# PERBANDINGAN TEKNIK INTERPOLASI BERBASIS R DALAM ESTIMASI DATA CURAH HUJAN BULANAN YANG HILANG DI SULAWESI

*COMPARISON OF R-BASED INTERPOLATION TECHNIQUES TO ESTIMATE MONTHLY PRECIPITATION MISSING DATA IN SULAWESI*

**Muflihah*, Rizky Yudha Pahlawan**
BBMKG Wilayah IV Makassar, Jl.Prof.Dr.H. Abdurrahman Basalamah No.4, Makassar, 90231
**E-mail:* muflihah_mat@yahoo.co.id

Naskah masuk: 31 Juli 2016; Naskah diperbaiki: 3 September 2017; Naskah diterima: 22 Desember 2017

***ABSTRAK***

*Data hilang seringkali ditemukan pada jenis data dalam jangka waktu yang panjang (runtun waktu), termasuk data curah hujan. Analisis akan sulit dilakukan jika terdapat banyak data hilang dan terletak di pertengahan runtun waktu. Oleh karena itu, kajian ini bertujuan untuk mengestimasi data hilang menggunakan beberapa metode interpolasi yang tersedia pada perangkat lunak RStudio (bahasa R) yaitu na.StructTs, na.loft, na.approx, dan na.interp. Data yang digunakan adalah data curah hujan bulanan dengan pola monsunal, lokal, dan ekuatorial. Data hilang disimulasikan secara acak yang terbagi dalam tiga kategori, yaitu 5%, 10%, dan 17%. Hasil menunjukan bahwa metode na.StructTS menghasilkan nilai estimasi data hilang dengan nilai RMSE terkecil dan koefisien korelasi terbesar. Nilai koefisien korelasi terbesar diperoleh dalam mengestimasi data hilang sebanyak 5% dengan tipe monsun yakni antara 0,78-0,86.*

Kata kunci: data hilang, curah hujan, interpolasi, RStudio

***ABSTRACT***

*Missing data is often found on the type of data in the long term (time series),including rainfall data. The analysis will be difficult to be done if there is a lot of missing data and is located in the middle of timeseries. Therefore, this study aimed to estimate those missing data using interpolation methods in software RStudio (R language) including na.StructTs, na.loft, na.approx, and na.interp. The used data are monthly rainfall data with type of the monsoonal, local and equatorial. Missing data are randomly simulated and divided into three categories, which is 5%, 10% and 17%. The results show that na.StructTS method produces the estimated missing data with the smallest RMSE value and the greatest correlation coefficient. The highest coefficient correlation is found in the estimated of 5% missing data monsoonal type with range of 0.78 to 0.86.*

Keywords: missing data, precipitation, interpolation, Rstudio

## 1. Pendahuluan

Masalah data hilang seringkali ditemukan dalam berbagai jenis data runtun waktu. Hal ini akan menimbulkan kesulitan analisis dan proses pengambilan keputusan, sehingga dibutuhkan metode estimasi yang akurat dan efisien. Masalah data hilang juga seringkali ditemukan dalam data curah hujan. Hal ini disebabkan antara lain karena kerusakan alat, kelalaian petugas, penggantian alat, pemindahan lokasi alat, bencana, dan bahkan oleh hal lain yang tidak diketahui secara pasti penyebabnya. Menurut Cryer [1], jika posisi data hilang terdapat pada awal atau akhir data runtun waktu, maka data hilang tersebut dapat dihilangkan sehingga data yang digunakan ialah data setelah data hilang tersebut dibuang. Namun jika posisi data hilang tersebut terletak di tengah data, maka perlu dilakukan pendugaan (estimasi) untuk mengisi data hilang tersebut. Perlakuan umum yang sering digunakan untuk menduga data hilang adalah mengisi data hilang dengan nilai rataan dari deret datanya. Metode ini sudah tidak layak lagi digunakan karena banyak kekurangannya. Diantaranya, menyebabkan berkurangnya keragaman data yang dapat berakibat korelasi dalam data menjadi bias.

