BAB III

MISSING DATA DAN PROSES RUNTUN WAKTU JANGKA PANJANG

3.1 *Missing Data*

Missing data merupakan hilangnya informasi atau data dalam suatu subjek. Terdapat banyak hal yang menyebabkan terjadinya *missing data*, yaitu dapat disebabkan salah penginputan, terkait respon dari perespon ataupun terdapat kendala pada alat pengumpulan data.

Adapun tipe dari *missing data* diantaranya:

1. *Missing Completely at Random* (MCAR) yang berarti bahwa *missing data* terjadi secara acak dari sampel lengkap,
2. *Missing not at Random* (MNAR) yang berarti bahwa probabilitas dari sebuah observasi yang hilang tidak derkaitan dengan hasil observasi lain. Sehingga nilainya tersebut berkaitan dengan dirinya sendiri, dan
3. *Missing at Random* (MAR) yang berarti bahwa probabilitas sebuah observasi dari *missing data* biasanya berkaitan dengan informasi yang diberikan responden dengan suatu alasan untuk tidak memberikan data.

(Donders, A.R.T *at al.*, 2006 : 1088)

Terdapat beberapa metoda imputasi yang biasa digunakan seperti imputasi rata-rata, imputasi maksimum maupun imputasi minimum dengan menggunakan bantuan *software* SPSS. Hanya saja menurut John W. Graham (2012:51) merekomendasikan untuk tidak pernah menggunakan imputasi rata-rata karena memasukan rata-rata pada *missing data* dapat mengurangi varians dari variabel yang bersangkutan dan dapat merusak kovarians dan autokovariansnya. Untuk mendapatkan estimasi yang baik pada *missing data* Jhon W.Graham merekomendasikan untuk menggunakan *Multiple Imputation* dengan Norm 2.03 atau menggunakan *Multiple Imputation* dan analisis dengan SAS. (2012:51)

Pada skripsi ini digunakan metode *Multiple Imputation* (MI) yang dapat menghasilkan inferensi valid untuk *missing data* pada *software* Norm 2.03, *Data Augmentatin* (DA) menganggap bahwa mekanisme *missing data* adalah *Missing at Random* (MAR) yang mana data yang hilang tidak tergantung pada nilai data yang hilang, tetapi tergantung pada nilai data yang teramati.

3.2 Proses runtun waktu jangka panjang

Sebuah kasus khusus dari proses runtun waktu adalah proses jangka panjang (*long memory)* atau *long-range dependent processes.* Terdapat berbagai definisi dari *long-range dependent,* namun pada intinya berdasarkan Hall (1997) dalam (Palma, Wilfredo, 2007:39) alasan semula konsep jangka panjang ini erat hubungannya dengan kestasioneran pada rata-rata.

(Haslet dan Raftery, 1989) mengatakan bahwa data yang dikategorikan sebagai data *long memory* ditandai dengan plot fungsi autokorelasi (fak) yang tidak turun secara eksponensial melainkan menurun secara sangat lambat. Fenomena *long memory* didalam data runtun waktu pertama kali diperkenalkan oleh Hurst (1951, 1956). Granger dan Joyeux (1980), serta Hosking (1981), mengembangkan model *Autoregressive Fractionally Integrated Moving Average* (ARFIMA) untuk memodelkan *long memory* pada data runtun waktu.

Model ARFIMA merupakan model terbaik yang dapat menjelaskan data deret waktu baik berupa *short memory* maupun *long memory* dengan *differencing d* yang dapat bernilai real (Moulines dan Soulier, 1999). Sehingga model ARFIMA dapat mengatasi kelemahan dari model ARIMA yang hanya dapat menjelaskan *short memory* dengan *differencing d* bernilai bilangan bulat. Berikut adalah definisi daripada proses jangka panjang dari McLeod and Hipel (1978)

Definisi 3.1

Sebuah proses dikatakan mengandung jangka panjang jika $\lim_{T\to\infty} \sum_{k=-T}^{T} |\rho_k|$ adalah tak hingga.

(Pfaff, 2008:40)

Menurut Hall pada tahun 1997 dalam Palma (2007: 39) mengatakan jika autokovarian suatu proses stasioner dapat dijumlahkan maka proses tersebut memiliki proses jangka pendek Sedangkan jika autokovarian suatu proses stasioner tidak dapat dijumlahkan (nilainya tidak terdefinisi) maka proses tersebut memiliki proses jangka panjang.

