Proceeding International Joint Conference on Science and Technology (IJCST) 2017

ISBN : 978-602-99806-3-9

Price of Rice Forecasting Based on Pattern Similarity

E.D. Wahyuni[1] and M.I. Afandi[2]

[1,2] Information System, Faculty of Computer Science, Pembangunan Nasional "Veteran" Jawa Timur. Surabaya, East Java, Indonesia

E-mail: ekawahyuni.si@upnjatim.ac.id, mohamad.afandi@gmail.com

Abstract. This paper implement PSF to forecast price of rice based on similarity of pattern sequences. First, clustering techniques are used with the aim of grouping and labeling the samples from a data set. Thus, the prediction of a data point is provided as follows: first, the pattern sequence prior to the day to be predicted is extracted. Then, this sequence is searched in the historical data and the prediction is calculated by averaging all the samples immediately after the matched sequence. Results from several price time series are reported and the performance from PSF is compared to that of statistic based techniques (AHW, MHW, ARIMA and ETS). From the result, AHW method shows better accuracy in prediction.

1. Introduction

Rice is one of the staple foods consumed by the people of Indonesia. The price of rice on the market is quite volatile, as evidenced by the data obtained from the siskaperbapo website which contains the average price data of basic ingredients at the consumer level for East Java province. People on the one hand, as consumers of rice, are very concerned with stable staple food prices. Stable in the sense of not experiencing a drastic increase for each month. This can not be separated from the fact that the volatility of a food price has a wide impact and often extends to other dimensions (Surachman, et al., 2009). The government, as the policymaker, needs to formulate and implement a price stabilization policy. In order to formulate this policy, information on market projections is required, in this case is price forecasting at the retail level.

Forecasting is a prediction for the future state (Mart'ınez-Alvarez et al). This forecasting can be short term, medium and long term. Forecasting that will be done in this study is forecasting short term (short term). To do this forecasting, there are many methods. Broadly speaking, this method can be categorized based on surveys (customer surveys, market surveys etc) and statistical based (MA, ARIMA, regression etc). Along with the number of data generated in the last 20 years, data mining-based forecasting have been developed. Several studies have proved that data mining-based methods have better accuracy than other methods.

This study aims to compare statistical forecasting methods with data mining-based methods for rice commodities. Of the many statistical methods, the preferred method is holt winters, ARIMA and ETS. Holt winters are better suited to predict the price of rice, corn, and soybeans because there are seasonal patterns in the data (Surachman, et al., 2009 and Ishaque & Ziblim, 2013). The model test results also show if holt winters are more appropriate, because the error value is smaller when compared to other methods. ARIMA and ETS used for comparison methods in this study to recreate the test environment already done by Bokde, et al., (2016). For data mining-based methods, this study uses the PSF algorithm. PSF stands for Pattern Sequence Forecasting algorithm. PSF is a successful forecasting technique based on the assumption that there exist pattern sequences in the target time series data. For the first time, it was proposed in Martínez-Álvarez et al. (2008), and an improved version was discussed in Martínez-Álvarez et al. (2011).

ISBN : 978-602-99806-3-9

2. Methodology

The PSF algorithm can be divided in two steps. The first step is clustering of data and the second step is forecasting based on clustered data in earlier step. The block diagram of PSF algorithm shown in Figure 1 was proposed by Martínez-Álvarez et al. (2008).

