Penanganan Data Hilang Menggunakan Metode MarkoviChain Monte Carlo (MCMC)

Sintia Arjabi Oktavianti*, Teti Sofia Yanti

Prodi Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Islam Bandung, Indonesia.

*sintiamobile001@gmail.com, tetisofiayanti@unisba.ac.id

Abstract. In an observation or research, there are often cases where the data observed or researched is incomplete because the measuring instrument used is inaccurate, damaged, not recorded and other technical problems. Incomplete data is commonly referred to as missing data. Missing data is an important problem in various studies because it can lead to bias and inaccuracy in predicting the response from observations. The method used to estimate missing data in the writing of this thesis is the *Multiple Imputation* method. *Multiple Imputation* used is the Markov Chain Monte Carlo (MCMC) method with Data Augmentation (DA) algorithm. The MCMC method is an algorithm for simulating conditional probability, which is suitable for any data pattern, where it is assumed that the underlying complete data follows a multivariate normal distribution. One case of incomplete data is the time-travel (Home – GSK) in Canada, the Province of Ontario in 2011 – 2012, for the variables *Distance*, *Maxspeed* and *FuelEconomy*. Based on the DA algorithm, complete data is formed for the three variables, so that these variables can be used for further analysis.

Keywords: *Missing Data, Multivariate Normal, MCMC Method, DA Algorithm, and Complete Data.*

Abstrak. Dalam sebuah pengamatan atau penelitian sering sekali terjadi kasus dimana data yang diamati atau diteliti tidak lengkap dikarenakan alat ukur yang digunakan kurang akurat, rusak, tidak tercatat dan masalah-masalah teknis lainnya. Data yang tidak lengkap biasa disebut sebagai data hilang (missing data). Data hilang merupakan suatu masalah penting dalam berbagai penelitian karena dapat menyebabkan terjadinya bias dan ketidakakuratan dalam memprediksi respon dari amatan. Metode yang digunakan untuk melakukan pendugaan data hilang pada penulisan artikel ini adalah metode Imputasi Ganda (*Multiple Imputation*). Imputasi Ganda (*Multiple Imputation*) yang digunakan adalah metode Markov Chain Monte Carlo (MCMC) dengan algoritma Data Augmentation (DA). Metode MCMC adalah algoritma untuk mensimulasikan peluang bersyarat, yang cocok untuk pola data apapun, di mana diasumsikan bahwa data lengkap yang mendasari mengikuti distribusi normal multivariat. Salah satu kasus data tidak lengkap yaitu time-travel (Home – GSK) yang ada di Negara Canada Provinsi Ontario pada tahun 2011 – 2012, untuk variabel *Distance*, *Maxspeed* dan *FuelEconomy*. Berdasarkan algoritma DA terbentuk data lengkap untuk ketiga variabel tersebut, sehingga variabel tersebut dapat digunakan untuk analisis lebih lanjut.

Kata Kunci: *Data Hilang, Normal Multivariat, Metode MCMC, Algoritma DA, dan Data Lengkap.*

Corresponding Author
Email: tetisofiayanti@unisba.ac.id

A. Pendahuluan

Dalam sebuah pengamatan atau penelitian sering sekali terjadi kasus dimana data yang diamati atau diteliti tidak lengkap dikarenakan alat ukur yang digunakan kurang akurat, rusak, tidak tercatat dan masalah-masalah teknis lainnya. Data yang tidak lengkap biasa disebut sebagai data hilang (missing data). Data hilang hampir muncul di semua kumpulan data (1).

Data hilang merupakan suatu masalah penting dalam berbagai penelitian karena dapat menyebabkan terjadinya bias dan ketidakakuratan dalam memprediksi respon dari amatan. Dalam penelitian yang dilakukan oleh Suhaimi, dkk (2017) (1) menyatakan bahwa tingkat data yang hilang kurang dari 1% akan dianggap sepele, 1 – 5% masih dapat dikelola dengan metode yang sederhana, 5 – 15% membutuhkan metode yang kuat untuk menangani data hilang tersebut dan lebih dari 15% dapat membahayakan hasil model penelitian.

Metode yang dapat digunakan untuk menangani data hilang adalah Metode Imputasi, yang terdiri dari metode imputasi tunggal (*Single Imputation*) dan imputasi ganda (*Multiple Imputation*). Dalam *Single Imputation*, data yang hilang diisi dengan nilai tunggal (*single value*), Masalah umum yang sering muncul dalam *Single Imputation* adalah mengganti nilai yang hilang dengan nilai tunggal dan kemudian memperlakukannya sebagai nilai sebenarnya. Ini adalah kelemahan dari *Single Imputation* (2).

