JOURNAL OF STATISTICAL METHODS AND DATA SCIENCE

Journal of Statistical Methods and Data Science
Volume 01 Nomor 1 June 2023
https://jurnalmipa.unri.ac.id/jsmds

KALMAN FILTER METHOD FOR HANDLING MISSING VALUES IN SOIL MOISTURE SENSOR DATA

Fika Reski Amaliah[1*], Rais[2], Iman Setiawan[3], Hartayuni Sain[4]
[1,2,3,4] Statistic Study Program, Tadulako University, Sulawesi Tengah, 94148, Indonesia

***e-mail**: npl.untad@gmail.com

Article Info:
Received: 02-06-2023
Accepted: 17-06-2023
Available Online: 30-06-2023

Keywords:
imputation
kalman filter
MAPE
soil moisture

Abstract: *Several imputation techniques have been developed specifically to deal with missing values. This research used data from soil moisture sensors for 34 days where there are missing values. Size of soil moisture sensor data to be quite large. So that with the missing value, it is difficult to determine how well the imputation technique is applied. Therefore, imputation technique is performed on the generated data based on the distribution of soil moisture sensor data so that an evaluation of the utilization of the imputation technique can be carried out on large data containing missing values. The method used in this study is Kalman Filter. Evaluation using the Mean Average Percentage Error (MAPE) after intentionally removed using the Missing Completely at Random (MCAR) technique with a missing rate of 5%, 10% and 15%. The results showed that soil moisture data has a Box-Cox power exponential distribution. It was found that for generating data, Kalman Filter method has not much different MAPE value for 100 and 1000 data with missing rate was 5%, 10%, and 15%. The results of estimating missing values with the Kalman Filter on the soil moisture sensor data are in line with the soil moisture sensor data.*

1. INTRODUCTION

Primary data often has poor quality, one of the problems is missing values. Missing values is information that is not available for a case. Missing values can occur because the information needed for something in one or several variables is not provided, it is difficult to find or indeed the information is not available. There are several reasons why the data may be lost, including because the equipment is not working, bad weather, the person being observed is sick, or the data is not entered correctly. Previous research by Setiawan et al. 2022, regarding the manufacture of an Arduino-based automatic watering device using a soil sensor to measure the moisture level in the soil, has missing values. The success of this automatic watering device depends on how the sensor detects soil moisture as seen from the growth of the red onion and the weight of the red onion after harvest.

Several methods have been developed specifically to deal with missing values. One such method is the imputation technique. Data imputation is carried out by filling in missing or problematic data based on the results of the process of re-checking the reasonableness of the data or the consistency between variables. In addition, imputation is also carried out as a treatment for outliers in an effort to improve data quality (Biemer & Lyberg, 2003).

There are several imputation methods that can be used for univariate data such as Kalman Filter. The Kalman Filter is a method that is part of the state space that can be applied in estimating statistical models. According to Wie, 2006, Kalman Filter uses a recursive technique in integrating the latest observational data into the model to correct previous predictions and optimally make further predictions based on past data information and current data information. Previous research on the use of Kalman Filters and sensor data is Distributed Filtering for Multi-sensor Systems with Missing Data by Jin & Sun, 2022, and utilization of the discrete Kalman Filter method to estimate air temperature by Tengger & Ropiudin, 2019 .

Soil moisture sensor data has a fairly large size. Missing values on sensor data is also randomly distributed. Therefore, this study uses generated data in order to be able to evaluate the use of the Kalman Filter method in dealing with missing values at a fairly large data size.

2. MATERIALS AND METHODS

2.1. Soil Moisture Sensor Data

Setiawan et al. 2022 conducted an experiment on an automatic plant watering system on red onion plants for 34 days. The watering system uses a soil moisture sensor that is stored for every second. However, the data from the soil moisture sensor contains a missing value as shown in Figure 1.