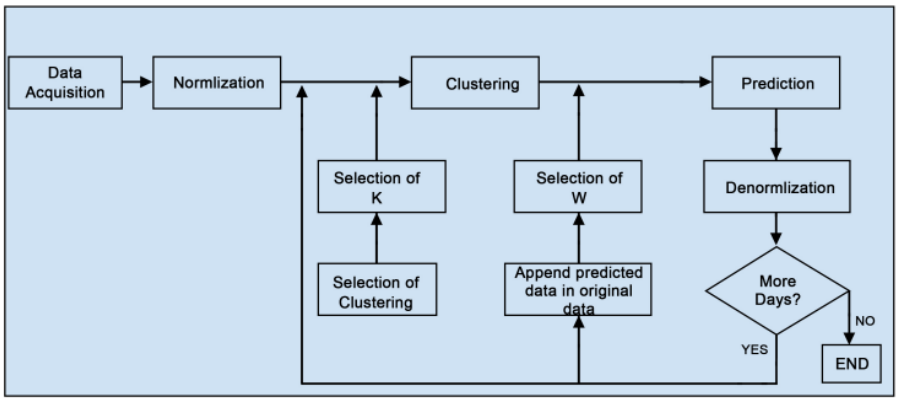

Figure 1. Block diagram of SPF algorithm

2.1. *Clustering*

Martínez-Álvarez et al., (2008) and Martínez-Álvarez et al., (2011) chose the k-means clustering technique for generating data clusters according to the time series data properties. The limitation of k-means clustering technique is that the adequate number of clusters must be provided by users. In Martínez-Álvarez et al. (2011), multiple indexes (Silhouette index, Dunn index and the Davies–Bouldin index) were considered to determine the optimum number of clusters. As output of the clustering process, the original time series data is converted into a series of labels, which is used as input in the prediction block of the second phase of the PSF algorithm.

2.2. *Prediction*

The prediction technique consists of window size selection, searching for pattern sequences and estimation processes. Let x(t) be the vector of time series data such that x(t) = [x1(t), x2(t), . . . , xN(t)]. After clustering and labeling, the vector converted to y(t) = [L1, L2, . . . , LN], where Li are labels identifying the cluster centers to which data in vector x(t) belongs to. Note that every xi(t) can be of arbitrary length and must be adjusted to the pattern sequence existing in every time series. For instance, in the original work, x(i) was composed of 24 values, representing daily patterns. Then the searching process includes the last W labels from y(t) and it searches for these labels in y(t). If this sequence of last W labels is not found in y(t), then the search process is repeated for last (W-1) labels.

In PSF, the length of this label sequence is named as window size. Therefore, the window size can vary from W to 1, although it is not usual that this event occurs. The selection of the optimum window size is very critical and important to make accurate predictions. The optimum window size selection is done in such a way that the forecasting error is minimized during the training process. Mathematically, the error function to be minimized is:

$$\sum_{t \in TS} \| \bar{X}(t) - X(t) \| \quad (1)$$

where $\bar{X}(t)$ are predicted values and $X(t)$ are original values of time series data. In practice, the window size selection is done with cross validation. All possible window sizes are tested on

ISBN : 978-602-99806-3-9

sample data and corresponding prediction errors are compared. The window size with minimum error considered as the optimum window size for prediction. Once the optimum window size is obtained, the pattern sequence available in the window is searched for in $y(t)$ and the label present just after each discovered sequence is noted in a new vector, called ES. Finally, the future time series value is predicted by averaging the values in vector ES.

3. Result and discussion

3.1. *Dataset*

This study used R Language, dataset obtained from siskaperbapo.com for IR64 of rice commodity from Pasar Pucang Anom in Surabaya, East Java. The data obtained is daily data from January 1, 2013 to December 31, 2015 which are stored in csv format. To represent the data in plot format, the plot() function from forecast package is used.

```
> train = read.csv("data.csv", colClasses = c("Date",
"integer"), header = FALSE)
> plot(x = train$V1, y = train$V2, type = "o", xlab =
"periode", ylab = "harga", col="brown", ylim = c(7000,
9000))
```

