The development rainfall forecasting using kalman filter

Mohammad Zulfi¹, Moh.Hasan¹, Kosala Dwidja Purnomo¹

¹Department of Mathematics, University of Jember, Indonesia

E-mail: mohammadzulfi94@yahoo.com, hasan@unej.ac.id, kosaladp.fmipa@unej.ac.id

Abstract. Rainfall forecasting is very interesting for agricultural planning. Rainfall information is useful to make decisions about the plan planting certain commodities. In this studies, the rainfall forecasting by ARIMA and Kalman Filter method. Kalman Filter method is used to declare a time series model of which is shown in the form of linear state space to determine the future forecast. This method used a recursive solution to minimize error. The rainfall data in this research clustered by K-means clustering. Implementation of Kalman Filter method is for modelling and forecasting rainfall in each cluster. We used ARIMA (p,d,q) to construct a state space for KalmanFilter model. So, we have four group of the data and one model in each group. In conclusions, Kalman Filter method is better than ARIMA model for rainfall forecasting in each group. It can be showed from error of Kalman Filter method that smaller than error of ARIMA model.

1. Introduction

Rainfall information play an important role in agricultural planning. Therefore, the success of an agricultural plan depends on the availability of good rainfall forecasting. Rainfall data can be analyzed using time series analysis. Time series forecasting is an important area of forecasting in which past observations of the same variable are collected and analyzed to develop a model describing the underlying relationship [5]. Kalman filter is one of several methods that can be used for time series forecasting.

The Kalman filter is essentially a set of mathematical equations that implement a predictor-corrector type estimator that is optimal in the sense that it minimizes the estimated error covariance when some presumed conditions are met[4].The Kalman Filter method is used to represent a time series model that is displayed in linear state space[1]. Kalman Filter is an estimator that serves to estimate the state of output. The advantages of the Kalman Filter method are easy to apply in various sciences because they are recursive [4]. While the disadvantages of the Kalman Filter method is the success in getting optimal predictive results depending on the accuracy in estimating the initial state on the latest observation data [5]. In this research we used ARIMA model to construct a state space model in Kalman Filter.

ARIMA forecasting model is a forecasting technique that does not require specific assumptions about historical data, but uses an iterative model to determine the best model. The selected model will be evaluated with historical data whether it has accurately described the data. The best models will produce relatively small residuals, independent and have random distributions.

This research aim to know ARIMA model and state space model in rainfall data. then we will see how the comparison of forecasting between Kalman Filter and ARIMA.
2. Basic Concept of Kalman Filter

Kalman Filter is a recursive procedure used to forecast state vectors. This method uses a recursive technique to integrate the latest observational data into a model to correct previous predictions and predict subsequently optimally based on past data information[3]. The Kalman Filter concept consists of two stages. The first stage is the forecasting stage. The second stage is the renewal stage. In the forecasting stage, an estimated value is generated for the current state and the value of the error covariance used as the initial estimation information for the next step. The renewal stage serves as a corrector [4]. The correction value is determined by how well the initial guess predicted the new observation [2].

The general equation of the Kalman Filter model is given as the vector of size (n x 1) variables observed at time t, the unobserved dynamic model for $\bar{X}_t$ can be explained $\bar{Z}_t$ on the vector of size (r x 1). The space state representation for $y$ is given in the following equation [2]:

\[ \bar{Z} = \begin{pmatrix} F \end{pmatrix} \bar{Z}_t + \begin{pmatrix} G \end{pmatrix} \bar{a}_{t+1} \]

\[ \bar{X}_t = H\bar{Z}_t + \bar{b}_t \]

where:
- $F = (3)$ and (4)
- $\bar{Z}$: state vector size (r x 1) in state equation
- $F$: transition matrix size (r x r)
- $G$: input matrix
- $H$: matrix parameters size (n x r) in observation equation
- $\bar{a}$: noise vector size (r x 1) in state equation
- $b$: noise vector size (n x 1) in observation equation
- $\bar{X}$: vector of the observed variables size (n x 1)