Prawaka, dkk [2] telah melakukan penelitian tentang analisis data curah hujan yang hilang dengan menggunakan metode *normal ratio, inversed squared distance,* dan rata-rata aljabar di Bandar Lampung

dengan nilai korelasi untuk data curah hujan harian 0,19-0,26 dan nilai korelasi untuk data curah hujan bulanan 0,67-0,72. Nilai korelasi estimasi curah hujan harian sangat lemah, sedangkan untuk curah hujan bulanan cukup kuat. Akan tetapi, model yang digunakan ini membutuhkan data curah hujan yang berada di sekitar lokasi data curah hujan yang hilang, sehingga jika lokasi data curah hujan yang hilang tersebut jaraknya berjauhan atau datanya kosong atau terletak di pulau-pulau kecil maka kemungkinan akan sulit dan tidak efisien dalam melakukan analisis, terlebih lagi bila tipe hujannya berbeda.

Pengisian data hilang juga telah dilakukan oleh Moritz,dkk [3] dengan membandingkan beberapa metode inputasi di program Rstudio untuk data runtun waktu univariat, menunjukkan hasil bahwa metode imputasi menggunakan interpolasi dengan kalman filter musiman dari *package "zoo"* dan interpolasi linier pada dekomposisi musiman dari *package"forecast"* paling efektif dalam menangani data hilang. Pada metode ini, data yang digunakan hanya data curah hujan yang mengandung data hilang, tidak perlu mencari data dari pengamatan di sekitarnya sehingga lebih efisien. Kalman filter dapat digunakan untuk estimasi data hilang karena kalman filter mengkombinasikan perhitungan model dan hasil ukuran [4]. Selain itu, berdasarkan penelitian yang telah dilakukan oleh Muflihah[5], estimasi data hilang pada data curah hujan tipe monsunal dengan menggunakan model *state space* melalui kalman fiter yang juga menggunakan *package "zoo"* di RStudio menghasilkan nilai korelasi yang kuat.

Berdasarkan distribusi data rata-rata curah hujan bulanan, umumnya wilayah Indonesia dibagi menjadi tiga pola hujan, yaitu pola ekuatorial, pola monsunal, dan pola lokal yang masing–masing pola mempunyai ciri khas yang berbeda. Pola ekuatorial berhubungan dengan pergerakan zona konvergensi ke arah belahan bumi utara dan selatan mengikuti pergeseran matahari melalui garis khatulistiwa. Pola ekuatorial dicirikan oleh tipe curah hujan dengan bentuk bimodial (dua puncak hujan) dimana terdapat curah hujan bulanan maksimum dua kali [6]. Pola monsunal memiliki distribusi berbentuk huruf 'V' atau 'U'. Pola ini berhubungan dengan angin monsun, yaitu saat monsun barat jumlah curah hujan berlimpah (musim hujan) sedangkan pada saat monsun timur jumlah curah hujan sangat sedikit (musim kemarau). Curah hujan dengan pola lokal sangat dipengaruhi oleh kondisi sirkulasi atmosfer atau kontur topografi setempat. Sebagai contoh hujan dengan pola lokal disebabkan oleh pemanasan lokal yang menyebabkan naiknya udara lembab dari aliran udara yang menuju ke dataran tinggi atau pegunungan. Wilayah dengan pola lokal memiliki distribusi hujan bulanan yang berkebalikan dengan pola monsunal [7]. Distribusi curah hujan bulanan dengan pola lokal ini berbentuk 'V' atau 'U' terbalik.

Kajian ini dilakukan untuk memperoleh nilai estimasi data curah hujan bulanan yang hilang dengan menggunakan beberapa metode interpolasi yang tersedia di RStudio (R) pada tiga pola curah hujan di Indonesia, yaitu pola monsunal, lokal, dan ekuatorial. Selanjutnya melakukan perbandingan metode interpolasi pada tiap pola curah hujan untuk menentukan hasil terbaik.

