JURNAL MATEMATIKA, STATISTIKA DAN KOMPUTASI e-ISSN: 2614-8811
Published by Departement of Mathematics, Hasanuddin University, Indonesia
https://journal.unhas.ac.id/index.php/jmsk/index
Vol. 21, No. 1, September 2024, pp. 307-320
DOI: 10.20956/j.v21i1.36468

Machine Learning-Based Univariate Time Series Imputation Method for Estimating Missing Values in Non-Stationary Data

Metode Machine Learning-Based Univariate Time Series Imputation Method untuk Estimasi Nilai Hilang pada Data Non-Stasioner

Dini Ramadhani[1*], Agus Mohamad Soleh[2], Erfiani[3]

[1,2,3] *Pogram Studi Statistika dan Sains Data, Fakultas Matematika dan Ilmu Pengetahuan Alam, Institut Pertanian Bogor, Indonesia*
Email: [1]*diniramadhani@apps.ipb.ac.id*,[2]*agusms@apps.ipb.ac.id*, [3]*erfiani@apps.ipb.ac.id*

Abstrak

Handling missing values in time series data is crucial because they can disrupt data analysis and interpretation. Sequentially missing values in time series often pose a more complex challenge compared to randomly missing values. One of the promising recent methods is Machine Learning-Based Univariate Time Series Imputation (MLBUI), although it is still not widely used and its accessibility is limited. MLBUI employs Random Forest Regression (RFR) and Support Vector Regression (SVR) algorithms. This study evaluates the performance of MLBUI in addressing missing data scenarios in non-stationary univariate time series data. The data used in this research is the average temperature data from Bogor Regency. The missing data scenarios considered include rates of 6%, 10%, and 14%. Besides MLBUI, five other comparison methods are used: Kalman StructTS, Kalman Auto-ARIMA, Spline Interpolation, Stine Interpolation, and *Moving Average*. The results show that MLBUI performs poorly for non-stationary data, although the obtained Mean Absolute Percentage Error (MAPE) is below 10%.
Keywords: Imputation, Non-Stationary Data, Machine Learning, Missing Values

Abstract

Penanganan nilai hilang dalam data deret waktu adalah krusial karena dapat menyebabkan gangguan dalam analisis dan interpretasi data. Khususnya, nilai-nilai hilang yang terjadi secara berurutan dalam deret waktu seringkali menjadi tantangan yang lebih kompleks dibandingkan dengan nilai-nilai yang hilang secara acak. Salah satu metode terbaru yang menjanjikan adalah *Machine Learning-Based Univariate Time Series Imputation* (MLBUI), meskipun masih belum banyak digunakan dan aksesibilitasnya masih terbatas. MLBUI menggunakan algoritma *Random Forest Regression* (RFR) dan *Support Vector Regression* (SVR). Dalam studi ini, evaluasi dilakukan terhadap kinerja MLBUI dalam mengatasi skenario data yang hilang pada deret waktu univariat yang non-stasioner. Data yang digunakan pada penelitian ini yaitu data suhu rata-rata dari Kabupaten Bogor. Skenario kehilangan data yang dipertimbangkan mencakup tingkat 6%, 10%, dan 14%. Selain MLBUI, lima metode perbandingan lainnya digunakan: Kalman StructTS,

Kalman Auto-ARIMA, Interpolasi Spline, Interpolasi Stine, dan *Moving Average*. Hasil penelitian menunjukkan bahwa MLBUI memberikan hasil yang kurang baik untuk data yang non-stasioner walaupun nilai *Mean Absolute Percentage Error* (MAPE) yang diperoleh berada dibawah 10%.

Kata kunci: Imputasi, Data Non-Stasioner, Pembelajaran Mesin, Nilai Hilang

1. PENDAHULUAN

Penanganan nilai hilang pada data deret waktu sangat penting karena nilai hilang dapat menyebabkan distorsi dalam analisis dan interpretasi data. Jika tidak ditangani dengan benar, nilai hilang dapat mengakibatkan kesimpulan yang salah dan model yang tidak akurat. Metode yang umum digunakan untuk menangani nilai hilang meliputi imputasi, interpolasi, dan penggunaan model statistik khusus yang dapat mengakomodasi nilai hilang. Dengan penanganan yang tepat, kita dapat meminimalkan dampak negatif dari nilai hilang dan meningkatkan akurasi analisis serta hasil prediksi.

Nilai hilang yang muncul secara berturut-turut dalam data deret waktu dapat menjadi tantangan yang lebih besar dibandingkan dengan nilai hilang yang tersebar secara acak. Nilai hilang secara beruntun pada time series univariate termasuk dalam kategori *Missing Not at Random* (MNAR). MNAR [10] terjadi ketika data yang hilang memiliki pola tertentu dan terkait dengan nilai data itu sendiri atau faktor lain yang tidak diamati. Identifikasi nilai hilang beruntun sebagai MNAR penting untuk memastikan pendekatan yang tepat dalam mengatasi data hilang tersebut. Sayangnya, baru sedikit penelitian yang membahas penanganan nilai hilang secara berturut-turut pada data deret waktu univariat. Sedangkan peneitian yang tidak membahas hal tersebut banyak seperti penelitian secara umum penanganan nilai hilang dalam konteks MNAR dalam uji klinis longitudinal [11], penelitian mengenai penanganan nilai hilang di berbagai variabel dalam studi epidemiologi dalam konteks *Missing Completely at Random* (MCAR), *Missing at Random* (MAR), dan *Missing Not at Random* (MNAR) [6], penelitian yang berfokus pada peningkatan metode analisis ketika berhadapan dengan kovariat yang hilang menurut mekanisme MNAR [2], penelitian yang berfokus pada penerapan model Heckman untuk mengatasi data hasil yang hilang secara MNAR, dengan prediktor yang mungkin hilang secara MAR. [4], dan lain-lain.

Phan *et al.* [13] mengusulkan suatu kerangka kerja yang berfokus pada penanganan nilai yang hilang dalam data deret waktu univariat dengan menggunakan pendekatan *Dynamic Time Wrapping* (DTW). Metode DTW didasarkan pada konsep bahwa nilai yang hilang dapat diestimasi dengan mencari dan menggantikan rangkaian data yang paling mirip dengan nilai yang hilang. Pendekatan tersebut memiliki kelebihan dalam mengatasi data deret waktu yang kompleks dengan kesenjangan besar antara nilai yang hilang, namun memiliki batasan tertentu. Salah satu batasan utamanya adalah asumsi bahwa data deret waktu berulang, yang berarti memiliki indikator korelasi silang yang tinggi antar observasi. Penggunaan DTW untuk interval yang hilang dengan jumlah data yang cukup besar dapat memerlukan waktu komputasi yang tinggi, yang dapat menjadi tantangan dalam analisis deret waktu yang skala besar.

