IJDSEA
International Journal of Data Science Engineering, and Analytics

e-ISSN : 2807-1689
p-ISSN : 2798-9208

Negative Binomial Time Series Regression – Random Forest Ensemble in Intermittent Data

Amri Muhaimin [1,*], Trimono [1], Prismahardi A.R [1] and Hendri Prabowo [2]

[1] Universitas Pembangunan Nasional "Veteran" Jawa Timur
[2] Institut Teknologi Sepuluh Nopember
* Correspondence: amri.muhaimin.stat@upnjatim.ac.id

Citation: Muhaimin, A., Trimono, T., Riyantoko., P.A., Prabowo., H. Negative Binomial Time Aeries Regression – Random Forest Ensemble in Intermittent Data *C* **2021**, *Volume 1, Number* 2, Page 36-42.

Received: November 16, 2021
Accepted: November 20, 2021
Published: November 25, 2021

Abstract: Intermittent dataset is a unique data that will be challenging to forecast. Because the data is containing a lot of zeros. The kind of intermittent data can be sales data and rainfall data. Because both sometimes no data recorded in a certain period. In this research, the model is created to overcome the problem. The approach that is used in this research is the ensemble method. Mostly the intermittent data comes from the Negative Binomial because the variance is over the mean. We use two datasets, which are rainfall and sales data. So, our approach is creating the base model from the time series regression with Negative Binomial based, and then we augmented the base model with a tree-based model which is random forest. Furthermore, we compare the result with the benchmark method which is The Croston method and Single Exponential Smoothing (SES). As the result, our approach can overcome the benchmark based on metric value by 1.79 and 7.18.

Keywords: Ensemble Model, Intermittent Data, Negative Binomial, Time Series, Random Forest.

1. Introduction

In certain cases, the rainfall data and sales data are not a continuous process. Sometimes the rain happens, sometimes not, the sales data also. Intermittent data is data that the data is not always available every time. Sometimes it recorded zeros value for a long period. And this case also needs a special method to forecast. Time series intermittent data is different from naive time series because intermittent time series have a lot of zeros value and that is quite challenging to forecast [1]. The characteristic of the intermittent data is the zeros data is quite high, so it causes the variance from the data is higher than the mean value. Then the over disperse might be happening.

There are several methods to forecast the intermittent dataset, such as the Croston method, Single Exponential Smoothing (SES), and maybe Autoregressive Integrated Moving Average (ARIMA). The Croston method is come from 1972 and was developed by [2], it was calculated by the mathematic algorithm and us the conditional statement. The algorithm just tries to figure out when the data is recorded and not, then uses the mean approximation to create some forecasting value. The performance of course is not well, because the pattern from it cannot represent the actual value. SES is also one of the methods that can forecast intermittent data, using the naive algorithm to forecast the data. The model generated from SES is $y_t = \beta_t + \varepsilon_t$, the parameter β is constant and evolves slowly over time t. ARIMA is a model-based method, it can't perform well to forecast the intermittent data. As we can see here, the model that ARIMA produces depends on the p, d, and q. Let's say $AR(p)$ model, the mathematical function from it is $y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon$, with that model, the fitted value from intermittent data will be worst because the calculation of the formula is not good for the data that contains too many zeros values.

Our approach is the ensemble method between distribution or probabilistic method and tree-based method. The base model comes from Negative Binomial time series regression. After we train the data using the Negative Binomial time series regression, we create the sub-model which is tree-based and then augment the result. The Negative Binomial time series regression might have an important rule to create the fitted value, while the tree-based is used to an approximation between the fitted value and real value.

This study focuses on the ensemble method and applies it to the data, which are rainfall data and sales data. The benchmark comes from an ad-hoc method which is SES, and a model-based method which is ARIMA. There are four sections left in this paper, section 2 is explained about related works, section 3 is experimental and analysis and section 4 is the conclusion.

2. Related Works

Mostly the method that is used to forecast intermittent data is an ad-hoc model. The model-based is rarely seen. With the ad-hoc method, the algorithm can define when the data is available (not zero) and when the data is not available (zero).

Ref [3] used some neural-network-based method, which is Recurrent Neural Network (RNN). Ref [3] compared the result with the conventional method such as Croston and SES method. The metric evaluation that is used is Mean Absolute Error (MAE) and Root Mean Squared Scaled Error (RMSSE). The result is not good enough. In certain data, RNN can overcome the conventional method, otherwise can't.