Karena kelemahan dari *Single Imputation* sehingga metode *Multiple Imputation* (imputasi ganda) dikembangkan. *Multiple Imputation* memiliki banyak keuntungan sebagai metode penanganan data hilang, karena dapat menggunakan lebih dari satu kemungkinan untuk mengisi nilai yang hilang, atau hingga m imputasi. Salah satu metode imputasi ganda adalah metode Markov Chain Monte Carlo (MCMC), yang memiliki beberapa algoritma yaitu : algoritma Data Augmentation (DA), Gibbs Sampling, dan Metropolis Hasting (MH).

Salah satu kasus data tidak lengkap yaitu *time travel* (Home – GSK) yang ada di Negara Canada Provinsi Ontario pada tahun 2011 – 2012. Ketidaklengkapan data *time travel* pada provinsi tersebut mencapai rata-rata 18.05% pada variabel "*Distance*", 8.78% pada variabel "*Maxspeed*" dan 9.27% pada variabel "*FuelEconomy*". Karena persentasenya yang cukup besar sehingga tidak dapat ditangani dengan cara sederhana. Oleh karena itu, diperlukan suatu metode lain untuk menduga nilai data yang hilang pada data tidak lengkap tersebut yaitu metode Imputasi.

Berdasarkan latar belakang diatas, maka rumusan masalah dalam penelitian ini yaitu "bagaimana menangani kasus data hilang pada data *time travel* dengan menggunakan algoritma Data Augmentation (DA) yang merupakan salah satu algoritma dari metode Markov Chain Monte Carlo (MCMC)?". Selanjutnya, tujuan dari penelitian ini adalah mencari nilai dugaan untuk melengkapi data yang hilang pada data *time travel* dengan menggunakan algoritma Data Augmentation (DA) yang merupakan salah satu algoritma dari metode Markov Chain Monte Carlo (MCMC).

B. Metodologi Penelitian

Peneliti menggunakan Metode yang digunakan untuk melakukan pendugaan data hilang pada penulisan skripsi ini adalah metode Imputasi Ganda (*Multiple Imputation*). Imputasi Ganda (*Multiple Imputation*) yang digunakan adalah metode Markov Chain Monte Carlo (MCMC) dengan algoritma Data Augmentation (DA).

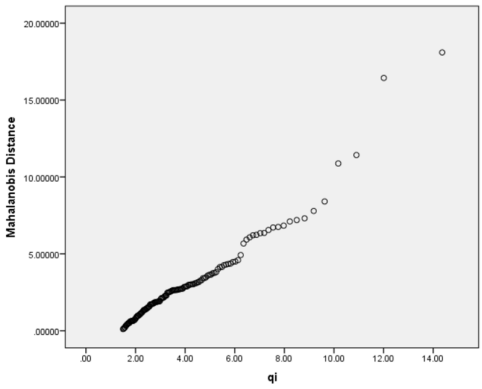
Data yang digunakan adalah Salah satu kasus data tidak lengkap yaitu *time travel* (Home – GSK) yang ada di Negara Canada Provinsi Ontario pada tahun 2011 – 2012. Data tersebut merupakan data sekunder dan variabel yang digunakan adalah variabel *Distance, Maxspeed* dan *FuelEconomy*.

C. Hasil Penelitian dan Pembahasan

Pengujian Distribusi Normal Multivariat

Pengujian tersebut dilakukan menggunakan SPSS dengan terlebih dahulu menghitung nilai square *Distance* (d_j^2) dan nilai quantil (q_i) dari distribusi chi-square $\left(\frac{j-0.5}{n}\right)$, dari hasil perhitungan tersebut dibuat sebuah Q-Q Plot untuk melihat apakah data tersebut berdistribusi

normal multivariat atau tidak. Dengan menggunakan SPSS diperoleh Q-Q Plot sebagai berikut :



Gambar 1. Plot Q-Q Normal Multivariat

Pada Q-Q plot diatas dapat dikatakan bahwa asumsi kenormalan multivariat pada data *time travel* terpenuhi karena titik-titik amatan dapat dikatakan cukup mengikuti arah garis lurus diagonal.