Penelitian lain diperkenalkan oleh Phan [12] mengusulkan pendekatan baru untuk mengisi nilai-nilai yang hilang dalam deret waktu univariat menggunakan Metode *Machine Learning-based Univariate Time Series Imputation* (MLBUI). Pendekatan ini mengubah data deret waktu univariat menjadi data multivariat. Metode yang diterapkan dalam MLBUI adalah *Random Forest Regression* (RFR) [14] dan *Support Vector Regression* (SVR) [5]. Pada penelitian tersebut Metode MLBUI dibandingkan dengan beberapa metode lain, seperti *e-Data Warehouse and Business Intelligence* (eDTWBI), Kalman, Interpolasi, dan *Last Observation Carried Forward* (LOCF).

Hasil penelitian menunjukkan bahwa metode MLBUI RFR mampu mengatasi data yang hilang lebih baik dibandingkan dengan metode-metode lain yang diujikan.

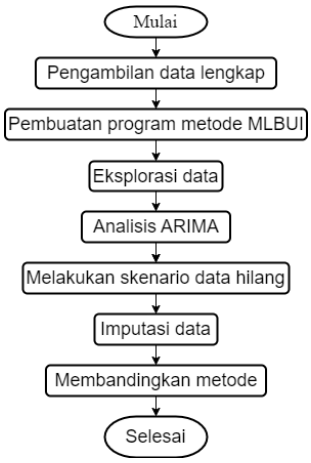
Penelitian lain yang diusulkan oleh Wongoutong [17] berfokus pada analisis deret waktu dengan tren ketika terdapat nilai yang hilang secara berturut-turut. Skenario yang digunakan pada penelitian Wongoutong ada tiga, yaitu data yang hilang dalam deret waktu sebesar 10%, 20%, dan 50%. Hasil dari penelitian tersebut yaitu metode Interpolasi, Kalman, dan *Linear Trend at Point* (LTP) memiliki kinerja yang jauh lebih baik daripada tiga metode imputasi lainnya.

Phan [12] mengarahkan perhatian pada teknik imputasi data hilang dalam deret waktu univariat menggunakan pendekatan pembelajaran mesin tanpa memperhitungkan kestasioneran data. Namun, kestasioneran data merupakan elemen krusial dalam analisis deret waktu yang dapat mempengaruhi efektivitas metode imputasi. Data yang tidak stasioner dapat menyebabkan hasil imputasi yang tidak akurat jika metode tidak dirancang untuk menangani ketidakstasioneran tersebut.

Dalam konteks ini, penulis berencana untuk menyelidiki lebih lanjut mengenai karakteristik metode *Machine Learning-based Univariate Imputation* (MLBUI) ketika diterapkan pada data non-stasioner. Penelitian ini akan menilai bagaimana MLBUI, yang secara umum dirancang dengan asumsi kestasioneran data, berfungsi ketika menghadapi data yang tidak stasioner. Dengan kata lain, penulis akan mengevaluasi sejauh mana metode ini dapat mengatasi tantangan yang ditimbulkan oleh tren dan pola musiman yang ada dalam data non-stasioner. Selain itu, Penelitian ini dimaksudkan untuk mempelajari metode MLBUI, yang belum umum digunakan serta program MLBUI belum dapat diakses dengan mudah.

2. METODE PENELITIAN

Prosedur analisis pada penelitian ini disajikan pada Gambar 2.1. Penelitian ini menggunakan data temperatur rata-rata di Kabupaten Bogor. Data diambil dari tanggal 15 Januari 2023 hingga 24 April 2023 dari laman https://dataonline.bmkg.go.id/data_iklim.



Gambar 2.1. Prosedur Penelitian

2.1 Pengambilan Data Lengkap

Pada tahap ini akan dilakukan pengambilan data lengkap sebanyak 100 data. Data yang diambil tidak boleh ada nilai hilang.

2.2. Pembuatan Program Metode MLBUI

Phan [12] mengembangkan Algoritma MLBUI diperkenalkan untuk memastikan peningkatan hasil imputasi secara teratur. Pembelajaran mesin pada MLBUI akan diterapkan pada data transformasi yang tersisa setelah nilai hilang jika nilai hilang berada di $3 \times T$ dari data. Pada posisi nilai hilang yang lain yaitu pada $N - (3 \times T)$dari data pembelajaran mesin pada MLBUI akan diterapkan pada data transformasi sebelum nilai hilang. Bila nilai hilang berada di tengah-tengah deret, yaitu di antara $3 \times T$ dan $N - (3 \times T)$ (N adalah jumlah deret waktu asli), metode pembelajaran mesin diterapkan pada data sebelum dan sesudah nilai hilang. Data yang akan diinputkan pada algoritma ini ada $X = \{x_1, x_2, \dots, x_N\}$ adalah deret waktu tidak lengkap). Data ini memiliki nilai hilang pada data ke t dan memiliki jumlah data hilang sebesar T. Metode ini akan menghasilkan data deret waktu lengkap (Y). Pendekatan ini terdiri dari lima fase disajikan pada Algoritma 1.

Algoritma 1. Algoritma MLBUI

Fase 1: Bagi X menjadi dua deret waktu yang terpisah Da, Db:

$$\boldsymbol{Da = X[N : t + T], Db = X[1 : t - 1]}$$

Fase 2: Perubahan deret waktu univariat ke deret waktu multivariat

$$M\ Da = \begin{Vmatrix} x_N & x_{N-1} & \cdots & x_{N-T+1} & x_{N-T} \\ x_{N-1} & x_{N-2} & \cdots & x_{N-T} & x_{N-T-1} \\ \cdots & & & & \\ x_{t+2T} & x_{t+2T-1} & \cdots & x_{t+T+1} & x_{t+T} \end{Vmatrix}$$

$$M\ Db = \begin{Vmatrix} x_1 & x_2 & \cdots & x_T & x_{T+1} \\ x_2 & x_3 & \cdots & x_{T+1} & x_{T+2} \\ \cdots & & & & \\ x_{t+T-1} & x_{t-T} & \cdots & x_{t-2} & x_{t-1} \end{Vmatrix}$$

