

ANALISIS *OUTLIER* DATA CURAH HUJAN BERDASARKAN *TROPICAL RAINFALL MEASURING MISSION* UNTUK WILAYAH JAWA-BALI (*OUTLIER ANALYSIS RAINFALL DATA BASED ON TROPICAL RAINFALL MEASURING MISSION FOR JAVA-BALI REGIONS*)

Ina Juaeni

Peneliti Pusat Sain dan Teknologi Atmosfer, LAPAN
e-mail: inajuaeni@yahoo.com, ina_j@bdg.lapan.go.id

ABSTRACT

The purpose of this study was to determine the statistical outlier in *Tropical Rainfall Measuring Mission* (TRMM) rainfall data having three hourly time resolution. Outliers in the rainfall data were defined as the rainfall data that is numerically distant from the rest of the data or is one that appears to deviate markedly from other members of the sample. The methods used in this study, to determine outliers, were statistical, that are arithmetic mean value plus deviation where deviation must be greater than three standard deviation. The result showed that outliers occur in an area of $2^{\circ} \times 8,5^{\circ}$ on the research region with an average frequency of 1,16%. Outlier values are very typical, implying its rain formation process.

Key words: *Outlier, Three hourly TRMM rainfall, Monthly mean*

ABSTRAK

Penelitian ini bertujuan menentukan *outlier* secara statistik dalam data curah hujan TRMM dengan resolusi waktu setiap 3 jam di wilayah Jawa-Bali. *Outlier* dalam data curah hujan didefinisikan sebagai data curah hujan yang secara numerik berbeda jauh dengan data lainnya atau data curah hujan yang sangat menyimpang dibandingkan data lain dalam suatu kelompok data. Pada penelitian ini, digunakan metode statistik untuk menentukan *outlier* yaitu rata-rata ditambah simpangan, dengan simpangan yang lebih besar dari 3 kali deviasi standar. Hasilnya menunjukkan bahwa *outlier* terjadi pada area $2^{\circ} \times 8,5^{\circ}$ di dalam wilayah penelitian dengan frekuensi rata-rata 1,16%. Nilai *outlier* sangat khas, secara tidak langsung menunjukkan proses pembentukan hujannya.

Kata kunci: *Outlier, Curah hujan setiap 3 jam TRMM, Rata-rata bulan*

1 PENDAHULUAN

Informasi curah hujan merupakan informasi elemen cuaca/iklim yang sangat penting. Pada kenyataannya, informasi curah hujan sangat sukar diperoleh akibat terbatasnya data tersebut. Munculnya data satelit TRMM, sebagai data spasial dengan cakupan luas disertai resolusi waktu yang cukup tinggi, menjadi solusi (sementara) untuk berbagai aktivitas kajian/penelitian atmosfer. Tersedianya data curah hujan TRMM dengan resolusi waktu 3 jam memberi

peluang untuk kajian/penelitian lebih mendalam tentang proses-proses atmosfer dalam rentang waktu tersebut. Namun sebelum digunakan untuk tujuan tersebut, analisis awal terhadap data perlu dilakukan agar diperoleh pemahaman terhadap karakter data secara umum. Dalam penelitian ini diuraikan salah satu tahap awal dalam metode analisis data yaitu pendeteksian *outlier*, sebagai tahap awal untuk analisis data selanjutnya. Data yang digunakan adalah data curah hujan setiap 3 jam TRMM type 3B42 rata-rata bulanan selama 13 tahun.

Penelitian ini bertujuan menentukan *outlier* secara statistik dalam data curah hujan TRMM dengan resolusi waktu 3 jam. Data *outlier* curah hujan setiap 3 jam setelah dievaluasi dengan melibatkan data kondisi permukaan seperti serapan atau daya tampung permukaan terhadap curah hujan, debit curah hujan, *runoff* dan analisis dampak bagi kehidupan, maka dapat dimanfaatkan untuk mengidentifikasi kejadian-kejadian ekstrim. *Outlier* dalam data curah hujan didefinisikan sebagai data curah hujan yang nilainya jauh diatas nilai rata-rata. Barnett dan Lewis (1994) membuat definisi bahwa *outlier* adalah objek yang secara numerik berbeda dengan data lainnya. Sedangkan Grubbs (1969) mendefinisikan *outlier* sebagai objek yang menunjukkan simpangan terhadap lainnya.