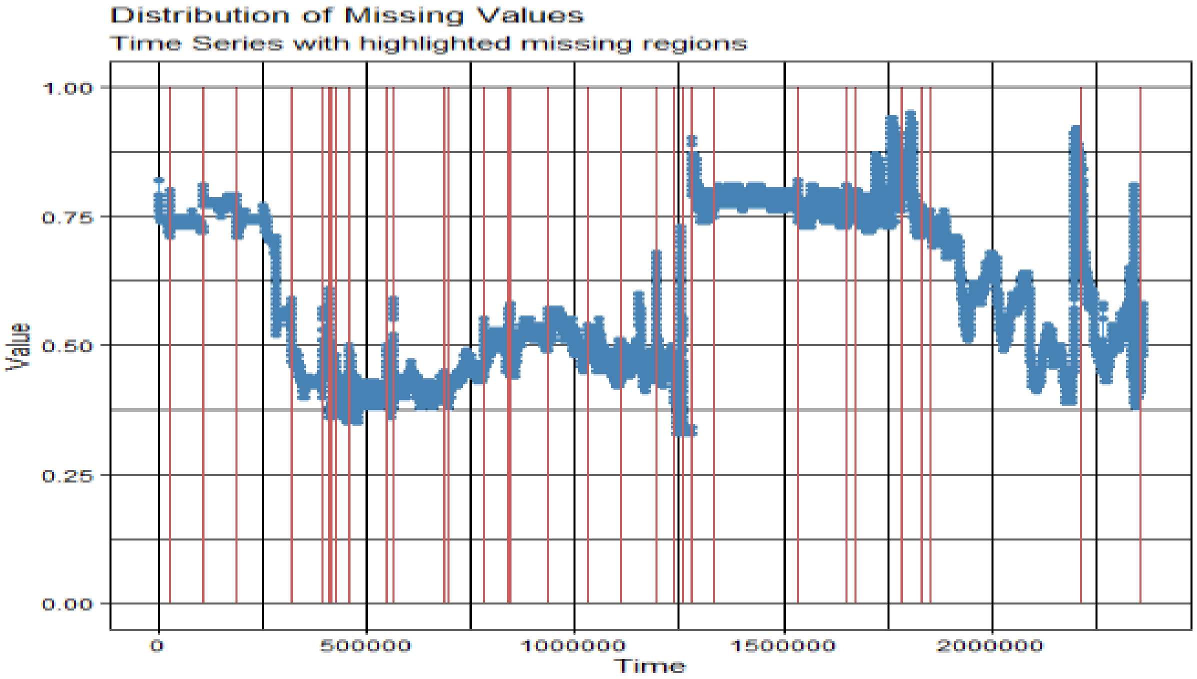

Figure 1. Distribution of Missing Values Soil Moisture Data Sensor

Figure 1 shows the size of the soil moisture sensor data is quite large with the amount of data reaching 2,363,201. The location of the missing values can be seen on the red line on the plot. Missing values need to be handled so that soil moisture sensor data can be further analyzed.

2.2. Kalman Filter

Kalman Filter is a mathematical method that can be used to calculate the optimal estimator of a situation based on available data. The advantage of the Kalman Filter is that it can estimate based on limited data. The latest measurement data is an important part of the

Kalman Filter algorithm, because the data will correct the data from the estimation results, so that the estimation results are always close to the actual conditions in the estimation (Setoodeh et al., 2022). In general, the algorithm for the Kalman Filter method is written as follows (Tengger & Ropiudin, 2019):

Prediction:

Prediction: $\hat{x}_{\bar{k}} = \hat{x}_{k-1}$ (1)

Covariance Error: $P_{\bar{K}} = P_{K-1}$ (2)

Updating

Kalman Gain: $G_k = \frac{P_{k.k-1}}{P_{k,k-1}+r_k}$ (3)

Update Prediction: $\hat{x}_k = \hat{x}_{\bar{k}} + G_k(y_k - \hat{x}_{\bar{k}})$ (4)

Update Covariance Error: $P_K = (1 - G_k)P_{\bar{K}}$ (5)

where:

$\hat{x}_k$	: Current prediction
$\hat{x}_{k-1}$	: Previous prediction
$P_{\bar{K}}$	: Uncertainty estimation
P_{K-1}	: Previous uncertainty estimation
G_k	: Kalman Gain
$P_{k.k-1}$	: Uncertainty estimation
r_k	: Measurement uncertainty
y_k	: Measurement value
$\hat{x}_{\bar{k}}$	: Estimated stage of predication
P_K	: Current uncertainty estimation

The value of *k* will always change with the Kalman Filter process.