Fase 3: Pelatihan model pembelajaran mesin

$$\hat{f}a = f(M\ Da) \text{ dan } \hat{f}b = f(M\ Db)$$

Fase 4: Estimasi data hilang: menerapkan 1 langkah peramalan

a. Data sebelum data hilang akan dilakukan langkah sebagai berikut:
   Langkah 1: $x_t = \hat{f}b(x_{t-T}, x_{t-T+1}, \dots, x_{t-1})$
   Langkah 2: $x_{t+1} = \hat{f}b(x_{t-T+1}, x_{t-T+2}, \dots, x_t)$
   …
   Langkah T: $x_{t+T-1} = \hat{f}b\ (x_{t-1}, x_t, \dots, x_{t+T-2})$

   sehingga didapatkan hasil imputasi yaitu $\boldsymbol{\hat{x}b}$. $\boldsymbol{\hat{x}b}$ disajikan pada persamaan berikut:

$$\boldsymbol{\hat{x}b = (x_t, x_{t+1}, \dots, x_{t+T-1})}$$

b. Data setelah data hilang akan dilakukan langkah 1 sampai T seperti pada point a dengan data M Da sehingga menghasilkan hasil imputasi yaitu

$$\boldsymbol{\hat{x}a = reverse(x_{t+T-1}, x_{t-T+2}, \dots, x_t)}$$

Fase 5: Melengkapi nilai hilang
Ganti nilai yang hilang pada posisi tersebut T dengan rata-rata vektor $\boldsymbol{\hat{x}b}$ dan $\boldsymbol{\hat{x}a}$ dengan demikian data deret waktu menjadi data yang lengkap.

2.3. Skenario Nilai Hilang

Pada penelitian ini dilakukan 3 skenario nilai data hilang. Skenario data hilang tersebut sebagai berikut:

Skenario 1. Nilai hilang di tengah secara berturut sebesar 6%.
Skenario 2. Nilai hilang di tengah secara berturut sebesar 10%.
Skenario 3. Nilai hilang di tengah secara berturut-turut sebesar 14%

Skenario tersebut dirancang sesuai ketentuan letak nilai hilang. Nilai hilang di bagian tengah deret waktu, yaitu antara indeks ke-$3 \times T$ dan $N - 3 \times T$, di mana T adalah jumlah nilai hilang dan N adalah total jumlah data [12].

2.4. Imputasi Data

Metode yang digunakan pada penelitian ini yaitu MLBUI SVR dan MLBUI RFR sebagai metode utama serta metode Kalman StructTS, Kalman *Auto-ARIMA* [3], Interpolasi Spline, Interpolasi Stine [8], *Moving Average* [1] sebagai metode pembanding.

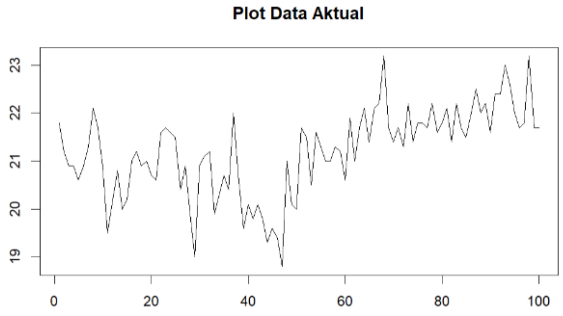
2.4. Membandingkan Metode

Pada tahap ini dilakukan perbandingan kinerja metode MLBUI dengan metode lainnya. Program metode Kalman StructTS, Kalman *Auto-ARIMA*, Interpolasi Spline, Interpolasi Stine, dan *Moving Average* terdapat pada paket imputeTS yang tersedia di CRAN [9] Kinerja metode akan dilihat berdasarkan metrik *Mean Squared Error* (MSE) dan *Mean Absolute Percentage Error* (MAPE) [7].

3. HASIL DAN PEMBAHASAN

3.1 Eksplorasi Data

Eksplorasi data digunakan untuk memahami karakteristik dasar sehingga pada tahap ini karakteristik dasar dari data aktual akan dilihat. Plot data aktual disajikan pada Gambar 3.1.



Gambar 3.1. Plot Data Aktual

Dapat dilihat dari Gambar 3.1. plot garis yang menunjukkan data aktual dengan sumbu horizontal (X) yang menampilkan indeks atau waktu, terdiri dari 100 titik mewakili 100 unit waktu yaitu harian. Sumbu vertikal (Y) menunjukkan nilai data aktual. Data menunjukkan variasi atau fluktuasi dengan beberapa periode di mana data meningkat secara bertahap dan beberapa periode di mana data menurun.

Meskipun terdapat fluktuasi, terlihat ada tren kenaikan sedikit pada nilai data dari kiri ke kanan, menunjukkan bahwa secara keseluruhan nilai data cenderung meningkat seiring waktu. Fluktuasi dalam data menunjukkan adanya volatilitas atau variasi yang signifikan dalam data yang direkam, mengindikasikan ketidakstabilan atau variabilitas dalam fenomena yang diukur. Hal tersebut dapat dikatakan bahwa ada indikasi bahwa data tidak stasioner baik dalam rata-rata maupun dalam ragam galat.