Ref [4] used the aggregated idea. The intermittent data can't be intermittent if the data is up-scaled. Let's say the data is daily, then we can up-scaled the data into monthly or annually. But the main information from it will fade away. Then Ref [4] uses the SES and Croston to compare, which one is better when the data is aggregated. The result said that the SES is better than the Croston Method.

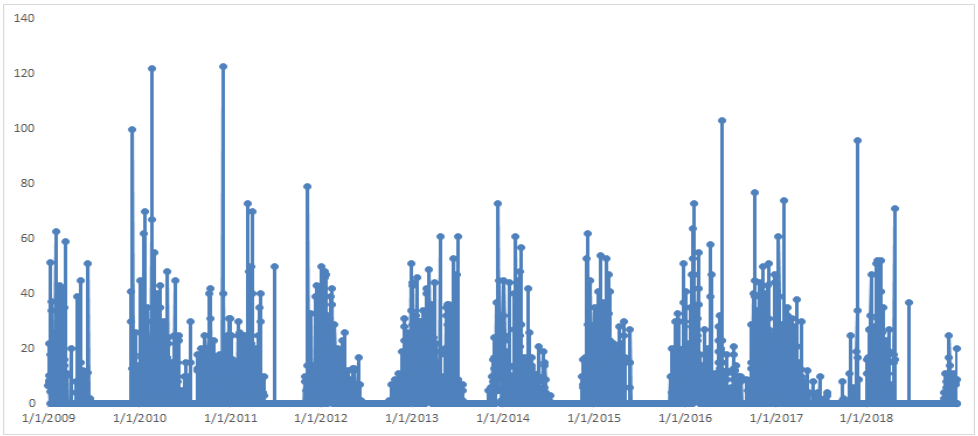
The ANN method is used in Ref [5] to forecast rainfall in the lake basin area. A feed-forward neural network was used as the ANN (FFNN). It was compared to ARIMA models in Ref [5]. The metrics used for evaluation are RMSE and mean absolute error (MAE). The testing data used is 72 months in advance. As a result, the ANN and ARIMA models do not differ significantly. However, in RMSE and MAE testing data, FFNN outperforms the ARIMA model.

3. Experiment and Analysis

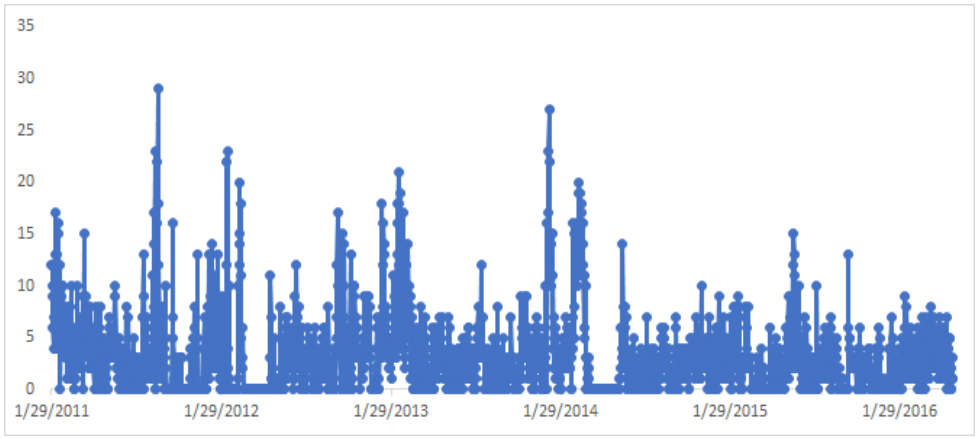
Three models are created to forecast the data. Whether linear model, ad-hoc model, or ensemble model. This data is univariate time-series, so the only variable used is the data itself, it means rainfall and sales. The metric score evaluation only used the validation score to provide that our approach is better than the baseline.

3.1. Dataset and Variable

The rainfall and sales data are daily recorded. The period of the rainfall data is 3653 days or equal to 10 years, and the period of the sales data is 1942 days equal to 5 years. We analyze the data still on the day scale, we are not aggregating the data. For the rainfall dataset, the source comes from a climatology agency in Indonesia, and for sales, the dataset is come from the M5 competition and can be downloaded via Kaggle. To further understand these data, we can see Figure 1 as represented below.



(a)



(b)

Figure 1. (a) Rainfall data (b) Sales data.