2.3. MAPE

Mean Absolute Percentage Error (MAPE) is the percentage size of the error from the prediction results. The smaller the MAPE value, the smaller the prediction error, conversely the greater the MAPE value, the greater the prediction error. MAPE can be calculated by the following formula:

$$MAPE = \frac{\sum_{i=1}^{n}\left(\frac{X_i - F_i}{X_i}\right)}{n} \times 100\% \tag{6}$$

Where X_i is the actual data for the *i*-period, F_i is the predicted result for the *i*-period and *n* is the number of time periods. Imputation results are very good if the MAPE value is <10%, while the imputation results are good if the MAPE value is between 10% and 20% (Malburg et al., 2023).

2.4. Data Analysis

The data used in this study is primary data on the percentage of soil moisture (%) taken in 2021 from the red onion watering experiment in the Greenhouse of the Faculty of Agriculture, Tadulako University. Data analysis used generated data based on the distribution of soil moisture data with Kalman Filter methods for handling missing values. The stages of analysis carried out were as follows:

i. Data Exploration. Data exploration is done by making graphs and identifying where the missing values is. Based on this analysis, testing the distribution of soil moisture sensor data was carried out. Distribution testing using the R package “imputeTS" (Moritz & Bartz-Beielstein, 2019).
ii. Generating data sets and missing value. In this case the generated data are 100 and 1000 data based on distribution of soil moisture sensor data. Generating missing values using missing completely at random (MCAR) with total number of missing values of 5%, 10%, and 15%.
iii. Perform a simulation of missing values handlers using the Kalman Filter method. Simulations are carried out for 100 and 1000 generated data for each level of missing values 5%, 10%, and 15%

3. RESULTS AND DISCUSSION

3.1. Exploration Data

Soil moisture is expressed in percentage form. Soil moisture sensor data has a fairly large data size and contains missing values. The number of missing values is in Table 1.

Table 1. Exploration of Missing Values

Length of time series	Number of Missing Values	Percentage of Missing Values
2363201	49	0.002%

Table 1 shows that the number of missing values is very small. The analysis is then continued by generating sample data. The population data that was tested for distribution was data on the level of soil moisture. The results of the data distribution test for population data obtained data with a Box-Cox Power Exponential distribution, with a coefficient value that can be seen in the Table 2.

Table 2. Box-Cox Power Exponential Coefficient

Mu	Nu	Tau
0.586661998	0.511422569	2.116226467

Table 2 shows that Box-Cox Power Exponential Coefficient for generating data sample of soil moisture data sensor.

3.2. Data Generation with Missing Value

The data is generated based on the distribution of data from the previously tested population, namely the Box-Cox Power Exponential distribution. The generated data analysis will be carried out through two simulations, the first simulation on data with a total of 100 data, and the second simulation on data with a total of 1000 data. Data of 100 and 1000 are intended so that analysis can be carried out on small and large amounts of data. Plot data for generating 100 data is as follows:

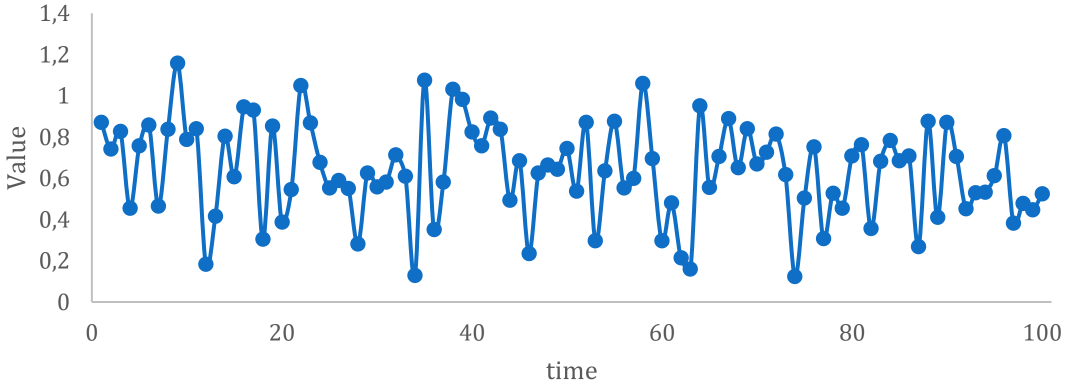

Figure 2. Generated 100 data with Box-Cox Power Exponential distribution

Formation of data sets containing missing values was carried out in Figure 2 using the Missing completely at random (MCAR) mechanism for 5%, 10% and 15% levels of missing values.