3.2 Analisis ARIMA

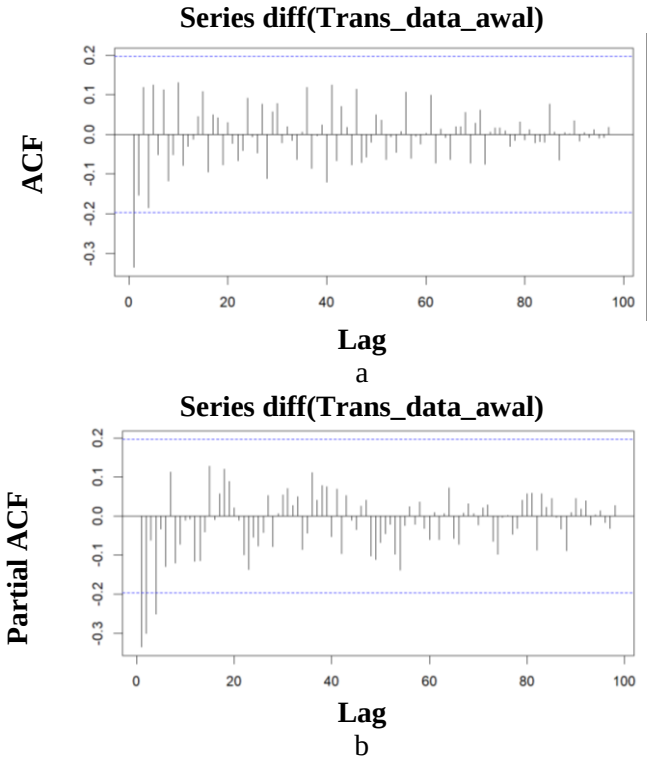
a) Uji Stasioneritas

Jika dilihat dari Gambar 3.1 dapat dilihat bahwa data mengandung pola tren, oleh karena itu ada indikasi bahwa data tidak stasioner terhadap variansi dan rata-rata. Pada pengujian Box-Cox transformasi diperoleh nilai λ sebesar 1.999, sehingga dapat disimpulkan bahwa data tidak stasioner secara varians karena nilai λ tidak mendekati 1, sehingga perlu dilakukan transformasi data. Selanjutnya jika menggunakan uji ADF, *P-Value* yang didapat sebesar 0.366. *P-Value* lebih dari 0.05 sehingga H_0 tidak ditolak jadi dapat disimpulkan bahwa data tidak stasioner terhadap rata-rata sehingga data akan dilakukan transformasi dan *differencing*.

Setelah dilakukan transformasi dan *differencing*, didapatkan nilai λ sebesar 1.999 sehingga dapat dikatakan bahwa data belum stasioner terhadap variansi. Adapun *P-Value* dari uji ADF didapatkan sebesar 0.01 sehingga H_0 ditolak dan disimpulkan bahwa data stasioner terhadap rata-rata. Karena data sudah stasioner walupun hanya pada rata-ratanya saja, maka selanjutnya akan dilakukan identifikasi model.

b) Indentifikasi Model

Pada tahap ini akan dilakukan identifikasi model dengan menggunakan Plot *Autocorrelation Function* (ACF) dan *Partial Autocorrelation Function* (PACF) yang diperlihatkan pada Gambar 3.2.



Gambar 3.2. Plot (a) ACF dan (b) PACF

Dari plot ACF dan PACF dapat dilihat bahwa, pada plot ACF lag terpotong pada lag ke-1 dan lag pada plot PACF terpotong pada lag ke-4 sehingga kombinasi model yang mungkin yaitu: ARIMA(1,1,0), ARIMA (2,1,0), ARIMA (3,1,0), ARIMA (4,1,0), ARIMA(1,1,1), ARIMA (2,1,1), ARIMA (3,1,1), dan ARIMA (4,1,1).

c) Pengujian Signifikansi Parameter

Pada tahap ini dilakukan pengujian signifikansi parameter dengan hipotesis sebagai berikut:

H_0 : Parameter tidak signifikan

H_1 : Parameter signifikan

Statistik uji yang didapatkan disajikan pada Tabel 3.1.

Tabel 3.1 Statistik Uji Signifikansi Parameter

Model		***P-Value***
ARIMA(1,1,0)	AR(1)	4.18^{-12}
ARIMA (2,1,0)	AR(1)	$< 2.2^{-16}$
	AR(2)	7.83^{-10}
ARIMA (3,1,0)	AR(1)	$< 2.2^{-16}$
	AR(2)	1.75^{-09}
	AR(3)	0.02578
ARIMA (4,1,0)	AR(1)	$< 2.2^{-16}$
	AR(2)	3.07^{-14}
	AR(3)	6.47^{-06}
	AR(4)	7.22^{-05}
ARMA(0,1,1)	MA(1)	$< 2.2^{-16}$
ARIMA(1,1,1)	AR(1)	0.0005626
	MA(1)	$< 2.2^{-16}$
ARIMA (2,1,1)	AR(1)	6.17^{-06}
	AR(2)	0.001616
	MA(1)	$< 2.2^{-16}$
ARIMA (3,1,1)	AR(1)	8.49^{-06}
	AR(2)	0.002069
	AR(3)	0.618096
	MA(1)	$< 2.2^{-16}$
ARIMA (4,1,1)	AR(1)	1.41^{-06}
	AR(2)	0.0001159
	AR(3)	0.1164809
	AR(4)	0.0089313
	MA(1)	$< 2.2^{-16}$

Pada uji ini digunakan tingkat signifikansi (α) sebesar 0.05. Daerah kritis yang digunakan adalah H_0 dilotak jika *P-Value* < 0.05. Dapat dilihat pada Tabel 3.1bahwa *P-Value* < 0.05 untuk model ARIMA(1,1,0), ARIMA(2,1,0), ARIMA(3,1,0), ARIMA(4,1,0), ARMA(0,1,1), ARIMA(1,1,1), dan ARIMA(2,1,1) dengan demikian dapat dikatakan bahwa parameter-parameter pada model tersebut signifikan. Model-model ini selanjutnya akan dilakukan diagnostik model.

d) Diagnostik Model

Pada tahap ini akan dilakukan uji autokorelasi, heteroskedastisitas, dan normalitas dengan hipotesis H_0 uji autokorelasi yaitu tidak ada autokorelasi dalam residu model, uji heteroskedastisitas yaitu tidak ada heteroskedastisitas dalam residu model, dan uji normalitas yaitu residu model memiliki distribusi normal. Statistik uji dignostik model disajikan pada Tabel 3.2.

Tabel 3.2. Statistik Uji Diagnostik Model

Model	*P-Value*		
	A	B	C
ARIMA(1,1,0)	0.003	0.481	0.463
ARIMA(2,1,0)	0.218	0.324	0.577
ARIMA(3,1,0)	0.368	0.491	0.967
ARIMA(4,1,0)	0.440	0.857	0.811
ARIMA(0,1,1)	0.001	0.513	0.577
ARIMA(1,1,1)	0.278	0.912	0.994
ARIMA(2,1,1)	0.780	0.893	0.577

A= Uji Autokorelasi, B= Uji Heteroskedastisitas, C= Uji Normalitas.

Pada uji ini digunakan tingkat signifikansi (α) sebesar 0.05 dan daerah kritisnya adalah H_0 dilotak jika *P-Value* < 0.05. Pada Tabel 3.2 dapat disimpulkan bahwa ARIMA(2,1,0), ARIMA(3,1,0), ARIMA(4,1,0), ARIMA(1,1,1), dan ARIMA(2,1,1) tidak ada autokorelasi dalam residu model, tidak ada heteroskedastisitas dalam residu model, dan residu model berdistribusi normal.

Model – model tersebut selanjutnya akan dipilih model terbaiknya dengan melihat nilai AIC (*Akaike Information Criterion*) dan BIC (*Bayesian Information Criterion*) model dengan nilai AIC atau BIC yang lebih rendah dianggap lebih baik [16]. AIC dan BIC disajikan pada Tabel 3.3.