Forecasting uses Kalman filter through two stages. The first stage is the prediction stage (time update) and the next stage is measurement update. All the steps can be seen in table 1

| Table 1. Kalman Filter Stages |
|-----------------------------|
| construction of state space | 1) state equation |
|                            | $\bar{Z}_{t+1} = F\bar{Z}_t + G\bar{a}_{t+1}$ |
|                            | 2) output equation |
|                            | $\bar{X}_t = H\bar{Z}_t + \bar{b}_t$ |
| prediction stage           | 1) estimation of state equations |
|                            | $\hat{Z}_{t+1} = F\bar{Z}_t + G\hat{a}_{t+1}$ |
|                            | 2) covarian error |
|                            | $R_{t+1} = F\Gamma_t F' + G\Sigma G'$ |
| Measurement update          | 1) determine kalman gain matrix |
|                            | $R_{t+1} = F\Gamma_t F' + G\Sigma G'$ |
|                            | 2) update estimation |
|                            | $\hat{Z}_{t+1} = \hat{Z}_{t+1} + K_{t+1} (\bar{X}_t - H\bar{Z}_{t+1})$ |
|                            | 3) covarian error |
|                            | $\Gamma_{t+1} = (I - K_{t+1}H)R_{t+1}$ |
3. Methods

In this research we used rainfall data of Kabupaten Jember which have been grouped based on K-mean clustering. In general, this data is divided into two parts, namely data insampel that began from January 2005 to December 2015 and data outsampel which starts from January 2016 until December 2016. The insampel data is the data used to construct the ARIMA model in each cluster, while the outsampel data is used to test how well the estimation results obtained. The stages in this research are as follows:

1. Plotting Data
   Plotting data is an important part of the analysis. In addition to viewing data patterns, plotting also provides an overview of stationary data in variance. The basic assumption that must be met is the stationary data in mean and variance. If the time series data plot does not show any apparent change in variance over time, it can be said that the time series data is stationary in its variance.

2. Augmented Dickey Fuller (ADF) test
   Augmented Dickey Fuller (ADF) test is used to see the stationarity of the data in the mean. If the time series data has been stationary in its mean and variance, then the data can be used to construct the ARIMA model. But if the data is not stationary then need to be done differencing transformation process.

3. ARIMA Forecasting Model
   After passing the stationary test, the insampel data can be used to construct the ARIMA forecasting model. Since the data is divided into four clusters, we will construct a model for each cluster. In the end we have four ARIMA models that may all be different.

4. Forecasting using Kalman Filter
   Calculation of Kalman Filter recursive equation includes:
   a. time update
   b. measurement update

5. Comparison of Forecasting Kalman Filter and ARIMA
   After the forecasting results are produced, the forecasting results are compared with the outsampel data. From this process we will get the RMSE value on each forecasting model. Small RMSE values indicate that the model is better than the other model.

4. Results

4.1 Plotting Rainfall Data
   Figure 1 is a graph of monthly rainfall time series data in 4 groups. The plot shows the amount of rainfall from January 2005 to December 2016. The amount of rainfall in the form of graph shows that there is an indication of season time series model with time series pattern that repeats at certain time and seen the pattern of up and down regularly formed from plot of rainfall data. In addition, the graph indicates a correlation that is time-sequence associated with the number of observations per season period. Therefore, before estimating the model, the data is tested for the mean and variance. To determine the stationarity of the data on the variance can be seen descriptively plot time series in Figure 1 which shows that the data over time have fixed or constant data fluctuations and are constantly changing on all clusters, so that the data on each cluster has been stationary variance.
Figure 1 graph of monthly rainfall time series data

Figure 1 is a graph of monthly rainfall time series data in 4 groups. The plot shows the amount of rainfall from January 2005 to December 2016. The amount of rainfall in the form of graph shows that there is an indication of season time series model with time series pattern that repeats at certain time and seen the pattern of up and down regularly formed from plot of rainfall data. In addition, the graph indicates a correlation that is time-sequence associated with the number of observations per season period. Therefore, before estimating the model, the data is tested for the mean and variance.

