IOP Conference Series: Materials Science and Engineering

PAPER • OPEN ACCESS

Forecasting the Amount of Pneumonia Patients in Jakarta with Weighted High Order Fuzzy Time Series

To cite this article: Sebastian Tricahya and Zuherman Rustam 2019 *IOP Conf. Ser.: Mater. Sci. Eng.* **546** 052080

Forecasting the Amount of Pneumonia Patients in Jakarta with Weighted High Order Fuzzy Time Series

Sebastian Tricahya,[1] **Zuherman Rustam**[1*]

[1]Department of Mathematics, University of Indonesia, Depok 16424, Indonesia

[*]Corresponding author: rustam@ui.ac.id

Abstract. Forecasting the amount of Pneumonia patients could help medical practitioners to prepare the required medicines, aid-workers, or even prevent it by sharing knowledge to parents, elders, and smokers. This problem poses great concerns on the lives of many people, therefore, adequate accuracy is required in forecasting. Fuzzy Time Series (FTS) is an alternative way to forecast data. By using ARIMA and Holt's Exponential Smoothing, there are some problems that are difficult to obtain the best model. Using our FTS method, we modified the Cheng algorithm by using higher order (using two or more historical data) to make the accuracy better by seeing the Mean Absolute Percentage Error (MAPE). Data was selected from the amount of Pneumonia Patients in Jakarta from 2008 to 2018. We use R to carryout ARIMA and Holt's Exponential Smoothing. Forecasting's accuracy will decrease if the timeframe between these occurrences is lengthy. As a result of this, we made use of 5 periods which are January until May 2019. The result obtained was compared against ARIMA and Holt's Exponential Smoothing, as well as the MAPE are 9.70%, 16.85%, and 18.55% respectively.

Keywords: Forecasting, Fuzzy Time Series, MAPE, Modified FTS, Pneumonia

1. Introduction

One of the diseases prioritized by the WHO and Lancet Global Burden is pneumonia [1]. In Indonesia, it has become one of the most dread diseases since 2010. Pneumonia is an infection of the lung due to bacterial, viral, or fungal and a significant cause of mortality worldwide [2]. Initially, it was common on toddlers and elderly, however, owing to the growing number of smokers especially in Indonesia [3], people of all ages are now prone to this ailment. Lack of information on its symptoms, as well as early and appropriate treatment, has led to many deaths. Medical practitioners should prepare medicines and aid-workers (including doctors, nurses, and other staffs) to help people diagnosed with pneumonia [4]. In addition, preventive measures need to be conducted by educating people on some of the causes of this disease, how to detect its diagnosis, first aid treatment, and preventive measures. This research aims to predict the amount of pneumonia patients to help reduce the number of deaths.

Forecasting helps businesses, risk managers, financial decision makers, and many more to predict the future. This same technique will be used to forecast and analyze the medical field. Charles C. Holt developed the Holt's method used in forecasting which gained inadequate popularity as a lot was needed to be improved [5]. Box and Jenkins introduced Auto Regressive Integrated Moving Average (ARIMA) used to analyze forecasting. ARIMA attempts to find the best data patterns using both old and new results [6]. There are lots of inconsistencies associated with forecasting, which led to further research. Song and Chissom introduced the fuzzy time series using its relation [7]. Thereafter, Chen updated the model by using enrollments to facilitate some arithmetic operations which will make it easier [8], and since then, this method has been greatly improved. Fuzzy time series was updated by adding weight and modifying the class intervals [9, 10]. Some forecasting analyzes have been carried out using Holt's method, ARIMA, and Fuzzy Time Series. With regards to the inconsistencies associated with pneumonia, some other researchers made use of ARIMA [2, 11] and MAPE to obtain the required results. In this research, we are going to carry out weighted high-order fuzzy time series to solve the problem and use the MAPE for checking the accuracy of the method. But first, we will explain the ARIMA, fuzzy time series (FTS), and Mean Absolute Percentage Error (MAPE).