## 2. Metode Penelitian

Data curah hujan bulanan yang lengkap diperoleh dari kantor Balai Besar Meteorologi, Klimatologi, dan Geofisika Wilayah IV Makassar berupa data curah hujan bulanan dengan periode 30 tahun dari awal data untuk tiga lokasi yaitu Pattallassang Kabupaten Takalar (tipe monsunal), Batukaropa Kabupaten Bulukumba (tipe lokal), dan Lalos Tolitoli (tipe ekuatorial). Selanjutnya, dilakukan simulasi dengan menghilangkan data yang lengkap secara acak atau random sebanyak 18 data (5%), 36 data (10%), dan 61 data (17%) dari 360 data. Semakin banyak data yang dianggap hilang, semakin banyak posisi data yang dihilangkan berurutan.

Proses estimasi data hilang dilakukan dengan menggunakan *function na.StructTS*, *na.locf*, dan *na.approx* dalam *package zoo* dan *na.interp* dalam *package forecast* pada program RStudio. Selain itu pengolahan data dan visualisasi grafik menggunakan Microsoft Excel.

Langkah awal yang dilakukan adalah menyimpan data dalam bentuk format *csv* untuk diolah dalam R. Data yang hilang (kosong) diberi simbol "NA". Kemudian data diubah bentuk menjadi runtun waktu (time series) dan dilakukan interpolasi atau estimasi data hilang dari time series tersebut. Ada empat metode interpolasi yang digunakan dalam penelitian ini, yaitu:

1. ***na.StructTS*** adalah metode dengan cara mengisi data hilang "NA" menggunakan Kalman Filter musiman. Cocok digunakan untuk data runtun waktu yang bersifat musiman. Perintah dalam R adalah "`na.StructTS(data)`" pada *package zoo [8]*. Rumus Kalman Filter terdiri atas dua persamaan, yaitu persamaan observasi (1) dan persamaan transisi (2).

$$X_t = H\theta_t + \varepsilon_t \quad (1)$$

$$\theta_t = G\theta_{t-1} + K\eta_t \quad (2)$$

Keterangan:

$X_t$ = vektor observasi
$H$ = matriks observasi
$\theta_t$ = vektor transisi
G = matriks transisi
K = matriks input
$\varepsilon_t$ = vektor noise pada persamaan observasi
$\eta_t$ = vektor noise pada persamaan transisi[9].

2. ***na.locf*** merupakan metode dengan cara mengganti setiap nilai "NA" dengan data observasi

sebelumnya. Metode ini digunakan jika ada hubungan yang kuat antara pengamatan saat ini dengan pengamatan sebelumnya. Metode ini memiliki kelemahan jika terdapat perbedaan yang besar antara pengamatan saat ini ($t_n$) dengan pengamatan sebelumnya($t_{n-1}$)[10]. Perintah dalam R adalah "`na.locf(data)`" pada *package zoo*[8].

3. ***na.approx*** merupakan sebuah fungsi yang mengganti nilai "NA" melalui interpolasi linier. Perintah dalam R adalah "`na.aprrox(data)`" pada *package zoo [8]*.
4. ***na.interp*** merupakan metode yang sangat baik dalam menduga data hilang yang bersifat musiman kuat. Metode ini menggunakan interpolasi liner untuk non musiman dan dekomposisi *periodic stl* dengan runtun musiman untuk mengganti nilai yang hilang [11]. Perintah dalam R adalah "`na.interp(data)`" pada *package forecast* [10].

