4}$$

  Here k is the estimated overdispersion coefficient.
  We thus obtain for Truncated NB-

$$E(Y_t|H_t) = (1-\pi_t)\frac{\mu_t}{1-(1+\frac{\mu_t}{k})^{-k}} \tag{5}$$

$$Var(Y_t|H_t) = (1\pi_t)Var(Y_t|H_t, Y_t > 0) + \pi_t(1-\pi_t)[E(Y_t|H_t, Y_t > 0)]^2 \tag{6}$$

  And we obtain for NB-

$$E(Y_t|H_t) = (1-\pi_t)\mu_t \tag{7}$$

$$Var(Y_t|H_t) = \lambda_t(1\pi_t)[1 + \lambda_t\pi_t + \lambda_t/k] \tag{8}$$

In the zero inflated model, there are two means, the NB mean μ_t, and the Bernoulli mean π_t, which are respectively modeled as -

$$log(\mu_t) = W_t \Rightarrow \mu_t = exp(W_t) \tag{9}$$

$$logit(\pi_t) = M_t \Rightarrow \pi_t = \frac{e^{M_t}}{1+e^{M_t}} \tag{10}$$

We further model the terms W_t and M_t. Firstly we address the issue of overdispersion and leave out the autocorrelation. For that we consider the following simple model -

- Model:

$$W_t = x_t^T\beta$$

$$M_t = u_t^T\delta$$

Here we have taken into account only the covariates $x_t = \{1, t\}$ and $u_t = \{1, t\}$. We can always improve the model by taking into account other covariates which also include the seasonality component.

We fit the above two parts model on our dataset to obtain the following result -

	Estimate
Intercept	2.785
time	0.005

TABLE 1: Count model coefficients (truncated negbin with log link)

	Estimate
Intercept	3.657
time	0.023

TABLE 2: Zero inflation model coefficients (binomial with logit link)

Note that the effect of time on the inflation parameter π_t is comparatively higher. Also the AIC value is 3772.915 which is significantly smaller to that was obtained in the NB model. Moreover we obtain the variance of the standardized residuals to be: $0.999 \approx 1$. Thus we have almost removed the overdispersion.

Now looking into the residuals of the fitted zero inflated NB model we see the following structure of the residuals.

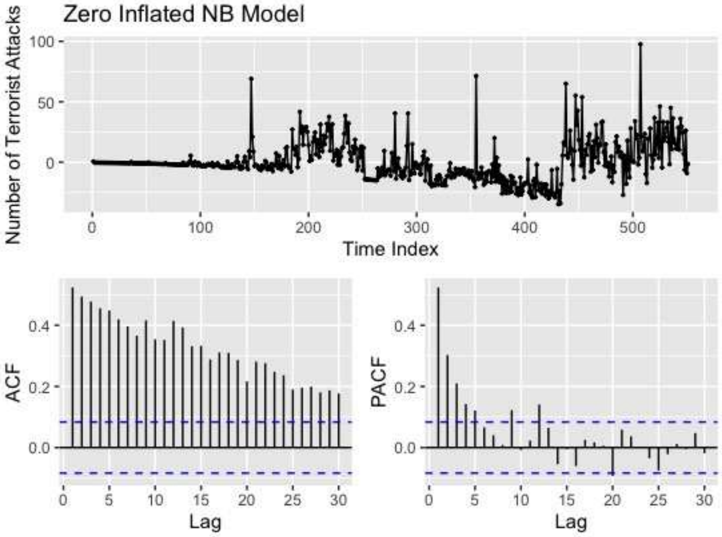

FIGURE 6: Visualization of Residuals of the zero inflated NB model

Although we somewhat resolved overdispersion in the data we see that there is high autocorrelation among the observations. The ACF plot shows that the acf value decreases slowly over time, while the PACF plot cuts off at 6 although there are a few significant pacf values at higher lags.

4.4 Zero-Inflated ARMA Model

4.4.1 Zero-Inflated Negative Binomial ARMA

We further modify the models of W_t and M_t-

- Model:

$$W_t = x_t^T \beta + Z_t$$
$$M_t = u_t^T \delta + V_t$$

Here along with the covariate terms x_t and u_t we also consider autoregressive and moving average terms Z_t and V_t. For simplicity we have not taken V_t into consideration.

$$Z_t = \sum_{i=1}^{p} \phi_i (Z_{t-i} + e_{t-i}) + \sum_{j=1}^{q} \theta_j e_{t-j} \tag{11}$$

$$e_t = \frac{(Y_t - \Lambda_t)}{\sqrt{\Psi_t}} \tag{12}$$

It can be proved that the residuals e_t has expectation 0 and variance 1 and are also uncorrelated. [3]

Since this is a two parts model we fit the models separately.

- Part 1: Bernoulli
  Here for simplicity we have not considered any autoregressive or moving average terms.

	Estimate
Intercept	3.657
time	0.022

TABLE 3: Zero inflation model coefficients (binomial with logit link)

- Part 2: Negative Binomial ARMA
  - The ACF plot of the original data tails off. Therefore it is not a pure MA model.
  - The PACF plot cuts off at lag 5. However there are some significant partial autocorrelations at higher lags - 9, 12, 20, 45.
  - So as an initial guess we start with -Model 1: AR(5). We further go for higher order ARMA models.