Figure 1 (a) represents the rainfall data plot. It contains a lot of zeros and it is intermittent. The patterns might be seasonal because in Indonesia there are only two seasons, summer and rainy season. For the (b) is sales data. The period is no longer than the rainfall data, but the zeros value is shown by the plot. With these data, we will find out that our approach can do well to create a model especially forecasting. For the zero value, in rainfall data, at least there are 2490 out of 3653 days, it is more than 50% of the data. For the sales data the zero value is around 546 out of 1941 days, it is less than 50%. We knew that the higher the zeros value in the data, the harder model can forecast the data.

In a forecasting case, the predictor variable might be difficult to define. Some methods are needed to define the variable. Some methods that can be used are the plot of the autocorrelation function and the plot of the partial autocorrelation function. With those methods, we can find which lag influence significantly the dataset. But the condition can be applied if the data is stationary distributed. As we can see here, those data contain a lot of zero, so the distribution is not stationary. So we just define the lag based on the previous research. Ref [3] use 28 of lag to forecast intermittent demand, and Ref [6] use lag 1, lag 2, and lag 12. Another factor that can be used to define the lag is the seasonal pattern. As we can see the rainfall data maybe have a seasonal pattern around 180 days

or equal to 6 months. If we look back that far, the Negative Binomial time series regression might be not good. So we just use up to 30 days or 1 month. After defining the variable, we feed the lag and the response variable to the model. Moreover, the predicted value or forecast value is calculated for 7 days ahead.

3.2 Methodology

The ensemble model is the combination of two models or more in one function. In our approach, we use Negative Binomial time series regression combine with the random forest as a tree-based model. The Negative Binomial time series regression is used to catch the zero – non-zero model, and a tree-based model such (Random Forest) used to catch the continued value of the data.

Y_t is a time series count or measurement data and assuming that on the index t of the data, is influenced by the previous values such as $Y_{\{t-1\}}, Y_{\{t-2\}}, \dots, Y_{\{t-n\}}$. If the mean is less than the variance of the data, then it's overdispersed. By that condition, the Negative Binomial approach can be used. Here is the following parameter-driven model [7]:

$$Y_t|\alpha_t \sim NegBin(r, p_t), \quad (1)$$

Where r is positive value and p_t is the logit model

$$-\log\left(\frac{p_t}{p_t}\right) = x_{nt}^T\beta + \alpha_t. \quad (2)$$

β is the vector of coefficient parameter, then the conditional density function of Y_t is:

$$p(Y_t = y_t|\alpha_t) = \binom{y_t + r - 1}{r - 1} p_t^r (1 - p_t)^{y_t}$$

for $y_t = 0, 1, \dots$. This will produce a non-negative value, and we assume that the error from this model is white noise with $i.i.d$. By equation (2) we can estimate the α. Thus this model will be applied to the sales data and rainfall data.

Random forest is a tree-based model proposed by [8]. In this case, Random Forest is an ensemble from a regression tree. The regression model is created one by one, then the value is ensembled together, either using aggregating or major voting. For some dimensional random variable $X = \left(X_1, X_2, \dots, X_p\right)^T$ is the input data from the lag that extracted, the joint distribution is assumed by $P_{XY}(X, Y)$, then the function to predict Y is built notated by $f(X)$. Moreover, the prediction is expected to minimize the loss function. The ensemble is constructed from the base learner or we can call it regression tree let's notate it with $h_1(x), \dots, h_j(x)$, and these base learner is combined either using aggregating, averaging, or majority voting [9]. Here is the picture for the full model [10].