2 STUDI PUSTAKA

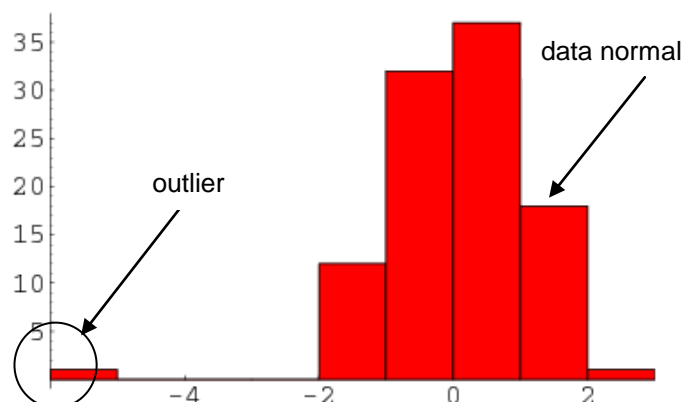
Pendeteksian anomali juga dianggap sebagai pendeteksian *outlier* (Hans *et al.*, 2010) yaitu pendeteksian pola yang tidak sesuai perilaku normalnya (Varun *et al.*, 2009). Pola yang terdeteksi tersebut disebut sebagai anomali dan sering diterjemahkan sebagai informasi yang kritis dan *aksinabel*. Anomali atau *outlier* juga diidentifikasi sebagai perubahan, deviasi, kejutan, penyimpangan, kekhasan, intrusi, dan lain-lain.

Hawkins (1980) dalam Hans *et al.* (2009) menyatakan bahwa *outlier* adalah observasi yang menyimpang banyak dari observasi lainnya sehingga dicurigai sebagai hasil dari mekanisme yang

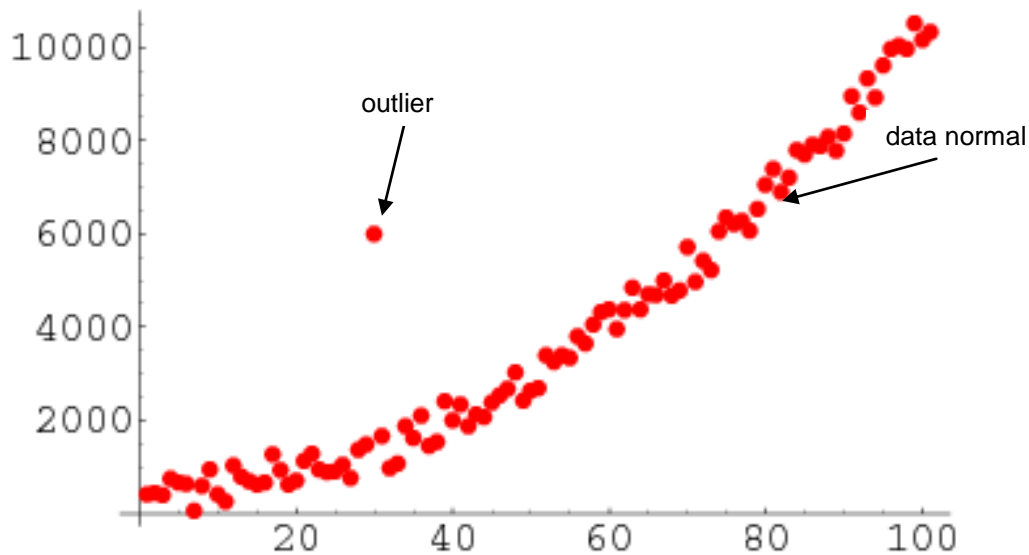
berbeda. Hans memberikan beberapa contoh yang disebut *outlier*, seperti wabah tetanus yang diakibatkan oleh penyebaran silang antar rumah sakit dalam suatu kota. Definisi ketidaknormalan suatu *outlier* sangat tergantung kepada aspek yang berkaitan dengan, misalnya, frekuensi dan korelasi spasial.