Model	AIC	BIC
Bin	265.8138	274.4372
NB-AR(5)	3676.223	3715.028
NB-ARMA(1,1)	3650.239	3676.109
NB-ARMA(2,1)	3652.278	3682.460

TABLE 4: Comparison Table of Fitted Models

We now combine the two parts to generate a zero inflated NB-ARMA: Binomial and NB-ARMA(1,1), since it has he least AIC and BIC values to obtain the final model ZINB-ARMA(1,1). The fitted plot is obtained as-

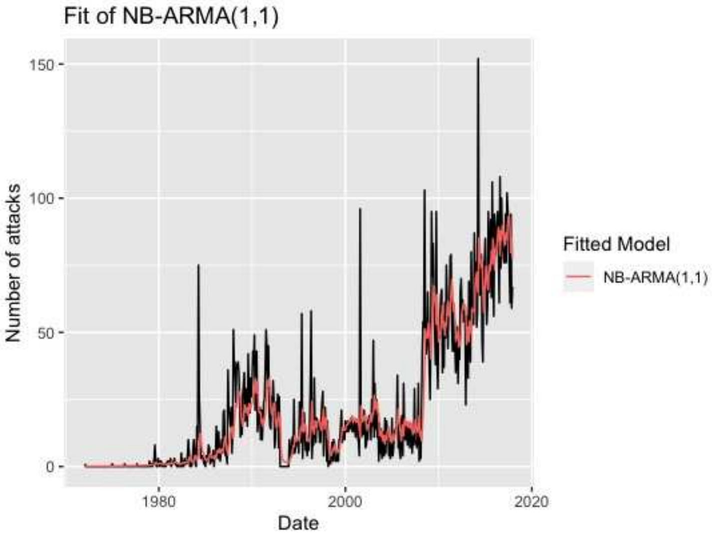

FIGURE 7: Fitted ZINB-ARMA(1,1)

4.4.2 Testing and Visualizing the Noise-Residuals

After fitting the ZINB-ARMA(1,1) we investigated the residuals in order to get an idea of how good is the fitted model. The ACF and PACF plot of the residuals were found to be -

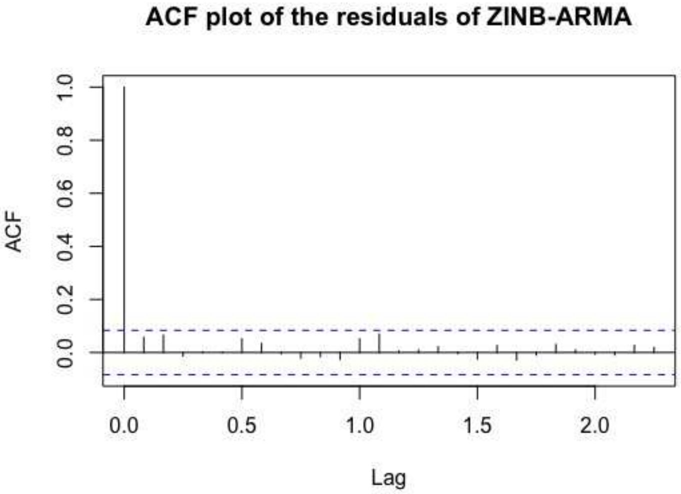

FIGURE 8: ACF of residuals

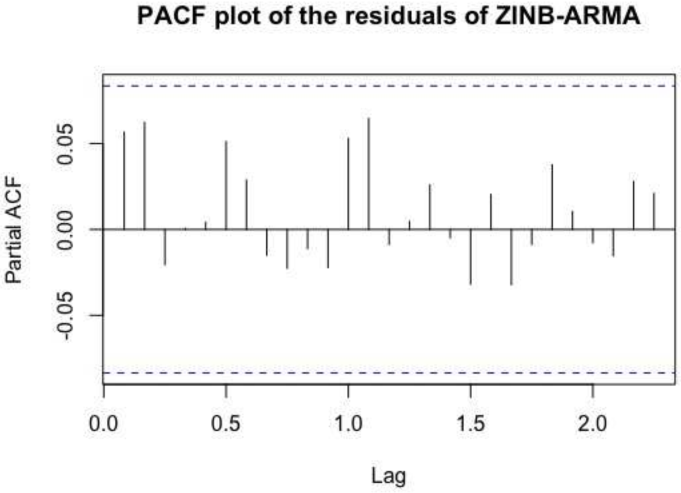

FIGURE 9: PACF of residuals

Clearly, the ACF and PACF plots indicates it is appropiate enough to assume that the residuals come from an independent and identical distribution (iid). We performed a few tests to verify this.

- Ljung-Box Test
  - X-squared = 1.7775, df = 1, p-value = 0.1825

  - Since p-value >0.05, we fail to reject the null hypothesis and conclude that the residuals are a sample from an iid sequence.

- Test for unit root
  - Dickey-Fuller = -7.1332, Lag order = 8, p-value <0.01 alternative hypothesis: stationary
  - In light of the given data we have sufficient evidence to reject the null hypothesis and conclude that there is no unit root. Thus the residuals are stationary.

- Test for normality
  - Shapiro R^2 test - W = 0.48005, p-value <2.2e-16
  - In light of the data as p-value <0.05 we have sufficient evidence to reject the null hypothesis and conclude that the residuals do not follow a normal distribution.

5 Conclusion

The ZINB-ARMA models considered here are simple and there is no account for seasonality component. The only covariate used for modeling the conditional mean is time. Several other covariates like those considered in [2] can be taken into account. Also there are sudden peaks in the data - February 1984, August 2008, April 2014 which can be addressed in the model. The variance of the residuals of NB-ARMA(1,1) turns out to be 0.982. Thus there is a slight underdispersion which can be improved.

References

[1] Yisu Jia. *Some Models for Count Time Series*. 2018.

[2] Daan Koopman. *Modelling Terrorist Attacks*. 2018.

[3] Siuli Mukhopadhyay Vurukonda Sathish and Rashmi Tiwari. *ARMA Models for Zero Inflated Count Time Series*. 2020.