Definisi 3.2

Misalkan $\gamma(k) = cov(z_t, z_{t+k})$ adalah fungsi autokovarian pada *lag* ke-k dari proses stasioner $\{z_t; t \in \mathbb{Z}\}$, *long memory* dapat didefinisikan sebagai:

$$\sum_{k=-\infty}^{\infty} |\gamma_k| = \infty \text{ ... (3.1)}$$

(Palma, 2007 : 40)

3.3 Proses *Autoregressive Fractionally Integrated Moving Average* (ARFIMA)

Suatu deret z_t dikatakan mengikuti model ARIMA jika penyelisihan ke-d yakni $w_t = \nabla^d z_t$ adalah proses ARMA. Jika w_t adalah ARMA (p, q), maka z_t adalah ARIMA(p, d, q). Dalam prakteknya nilai d yang digunakan pada umumnya bernilai 1 atau paling banyak 3 (Wei, 1994).

Model *Autoregresive Fractionally Integrated Moving Average* ditulis ARFIMA (p, d, q) dapat memodelkan proses ketergantungan jangka pendek dan jangka panjang. Model ini memiliki tiga parameter sebagaimana model ARIMA yaitu p, d dan q, dimana p adalah parameter AR, q adalah parameter MA, sedangkan d parameter pembeda berupa bilangan pecahan d, yang menyebabkan nilai-nilai fak turun secara hiperbolik. Model ARFIMA (p, d, q) yang dikenalkan oleh Granger dan Joyeux (1980) adalah sebagai berikut,

$$\phi(B)(1-B)^d z_t = \theta(B)a_t \quad \ldots (3.2)$$

Dimana,

t : indeks dari pengamatan

d : parameter pembeda (bilangan pecahan)

μ : rata-rata dari pengamatan

∇^d : operator fraksional diferensi

a_t berdistribusi identik independen $N(0, \sigma_a^2)$

$\phi(B) = 1 + \phi_1 B + \cdots + \phi_p B^p$ adalah operator $AR(p)$

$\theta(B) = 1 + \theta_1 B + \cdots + \theta_q B^q$ adalah operator $MA(q)$

Untuk suatu nilai d bernilai pecahan, operator fraksional diferensi $(1-B)^d$ didefinisikan sebagai berikut:

$$(1-B)^d = 1 + \sum_{k=1}^{\infty} \frac{\Gamma(-d+k)}{\Gamma(-d)k!} B^k \quad \ldots (3.3)$$

Pada persamaan (3.3) untuk berbagai nilai k, ekuivalen dengan:

$$(1-B)^d = 1 + \sum_{k=1}^{\infty} \frac{\Gamma(-d+k)}{\Gamma(-d)k!} B^k$$

$$= 1 + \frac{\Gamma(-d+1)}{\Gamma(-d)1!} B^1 + \frac{\Gamma(-d+2)}{\Gamma(-d)2!} B^2 + \frac{\Gamma(-d+3)}{\Gamma(-d)3!} B^3 + \cdots$$

$$= 1 + \frac{(-d)!}{(-d-1)!1!} B + \frac{(-d+1)!}{(-d-1)!2!} B^2 + \frac{(-d+2)!}{(-d-1)!3!} B^3 + \cdots$$

$$= 1 - dB - \frac{d(1-d)B^2}{2} - \frac{d(1-d)(2-d)B^3}{6} - \cdots \quad \ldots (3.4)$$

Spectral density adalah sebuah fungsi real positif yang variabel frekuensi nya dihubungkan dengan fungsi deterministik dari waktu. Dalam (Palma, 2007 : 49), *spectral density* dari (3.2) adalah sebagai berikut:

$$f(\lambda_j) = f_0(\lambda_j)\left[2\sin\frac{\lambda}{2}\right]^{-2d}$$

$$= \frac{\sigma^2}{2\pi}\left[2\sin\frac{\lambda}{2}\right]^{-2d}\frac{|\theta(e^{-i\lambda})|^2}{|\phi(e^{-i\lambda})|^2}$$

Dengan $\frac{\sigma^2}{2\pi}\frac{|\theta(e^{-i\lambda})|^2}{|\phi(e^{-i\lambda})|^2}$ adalah *spektral density* dari proses ARMA(p,q) dan λ adalah frekuensi dari periodogram.