Pendeteksian *outlier* (Rousseeuw dan Leroy, 1996) sudah digunakan berabad-abad yang lalu untuk mendeteksi dan jika perlu menghilangkan anomali tersebut dari kumpulan data. Pendeteksian *outlier* berguna untuk mengidentifikasi sistem yang salah sebelum menimbulkan akibat yang fatal. Sebelumnya, metode pendeteksian *outlier* bisa suka-suka karena tidak ada standar, sangat tergantung waktu dan tempat. Namun sekarang dapat digunakan teknik yang mendasar dan sistematis karena perkembangan sains komputer dan statistik.

Definisi lain dari *outlier* adalah objek yang terletak jauh atau berbeda jauh dari pola distribusinya (Moore dan McCabe, 1999). Biasanya, keberadaan *outlier* mengindikasikan adanya masalah dalam model yang dirancang atau karena adanya kesalahan dalam pengukuran. *Outlier* dapat diamati dengan mudah melalui histogram (Gambar 2-1) atau diagram *scatter* (Gambar 2-2). Ketika membuat plot data, kadangkala ditemukan beberapa titik yang sangat berbeda dengan data lainnya atau nilai data tersebut tidak konsisten dengan data yang lain. Jika terjadi demikian, maka titik/data tersebut dianggap sebagai *outlier*.



Gambar 2-1: *Outlier* dan kumpulan data normal (Renze, 2011)



Gambar 2-2: *Outlier* dan kumpulan data normal (Renze, 2011)

Dengan beberapa definisi *outlier* seperti yang disebutkan di atas, maka bermunculan berbagai metode untuk mengidentifikasi *outlier* atau menghilangkan *outlier* dari kumpulan data.

Outlier adalah nilai yang jauh lebih kecil dari 1,5 x range interkuartil, lebih besar dari kuartil ketiga atau lebih kecil dari kuartil pertama (Renze, 2011). Kuartil didefinisikan sebagai persentil ke 25 dan ke 75. Jika kuartil terbawah adalah Q1 dan kuartil teratas adalah Q2, maka perbedaan antara Q2 dan Q1 disebut sebagai *range* interkuartil/IQ (Weisstein, 2011).

Hans *et al.* (2009) memberikan rumusan bahwa *outlier* adalah nilai yang simpangannya lebih besar dari 3x deviasi standar. Batasan seperti itu juga disampaikan dalam http://www.mathworks.com/help/techdoc/data_analysis/ (2011).

Jika distribusi data mengikuti atau mendekati distribusi normal, maka 68% nilai data berada dalam wilayah satu deviasi standar dari rata-rata (secara matematik ditulis, $\mu \pm \sigma$, dimana μ adalah rata-rata aritmatik), atau 95 % berada dalam wilayah dua deviasi standar dari rata-rata ($\mu \pm 2\sigma$), dan sekitar 99,7% terletak dalam wilayah tiga deviasi standar dari rata-rata ($\mu \pm 3\sigma$). Ini

dikenal dengan kaidah 68-95-99,7 atau *kaidah empirik*. Maka di luar wilayah tersebut atau dalam wilayah $> (\mu \pm 3\sigma)$ dapat dianggap sebagai *outlier*.