4.2 Augmented Dickey Fuller (ADF) Test

The stationarity in mean test is done by using Augmented Dickey Fuller (ADF) test. Based on the results of the ADF test (table 2), it can be seen that the four locations have been stationary without going through differencing process, which is indicated by the \( p \)-value is less than 0.05.

| Cluster | ADF-Test | \( p \)-value |
|---------|----------|--------------|
| 1       | -7.6688  | 0.01         |
| 2       | -7.3469  | 0.01         |
| 3       | -7.2546  | 0.01         |
| 4       | -8.0086  | 0.01         |

4.3 ARIMA Forecasting Model

In this step we will construct the ARIMA model for each cluster. The ARIMA model is required to construct the state space model before making a forecasting using Kalman Filter. \( \hat{Y}_{C1}, \hat{Y}_{C2}, \hat{Y}_{C3} \) and \( \hat{Y}_{C4} \) respectively are ARIMA models in the first cluster, second cluster, third cluster and fourth cluster. The ARIMA equation is as follows:

\[
\begin{align*}
Y_{C1} &= 0.917Y_{t-1} - 0.145Y_{t-2} - 0.12Y_{t-3} - 0.046Y_{t-4} - 0.201Y_{t-5} + a_t + 0.502a_{t-1} \\
Y_{C2} &= 1.5538Y_{t-1} - 0.7702Y_{t-2} + a_t + 0.8567a_{t-1} \\
Y_{C3} &= 1.102Y_{t-1} - 0.377Y_{t-2} - 0.308Y_{t-3} - 0.028Y_{t-4} - 0.195Y_{t-5} + a_t + 0.56a_{t-1} \\
Y_{C4} &= 0.99Y_{t-1} - 0.279Y_{t-2} - 0.080Y_{t-3} + 0.0197Y_{t-4} - 0.252Y_{t-5} + a_t + 0.61a_{t-1}
\end{align*}
\]
In this section we used auto.arima to get the best arima model on every cluster. Auto.arima can be used if the data has been stationary in mean and variance. Equation (10), (12) and (13) shows that there is a similarity of orders between models in the first cluster, third cluster and fourth cluster. On the other hand, in the second cluster obtained the best model of ARIMA (2.0,1). The equation (10), (11), (12) and (13) will be used to construct the state space model in Kalman Filter. However, this equation is actually a model forecasting of time series data for each cluster. Then we have two forecasts on each cluster that is Kalman Filter forecasting and ARIMA forecasting.

4.4 Forecasting using Kalman Filter

Forecasting using Kalman Filter and ARIMA method is shown in table 3 below.

| Time  | First Cluster | Second Cluster | Third Cluster | Fourth Cluster |
|-------|---------------|----------------|---------------|---------------|
| Jan-16 | 174.25        | 218.37454      | 323           | 256           | 327.7332       | 218.37454 | 264.07617 |
| Feb-16 | 230.076       | 242.66848      | 312.536       | 398.0193      | 355.44275      | 242.66848 | 279.33621 |
| Mar-16 | 199.46937     | 241.58537      | 342.427       | 320.5746      | 351.4885       | 241.58537 | 270.94149 |
| Apr-16 | 112.456       | 215.38891      | 290.3753      | 248.9465      | 299.27477      | 215.38891 | 241.23351 |
| May-16 | 181.3746      | 162.20481      | 232.475       | 168.1265      | 216.018        | 162.20481 | 174.63679 |
| Jun-16 | 139.2501      | 107.33191      | 214.572       | 148.5362      | 143.50329      | 143.50329 | 106.1033  |
| Jul-16 | 78.46288      | 63.026         | 185.273       | 79.3547       | 81.34673       | 79.3547   | 54.62466  |
| Aug-16 | 36.142        | 38.1777        | 174.8274      | 128.7009      | 28.599         | 22.275925 | 162.20481 |
| Sep-16 | 51.284        | 36.15553       | 290.3753      | 137.5041      | 74.9736        | 137.5041  | 30.03783  |
| Oct-16 | 136.0947      | 56.49987       | 237.309       | 158.3284      | 103.647        | 103.647   | 57.52409  |
| Nov-16 | 190.2093      | 91.55585       | 473.5847      | 183.9056      | 259.4637       | 183.9056  | 103.10807 |
| Dec-16 | 285.394       | 131.08579      | 383.4096      | 207.6094      | 222.301        | 207.6094  | 153.40549 |