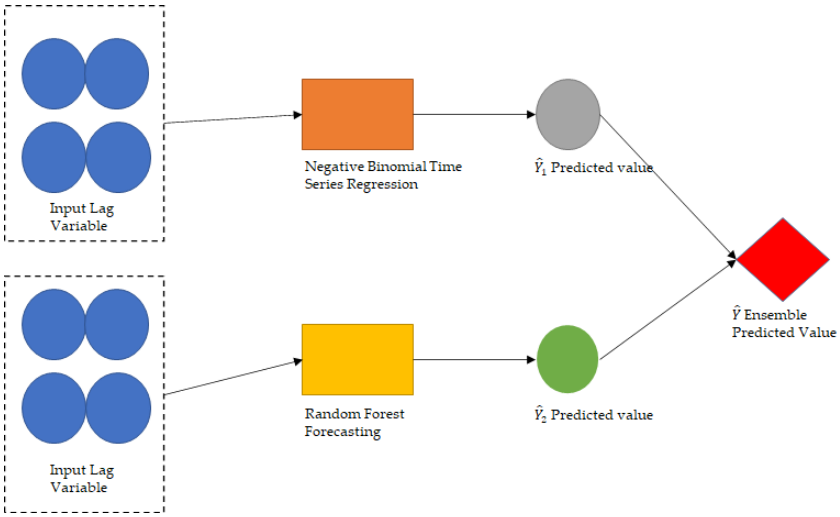


Figure 2. Full Model of Ensemble Negative Binomial and Random Forest

To evaluate intermittent data special metric evaluation is needed. Because some metric evaluations cannot evaluate the intermittent data. Especially Mean Absolute Percentage Error (MAPE), because the data contain a lot of zeros, thus MAPE can't work. Root Mean Squared Error (RMSE) is one of the metrics that can evaluate intermittent data. The data that will be forecast in this case is seven days ahead and the rest is the training or in sample data.

3.3 Forecasting Data with Negative Binomial Time Series Regression – Random Forest

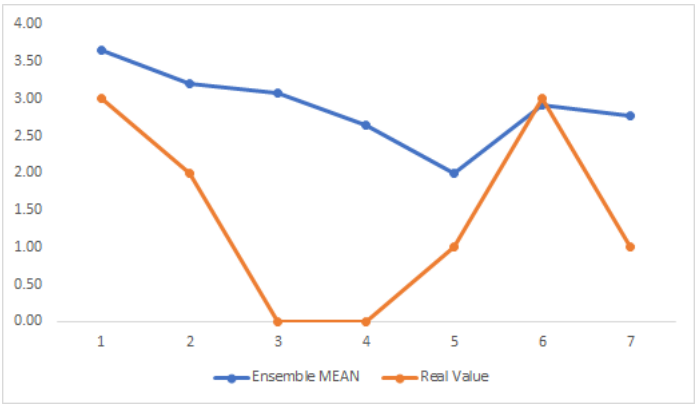
The lag variable that was used is 30 lag, and the testing period is seven days ahead. After training the model using the random forest and negative binomial we ensemble the result and compare it with the conventional method such as Croston Method and SES. The parameter that is used in the Random Forest is default by the package with the n-tree equal to 1000, and for the negative binomial we also use the default parameter. In the SES method, the smoothing value used is 0.2, and for the Croston method, we use Croston that was developed by [2]. After that, we compare the result within in sample (training data) and out the sample (testing data) using the RMSE value.

The result from the negative binomial parameter is some of the lag variables are significant in sales data and rainfall data. But we keep the model without remodeling it using only significant variables. The fitted value is then produced to create the prediction. For the negative binomial time series regression, we use K – step forecasting with K = 7, for the random forest also. After we got the forecasting value either from the random forest and negative binomial time series regression, we ensemble the result using the average method.

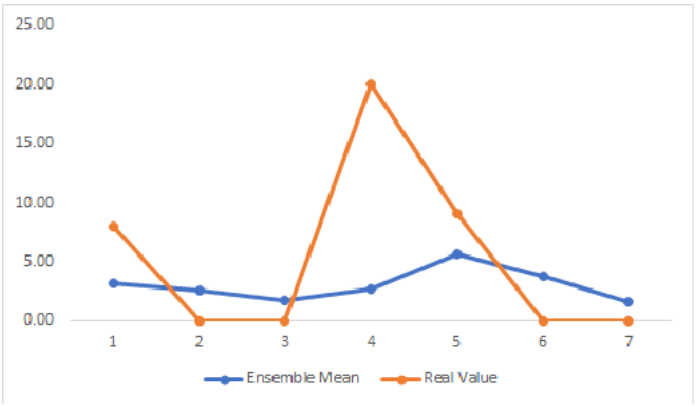
Table 1. RMSE In Sample and RMSE Out Sample Each Method

Method	RMSE Out Sample Sales Data	RMSE Out Sample Rainfall Data	RMSE In Sample Sales Data	RMSE In Sample Rainfall Data
Croston	2.24	8.04	2.61	9.66
SES	3.04	7.47	3.02	10.76
Ensemble	1.79	7.18	2.62	23.58

Here is the result for the in-sample and out-sample, represented in Table 1. The ensemble method overcomes both baseline methods, either SES and Croston. Even in rainfall data and sales data. But when we look at RMSE in the sample the conventional overcome the proposed method. We can say that the conventional data is overfitting while the proposed method is underfitting in rainfall data. To more understand the result, the plot is represented in Figure 2.