Figure 2. R code snippet for create and plotting train data

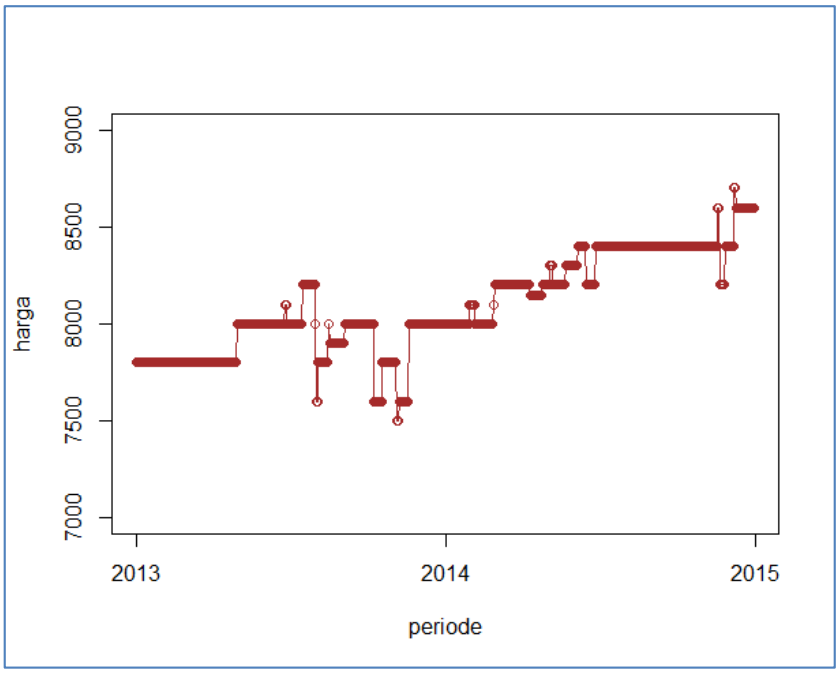

Figure 3. Result of plot() from train data

By plotting the data, it clearly shows that the price of rice from the years 2013-2015 tends to increase every year.

3.2 Forecast using PSF

ISBN : 978-602-99806-3-9

This process is performed by calling function psf () in the PSF library. This function receives univariate ti me series as input. The most optimal parameters for k and w are used for this model is 4 and 10

```
> psf_model <- psf(train$v2,
k = 8, w = 10 )
> a <-
(psf_model, n.ahead = 48)
```

Figure 4. R code snippet for forecast using PSF

3.3 Comparison Forecast Method

In order to produce accurate predictions, test on many methods are required. Each method will predict the price of rice for n period ahead, in this case for 48 days ahead. This study tested 5 different methods, ie Pattern Sequence Forecast (PSF), Multiplicative Holt Winter (MHW), Additive Holt Winter (AHW), ARIMA and Exponential Smoothing (ETS). To execute those methods, PSF and Forecast package are required. Result from prediction will be compared with the actual data to calculate the accuracy of the method. To calculate the accuracy, this study used accuracy function of library MLmetrics.

```
> accuracy(a,test$v2)
```

Figure 5. R code snippet for calculating accuracy

The same function is used to calculate the accuracy of MHW, AHW, ARIMA and ETS. The results of these accuracy calculations are shown in the following table.

Table 1. Accuracy table

	PSF	MHW	AHW	ARIMA	ETS
ME	963.9583	517.6445	509.7192	547.1633	531.25
RMSE	998.1256	578.605	569.6469	606.3455	593.5416
MAE	963.9583	517.6445	510.0372	547.1633	531.25
MPE	10.48152	5.590236	5.504617	5.912042	5.736513
MAPE	10.48152	5.590236	5.508314	5.912042	5.736513

From the table above, it is clearly shown that the AHW method has good accuracy. This is indicated by lower value of ME, RMSE, MAE, MPE and MAPE compared to other methods. In this study it is also shown that the accuracy of PSF is the worst compared to other methods. This result is clearly contrary to the result of the experiment conducted by Mart'ınez-Alvarez et al (2008), which states that the PSF algorithm outperforms the other algorithm, because its accuracy is the best among other methods. The comparison of each prediction results, is more clearly shown in the following figure