Fungsi autokovarian dari proses ARFIMA $(0,d,0)$ adalah

$$\gamma_0(k) = \sigma^2\frac{\Gamma(1-2d)}{\Gamma(1-d)\Gamma(d)}\frac{\Gamma(k+d)}{\Gamma(1+k-d)} \text{ ... (3.5)}$$

Dimana $\Gamma(.)$ adalah fungsi gamma, dan fungsi autokorelasi (fak) adalah:

$$\rho_0(k) = \frac{\Gamma(1-d)}{\Gamma(d)}\frac{\Gamma(k+d)}{\Gamma(1+k-d)} \text{ ... (3.6)}$$

Dan fungsi autokorelasi parsial (fakp) adalah sebagai berikut:

$$\phi_{nn} = \frac{d}{(n-d)} \text{ ... (3.7)}$$

Dengan $\phi_{nn} \approx \frac{d}{n}$ untuk ukuran *n* yang besar.

Menurut Boutahar dan Khalfaoui (2011), karakteristik utama dari sebuah model ARFIMA (p,d,q) adalah sebagai berikut,

1. Jika $d > 0{,}5$ maka Z_t adalah *invertible.*
2. Jika $d < 0{,}5$ maka Z_t adalah stasioner.
3. Jika $-0{,}5 < d < 0$ maka fungsi autokorelasi menurun lebih cepat daripada kasus $0 < d < 0{,}5$, model ini disebut *anti-persistent* atau *intermediate memory.*
4. Jika $0 < d < 0{,}5$ maka Z_t adalah sebuah model *long memory* yang stasioner dimana fungsi autokorelasi menurun secara hiperbolik menuju nol.
5. Jika $d = 0{,}5$ maka *spektral density* tidak terbatas pada frekuensi nol.

3.4 Pemodelan ARFIMA

Tahapan pemodelan ARFIMA melalui metode Box-Jenkins dilakukan melalui tahap identifikasi, estimasi, verifikasi dan peramalan. Berikut akan dijelaskan tiap tahapan Pemodelan ARFIMA.

3.4.1 Tahap Identifikasi Model ARFIMA

Identifikasi untuk model ARFIMA dilakukan dengan memperhatikan plot data runtun waktu untuk melihat pola data, plot fungsi autokorelasi (fak), plot fungsi autokorelasi parsial (fakp) dan transformasi Box-Cox untuk data yang tidak stasioner dalam varians (Wei, 1990).

Proses jangka panjang diidentifikasi bukan hanya melalui fungsi autokorelasi (fak) dan fungsi autokorelasi parsial (fakp). Namun diperlukan juga fungsi *spectral density* dari proses z_t yang diestimasi oleh *periodogram.*

Pada saat kondisi rata-rata tidak stasioner dilakukan *differencing* $((1-B)^d)$ untuk menstasionerkan data, untuk proses *short memory* dilakukan dengan *d* bernilai bilangan bulat, sedangkan untuk proses *long memory* dilakukan dengan $0 < d < 0{,}5$. Pada saat kondisi varians yang tidak stasioner dilakukan transformasi data, salah satunya dengan transformasi *Box-Cox (power transformation).*

3.4.2 Tahap Estimasi Parameter Model ARFIMA

Metode estimasi parameter d yang akan digunakan adalah metode GPH (*Gawake and Poter-Hudak*). Metode GPH pertama kali diusulkan oleh Geweke dan Poter-Hudak pada tahun 1983, dimana parameter *differencing* (d) dapat diestimasi secara konsisten dari regresi kuadrat terkecil yang diperoleh dari sebuah penaksiran persamaan regresi logaritma *spectral density.*

Kelebihan metode GPH dibandingkan dengan yang lainnya seperti metode Maksimum Likelihood (Sowell, 1992) dan Metode *Nonlinear Least Square* (Beran, 1995) adalah fleksibilitas dalam penaksiran parameternya. Penaksiran parameter pembeda d pada metode GPH dapat dilakukan secara langsung tanpa mengetahui nilai parameter p dan q terlebih dahulu. Pendekatan dengan *maksimum likelihood* berkendala pada penurunan fungsi autokovarians dari model ARFIMA (Darmawan,2008).