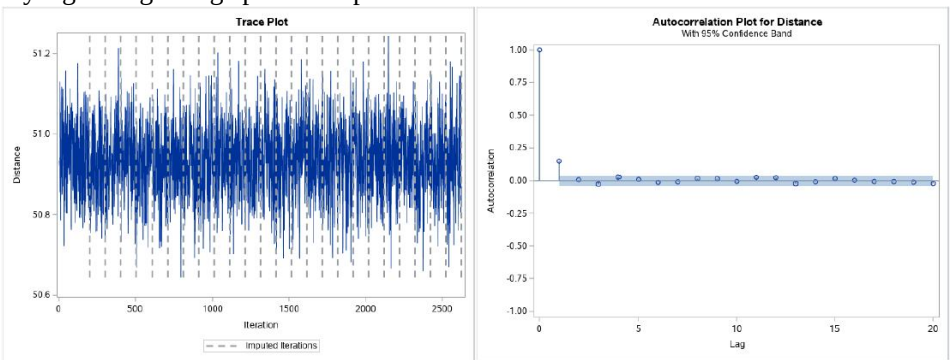
Dalam melakukan pendugaan tehadap data hilang, pada tahap imputasi (step-I), diperoleh nilai awal untuk vektor rata-rata dan matriks kovarians pada program SAS dari nilai maksimal posterior berdasarkan algoritma Expectation Maximization (EM). Nilai awal tersebut adalah

$$\mu_0 = \begin{pmatrix} 50.9328 \\ 127.4341 \\ 8.6906 \end{pmatrix} \text{dan } \Sigma_0 = \begin{pmatrix} 1.2491 & 0.8399 & 0.0983 \\ 0.8399 & 16.8923 & -0.1870 \\ 0.0983 & -0.1870 & 0.2530 \end{pmatrix}$$

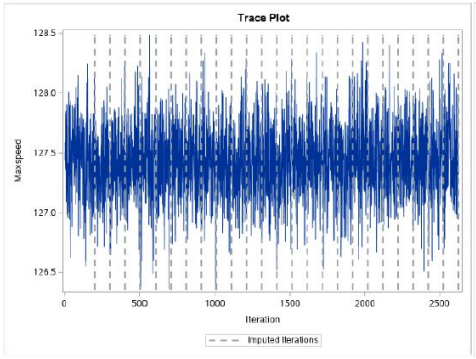
Pendugaan Nilai Data Hilang menggunakan Algoritma DA

Dari hasil proses iterasi, pada output SAS diperoleh plot *time series* dan plot autokorelasi yang menampilkan nilai rata-rata setiap variabel pengamatan pada setiap iterasi. Plot *time series* dan plot autokorelasi berguna untuk memantau konvergensi dari proses data augmentation.

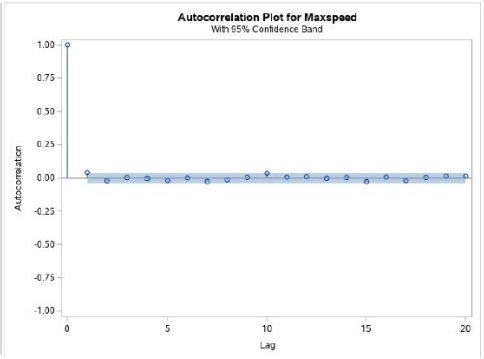
Plot *time series* dan plot autokorelasi untuk rata-rata masing-masing variabel setelah nilai data yang hilang diduga pada setiap iterasi adalah:



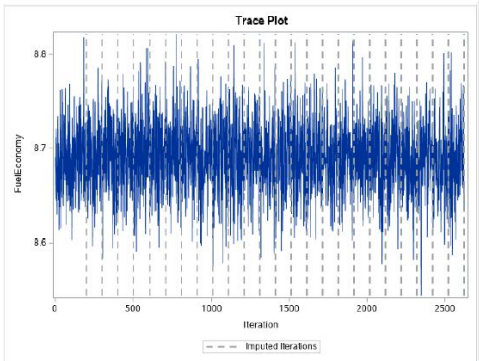
(a) Plot *time series Mean Distance* (b) Plot Autokorelasi *Mean Distance*