ISBN : 978-602-99806-3-9

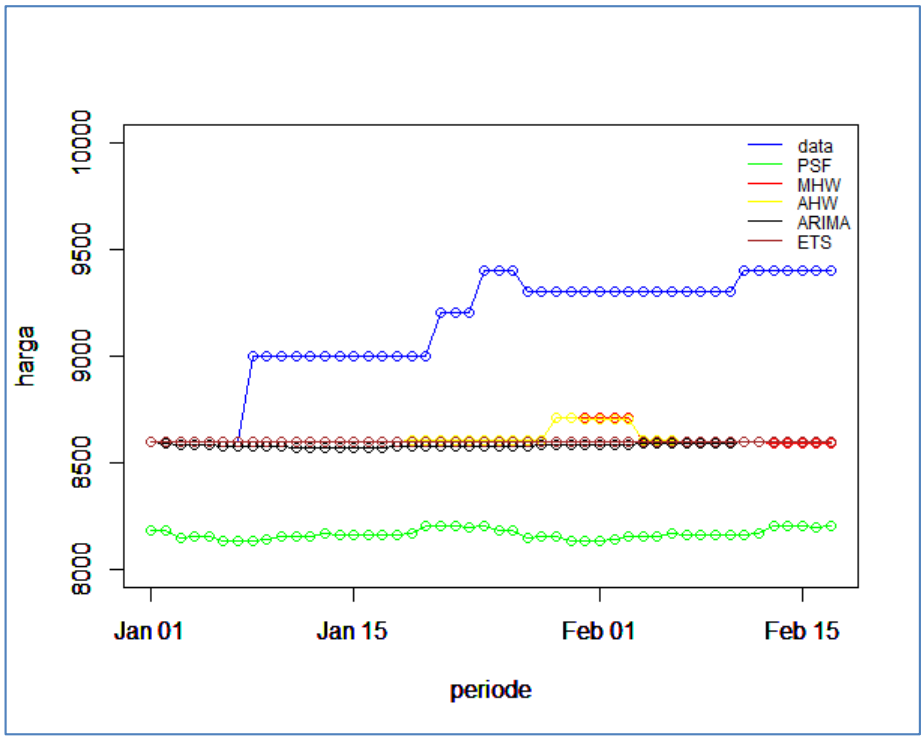

Figure 6. Plot of prediction result comparison

From the plot above, it is clear that forecasting from the PSF has the worst performance compared to the others, and the forecasting of AHW is close to reality. The cause of different results between this study and other research is probably due to the lack of data used for training (using only daily data for 2 years), the second cause because, this forecast still does not consider any local holidays in Indonesia. From the test result, the method that will be implemented to the system is AHW method, because its accuracy is the best compared to other methods.

4. Conclussion

From the above experiments, it can be concluded that, a forecast method, may not show the same performance in different cases, as in the example above, PSF has good accuracy, if it's applied to determine the daily electricity tariff rates in certain countries, done by Martínez-Álvarez et al (2007). The same method, when applied to different data and cases, ie to forecast rice prices, did not result the expected performance, in this study, PSF shown the worst performance. The results of this study support the results of research conducted by Surachman, et al., (2009) and Ishaque & Ziblim, (2013) which concluded that holt winter is suitable to predict rice commodity prices.

5. References

[1] Martínez Álvarez, F., Troncoso, A., Riquelme, J. C. & Riquelme, J. M., 2007. Discovering Patterns in Electricity Price Using Clustering Techniques. *RE&PQJ*, 1(5).

[2] Bokde, N. D., Martínez-Álvarez, F., Cortés, G. A. & Kulat, K. D., 2016. PSF: Introduction to R Package for Pattern Sequence Based Forecasting Algorithm. *arXiv preprint arXiv*, p. 1606.05492.

ISBN : 978-602-99806-3-9

[3] Ishaque, M. & Ziblim, S.-D., 2013. Use of some exponential smoothing models in forecasting some food crop prices in the upper east region of Ghana.. *Mathematical Theory and Modeling*, pp. 2224-5804.

[4] Kumar, M. & Patel, N. R., 2010. Using clustering to improve sales forecasts in retail merchandising. *Ann Oper Res* , Volume 174, p. 33–46.

[5] Surachman, H. et al., 2009. *Kajian Pengembangan Model Proyeksi Jangka Pendek Pasar Komoditas Pangan Pokok*, Jakarta: Pusat Penelitian dan Pengembangan Perdagangan Dalam Negeri, Badan Penelitian dan Pengembangan Perdagangan, Kementerian Perdagangan.