2. Methodology

ARIMA is a combination of Auto-regressive and Moving Average, denoted with p, d, and q. Where p is the degree of AR, d the stationery data, and q the degree of MA [6]. The ARMA (p, q) model given in Equation (1) below:

$$Y_t = \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + e_t - \theta_1 e_{t-1} - \cdots - \theta_q e_{t-q} \tag{1}$$

Where Y_t is the observed time series, each ϕ_p and θ_q are the parameters of the model, and e_t is white noise.

The procedure used in analyzing a time series data with ARIMA is as follows [12],

a) Identification of the model
By making the data fulfill the stationary assumption, use Auto Correlation Function (ACF) and Partial Auto Correlation Function (PACF) plot to identify the number of p and q

b) Estimation parameters
Use one of three most popular methods to estimate the parameters. Most of them are Moment, Least Square, and Maximum Likelihood method.

c) Diagnostic the model
This step will check the estimated parameters and residuals to determine if they are significant or not. Use Ljung-Box test while the hypotheses being tested is
H_0: There is no correlation between residuals
H_1: There is correlation between residuals
If the p-value < α, the ARIMA model is feasible for forecasting. Then, we can check the MAPE, MSE, and so on.

d) Forecasting
If we want to forecast for long period, the result will be a mean value, hence short time is more ideal.

Fuzzy is a kind of classification method consisting of numerous types one of which is the intrusion detection systems designed by Rustam and Talita [13]. In the fuzzy time series we classify each data against time as behavior-based to solve some seasonal factors in it. As discussed above its and definition remains unchanged. Here is the explanation about fuzzy time series [7, 8, 9, 10].

The universe U is defined = [smallest and, biggest number] to make our fuzzy set classification A_i.

Definition 1: Let $Z(t)$ be the universe of fuzzy sets $f_i(t)$. $F(t)$ is called fuzzy time series defined by $Z(t)$.

Definition 2: In case $F(t)$ is made by $F(t-1)$. Then, $F(t) = F(t-1) \bullet R(t, t-1)$, where $R(t, t-1)$ is the relation amongst $F(t)$ and $F(t-1)$. • means an operator

Algorithm 1

1) Define the universe U
2) Partition U into equal length intervals and then the total of intervals is corresponding to the total of linguistic variables (A_i)

3) Define each time series data will go to which A_i

$$A_i = \frac{m_{A_i}(u_1)}{u_1} + \frac{m_{A_i}(u_2)}{u_2} + \cdots + \frac{m_{A_i}(u_n)}{u_n} \quad (2)$$

Where m_{A_i} the membership function of fuzzy set A_i, therefore $m_{A_i} \rightarrow [0,1]$, u_i is the interval partition from the universal U, and n is the amount of intervals. Then, check the frequencies of each interval, if it is bigger than average.

4) Define fuzzy logic relations and fuzzy logic relations group

For example, the fuzzy relations is like this $A_i \rightarrow A_2, A_i \rightarrow A_1, A_i \rightarrow\ A_3, A_i \rightarrow A_1$, so the fuzzy relations group will be $A_i \rightarrow A_1, A_1, A_2, A_3$ (Example 1.1) and there are only 3 intervals.

5) Define the middle value of each interval and note as $\boldsymbol{t} = [t_1, t_2, \dots, t_n]$.

6) Normalized Weighting $\boldsymbol{W}(t)$

7) Forecasting $F(t) =\ \boldsymbol{t} \bullet (\boldsymbol{W}(t))'$

2.1 The proposed method

In this section, a modified fuzzy time series was proposed. From literatures, it can be seen that better methods are obtained through adequate improvements. The Algorithm 1 on step 3) and 4) will be modified by splitting the interval which its frequency is more than the average of the number of data and adding some others historical data to make forecasting.

From **Algorithm 1**, some steps are the same. But we define the different $F(t)$. $F(t)$ is caused by $\mathrm{F(t-1)},\ \mathrm{F(t-2)}, \dots,\ \mathrm{F(t-s)}$ and will be estimated as $\mathrm{F(t) = F(t-1) \bullet R(t-1,\ t-2, \dots,\ t-s)}$ where s is how many historical data that we want to use.