Salah satu metode untuk mengidentifikasi *outlier* yang kemudian disebut model *slippage*, yaitu:

$$g(x) = N(\mu + a\sigma, \sigma^2) \quad (2-1)$$

dengan $g(x)$ adalah distribusi *outlier*, μ adalah nilai rata-rata dan σ adalah deviasi standar (Hans *et al.*, 2009). Sedangkan yang menggunakan metode statistik contohnya adalah tes perbandingan *maximum-likelihood* (tes Likelihood-ratio), yaitu:

$$T = (Y_{n:n} = \bar{Y}_n) / S \quad (2-2)$$

dengan $Y_{n:n}$ adalah data terbesar, \bar{Y}_n rata-rata dan S^2 adalah variansi

Dengan metode ini, data terbesar menjadi *outlier* (Barnett dan Lewis, 1994).

Telah disebutkan di atas bahwa *outlier* adalah data dengan nilai simpangan $> 3\sigma$. *Outlier* dapat dihilangkan dari kumpulan data, akan tetapi harus hati-hati karena menghilangkan *outlier* mempunyai dampak yang besar terhadap deviasi standar dibandingkan terhadap

rata-rata. Dalam termin atmosferis, menghilangkan *outlier* harus lebih hati-hati lagi karena akan berpengaruh terhadap analisis kondisi cuaca secara langsung.

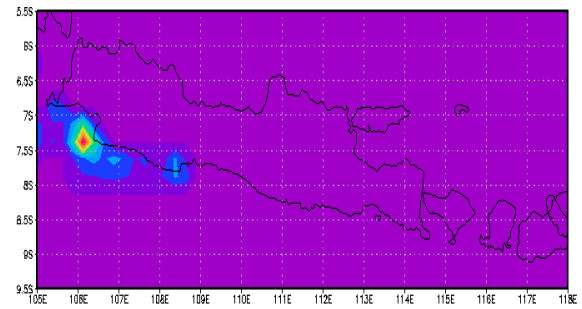
Pendeteksian *outlier* dalam data multivariat, dapat dilakukan dengan menggunakan jarak Mahalanobis (D^2) kemudian membaginya dengan derajat kebebasan (df) yang bernilai sama dengan jumlah variabel. Sehingga nilai (D^2/df) mengikuti nilai distribusi t . Kemudian dihitung nilai peluang (signifikansi) dari nilai (D^2/df) tersebut. Data yang signifikansinya lebih kecil dan sama dengan 0,001 dianggap sebagai *outlier* (Juani, 2010). Data yang diidentifikasi sebagai *outlier* harus dianalisis apakah merepresentasikan populasi atau tidak merepresentasikan populasi. Jika dinilai tidak merepresentasikan populasi, data harus dibuang. Namun, perlu diperhatikan apakah penghapusan *outlier* dapat mengubah struktur data yang sebenarnya.

3 DATA DAN METODOLOGI

Dalam penelitian ini ada beberapa batasan yang digunakan, yaitu:

- Data curah hujan setiap 3 jam (waktu sinoptik/Z/UTC/GMT) berasal dari satelit TRMM resolusi $0,25 \times 0,25^\circ$ yang diunduh secara bebas dari http://disc2.nascom.nasa.gov/Giovanni/tovas/TRMM_V6.3B42.2.shtml.
- Data yang digunakan adalah data rata-rata bulan periode tahun 1998 sampai dengan 2010 yang distribusinya mendekati distribusi normal.
- Lokasi penelitian meliputi Pulau Jawa, Bali dan Lombok ($5,5^\circ \text{LS} - 9,5^\circ \text{LS}$, $105^\circ \text{BT} - 118^\circ \text{BT}$) seperti tampak pada Gambar 3-1.