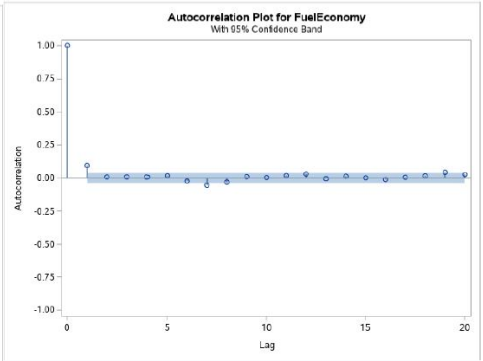
(c) Plot *time series Mean Maxspeed*



(*d*) Plot Autokorelasi *Mean Maxspeed*



(e) Plot *time series Mean FuelEconomy*



(f) Plot Autokorelasi *Mean FuelEconomy*

Gambar 2. Plot *time series* dan plot autokorelasi masing-masing variabel

Secara keseluruhan plot *time series* untuk variabel *Distance*, *Maxspeed*, dan *FuelEconomy* menunjukkan tidak adanya trend naik atau turun untuk masing-masing variabel tersebut. Lalu pada plot autokorelasi untuk variabel *Distance*, *Maxspeed*, dan *FuelEconomy* menunjukkan tidak adanya autokorelasi positif atau negatif yang signifikan, yang berarti bahwa rata-rata hasil dugaan data hilang untuk setiap variabel adalah tidak berautokorelasi. Sehingga dapat disimpulkan bahwa rata-rata setiap variabel bergerak stabil dan konvergen kepada distribusi posterior yang dituju.

Selanjutnya masuk ke tahap posterior (step-P). Secara baku, prosedur pada SAS menggunakan sebuah rantai tunggal untuk menciptakan 5 imputasi. Maka pada tahap ini, jumlah imputasi yang akan dijalankan untuk proses pendugaan nilai data hilang sebanyak $m = 5$ imputasi. Untuk setiap variabel akan diambil masing-masing satu nilai $Y_{i(mis)}$ yang terletak di urutan terakhir iterasi pada setiap imputasi. Maka akan diperoleh 5 nilai $Y_{i(mis)}$ yang kemudian akan dirata-ratakan dan dijadikan sebagai nilai dugaan akhir pada setiap kasus data hilang yang terjadi di setiap variabel.

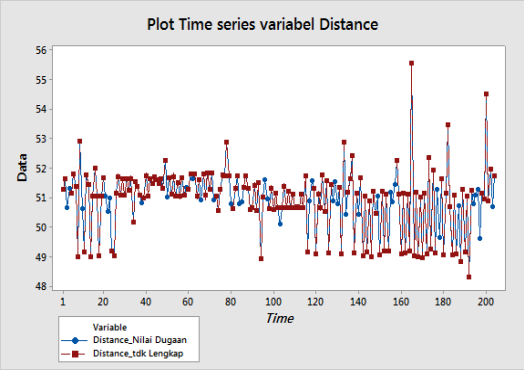
Dengan diperolehnya nilai dugaan untuk masing-masing data hilang, maka data *time travel* (Home – GSK) yang ada di Negara Canada Provinsi Ontario pada tahun 2011 – 2012 telah menjadi data yang lengkap. Data lengkap tersebut disajikan dalam Tabel 1.

Tabel 1. Data *time travel* (Home - GSK) pada tahun 2011 - 2012 sudah lengkap

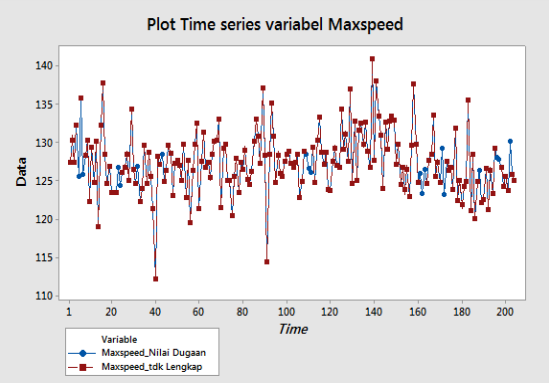
Date	**Start Time**	***Distance* (km)**	***Maxspeed* (km/jam)**	***FuelEconomy* (km/l)**	**J**
1/6/212	16:37	51.29	127.40	8.31	1
1/6/212	08:02	51.63	130.30	8.13	2
1/4/212	16:17	50.65	127.40	9.27	3