To define our universe formula (3) is used by adding D_1 and D_2, expand the Universe to give some possibilities that the forecasting results will go lower or higher than our minimum or maximum data.

$$U = [D_{min} - D_1, D_{max} + D_2] \quad (3)$$

Then we use the Sturges formula (eq.4) to find out how many partition we are going to make from our Universe U.

$$n = 1 + 3.322 log_{10} N \quad (4)$$

Where N is the number of data, the length of intervals l will use formula (5)

$$l = \frac{[(D_{min}-D_1)-(D_{max}+D_2)]}{-n} \quad (5)$$

Then, follow the next step from **Algorithm 1**.

The summary can be seen in **Figure 1**.

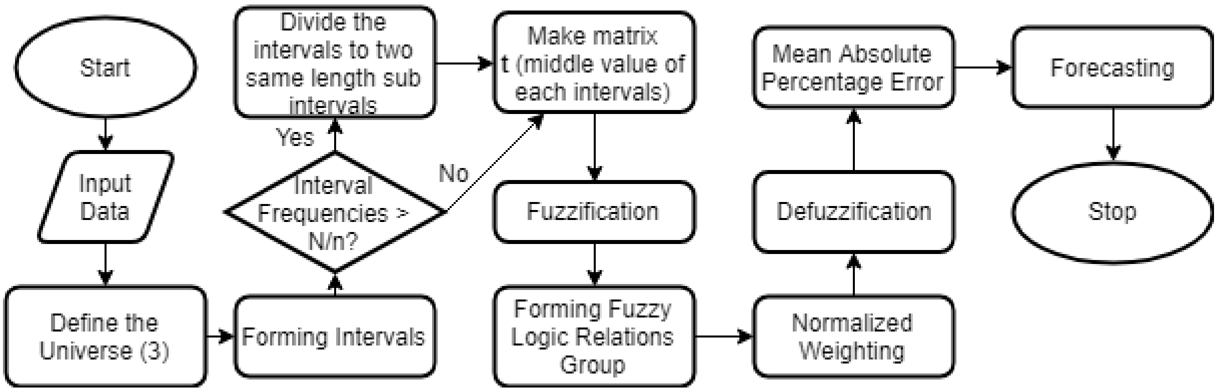
Start
Input Data
Define the Universe (3)
Forming Intervals
Interval Frequencies > N/n?
Yes
No
Divide the intervals to two same length sub intervals
Make matrix t (middle value of each intervals)
Fuzzification
Forming Fuzzy Logic Relations Group
Normalized Weighting
Defuzzification
Mean Absolute Percentage Error
Forecasting
Stop

Figure 1: Proposed Flow Chart

3. Experiment results

3.1. Data set

To examine the method, data was taken from the Jakarta Public Health Office and was based on the results achieved from 2008 to 2018 http://surveilans-dinkesdki.net. There are 132 time series data of number of pneumonia patients per month (see table 1).

Table 1: The Amount of Pneumonia Patients in Jakarta Based on Hospitals from 2008 to 2018

2008	X_i	2009	X_i	And so on to	2018	X_i
January	309	January	199	January	January	910
February	496	February	171	February	February	976
March	445	March	210	March	March	1065
April	276	April	193	April	April	906
May	234	May	160	May	May	757
June	257	June	147	June	June	647
July	212	July	264	July	July	685
August	253	August	290	August	August	693
September	254	September	215	September	September	554
October	260	October	279	October	October	682
November	257	November	172	November	November	730
December	210	December	189	December	December	709

By using **Algorithm 1**, (3), (4), and (5) or **Figure 1**, forecasted from January to May 2019 by conducting calculations and using third order fuzzy time series, as detailed below:

$X_4 = (X_1, X_2, X_3)$ → estimated for April 2008 by using January, February, and March 2008 data
$X_5 = (X_2, X_3, X_4)$ → estimated for May 2008 by using February, March, and April 2008 data
$X_6 = (X_3, X_4, X_5)$ → estimated for June 2008 by using March, April, and May 2008 data
$X_7 = (X_4, X_5, X_6)$ → estimated for July 2008 by using April, May, and June 2008 data
⋮
$X_{132} = (X_{129}, X_{130}, X_{131})$ → estimated for December 2018 by using Sept, Oct, and Nov 2018 data
$X_{133} = (X_{130}, X_{131}, X_{132})$ → forecasted for January 2019
$X_{134} = (X_{131}, X_{132}, X_{133})$ → forecasted for February 2019
$X_{135} = (X_{132}, X_{133}, X_{134})$ → forecasted for March 2019

By using the process from Equation (4) and (5), eight different intervals which are made from our Universe$[147 - D_1, 1065 + D_2]$, with D_1=D_2=3 (random constant) were obtained. However, there are some interval frequencies more than 132/8. Therefore, the intervals were divided into ten different sections.