Tabel 3.3. Kriteria AIC dan BIC

Model	AIC	BIC
ARIMA(2,1,0)	983.546	991.301
ARIMA(3,1,0)	980.714	991.054
ARIMA(4,1,0)	968.319	981.244
ARIMA(1,1,1)	951.146	958.901
ARIMA(2,1,1)	943.780	954.120

Dari Tabel 3.3 didapatkan bahwa model ARIMA(2,1,1) adalah model yang paling optimal berdasarkan kedua kriteria (AIC dan BIC) dibandingkan dengan model-model lainnya dalam tabel. AIC dan BIC yang lebih rendah menunjukkan bahwa model ini paling efisien dalam menangkap pola data tanpa *overfitting*. Dengan menggunakan ARIMA(2,1,1), bisa didapatkan keseimbangan terbaik antara kompleksitas model dan kecocokan data.

Hasil analisis ARIMA(2,1,1) pada data temperatur rata-rata menunjukkan bahwa data tersebut awalnya bersifat non-stasioner, artinya data memiliki tren atau pola tertentu yang tidak tetap seiring waktu. Dalam analisis *time series*, stasioneritas adalah sifat penting yang harus dipenuhi agar model dapat secara efektif menangkap pola dan membuat prediksi yang akurat. Oleh karena itu, pada model ini, dilakukan proses *differencing* sebanyak satu kali (I=1) untuk menghilangkan tren jangka panjang dari data. Proses *differencing* ini melibatkan pengurangan nilai temperatur saat ini dengan nilai temperatur sebelumnya, sehingga data menjadi lebih stasioner dan memungkinkan model untuk lebih baik dalam mengidentifikasi pola musiman atau fluktuasi lainnya dalam data.

Komponen *Autoregressive* (AR) dengan order 2 dalam model ARIMA(2,1,1) menunjukkan bahwa nilai temperatur pada suatu waktu tertentu dipengaruhi oleh dua nilai temperatur sebelumnya. Ini berarti, model memperhitungkan pengaruh dari dua periode sebelumnya untuk memprediksi nilai saat ini. Jika ada pola yang berulang dalam data, seperti suhu yang cenderung meningkat atau menurun secara musiman, model ini akan dapat menangkap pola tersebut dengan mengandalkan dua lag sebelumnya.

Selain itu, komponen *Moving Average* (MA) dengan order 1 menambahkan lapisan kompleksitas lebih lanjut dalam model dengan memasukkan pengaruh dari kesalahan prediksi sebelumnya. Dengan kata lain, selain memperhitungkan nilai temperatur dari periode sebelumnya, model ini juga mempertimbangkan kesalahan yang dibuat saat memprediksi nilai sebelumnya. Hal ini berguna dalam memperbaiki prediksi saat ini dengan mengoreksi kesalahan yang terjadi sebelumnya.

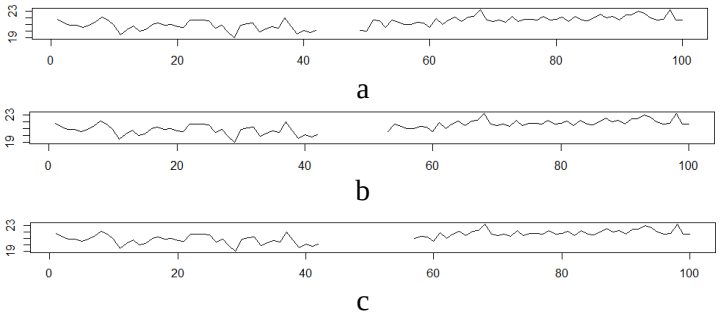
Namun, karena model ARIMA(2,1,1) ini sangat bergantung pada data sebelumnya dan kesalahan prediksi masa lalu, nilai yang hilang (*missing values*) dalam data temperatur rata-rata dapat memiliki dampak signifikan pada hasil prediksi. Jika terdapat nilai yang hilang dalam dua lag sebelumnya atau dalam kesalahan prediksi yang diandalkan oleh komponen *Moving Average*, model dapat menghasilkan prediksi yang kurang akurat atau bahkan bias. Oleh karena itu, penting untuk menangani nilai yang hilang dengan tepat.

Secara keseluruhan, pengenalan data awal, apakah stasioner atau tidak stasioner, berpengaruh besar terhadap pemilihan metode imputasi dalam analisis deret waktu. Data stasioner memungkinkan penggunaan metode imputasi yang lebih sederhana dan efektif, sedangkan data non-stasioner memerlukan metode yang lebih adaptif untuk menangani tren atau pola musiman. Pemilihan metode imputasi harus disesuaikan dengan karakteristik data untuk mengurangi risiko prediksi yang tidak akurat. Selanjutnya, penelitian ini akan menyelidiki karakteristik metode MLBUI pada data non-stasioner, mengingat bahwa penelitian sebelumnya [12] belum menekankan aspek ini.

3.3 Skenario Data Hilang pada Data Aktual

Pada data aktual, akan diterapkan skenario nilai hilang sebesar 6%, 10%, dan 14% untuk menganalisis dampaknya. Nilai hilang ini ditempatkan di bagian tengah deret waktu, yaitu antara indeks ke-$3 \times T$ dan $N - (3 \times T)$, di mana T adalah jumlah nilai hilang dan N adalah total jumlah data. Jumlah data yang digunakan pada penelitian ini yaitu sebesar 100 data.

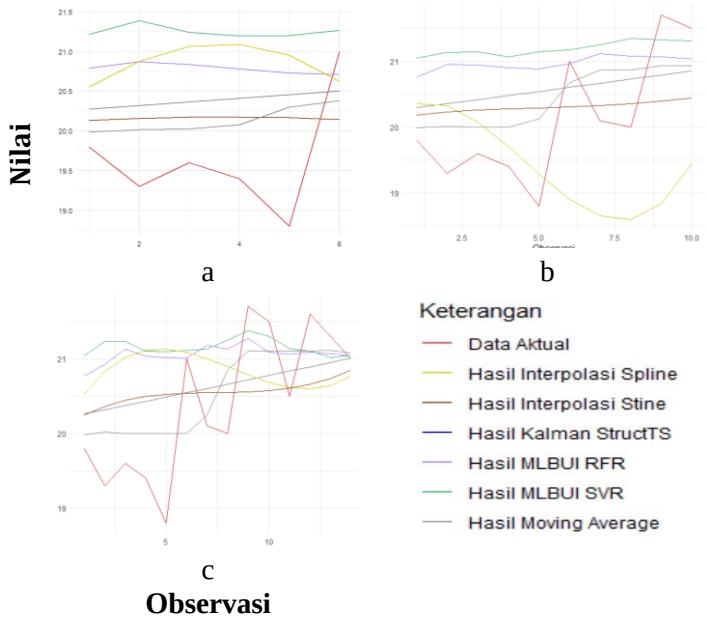
Plot skenario nilai hilang pada data simulasi, yang ditampilkan dalam Gambar 3.4, memberikan visualisasi bagaimana nilai hilang tersebut didistribusikan dalam deret waktu. Pemilihan skenario ini didasarkan pada total jumlah data yang tersedia, di mana 14% adalah batas maksimum nilai hilang yang masih memungkinkan untuk dianalisis secara efektif. Dengan menetapkan nilai hilang di bagian tengah deret, analisis dapat memeriksa bagaimana metode imputasi menangani berbagai tingkat kekurangan data.