(a) Sales Data



(b) Rainfall Data

Figure 3. (a) Plot Between Predicted and Real Value of Sales Data, (b) Plot Between Predicted and Real Value of Rainfall Data

The forecast period is only 7 days ahead. The model is built from the 30 lag variable. Figure 3 shows that the ensemble model can follow the bottom value of the real data. But when the peak happens, the ensemble can't follow it optimally. But this result is better than the Croston and SES. Because both of them only create a static forecast for the predicted value.

4. Conclusions

The negative binomial time series regression is a parametric method to forecast count data. Especially when the mean value is lower than the variance value, it calls overdispersed. Our data is intermittent data, most of the value is zero. Negative binomial time series regression expects that the model can capture the dynamic value from the data. Because when the non-zero value comes accidentally, the conventional method such as ARIMA can't deal with it. But with the negative binomial time series regression, it is in the core of it, the model can follow the pattern of intermittent data. We combine with the random forest method because we want to capture non-discrete data. So,

this ensemble can forecast either discrete or continuous datasets. Based on the result, the ensemble method overcome the baseline in out-sample data. Furthermore, if we see the in-sample model evaluation, it is a little underfitting on rainfall data. Yet, if we talk about forecasting, the out sample accuracy is better to choose than the in-sample accuracy.

For future works, we want to create the hybrid model with negative binomial as a based model. And combine the error from the model with non-parametric machine learning models such as tree-based model, neural network model, or maybe support-vector. In short, our approach can overcome the baseline in the out sample data, and the predicted value can follow the pattern of true data.

Author Contributions: The concept comes from Muhaimin, Muhaimin also writes this paper. For negative binomial time series regression, Prabowo did it. Prisma creates the random forest model, and Muhaimin compiles it to become the ensemble. The data used is from the M5 competition and Indonesia Climatology Agency.
Funding: This research received no external funding
Data Availability Statement: The data can be accessed in kaggle.com and bmkg.go.id/akses_data
Acknowledgments: Many thanks to Allah SWT, my parents, my wife, and my friends. Whatever happens, never give up.
Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kourentzes, Nikolaos. (2013). Intermittent demand forecasts with neural networks. International Journal of Production Economics. 143. 198-206. 10.1016/j.ijpe.2013.01.009.
2. Croston, J. D. Forecasting and Stock Control for Intermittent Demands. *Operational Research Quaterly* **1972**, *23*, 289 - 303.
3. A. Muhaimin, D. D. Prastyo and H. Horng-Shing Lu, "Forecasting with Recurrent Neural Network in Intermittent Demand Data," 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 2021, pp. 802-809, doi: 10.1109/Confluence51648.2021.9376880.
4. Nikolopoulos, Konstantinos & Syntetos, Aris & Boylan, John & Petropoulos, Fotios & Assimakopoulos, Vassilis. (2011). An Aggregate-Disaggregate Intermittent Demand Approach (ADIDA) to Forecasting: An Empirical Proposition and Analysis. JORS. 62. 544-554. 10.1057/jors.2010.32.
5. Farajzadeh, J., Fard, A.M., Lotfi, S. Modeling of monthly rainfall and runoff of Urmia lake basin using "feed-forward neural network" and "time series analysis" model. Water Resources and Industry 7-8 (2014) 38–48
6. Suhartono, Prabowo H., Prastyo D.D., Lee M.H. (2019) New Hybrid Statistical Method and Machine Learning for PM_{10} Prediction. In: Berry M., Yap B., Mohamed A., Köppen M. (eds) Soft Computing in Data Science. SCDS 2019. Communications in Computer and Information Science, vol 1100. Springer, Singapore. https://doi.org/10.1007/978-981-15-0399-3_12.
7. DAVIS, R. A., & WU, R. (2009). A negative binomial model for time series of counts. *Biometrika*, *96*(3), 735–749. http://www.jstor.org/stable/27798860.
8. Cutler, Adele & Cutler, David & Stevens, John. (2011). Random Forests. 10.1007/978-1-4419-9326-7_5.
9. Gawthorpe, Katerina. (2021). Random Forest as a Model for Czech Forecasting. Prague Economic Papers. 10.18267/j.pep.765.
10. Akbar, M.S et al. A Generalized Space-Time Autoregressive Moving Average ({GSTARMA}) Model for Forecasting Air Pollutant in Surabaya. *Journal of Physics: Conference Series* **2020**, 1490, 012-022. doi: 10.1088/1742-6596/1490/1/012022