Let's define the fuzzy sets of linguistic variables membership with Equation (6) [10] and after the process began, the result is shown in **Table 3** Fuzzy Logic Relations Group.

$$A_1 = {}^{1}/_{l_1} + {}^{0.5}/_{l_2} + {}^{0}/_{l_3} + {}^{0}/_{l_4} + {}^{0}/_{l_5} + {}^{0}/_{l_6} + {}^{0}/_{l_7} + {}^{0}/_{l_8} + {}^{0}/_{l_9} + {}^{0}/_{l_{10}}$$

$$A_2 = {}^{0.5}/_{l_1} + {}^{1}/_{l_2} + {}^{0.5}/_{l_3} + {}^{0}/_{l_4} + {}^{0}/_{l_5} + {}^{0}/_{l_6} + {}^{0}/_{l_7} + {}^{0}/_{l_8} + {}^{0}/_{l_9} + {}^{0}/_{l_{10}}$$

$$A_3 = {}^{0}/_{l_1} + {}^{0.5}/_{l_2} + {}^{1}/_{l_3} + {}^{0.5}/_{l_4} + {}^{0}/_{l_5} + {}^{0}/_{l_6} + {}^{0}/_{l_7} + {}^{0}/_{l_8} + {}^{0}/_{l_9} + {}^{0}/_{l_{10}}$$

$$
\begin{aligned}
A_4 &= {}^{0}/_{l_1} + {}^{0}/_{l_2} + {}^{0.5}/_{l_3} + {}^{1}/_{l_4} + {}^{0.5}/_{l_5} + {}^{0}/_{l_6} + {}^{0}/_{l_7} + {}^{0}/_{l_8} + {}^{0}/_{l_9} + {}^{0}/_{l_{10}} \\
A_5 &= {}^{0}/_{l_1} + {}^{0}/_{l_2} + {}^{0}/_{l_3} + {}^{0.5}/_{l_4} + {}^{1}/_{l_5} + {}^{0.5}/_{l_6} + {}^{0}/_{l_7} + {}^{0}/_{l_8} + {}^{0}/_{l_9} + {}^{0}/_{l_{10}} \\
A_6 &= {}^{0}/_{l_1} + {}^{0}/_{l_2} + {}^{0}/_{l_3} + {}^{0}/_{l_4} + {}^{0.5}/_{l_5} + {}^{1}/_{l_6} + {}^{0.5}/_{l_7} + {}^{0}/_{l_8} + {}^{0}/_{l_9} + {}^{0}/_{l_{10}} \\
A_7 &= {}^{0}/_{l_1} + {}^{0}/_{l_2} + {}^{0}/_{l_3} + {}^{0}/_{l_4} + {}^{0}/_{l_5} + {}^{0.5}/_{l_6} + {}^{1}/_{l_7} + {}^{0.5}/_{l_8} + {}^{0}/_{l_9} + {}^{0}/_{l_{10}} \\
A_8 &= {}^{0}/_{l_1} + {}^{0}/_{l_2} + {}^{0}/_{l_3} + {}^{0}/_{l_4} + {}^{0}/_{l_5} + {}^{0}/_{l_6} + {}^{0.5}/_{l_7} + {}^{1}/_{l_8} + {}^{0.5}/_{l_9} + {}^{0}/_{l_{10}} \\
A_9 &= {}^{0}/_{l_1} + {}^{0}/_{l_2} + {}^{0}/_{l_3} + {}^{0}/_{l_4} + {}^{0}/_{l_5} + {}^{0}/_{l_6} + {}^{0}/_{l_7} + {}^{0.5}/_{l_8} + {}^{1}/_{l_9} + {}^{0.5}/_{l_{10}} \\
A_{10} &= {}^{0}/_{l_1} + {}^{0}/_{l_2} + {}^{0}/_{l_3} + {}^{0}/_{l_4} + {}^{0}/_{l_5} + {}^{0}/_{l_6} + {}^{0}/_{l_7} + {}^{0}/_{l_8} + {}^{0.5}/_{l_9} + {}^{1}/_{l_{10}} \qquad (6)
\end{aligned}
$$