Gambar 3.4. Plot Skenario Data Hilang Sebesar (a) 6% (b) 10% (c) 14% pada Data Aktual

3.3 Hasil Imputasi pada Data Aktual

Nilai Hilang yang telah diimputasi selanjutnya dibuat grafik untuk mengetahui secara deskriptif. Hal yang akan diketahui disini yaitu kesamaan pola hasil imputasi dari metode-metode yang digunakan dengan nilai aktual. Grafik ini disajikan pada Gambar 3.5.



Gambar 3.5. Plot Perbandingan Data Awal dengan Hasil Imputasi Data Hilang sebesar (a) 6% , (b) 10% , dan (c) 14%

Dapat dilihat pada Gambar 3.5 semua metode belum dapat menghasilkan nilai imputasi yang mengikuti pola data awal. Hasil imputasi *Moving Average* paling dekat dengan data aktual sedangkan metode MLBUI cukup jauh. Metode Interpolasi Stine menempati posisi kedua dan Kalman StructTS menempati posisi ketiga untuk metode yang mendekati data aktual. Plot hasil imputasi metode Interpolasi Spline paling jauh posisinya dengan data aktual.

Hasil analisis pada Gambar 3.5 menunjukkan bahwa semua metode imputasi yang diuji belum sepenuhnya mampu menghasilkan pola data awal secara akurat. Dari grafik tersebut, jelas bahwa *Moving Average* memberikan hasil imputasi yang paling mendekati data aktual, sementara metode MLBUI menunjukkan jarak yang cukup jauh dari pola asli. Beberapa faktor dapat menjelaskan mengapa metode-metode ini menghasilkan hasil yang berbeda dalam hal akurasi imputasi.

Pertama, metode *Moving Average* bekerja dengan cara menghaluskan data [1] yang efektif dalam mengatasi data dengan tren yang relatif stabil dan *noise* yang tidak terlalu besar. Kemampuan metode ini untuk memberikan hasil yang mendekati data aktual bisa jadi disebabkan oleh kesederhanaannya dalam menyaring fluktuasi jangka pendek dan menangkap pola umum data. Sebaliknya, MLBUI dengan teknik RFR dan SVR tampaknya tidak berhasil mereplikasi pola data dengan baik, terutama jika data tersebut mengandung pola yang kompleks atau non-stasioner. RFR dan SVR cenderung lebih efektif untuk data yang lebih terstruktur dan mungkin tidak cukup adaptif dalam situasi di mana data hilang mengikuti pola yang lebih variatif atau tidak teratur.

Metode Interpolasi Stine dan Kalman StructTS, meskipun tidak sebaik *Moving Average*, memberikan hasil yang lebih baik dibandingkan beberapa metode lain. Interpolasi Stine, dengan pendekatannya yang menghaluskan data menggunakan fungsi polinomial [8], dan Kalman

StructTS, yang menggunakan model dinamis untuk memperkirakan nilai yang hilang, keduanya menunjukkan kemampuan yang lebih baik dalam mendekati data aktual dibandingkan dengan metode MLBUI. Namun, kedua metode ini mungkin tidak seoptimal *Moving Average* dalam mengatasi semua variasi dalam data yang hilang. Metode Interpolasi Spline menunjukkan hasil yang paling jauh dari data aktual, yang bisa jadi disebabkan oleh kemampuannya yang terbatas dalam menangani data dengan fluktuasi besar atau pola non-linear yang tidak dapat ditangkap dengan baik oleh spline polinomial.

Oleh karena itu, perbedaan dalam hasil imputasi ini menunjukkan bahwa efektivitas metode imputasi sangat bergantung pada karakteristik data yang hilang dan pola yang ada dalam data asli. Penyesuaian metode terhadap pola data yang ada dan pemilihan metode yang sesuai dengan jenis data adalah kunci untuk mencapai hasil imputasi yang lebih akurat. Tahap selanjutnya akan dilihat metrik MSE dan MAPE. Berikut ini adalah nilai metrik MSE yang disajikan pada Tabel 3.4.

Tabel 3.4. Nilai MSE

Metode	MSE Data Aktual		
	6% Nilai Hilang	10% Nilai Hilang	14% Nilai Hilang
MLBUI RFR	1.781	1.482	1.209
MLBUI SVR	3.020	1.897	1.402
Kalman StructTS	0.975	0.848	0.635
Kalman *Auto-ARIMA*	1.346	1.071	0.807
Interpolasi Spline	2.137	2.284	1.275
Interpolasi Stine	0.729	0.794	0.738
Moving Average	0.636	0.518	0.387

Berdasarkan nilai metrik MSE yang tercantum dalam Tabel 3.4, terlihat bahwa metode MLBUI dengan teknik RFR menunjukkan kinerja yang relatif baik dalam mengimputasi data hilang pada berbagai proporsi hilang (6%, 10%, dan 14%). Secara umum, MLBUI RFR unggul dibandingkan dengan Interpolasi Spline dan MLBUI SVR dalam semua skenario data hilang. Metode MLBUI SVR, di sisi lain, menunjukkan performa terburuk secara konsisten di seluruh skenario, jika dibandingkan dengan metode-metode lainnya seperti MLBUI RFR, Kalman StructTS, Kalman Auto-ARIMA, Interpolasi Stine, dan *Moving Average*. Meskipun MLBUI SVR menunjukkan hasil yang lebih baik dibandingkan Interpolasi Spline pada skenario 10% data hilang, secara keseluruhan performa SVR kurang memuaskan.