Langkah selanjutnya adalah menghitung error dan korelasi dari nilai estimasi dari keempat metode yang telah digunakan. Nilai error diperoleh dengan membandingkan nilai observasi yang lengkap dengan nilai estimasi. Nilai error dihitung dengan mencari nilai *root mean square error* (RMSE) antara data observasi dengan data estimasi model seperti pada persamaan (3).

$$RMSE(\bar{y}, y) = \sqrt{\frac{\sum_{t=1}^{n}(\bar{y}_t - y_t)^2}{n}} \quad (3)$$

Keterangan:
$\bar{y}t$= data prediksi ke-t
$yt$= data observasi ke-t
n = jumlah data

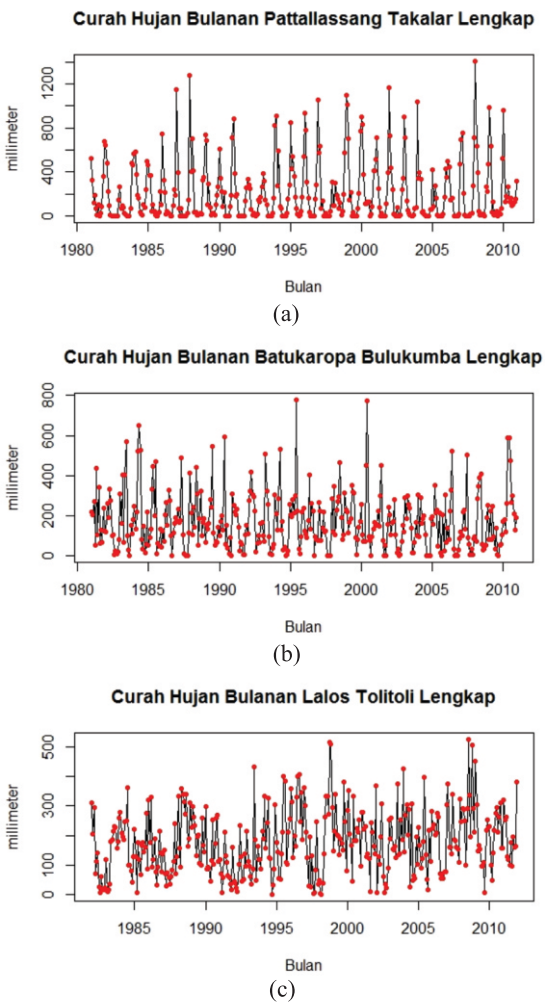
## 3. Hasil dan Pembahasan

Estimasi data hilang dengan menggunakan metode pada R pada tipe hujan monsunal, lokal dan ekuatorial, terlebih dahulu dilakukan dengan menampilkan plot data curah hujan masing-masing tipe (Gambar 1).

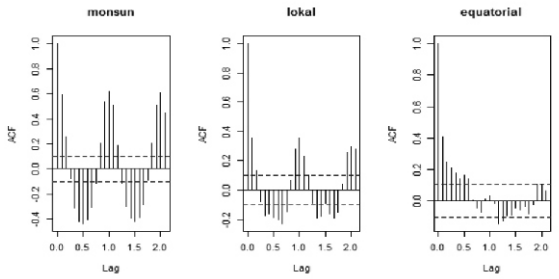
Plot fungsi autokolerasi (ACF) untuk tiap tipe curah hujan diberikan pada Gambar 2. ACF menunjukkan bahwa tipe monsunal dan lokal menunjukkan dipengaruhi faktor musiman, sedangkan tipe ekuatorial tidak. ACF ini berfungsi untuk melihat metode yang lebih cocok untuk estimasi data hilang.

Berdasarkan grafik pada Gambar 3a-c untuk data hilang sebanyak 5% terlihat bahwa nilai curah hujan yang tinggi sulit diestimasi oleh keempat metode, kecuali nilai estimasi dengan metode *na.locf* yang mendekati nilai aktual pada tipe monsunal. Estimasi data hilang pada tipe monsunal ada yang bernilai negatif yang kemudian diberi nilai 0, yakni pada metode *na.StrucTS* dan *na.interp*. Selain itu, estimasi data hilang pada tipe lokal juga ada yang bernilai negatif pada metode *na.interp*. Pada tipe lokal dan ekuatorial, ada nilai estimasi yang tepat sama dengan nilai aktual yaitu pada metode *na.StructTS*. Pada tipe monsunal, nilai estimasi yang tepat sama dengan nilai aktual yaitu pada metode *na.locf*.