3.3. Data Imputation

Imputation is a method that is carried out by filling in missing values with a value that is estimated to be quite feasible. The Kalman Filter method was carried out for 5%, 10% and 15% levels of missing values. The results of the estimation of the Kalman Filter at $n = 100$ and the level of missing values of 15% are as follows:

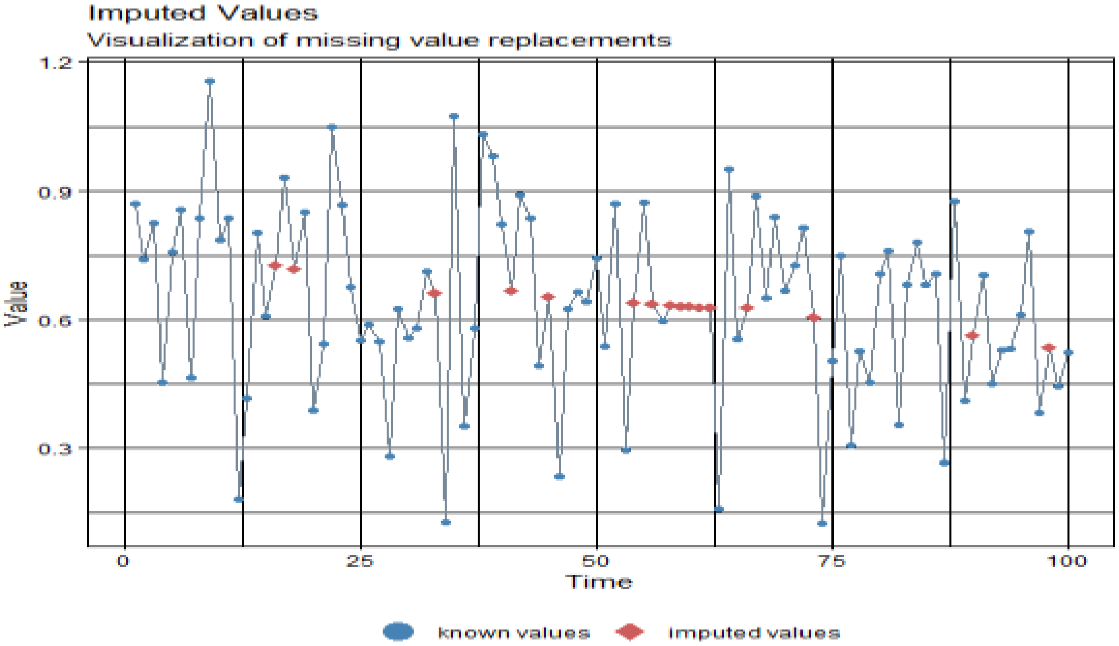

Figure 3. Imputed values for $n = 100$ and 15% level of missing values

Figure 3 shows that imputed values in line with the generated data pattern. So that the actual data value with the imputed results is not different. To find out how far the difference is, evaluation is then carried out using MAPE.

3.4. MAPE

The Mean Absolute Percentage Error (MAPE) method provides information on how much the prediction error is compared to the actual value of the series. The smaller the value of the percentage error in MAPE, the more accurate the prediction results will be. The results of MAPE are as follows:

Table 3. Kalman Filter MAPE

Level of Missing values	***n* = 100**	***n* = 1000**
5%	19.39%	15,30%
10%	20.47%	17,85%
15%	20.20%	22,15%

Table 3 shows that the imputation results used the Kalman Filter method for $n = 100$ and $n = 1000$ solid data to increase the level of missing values. In increasing the sample data from 100 to 1000 data, the MAPE value did not change significantly. This shows that the Kalman Filter method can be applied to large data containing missing values.

3.5. Discussion

The results of imputation on the generated data obtained MAPE < 23% indicating that the estimation of missing values can be done using the Kalman Filter method. The application of the Kalman Filter method to soil moisture sensor data can be as follows:

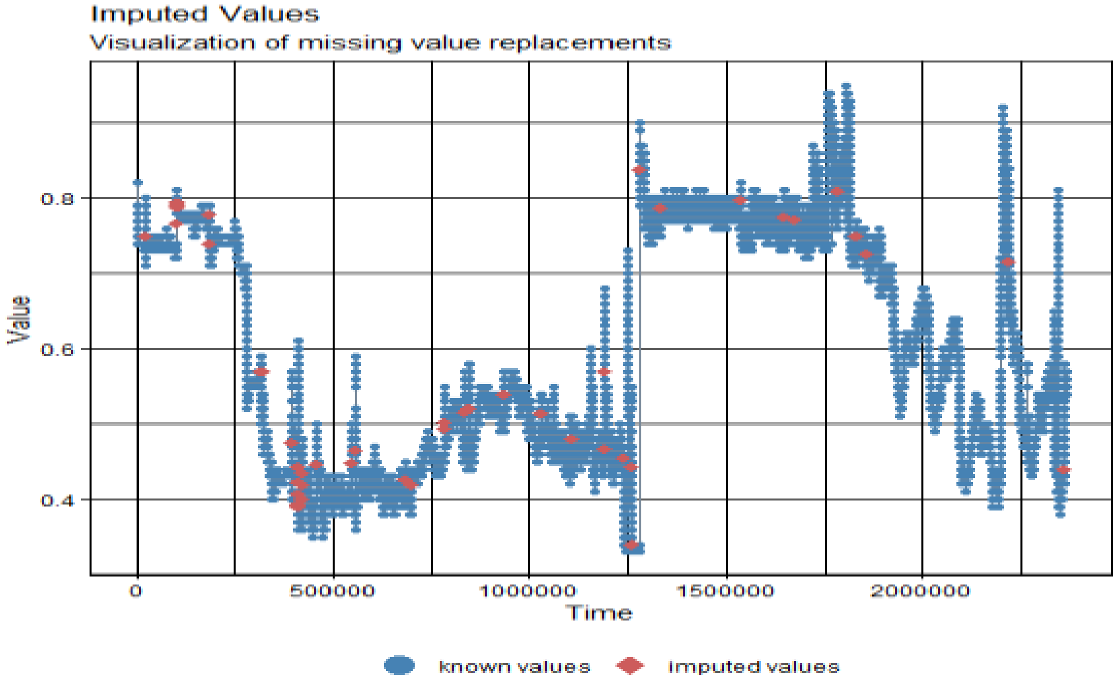

Figure 4. Imputed Values for soil moisture data sensor with Kalman Filter

Figure 4 shows that the imputed values are in line with the soil moisture sensor data pattern. However, it should be considered that the soil moisture sensor data is real time data. The results of this study only succeeded in showing that the process of inputting missing values can be carried out when the soil moisture sensor does not store data in real time. Of course, this will be interesting study material for further research on how to implement missing data estimation in real time.

4. CONCLUSION

Estimation of missing values using the Kalman Filter method can be carried out on soil moisture sensor data that has a large enough size. The simulation results show that the Kalman Filter method is robust to the increase in the amount of data and the level of missing values with MAPE <23%.

REFERENCES

Biemer, P. P., & Lyberg, L. E. (2003). Introduction to Survey Quality. *Introduction to Survey Quality*. https://doi.org/10.1002/0471458740

Jin, H., & Sun, S. (2022). Distributed Filtering for Multi-sensor Systems with Missing Data. *Information Fusion*, *86–87*, 116–135. https://doi.org/10.1016/J.INFFUS.2022.06.007

Malburg, L., Hoffmann, M., & Bergmann, R. (2023). Applying MAPE-K control loops for adaptive workflow management in smart factories. *Journal of Intelligent Information Systems*. https://doi.org/10.1007/S10844-022-00766-W

Moritz, S., & Bartz-Beielstein, T. (2019). *imputeTS: Time Series Missing Value Imputation in R*.

Setiawan, I., Junaidi, J., Fadjryani, F., & Amaliah, F. R. (2022). Automatic Plant Watering System for Local Red Onion Palu using Arduino. *Jurnal Online Informatika*, *7*(1), 28–37. https://doi.org/10.15575/JOIN.V7I1.813

Setoodeh, P., Saeid, H., & Haykin, S. (2022). Kalman Filter. *Nonlinear Filters*, 49–70. https://doi.org/10.1002/9781119078166.CH5

Tengger, B. A., & Ropiudin, R. (2019). Pemanfaatan Metode Kalman Filter Diskrit untuk Menduga Suhu Udara. *Square : Journal of Mathematics and Mathematics Education*, *1*(2), 127. https://doi.org/10.21580/SQUARE.2019.1.2.4202

Wie, W. W. S. (2006). Time Series Analysis: Univariate and Multivariate Methods Second Edition. *Pearson Education, Inc.*, *SFB 373*(Chapter 5), 837–900. http://books.google.com.au/books?id=B8_1UBmqVUoC