Forecasting with the Kalman Filter method is obtained after we construct the state space model as in (1). For example we use ARIMA model in second cluster. Based on the equation (11) then obtained parameters of the state space model as follows:

\[ \tilde{Z}_{t+1} = \begin{bmatrix} \frac{1}{1.5538} & 1 & 0 & \frac{Z_{t-2}}{Z_{t-1}} \\ -0.7702 & 1 & 0 & \frac{Z_{t-2}}{Z_{t-1}} \\ 0 & 1 & 0 & \frac{Z_{t-2}}{Z_{t-1}} \\ 0 & 0 & 1 & \frac{Z_{t-2}}{Z_{t-1}} \end{bmatrix} \begin{bmatrix} \frac{a_{t+1}}{a_t} \\ -0.8567 \\ b_{t-2} \\ b_{t-1} \end{bmatrix} \]  

\[ \tilde{X}_t = \begin{bmatrix} 1 & 0 & 0 & \frac{Z_{t-2}}{Z_{t-1}} \\ 0 & 1 & 0 & \frac{Z_{t-2}}{Z_{t-1}} \\ 0 & 0 & 1 & \frac{Z_{t-2}}{Z_{t-1}} \end{bmatrix} + \begin{bmatrix} b_{t-2} \\ b_{t-1} \end{bmatrix} \]  

This process also applies to all clusters. The state space model depends on the ARIMA order size and the value of each parameter.

4.5 Comparison of Forecasting Kalman Filter and ARIMA

To see the accuracy of forecasting using Kalman Filter and ARIMA, then we compare the forecast result in table 3 with outsampel data. One of the indicators we use is Root Mean Square Error (RMSE) which is expressed by the following equation.

\[ RMSE = \sqrt{\frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{n}} \]
Based on the forecast result on each cluster, we get the RMSE value shown in table 4.

| Number Of Cluster | METHOD | RMSE  |
|-------------------|--------|-------|
| 1                 | ARIMA  | 106.9148 |
|                   | KF     | 62.38718 |
| 2                 | ARIMA  | 144.1151 |
|                   | KF     | 89.61118 |
| 3                 | ARIMA  | 107.6135 |
|                   | KF     | 70.55476 |
| 4                 | ARIMA  | 128.6514 |
|                   | KF     | 86.30452 |

The RMSE value forecasting using the Kalman Filter method is smaller than the ARIMA model. This is shown in table 4. In general, forecasting using the Kalman Filter method is better than ARIMA. Table 4 shows the relationship between RMSE value in Kalman Filter and RMSE value in ARIMA. the RMSE value in Kalman Filter depends on the RMSE value of the ARIMA model. Forecasting using Kalman Filter in the first cluster produces a smaller RMSE value than the other clusters. Similarly, forecasting using ARIMA. Forecasting using ARIMA model also applies the same thing.

![Figure 2. Plotting of forecasting results](image)

5. Conclusion
In general, this study concluded that:

1. The Kalman Filter method gives better results compared to the ARIMA model. This is indicated by the RMSE value in Kalman Filter is smaller than ARIMA model. Based on Figure 2 we can conclude that the error of forecasting using Kalman Filter is relatively
unchanged significantly at any time. This is due to the measurement update process at each time of forecasting. On the other hand, the error of ARIMA model is getting bigger.

2. Forecasting results using the Kalman Filter method depends on the ARIMA model used as the state space model. The result is an error of the Kalman Filter method depending on the ARIMA model.

References
[1] Brocwell P J and Davis R A 1991 Time Series: Theory and Methods Second Edition (New York: Springer-Verlage, Inc)
[2] Meinhold R J and Singpurwala N D 1983 Understanding The Kalman Filter American Statistical Association 37 (2) 123-127
[3] Wei W W S 1989 Time Series Analysis: Univariate and Multivariate Methods (Canada. Addison Wesley Publishing Company)
[4] Welch G and Bishop G 2006 An Introduction to the Kalman Filter (ACM In)
[5] Zhang G P 2001 Time Series using Hybrid ARIMA and Neural Network Model Neurocomputing