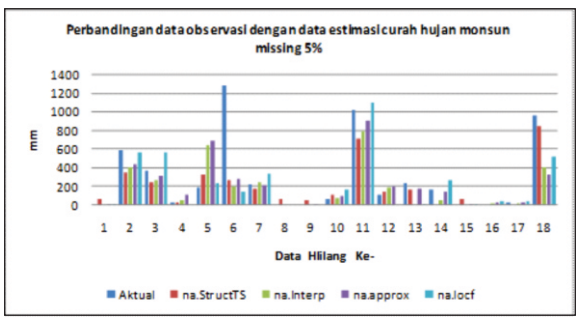
Berdasarkan Tabel 1, Tabel 2, dan Tabel 3, nilai RMSE terkecil dari keempat metode adalah *na.StructTS* kecuali untuk tipe monsun pada data hilang sebanyak 17% terendah adalah *na.approx* tetapi hanya selisih 0,74. Nilai korelasi terbesar dari keempat metode adalah *na.StructTS*, terutama untuk tipe monsun antara 0,78-0,86. Nilai korelasi terkecil dari keempat metode adalah *na.locf*.

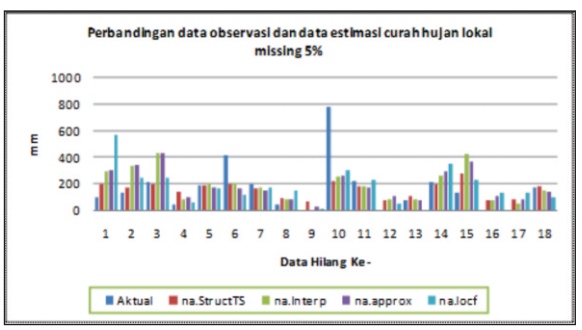


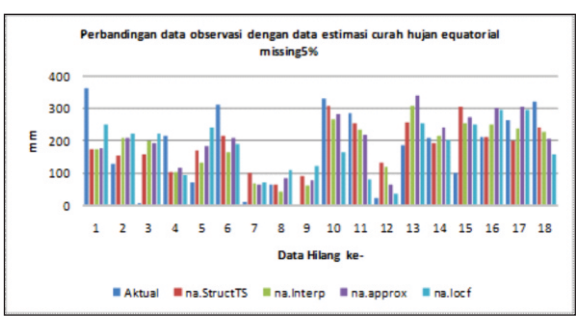

**Gambar 1. (a) Curah hujan bulanan tipe monsunal. (b) Curah hujan bulanan tipe lokal. (c) Curah hujan bulanan tipe ekuatorial**




**Gambar 2. Plot ACF untuk tiap tipe curah hujan**




**Gambar 3. Hasil estimasi curah hujan dengan data hilang 5% terhadap data aktual pada tipe (a) monsunal, (b) lokal dan (c) ekuatorial**

Secara umum, pengisian data hilang curah hujan bulanan dengan menggunakan empat metode interpolasi, metode *na.StructTS* yang paling baik karena mempunyai nilai RMSE terkecil dan korelasi terbesar. Meskipun pada tipe monsunal nilai RMSE metode *na.approx* paling kecil, tetapi dapat diabaikan karena memiliki selisih yang sedikit dari metode *na.StructTS*.