Analisis berdasarkan skenario proporsi data hilang menunjukkan adanya pola peningkatan akurasi pada semua metode, kecuali Interpolasi Spline dan Interpolasi Stine, seiring dengan meningkatnya proporsi data hilang. Ini bisa disebabkan oleh beberapa faktor. Pertama, dengan meningkatnya proporsi data hilang, metode imputasi mungkin menjadi lebih terfokus pada pola yang lebih besar dalam data yang tersedia, yang memungkinkan perbaikan dalam akurasi imputasi. Kedua, beberapa metode mungkin memanfaatkan informasi tambahan dari data yang tersisa secara lebih efektif ketika proporsi data hilang lebih besar, memperbaiki kemampuan untuk menangkap pola yang relevan [15]. Namun, peningkatan akurasi ini juga bisa bergantung pada bagaimana metode imputasi mengatasi data yang hilang dan bagaimana metode-metode itu mengadaptasi model mereka berdasarkan jumlah data yang tersedia.

Dengan demikian, pemilihan metode imputasi yang tepat harus mempertimbangkan proporsi data hilang dan karakteristik spesifik data yang dianalisis. Hasil ini menunjukkan bahwa meskipun beberapa metode menunjukkan peningkatan akurasi dengan meningkatnya proporsi data hilang, metode yang berbeda mungkin lebih atau kurang efektif tergantung pada konteks dan sifat data

yang hilang. Pada tahap selanjutnya akan dilihat kinerja berdasarkan metrik MAPE yang disajikan pada Tabel 3.5.

Tabel 3.5. Nilai MAPE

Metode	MAPE Data Aktual (%)		
	6% Nilai Hilang	10% Nilai Hilang	14% Nilai Hilang
MLBUI RFR	6.369	5.468	4.476
MLBUI SVR	8.274	6.071	4.779
Kalman StructTS	4.657	4.252	3.451
Kalman *Auto-ARIMA*	5.455	4.768	3.892
Interpolasi Spline	6.897	6.173	4.773
Interpolasi Stine	4.063	3.976	3.705
Moving Average	3.537	3.271	2.646

Berdasarkan nilai metrik MAPE yang tercantum dalam Tabel 3.5, dapat disimpulkan bahwa metode *Moving Average* menunjukkan kinerja terbaik dalam hal akurasi imputasi. Metode Kalman StructTS dan Interpolasi Stine juga menunjukkan kinerja yang baik dengan MAPE yang relatif rendah. Sebaliknya, metode MLBUI RFR menunjukkan kinerja yang kurang baik dengan MAPE yang relatif lebih tinggi, sementara MLBUI SVR adalah yang paling buruk meskipun nilai MAPE-nya masih di bawah 10%. Kinerja MLBUI yang kalah dibandingkan dengan metode *Moving Average* mungkin disebabkan oleh ketidakstasioneran data.

Berbeda dengan hasil pada metrik MSE, metrik MAPE menunjukkan adanya pola peningkatan akurasi pada semua metode seiring dengan meningkatnya proporsi data hilang. MSE, atau *Mean Squared Error*, mengukur kesalahan kuantitatif dengan menghitung rata-rata kuadrat dari perbedaan antara nilai yang diimputasi dan nilai aktual [7]. Pada proporsi data hilang sebesar 10%, nilai MSE menunjukkan bahwa kedua metode Interpolasi Spline dan Interpolasi Stine mengalami penurunan akurasi dibandingkan dengan proporsi 6%, mengindikasikan bahwa kesalahan kuadrat antara nilai imputasi dan nilai aktual lebih besar saat proporsi data hilang meningkat. Hal ini mungkin disebabkan oleh kesulitan metode interpolasi dalam menangani fluktuasi tambahan dan kompleksitas yang muncul ketika proporsi data hilang lebih besar.

Sebaliknya, MAPE, atau *Mean Absolute Percentage Error*, mengukur kesalahan relatif dengan menghitung rata-rata kesalahan absolut relatif terhadap nilai aktual [7]. Peningkatan akurasi yang terlihat pada MAPE saat proporsi data hilang meningkat menunjukkan bahwa meskipun kesalahan absolut mungkin meningkat, kesalahan relatif per unit nilai aktual cenderung lebih rendah. Ini mungkin terjadi jika metode interpolasi lebih mampu mengurangi kesalahan relatif dalam konteks nilai yang hilang yang lebih tinggi, menghasilkan kesalahan absolut yang lebih kecil relatif terhadap nilai aktual.

Perbedaan ini mencerminkan bagaimana MSE dan MAPE menangani kesalahan dengan cara yang berbeda. MSE lebih sensitif terhadap kesalahan besar karena menghitung kuadrat dari perbedaan, sehingga bisa menunjukkan penurunan akurasi yang signifikan ketika kesalahan besar muncul lebih sering. Sebaliknya, MAPE, dengan fokus pada kesalahan relatif, mungkin menunjukkan bahwa metode interpolasi lebih efektif dalam hal mengurangi kesalahan relatif meskipun kesalahan absolut meningkat.

Secara keseluruhan jika dibandingkan dengan penelitian sebelumnya [12] yaitu penelitian sebelumnya lebih fokus pada data dari berbagai domain tanpa terlalu mempertimbangkan kestasioneran data. Meski MLBUI telah terbukti unggul dalam penelitian sebelumnya, penelitian ini menunjukkan bahwa kinerja MLBUI tidak demikian jika dibandingkan dengan metode yang

digunakan. Namun, metode MLBUI konsisten dalam penanganan nilai hilang diberbagai skenario. Hal ini berkesesuaian dengan penelitian yang dilakukan Phan [12].

Penelitian ini juga mengindikasikan adanya pola peningkatan akurasi seiring dengan meningkatnya proporsi nilai hilang, yang tidak ditemukan dalam penelitian sebelumnya. Hal ini menunjukkan bahwa MLBUI lebih efektif dalam mengatasi nilai hilang yang lebih besar, namun efektivitasnya mungkin memiliki batasan tertentu. Kenaikan akurasi ini bisa jadi karena metode imputasi menjadi lebih baik dalam menangkap pola besar dari data yang tersedia ketika proporsi data hilang meningkat, tetapi ada batas di mana akurasi tidak meningkat lebih lanjut. Alasan perbedaan ini terletak pada variasi dalam karakteristik data yang digunakan dalam penelitian ini dibandingkan dengan penelitian sebelumnya, serta perbedaan dalam cara evaluasi dan metrik yang diterapkan.