1/4/212	07:53	51.32	132.30	8.33	4
1/3/212	18:57	51.15	125.62	8.56	5
1/3/212	07:57	51.80	135.80	8.65	6
1/2/212	17:31	51.37	125.89	8.72	7
1/2/212	07:34	49.01	128.30	8.39	8
12/23/211	08:01	52.91	130.30	8.89	9
12/22/211	17:19	50.63	122.30	8.89	10
12/22/211	08:16	49.15	129.40	8.89	11
12/21/211	07:45	51.77	124.80	8.89	12
12/2/211	16:05	51.45	130.10	8.89	13
12/2/211	06:04	49.01	119.00	8.89	14
12/19/211	16:18	51.04	132.20	8.89	15
12/19/211	07:34	52.00	137.80	8.89	16
12/16/211	12:22	51.05	128.40	9.08	17
12/16/211	07:21	49.04	124.60	9.08	18
12/15/211	16:14	51.06	126.90	9.08	19
...	...	...	...	...	...
7/26/211	08:11	49.16	122.60	8.45	190
7/25/211	16:59	51.05	126.60	8.45	191
7/25/211	08:06	48.32	121.20	8.45	192
7/22/211	16:47	51.24	126.30	8.28	193
7/22/211	08:28	50.78	123.30	8.28	194
7/21/211	07:59	51.09	129.30	8.28	195
7/2/211	17:17	51.29	128.12	7.89	196
7/2/211	08:24	49.63	127.74	7.89	197
7/19/211	17:17	51.16	126.70	7.89	198
7/19/211	08:11	50.96	124.30	7.89	199
7/18/211	08:09	54.52	125.60	7.89	200
7/14/211	08:03	50.90	123.70	7.89	201
7/13/211	17:08	51.96	130.13	9.10	202
7/12/211	17:51	50.70	125.80	8.41	203
7/11/211	16:56	51.73	125.00	8.62	204

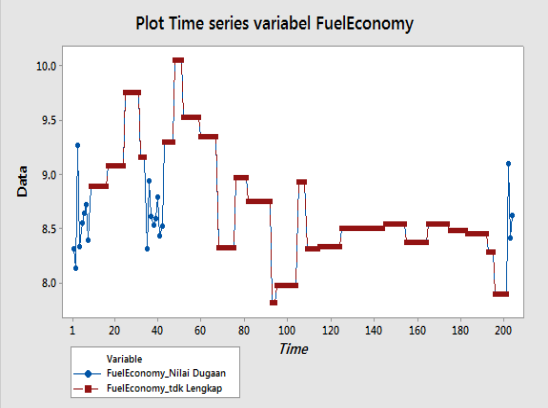
Pada Gambar 2, 3 dan 4 disajikan plot *time series* untuk variabel *Distance*, *Maxspeed*, dan *FuelEconomy* yang belum dilakukan pendugaan (Data Tidak Lengkap) dan sudah dilakukan pendugaan (Data Lengkap).



Gambar 2. Plot *Time series* variabel *Distance*



Gambar 3. Plot *Time series* variabel *Maxspeed*



Gambar 4. Plot *Time series* variabel *FuelEconomy*

D. Kesimpulan

Berdasarkan pembahasan dalam penelitian ini, peneliti menyimpulkan beberapa hasil penelitian sebagai berikut:

1. Data *time travel* berdistribusi normal multivariat untuk variabel *Distance*, *Maxspeed* dan *FuelEconomy* karena titik-titik amatan dapat dikatakan cukup mengikuti arah garis lurus diagonal.
2. Tahap Imputasi (Step-I), diperoleh nilai awal untuk vector rata-rata dan matriks kovarians sebagai berikut :

$$\mu_0 = \begin{pmatrix} 50.9328 \\ 127.4341 \\ 8.6906 \end{pmatrix} \text{dan } \Sigma_0 = \begin{pmatrix} 1.2491 & 0.8399 & 0.0983 \\ 0.8399 & 16.8923 & -0.1870 \\ 0.0983 & -0.1870 & 0.2530 \end{pmatrix}$$

3. Plot *time series* dan plot autokorelasi menunjukkan bahwa rata-rata setiap variabel bergerak stabil dan konvergen kepada distribusi posterior yang dituju.
4. Tahap Posterior (Step-P), dilakukan sebanyak $m = 5$ imputasi sehingga diperoleh 5 nilai $Y_{i(mis)}$ yang kemudian akan dirata-ratakan dan dijadikan sebagai nilai dugaan akhir pada setiap kasus data hilang yang terjadi di setiap variabel, sehingga diperoleh data yang lengkap untuk variabel tersebut mulai tahun 2011 sampai 2012.

Daftar Pustaka

[1] Suhaimi, N., Ghazali, N. A., Nasir, M. Y., Mokhtar, M. I., & Ramli, N. A. (2017). Markov Chain Monte Carlo Method For Handling Missing Data In Air Quality Datasets. Malaysian Journal Of Analytical Sciences, Xxi(3), 552-559.

[2] Nova, M., & Mukid, M. A. (2011). Pendugaan Data Hilang Dengan Menggunakan Data Augmentation. Media Statistika, Iv(2), 73-87.

[3] Susilawati, Ria, Sunendiari, Siti. (2022). *Peramalan Jumlah Penumpang Kereta Api Menggunakan Metode Arima dan Grey System Theory*, Jurnal Riset Statistika, 2(1), 1-13.