Irhamah (2007) telah melakukan penelitian mengenai perbandingan metode-metode pendugaan parameter model ARFIMA, yaitu metode Geweke and Poter Hudak (GPH), Estimasi Maksimum Likelihood (EML) dan *Nonlinear Least Square* (NLS). Hasil dari studi ini menyatakan bahwa penduga GPH meminimumkan bias dan AIC tetapi memaksimumkan MSE, sedangkan penduga EML untuk $d > 0$ adalah efisien namum memberikan bias dan AIC maksimum. Sebaliknya penduga NLS paling efisien untuk $d < 0$.

Sebagaimana yang telah diketahui sebelumnya, fungsi *spectral density* dari sebuah model stasioner z_t, dengan $t = 1,2,\dots,T$ adalah sebagai berikut:

$$f(\lambda_j) = f_0(\lambda_j)\left[2\sin\frac{\lambda_j}{2}\right]^{-2d}$$

Dengan $f_0(\lambda_j) = \frac{\sigma^2}{2\pi}\frac{|\theta(e^{-i\lambda})|^2}{|\phi(e^{-i\lambda})|^2}$ adalah *spektral density* dari proses ARMA(p,q) dan λ adalah frekuensi dari *periodogram.*

$$f(\lambda_j) = f_0(\lambda_j)\left[2\sin\left(\frac{\lambda_j}{2}\right)\right]^{-2d}$$

$$\frac{f(\lambda_j)}{f_0(0)} = \frac{f_0(\lambda_j)\left[2\sin\left(\frac{\lambda_j}{2}\right)\right]^{-2d}}{f_0(0)}$$

$$ln\left(\frac{f(\lambda_j)}{f_0(0)}\right) = ln\left(\frac{f_0(\lambda_j)\left[2\sin\left(\frac{\lambda_j}{2}\right)\right]^{-2d}}{f_0(0)}\right)$$

$$\ln f(\lambda_j) - \ln f_0(0) = \ln\left(\left[2\sin\left(\frac{\lambda_j}{2}\right)\right]^{-2d} . \frac{f_0(\lambda_j)}{f_0(0)}\right)$$

$$\ln f(\lambda_j) = \ln f_0(0) - d\ln\left[2\sin\left(\frac{\lambda_j}{2}\right)\right]^2 + \ln\left(\frac{f_0(\lambda_j)}{f_0(0)}\right)$$

Misalkan $\ln I(\lambda_j) = \ln f(\lambda_j) + \ln\left[\frac{I(\lambda_j)}{f(\lambda_j)}\right]$

$$\ln I(\lambda_j) = \ln f_0(0) - d\ln\left[2\sin\left(\frac{\lambda_j}{2}\right)\right]^2 + \ln\left[\frac{f_0(\lambda_j)}{f_0(0)}\right] + \ln\left[\frac{I(\lambda_j)}{f(\lambda_j)}\right]$$

Karena $\lambda_j \leq \lambda_m$ maka $\ln\left[\frac{f_0(\lambda_j)}{f_0(0)}\right]$ dapat diabaikan, sehingga persamaan di atas dapat ditulis kembali menjadi:

$$\ln I(\lambda_j) = \ln f_0(0) - d\ln\left[2\sin\left(\frac{\lambda_j}{2}\right)\right]^2 + \ln\left[\frac{I(\lambda_j)}{f(\lambda_j)}\right]$$

Dengan $x_j = \ln\left[2\sin\left(\frac{\lambda_j}{2}\right)\right]^2$ dan $y_j = \ln I(\lambda_j)$.