Pada Tabel 3.5 terlihat bahwa meskipun metode MLBUI tidak menunjukkan kinerja sebaik metode *Moving Average* dengan nilai MAPE 10%. Namun berdasarkan nilai MAPE, MLBUI tetap merupakan metode dengan tingkat prediksi yang sangat andal. Hal ini menunjukkan bahwa MLBUI dapat menjadi pilihan yang layak untuk digunakan dalam situasi tertentu, meskipun mungkin ada metode lain yang menunjukkan kinerja yang lebih baik dalam skenario yang berbeda. Berdasarkan pembahasan-pembahasan ini, kita dapat memahami kinerja MLBUI dalam menangani nilai hilang pada data non-stasioner.

4. KESIMPULAN

Hasil penelitian ini menunjukkan bahwa metode MLBUI RFR kurang efektif pada data non-stasioner, meskipun nilai MAPE-nya tetap di bawah 10%. Sebaliknya, metode *Moving Average* memberikan akurasi imputasi terbaik, diikuti oleh Kalman StructTS dan Interpolasi Stine. Metode MLBUI, baik dengan teknik RFR maupun SVR, menunjukkan performa yang kurang memuaskan, mungkin karena ketidakstasioneran data yang mempengaruhi efektivitasnya. MLBUI memerlukan penyesuaian khusus untuk menangani data dengan pola non-stasioner. Selain itu, akurasi metode MLBUI cenderung meningkat seiring dengan bertambahnya nilai hilang tapi dengan batasan tertentu. Dapat disimpulkan juga bahwa pemilihan metode imputasi harus mempertimbangkan karakteristik data dan proporsi nilai hilang untuk mencapai hasil yang optimal.

KONFLIK KEPENTINGAN

Penulis menyatakan tidak ada konflik kepentingan.

REFERENCES

[1] Arai, K., Kapoor, S. & Bhatia, R., 2020. Advances in Intelligent Systems and Computing. In *Advances in Intelligent Systems and Computing: Vol. 1130 AISC*, Vol. 1130. https://doi.org/10.1007/978-3-030-39442-4_18.

[2] Bartlett, J. W., Carpenter, J. R., Tilling, K. & Vansteelandt, S., 2014. Improving upon the Efficiency of Complete Case Analysis when Covariates are MNAR. *Biostatistics*, Vol. 15, No. 4, 719–730. https://doi.org/10.1093/biostatistics/kxu023.

[3] Denhard, A., Bandyopadhyay, S., Habte, A. & Sengupta, M., 2021. A Comparison of Time Series Gap-Filling Methods to Impute Solar Radiation Data. *Proceedings - ISES Solar World Congress 2021*, Vol. 2021, 1049–1057. https://doi.org/10.18086/swc.2021.38.03.

[4] Galimard, J. E., Chevret, S., Curis, E, & Resche-Rigon, M., 2018. Heckman Imputation Models for Binary or Continuous MNAR Outcomes and MAR Predictors. *BMC Medical Research Methodology*, Vol. 18, No. 1, 1–13. https://doi.org/10.1186/s12874-018-0547-1.

[5] Keller, A. C. & Evans, J. M., 2019. Application of Random Forest Regression to the Calculation of Gas-Phase Chemistry within the GEOS-Chem Chemistry Model v10.

Geoscientific Model Development, Vol. 12, No. 3, 1209–1225. https://doi.org/10.5194/gmd-12-1209-2019.

[6] Lee, K. J., Carlin, J. B., Simpson, J. A. & Moreno-Betancur, M., 2023. Assumptions and Analysis Planning in Studies with Missing Data in Multiple Variables: Moving beyond the MCAR/MAR/MNAR Classification. *International Journal of Epidemiology*, Vol. 52, No. 4, 1268–1275. https://doi.org/10.1093/ije/dyad008.

[7] Mir, A. A., Kearfott, K. J., Çelebi, F. V. & Rafique, M., 2022. Imputation by Feature Importance (IBFI): A Methodology to Envelop Machine Learning Method for Imputing Missing Patterns in Time Series Data. *PloS one*, Vol. 17, No 1, e0262131.

[8] Mohamad, N. B., Lim, B. H. & Lai, A. C., 2021. Imputation of Missing Values for Solar Irradiance Data under Different Weathers using Univariate Methods. *IOP Conference Series: Earth and Environmental Science*, Vol. 721, No. 1. https://doi.org/10.1088/1755-1315/721/1/012004.

[9] Moritz, S. & Bartz-beielstein, T., 2017. imputeTS : Time Series Missing Value Imputation in R. *The R Journal*, Vol. 9, No. 1, 1–12. https://doi.org/10.32614/RJ-2017-009.

[10] Newman, D. A., 2014. Missing Data: Five Practical Guidelines. *Organizational Research Methods*, Vol. 17. No. 4, 372–411. https://doi.org/10.1177/1094428114548590.

[11] Peugh, J. L., Toland, M. D. & Harrison, H., 2023. A Tutorial for Handling Suspected Missing Not at Random Data in Longitudinal Clinical Trials. *The Quantitative Methods for Psychology*, Vol. 19, No. 4, 347–367. https://doi.org/10.20982/tqmp.19.4.p347.

[12] Phan, T. T. H., 2020. Machine Learning for Univariate Time Series Imputation. *2020 International Conference on Multimedia Analysis and Pattern Recognition (MAPR)*, 1–6. https://doi.org/10.1109/MAPR49794.2020.9237768.

[13] Phan, T. T. H., Caillault, É. P., Lefebvre, A. & Bigand, A., 2020. Dynamic Time Warping-based Imputation for Univariate Time Series Data. *Pattern Recognition Letters*, Vol. 139, 139–147. https://doi.org/10.1016/j.patrec.2017.08.019.

[14] Riyani, D., Prastyo, D. D. & Suhartono., 2019. Input Selection in Support Vector Regression for Univariate Time Series Forecasting. *AIP Conference Proceedings*, Vol. 2194, No. 1, 020105. https://doi.org/10.1063/1.5139837.

[15] Welch, G. & Bishop, G., 2006. An Introduction to the Kalman Filter. *Asian J Control*, 2.

[16] Wong, W. M., Lee, M. Y., Azman, A. S. & Rose, L. A. F., 2021. Development of Short-term Flood Forecast using ARIMA. *International Journal of Mathematical Models and Methods in Applied Sciences*, Vol. 15, 68–75. https://doi.org/10.46300/9101.2021.15.10.

[17] Wongoutong, C., 2021. Imputation Methods in Time Series with a Trend and a Consecutive Missing Value Pattern. *Thailand Statistician*, Vol. 19, No. 4, 866–879.