Metode identifikasi *outlier* yang digunakan dalam penelitian ini adalah metode statistik. *Outlier* adalah data yang simpangannya dari rata-rata lebih besar dari tiga kali deviasi standar atau $X > \mu \pm 3\sigma$ (3-1)



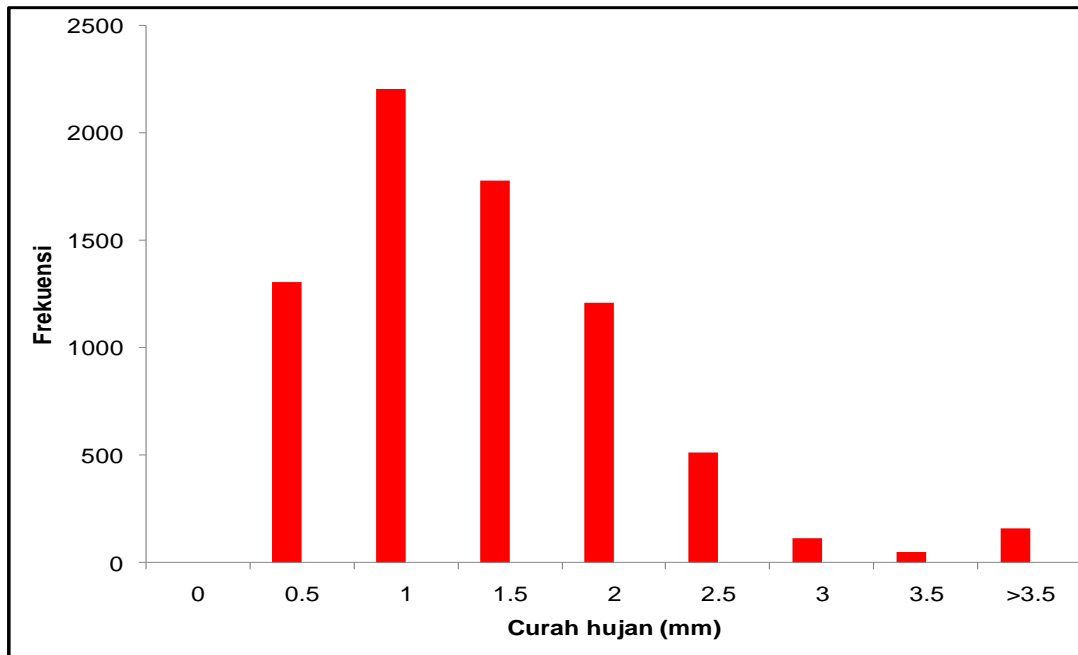
Gambar 3-1: Lokasi penelitian

4 HASIL DAN PEMBAHASAN

4.1 Hasil

Sebagai syarat awal diberlakukannya metode deviasi standar untuk deteksi *outlier* adalah bahwa data harus mempunyai distribusi normal atau mendekati distribusi normal. Dengan demikian terlebih dulu data ditentukan distribusinya. Dengan teknik yang sederhana yaitu dengan membuat histogram. Dari histogram diperoleh kesimpulan bahwa seluruh data curah hujan setiap 3 jam mendekati distribusi normal. Sebagai contoh, pada Gambar 4-1 diperlihatkan histogram data curah hujan rata-rata bulan Januari yang menunjukkan distribusi yang mendekati distribusi normal. Data curah hujan rata-rata bulan Januari memiliki frekuensi tertinggi pada intensitas curah hujan 1 mm, kemudian 1 mm, kemudian diikuti oleh intensitas 1,5 mm dan 2 mm (0,5 mm). Curah hujan dengan intensitas lebih besar dari 2 mm mempunyai frekuensi yang rendah. Meskipun distribusi tidak simetri, namun distribusi curah hujan setiap 3 jam rata-rata bulan Januari bisa dianggap distribusi normal. Histogram untuk data lain (rata-rata Februari sampai dengan rata-rata Desember ditunjukkan pada Gambar 4-2.

Selanjutnya, untuk menentukan *outlier* digunakan parameter simpangan data. Simpangan data ditentukan dari deviasi standar. Jika diterapkan sebuah distribusi dengan simpangan 3 kali σ (distribusi normal), maka ada sejumlah data yang dapat dianggap sebagai *outlier*, yaitu $X > \mu \pm 3\sigma$ ($X = \text{outlier}$, $\mu =$ rata-rata, $\sigma =$ deviasi standar). Prosentase jumlah data *outlier* rata-rata bulan yang sama diperlihatkan pada Tabel 4-1.