Notation "/" does not mean divided, but it means membership of interval l. This is our intervals and the values of t for the middle value in **Table 2**.

Table 2: Intervals and Middle Value

	Lower Bound	Upper Bound	t
l_1	144	201.75	172.875
l_2	201.75	259.5	230.625
l_3	259.5	317.25	288.375
l_4	317.25	375	346.125
l_5	375	490.5	432.75
l_6	490.5	606	548.25
l_7	606	721.5	663.75
l_8	721.5	837	779.25
l_9	837	952.5	894.75
l_{10}	952.5	1068	1010.25

From **Table 3**, 87 different groups were obtained, that will be used to estimate each real data, and to predict the next 5 periods.

Table 3: Fuzzy Logic Relations Group

T	X_T	FLRG
1	309	-
2	496	-
3	445	-
4	276	A3, A6, A5 → A3
5	234	A6, A5, A3 → A2
6	257	A5, A3, A2 → A2
7	212	A3, A2, A2 → A2
8	253	A2, A2, A2 → A2
⋮	⋮	⋮
129	554	A7, A7, A7 → A6
130	682	A7, A7, A6 → A7
131	700	A7, A6, A7 → A7
132	709	A6, A7, A7 → A7

3.2. Graphs

After the entire procedure, an estimation graph and the real data using R and Microsoft Excel as seen in **Figure 2**. It can be seen that the estimated points are closed enough with the real data. To prove it, values of the Mean Absolute Percentage Error (MAPE) with Equation (7) and Root Mean Square Error (RMSE) with Equation is seen (8).