Estimasi dari parameter *long memory* d, disimbolkan dengan $\widehat{d}_{GPH}$, didefinisikan sebagai berikut:

$$\widehat{d}_{GPH} = -\frac{\sum_{j=1}^{m}(x_j - \bar{x})(y_j - \bar{y})}{\sum_{j=1}^{m}(x_j - \bar{x})^2} \quad \text{... (3.8)}$$

Dengan $\bar{x} = \frac{1}{m}\sum_{j=1}^{m} x_j$ dan $\bar{y} = \frac{1}{m}\sum_{j=1}^{m} y_j$,

3.4.3 Tahap Verifikasi

Tahap erifikasi meliputi pengujian terhadap *residual* diantaranya dilakukan pengujian *residual* saling bebas, mempunyai rata-rata nol dan varians konstan. Uji yang digunakan untuk asumsi ini adalah uji *Ljung-Box.* Selanjutnya akan dilakukan pengujian *residual* berdistribusi normal dengan menggunakan uji *kolmogorov smirnov.*

3.4.4 Tahap Peramalan Model ARFIMA

Tahapan analisis data runtun waktu selanjutnya adalah melakukan peramalan. Penaksir terbaik dari Z_{t+k} adalah $\widehat{Z}_{t+k}$.

Teorema Dekomposisi Wold merupakan sebuah alat pokok untuk menganalisis proses stasioner (Wold, 1938 dalam Palma, 2007).

Teorema 3.3

Beberapa proses stasioner merupakan penjumlahan dari sebuah proses stokastik dan sebuah proses deterministik; kedua proses ini adalah orthogonal dan dekomposisi unik.

(Palma, 2007 : 5)

Berdasarkan representasi teorema Wold, sebuah proses stokastik stasioner dapat dituliskan sebagai berikut:

$$Z_t = \sum_{j=0}^{\infty} \psi_j \, a_{t-j} = \psi(B) a_t \; \dots (3.9)$$

Proses (1.1) dikatakan invertible, jika terdapat barisan koefisien π_j sedemikian sehingga:

$$a_t = -\sum_{j=0}^{\infty} \pi_j \, Z_{t-j} \dots (3.10)$$

Asumsikan $\pi_0 = -1$ pada persamaan (3.10), proses Z_t dapat ditulis sebagai:

$$Z_t = a_t + \sum_{j=1}^{\infty} \pi_j \, Z_{t-j} \; \dots (3.11)$$

Berdasarkan persamaan (3.11), Z_t merupakan sebuah proses invertible. Prediksi linier terbaik dari observasi Z_t adalah sebagai berikut:

$$\widehat{Z}_t = \sum_{j=1}^{\infty} \pi_j \, Z_{t-j} \; \dots (3.12)$$

Dengan konsekuensi $a_t = y_t - \hat{y}_t$ merupakan proses dengan rata-rata nol dan varians konstan. Sehingga dapat diketahui untuk $\hat{Z}_{t+k}$ adalah sebagai berikut:

$$\hat{Z}_{t+k} = \sum_{j=1}^{\infty} \pi_j Z_{t+k-j} \text{ ... (3.13)}$$

Dari persamaan (3.13) diperoleh hasil sebagai berikut,

$$\hat{Z}_{t+1} = -\sum_{j=1}^{t} \pi_j Z_{t+1-j},$$

$$\hat{Z}_{t+2} = -\pi_1 Z_{t+1} - \sum_{j=2}^{t+1} \pi_j Z_{t+2-j},$$

Dan seterusnya.

3.4.5 Tahap Pemilihan Model Terbaik

3.4.5.1. *Mean Square Error* (MSE)

Kriteria penentuan model terbaik berdasarkan *residual* digunakan persamaan *Mean Square Error* (MSE)

$$MSE = \frac{SSE}{n-n_p} \text{ ... (3.14)}$$

Dengan $SSE = \sum_{i=1}^{n} e_i^2$, n merupakan banyaknya observasi dan n_p merupakan banyaknya parameter yang diestimasi.

3.4.5.2. *Akaike's Information Criterion* (AIC)

Untuk menilai suatu kualitas dari pemilihan model, Akaike pada tahun 1973 memperkenalkan kriteria informasi yangg mempertimbangkan banyaknya parameter. Kriteria tersebut dinamakan *Akaike's Information Criterion* (AIC). Dirumuskan sebagai berikut:

$$AIC = n \ln \hat{\sigma}_a^2 + 2M \text{ ... (3.15)}$$

Dimana,

M : banyaknya parameter dalam model

n : banyaknya observasi

$\hat{\sigma}_a^2$: estimasi dari *Mean Square Error*

Dengan kriteria memilih AIC yang paling kecil.