Nilai estimasi yang dihasilkan metode *na.StructTS* untuk data hilang sebanyak 5% pada tipe monsun, terdapat dua nilai estimasi yang negatif. Menurut Muflihah[5], salah satu kekurangan metode ini adalah menghasilkan nilai prediksi yang negatif, padahal data curah hujan tidak ada yang bernilai negatif (terendah 0). Jika nilai negatif yang diprediksi oleh metode *na.StructTS* terletak pada musim kemarau dan kondisi atmosfer dalam keadaan normal, maka bisa diganti dengan angka 0. Akan tetapi, jika nilai prediksi yang negatif terletak pada musim hujan atau musim kemarau dengan kejadian la nina, maka harus dicari metode lain yang cocok untuk mengestimasi data hilang tersebut. Begitu pula halnya dengan metode *na.interp* yang juga menghasilkan dua nilai prediksi yang negatif. Akan tetapi, pada metode *na.locf* yang memiliki nilai korelasi terendah, mampu mengestimasi nilai dengan tepat yaitu 0 dengan dua posisi data hilang, ini disebabkan karena metode ini mengganti nilai hilang dengan nilai data sebelumnya yang bernilai 0 pada musim kemarau.

Menurut Soewarno [12], analisis hidrologi memang tidak selalu diperlukan untuk pengisian data yang kosong atau hilang. Misalnya terdapat data kosong pada musim kemarau sedangkan analisis data hidrologi tersebut menghitung debit banjir musim penghujan maka dipandang tidak perlu melengkapi data pada periode kosong musim kemarau tersebut, tetapi bila untuk analisis kekeringan maka data kosong pada musim kemarau tersebut harus diusahakan untuk dilengkapi. Selain itu, apabila nilai estimasi negatif terletak pada posisi data hilang musim kemarau dalam kondisi atmosfer normal, maka bisa digantikan dengan 0.

**Tabel 1. Perbandingan nilai RMSE dan korelasi pada data tipe monsunal**

| Jumlah data/ Metode | ***na.StrucTS* tipemonsun** | ***na.locf*tipe monsun** | ***na.approx* tipe monsun** | ***na.interp* tipemonsun** |
|---|---|---|---|---|
| **n=342(5%)** | | | | |
| *RMSE* | **270,18** | 307,23 | 308,71 | 322,21 |
| *r* | **0,78** | 0,67 | 0,63 | 0,62 |
| **n=324(10%)** | | | | |
| *RMSE* | **231,50** | 285,98 | 289,40 | 247,52 |
| *r* | **0,81** | 0,41 | 0,60 | 0,73 |
| **n=299(17%)** | | | | |
| *RMSE* | 342,26 | 346,98 | **341,52** | 342,25 |
| *r* | **0,86** | 0,41 | 0,60 | 0,73 |

**Tabel 2. Perbandingan nilai RMSE dan korelasi pada data tipe lokal**

| Jumlah data/ Metode | ***na.StrucTS* tipelokal** | ***na.locf* tipelokal** | ***na.approx* tipelokal** | ***na.interp* tipelokal** |
|---|---|---|---|---|
| **n=342(5%)** | | | | |
| *RMSE* | **153,90** | 175,10 | 176,24 | 189,36 |
| *r* | **0,62** | 0,32 | 0,38 | 0,41 |
| **n=324(10%)** | | | | |
| *RMSE* | **120,92** | 143,72 | 148,25 | 160,13 |
| *r* | **0,59** | 0,33 | 0,38 | 0,43 |
| **n=299(17%)** | | | | |
| *RMSE* | **111,51** | 131,64 | 146,86 | 158,86 |
| *r* | **0,65** | 0,37 | 0,36 | 0,52 |

**Tabel 3. Perbandingan nilai RMSE dan korelasi pada data tipe ekuatorial**

| Jumlah data/ Metode | ***na.StrucTS*** **tipe ekuatorial** | ***na.locf*** **tipe ekuatorial** | ***na.approx*** **tipe ekuatorial** | ***na.interp*** **tipe ekuatorial** |
|---|---|---|---|---|
| **n=342(5%)** | | | | |
| *RMSE* | **99,44** | 124,46 | 106,24 | 102,41 |
| *r* | **0,58** | 0,28 | 0,54 | 0,54 |
| **n=324(10%)** | | | | |
| *RMSE* | **107,75** | 138,17 | 134,31 | 125,72 |
| *r* | **0,50** | 0,22 | 0,39 | 0,35 |
| **n=299(17%)** | | | | |
| *RMSE* | **112,37** | 147,70 | 134,23 | 123,89 |
| *r* | **0,47** | 0,07 | 0,32 | 0,35 |

Nilai estimasi yang dihasilkan metode *na.StrucTS* pada tipe hujan lokal pada data hilang sebanyak 5%, terdapat satu nilai estimasi yang sama dengan nilai aktualnya. Tidak ada nilai negatif yang dihasilkan dari metode *na.StructTS* ini, kecuali pada metode *na.interp* terdapat satu nilai estimasi yang negatif.