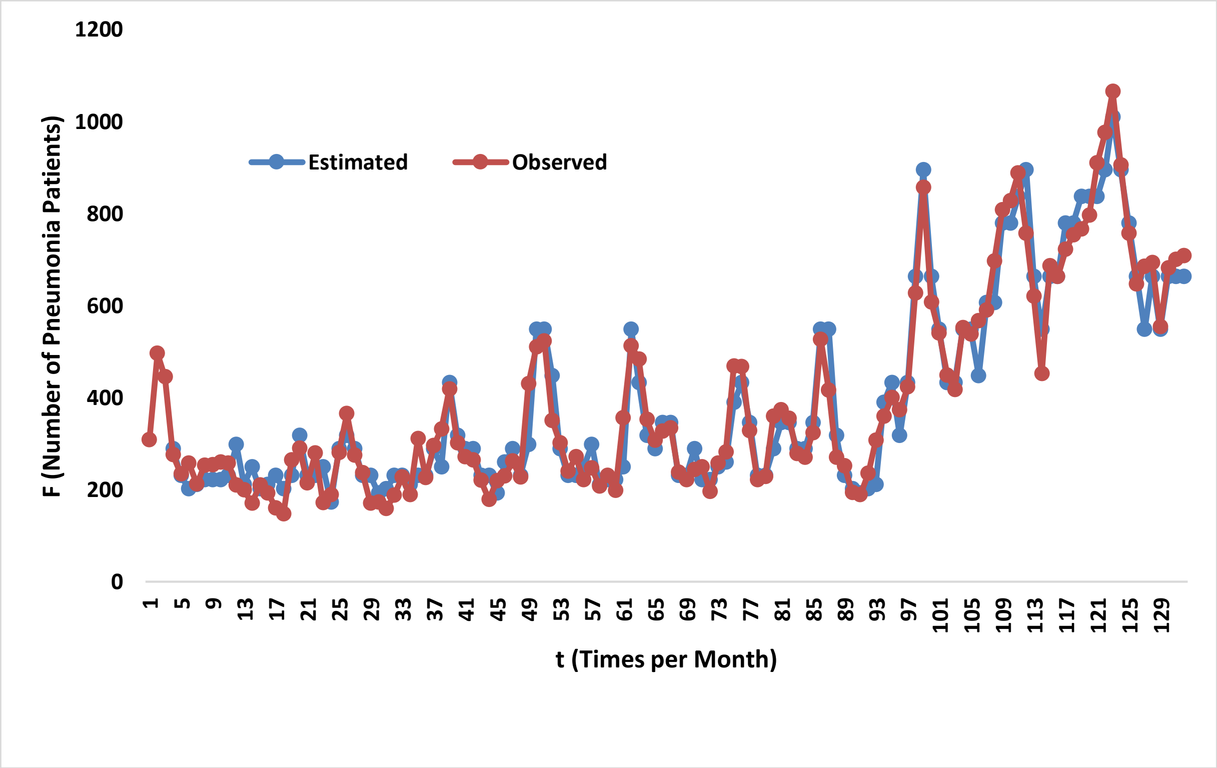

Figure 2: Estimated data (blue line) and real data (red line) plot Time Series

$$MAPE = \frac{100\%}{N} \sum_{i=1}^{N} \left| \frac{X_i - \hat{X}_i}{X_i} \right| \quad (7)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (X_i - \hat{X}_i)^2}{N}} \quad (8)$$

Our MAPE is 9.70% and RMSE is 39.90.

Table 4: MAPE Classification [14]

MAPE VALUE	Accuracy
Below 10%	Highly accurate forecasting
Between 10% to 20%	Good forecasting
Between 20% to 50%	Reasonable
Higher than 50%	Inaccurate forecasting

From **Table 4** it concludes that this method gives high accuracy on forecasting. Now, for forecasting next period, we are ready to use this technique.

3.3. Results

After all the process, the number of pneumonia patients in Jakarta in January, February, March, April, and May 2019 was predicted. The result is as shown in **Table 5**.

Table 5: Forecasting Pneumonia Patients for 5 periods

Time	*Forecasted Pneumonia Patients*
January 2019	548
February 2019	663
March 2019	664
April 2019	663
May 2019	549

However, obtained results in **Table 5** were rounded to the nearest whole number.

3.4. ARIMA and Holt's exponential smoothing results

Convinced with the results, other method used by the researcher will be compared.

3.4.1. ARIMA

ARIMA method will give the result of ARIMA (2, 1, 1) and the model is shown on Equation (9) [6].

$$W_t = Y_t - Y_{t-1} \tag{9}$$

where Y_t is taken from Equation (10)

$$Y_t = 0.8576Y_{t-1} - 0.2923Y_{t-2} + e_t + 0.8395e_{t-1} \tag{10}$$

3.4.2. Holt's Exponential Smoothing

Holt's Exponential Smoothing gave the result in **Table 6**.

Table 6: Holt's Exponential Smoothing

Parameters		Initial States		Sigma
α	β	l	b	
0.9999	0.0001	406.8813	2.2981	90.3001

If α is nearly 1, it means the model is fast learning in the day-to-day movements, while low β means it is slow. Therefore, as mentioned earlier, time series models need some modification, with regards to β. l is the estimation level of the time series data and b of the trend (slope). Sigma is the root of variance.

Therefore, the accuracy of these three methods will be compared with MAPE and RMSE. As shown in **Table 7.**

Table 7: Accuracy Comparison

Method	*MAPE*	*RMSE*
Holt's Exponential Smoothing	18.55%	88.92
ARIMA	16.85%	82.39
Weighted High Order FTS	9.70%	38.88

By the results, we can see that our MAPE and RMSE are smaller than the other method's.