3.5 Metodologi Penelitian

3.5.1 Pengambilan Data

Data yang digunakan pada skripsi ini adalah data sekunder yang diperoleh dari hasil penelitian di BATAN Bandung. (Lestani, *et al.*,2007:331) menjelaskan bahwa Pengambilan sampel partikulat udara dilakukan seminggu dua kali selama 24 jam menggunakan *Gent stacked filter unit sampler* di lokasi sampling stasiun Badan Meterologi dan Geofisika-BMG Lembang. Filter yang digunakan adalah filter jenis Nuclepore polikarbonat yang berukuran dua macam yaitu filter halus (berpori-pori 0,4 μm) dan filter kasar (berpori-pori 8 μm). Penentuan konsentrasi $PM_{2.5}$ dilakukan menggunakan metode gravimetri yang diperoleh dari pengurangan hasil penimbangan berat sampel pada filter halus dengan berat filter halus kosong. Sebelum dilakukan penimbangan, filter dikondisikan pada ruang bersih dengan temperatur $18 - 25°C$ dan kelembaban maksimum kurang dari 55%. Penentuan reflektansi dari filter sampel dilakukan menggunakan alat EEL *Smoke Stain Reflectometer, Diffusion System, Ltd, Model 43D.*

Tata cara pengukuran reflektans BC menggunakan EEL *smoke stain reflectometer* adalah sebagai berikut:

1. Sampel yang akan diukur harus disimpan (dikondisikan) minimal 12 jam pada kondisi yang sama dengan alat EEL *Smoke Stain Reflectometer.*
2. Sampel yang akan diukur harus ditangani menggunakan pinset yang bersih.
3. Alat dihubungkan dengan tegangan jala-jala 220-240V, tombol ON ditekan lalu dibiarkan minimal selama ½ jam agar kondisi alat stabil.

4. Angka pada *display* diatur hingga menunjukan angka 00,0 dengan memutar tombol ZERO tanpa memasang *Reflectometer Lead* (RL) pada soket INPUT.
5. Kabel RL dipasang pada soket INPUT, kemudian *Reflectometer Lead* (RL) diletakan di atas standar putih. Tombol COARSE atau tombol FINE diputar hingga angka pada *display* menunjukan angka 100.
6. Untuk pengukuran filter halus sampel partikulat udara, *Reflectometer Lead* (RL) diletakan di atas standar abu-abu, kemudian tombol COARSE dan FINE diputar hingga angka pada *display* menunjukan angka yang sesuai dengan nilai yang didapatkan dari nilai pengukuran 5 filter halus kosong pada standar abu-abu.
7. Untuk pengukuran filter kasar sampel partikulat udara, *Reflectometer Lead* (RL) diletakan di atas standar abu-abu, kemudian tombol COARSE dan FINE diputar hingga angka pada *display* menunjukan angka yang sesuai dengan nilai yang didapatkan dari nilai pengukuran 5 filter kasar kosong pada standar abu-abu.
8. Sampel partikulat udara diletakan pada standar putih dengan posisi sampel (debu) di atas, kemudian RL diletakan di atas sampel tersebut.
9. Pengukuran dilakukan sebanyak 3 kali untuk masing-masing sampel.

3.5.2 Statistika Deskriptif

3.5.3 Estimasi *Missing Data*

Metode yang digunakan untuk mengatasi *missing data* pada data karakteristik *black carbon* partikulat udara halus $PM_{2.5}$ di Lembang Bandung adalah menggunakan metode *multiple imputation* dengan bantuan *software* Norm 2.03. Sebagaimana telah dijelaskan pada bagian 3.1.

3.5.4 Aplikasi Pemodelan ARIMA

Langkah-langkah pemodelan ARIMA adalah sebagai berikut:

1. Tahap Identifikasi Model
2. Tahap Estimasi Parameter
3. Tahap Verifikasi (uji keberartian koefisien dan uji *lack of fit*)
4. Tahap Pemilihan Model Terbaik

3.5.5 Aplikasi Pemodelan ARFIMA

Langkah-langkah pemodelan ARFIMA hampir sama dengan langkah-langkah pemodelan ARIMA. Perbedaannya hanya pada tahap identifikasi, dimana untuk pemodelan ARFIMA melakukan identifikasi *long memory*.

3.5.6 Pemilihan Model Terbaik

Memilih model terbaik antara model ARIMA dan ARFIMA.