Nilai estimasi yang dihasilkan metode *na.StructTS* pada tipe ekuatorial terdapat dua nilai yang sama dengan data aktual yaitu pada data ke-8 dan data ke-16 pada jumlah data hilang sebanyak 5%. Pada tipe ini, tidak ada nilai estimasi yang negatif.

## 4. Kesimpulan

Dari keempat metode interpolasi data hilang tersebut, metode *na.StructTS* merupakan metode paling baik berdasarkan nilai RMSE terkecil dan nilai korelasi terbesar, khususnya untuk pola hujan monsunal yang memiliki korelasi kuat antara 0,78-0,86. Faktor musiman berpengaruh terhadap hasil estimasi pada metode *na.StructTS* dan *na.interp* yang dalam modelnya memperhitungkan pengaruh musiman. Keempat metode ini tidak memerlukan data yang berada di sekitar lokasi, cukup data yang ada diolah untuk mendapatkan nilai estimasi dari data hilang.

**Saran.** Data estimasi yang dihasilkan dari keempat metode yang digunakan pada R, belum bisa menghasilkan nilai curah hujan yang cukup tinggi. Sehingga diperlukan kajian lebih lanjut. Begitu pula untuk nilai negatif yang dihasilkan terutama pada tipe hujan monsunal.

## Daftar Pustaka


[1]J. . Cryer, Time series analysis. Boston: PWS-KENT Publishing Company, 1986.

[2]F. Prawaka, A. Zakaria, and S. Tugiono, "Analisis data curah hujan yang hilang dengan menggunakan metode Normal Ratio, Inversed Squared Distance, dan Rata-rata Aljabar (Studi kasus curah hujan beberapa stasiun hujan daerah bandar lampung," JRSDD, vol. 4, no. 3, pp. 397406, 2016.

[3]S. et. a. Moritz, "Comparison of different Methods for Univariate Time Series Imputation in R," Col. Univ. Appl. Sci., 2015.

[4]D. Kusnandar, M. N. Mara, Yundari, N. Satyahadewi, and N. N. Debataraja, "Mengatasi missing data hasil pengukuran satelit Altimetri Topex, Jason1 dan Jason2 dengan menggunakan metode kalman filter," Pros. Semin. Nasionar Mat. dan Pendidik. Mat. UNY, pp. 3740, 2013.

[5]Muflihah, "Prediksi curah hujan melalui model State Space untuk data hilang," Universitas Hasanuddin, 2015.

[6]E. Kristantri, "Prediksi curah hujan triwulanan di wilayah Sulawesi Selatan bagian barat dengan metode regresi komponen utama," IPB, 2014.

[7]Nuryadi, "Evaluasi dampak El Nino terhadap curah hujan dan masa tanam padi wilayah Sulawesi Selatan," IPB, 1998.

[8]P. Ratnasekera, "Introduction to R package - zoo," Stat. Comput., 2013.

[9]D. Sheung Chi Fung, "Methods for the Estimation of Missing Observation in Nonlinier Time Series Model using State Space Reprecentation," Edith Cowan University, 2006.[10]S. Morits, A. Sarda, T. Bartz-Beielstein, M. Zaefferer, and J. Storks, "Comparison of different methods for univariate time series imputation in R," Col. Univ. Appl. Sci., 2015.

[11]H. R.J., "forecast: Forecasting function for time series and linier models," R Packag. version 7.1, 2016.