4. Discussion

The increase in the number of smokers and some pollutant sources without adequate regulation will increase the possibility of pneumonia attack. Therefore, it is imperative for the number of pneumonia incidents to be forecasted. There is quite a number of prediction techniques, with most of them used in the financial sector. Some literatures use the ARIMA [16] in forecasting an incidence rate, however, in this research some other method were used in order to obtain a better forecasting. The use of the fuzzy time series by previous scholars inspired the researchers. Holt's method was also utilized because it is mostly used for some trend data. Improvement is recommended because it will produce enhanced accuracy. However, after the third fuzzy times series, it was terminated, because the fourth order will most likely be undefined. The forecasting process is being enhanced because health issues are growing fast [15]. The time series problem cannot always be solved with one method, therefore, using different data, and method, it can be handled adequately. Lastly, this method also doesn't consider causes of the problem, maybe there will be some researches that would carry out studies in this area.

5. Conclusion

From Table 4 and Table 7, it can be seen that the Weighted High Order Fuzzy Time Series method forecasts higher accuracy than the two other methods. However, there are also some strengths and weaknesses associated with this method. The reason it was compared with ARIMA because previous researchers did. Forecasting is an important thing in our lives, and in the medical industry, however, it doesn't focus on medicines alone. Forecasting the amount of patients or incidents can be a benchmark for medical practitioners or even a country to prepare, prevent, and treat health related problems. There are still a good number of new health-related issues that needs to be studied and predicted in order to improve our medical world and help lower mortality rate. With the results, people will be educated on the health implications and danger of pneumonia, and try to prevent it as it tends to be more common in toddlers and the elderly. For educational purposes, it is believed that some researchers will learn more about this method and use it for other good things or compare it with other techniques.

Acknowledgement

This research was financially supported by University of Indonesia, with PITTA B 2019 research grant scheme (ID number NKB-0688/UN2.R3.1/HKP.05.00/2019)

References

[1] Tong, Nga.Priority Medicines for Europe and the World “A Public Health Approach to Innovation.BP 6.22 Pneumonia 1-55 (2013)
[2] Lim, C and Chen, M. Forecasting Emergency Department Admissions for Pneumonia in Tropical Singapore.Online J Public Health Inform 10(1),e12 (2018)
[3] WHO report on the global tobacco epidemic Indonesia (2017)
[4] Sriwattanapongse, Wattanavadee, et al. Forecasting the Monthly Incidence Rate of Pneumonia in Mae Hong Son Province, Thailand. Chiang Mai Medical Jorunal 48(3) 85-94 (2009)
[5] Holt, C.C. Forecasting Seasonals and Trends by Exponentially Weighted Moving Averages. International Journal of Forecasting 20(1) 5-10 (2004)
[6] Box, G.E., Jenkins, G.M., Reinsel, G.C., Ljung, G.M.: Time Series Analysis: Forecasting and Control. Wiley, Hoboken (2015)
[7] Song, Q., Chissom, B.S. Forecasting enrollments with fuzzy time series-part i. Fuzzy Sets Syst. 54(1), 1–9 (1993)
[8] Chen, S.: Forecasting enrollments based on fuzzy time series. Fuzzy Sets Syst. 81(3), 311–319 (1996)
[9] Yu, H.K.Weighted fuzzy time series models for TAIEX forecasting. Physica A 349(3-4): 609-624 (2005)
[10] Cheng,C.H., et al. Fuzzy time-series based on adaptive expectation model for TAIEX forecasting. Expert Systems with Applications 34(2): 1126-1132 (2008)
[11] Ruchiraset, A. and Tantrakarnapa, K. Time series modeling of pneumonia admissions and its association with air pollution and climate variables in Chiang Mai Province, Thailand.Eviromental Science and Pollution Research 25(33) 33277-33285 (2018)
[12] Cryer, J.D. and Chan, K-S.Time Series Analysis with Applications in R Second Edition (2008)

[13] Rustam, Z. and Talita, A.S. Fuzzy Kernel C-Means Algorithm for Intrusion Detection Systems. JATIT&LLS, Vol. 81. 10th November, (2015)

[14] Moreno, J.J.M., et al. Using the R-MAPE index as a resistant measure of forecast accuracy. Psicothema 25(4): 500-506 (2013)

[15] Teel, Prinsez.Improving the Accuracy of Healthcare Forecasting. (2018)

[16] Soleh, A.M., et al. A Comparative Simulation Study of ARIMA and Fuzzy Time Series Model for Forecasting Time Series Data. IJRSET 4(11) 49-56 (2018)