Gambar 4-1: Histogram data curah hujan setiap 3 jam rata-rata bulan Januari

Tabel 4-1: JUMLAH *OUTLIER* PER 3 JAM DAN PROSENTASE JUMLAH *OUTLIER* TOTAL PERBULAN RATA-RATA SELAMA 13 TAHUN (1998 SAMPAI DENGAN 2010)

Bulan	Jumlah <i>outlier</i> per 3 jam								Prosentase jumlah <i>outlier</i> terhadap jumlah seluruh data
	00	03	06	09	12	15	18	21	
Januari	0	1	1	19	21	7	1	0	0,68
Februari	0	0	4	22	16	4	11	3	0,82
Maret	0	1	10	21	18	4	9	0	0,86
April	1	0	11	13	25	6	6	0	0,84
Mei	7	0	5	15	22	13	10	13	1,16
Juni	3	1	1	16	22	19	10	6	1,06
Juli	11	14	11	19	20	16	20	16	1,73
Agustus	26	18	26	24	22	20	16	18	2,31
September	8	13	12	28	27	20	23	6	1,87
Oktober	4	0	9	21	19	13	10	9	1,16
November	0	0	15	18	10	11	8	8	0,95
Desember	0	0	1	24	5	7	0	0	0,50
Rata-rata									1,16

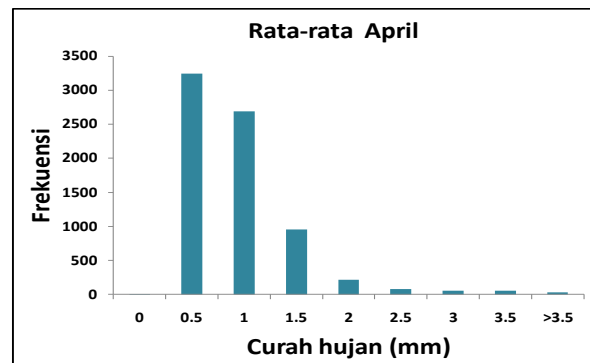
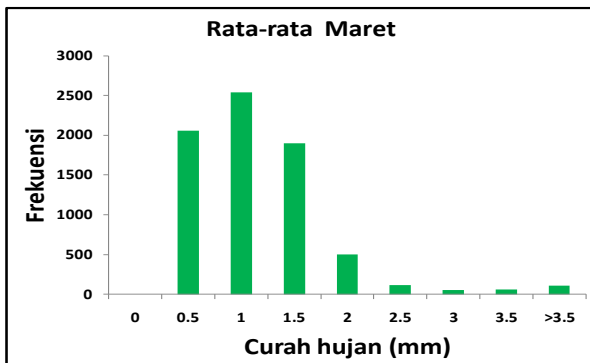
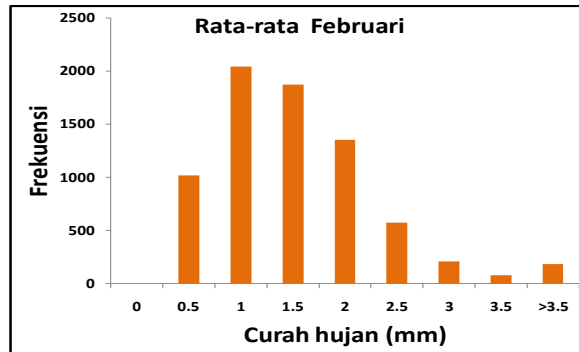
Catatan: data TRMM dalam satuan waktu sinoptik, jadi untuk konversi ke waktu setempat harus ditambah 7 untuk waktu Indonesia barat/WIB, ditambah 8 untuk waktu Indonesia tengah/WITA dan ditambah 9 untuk waktu Indonesia timur/WIT

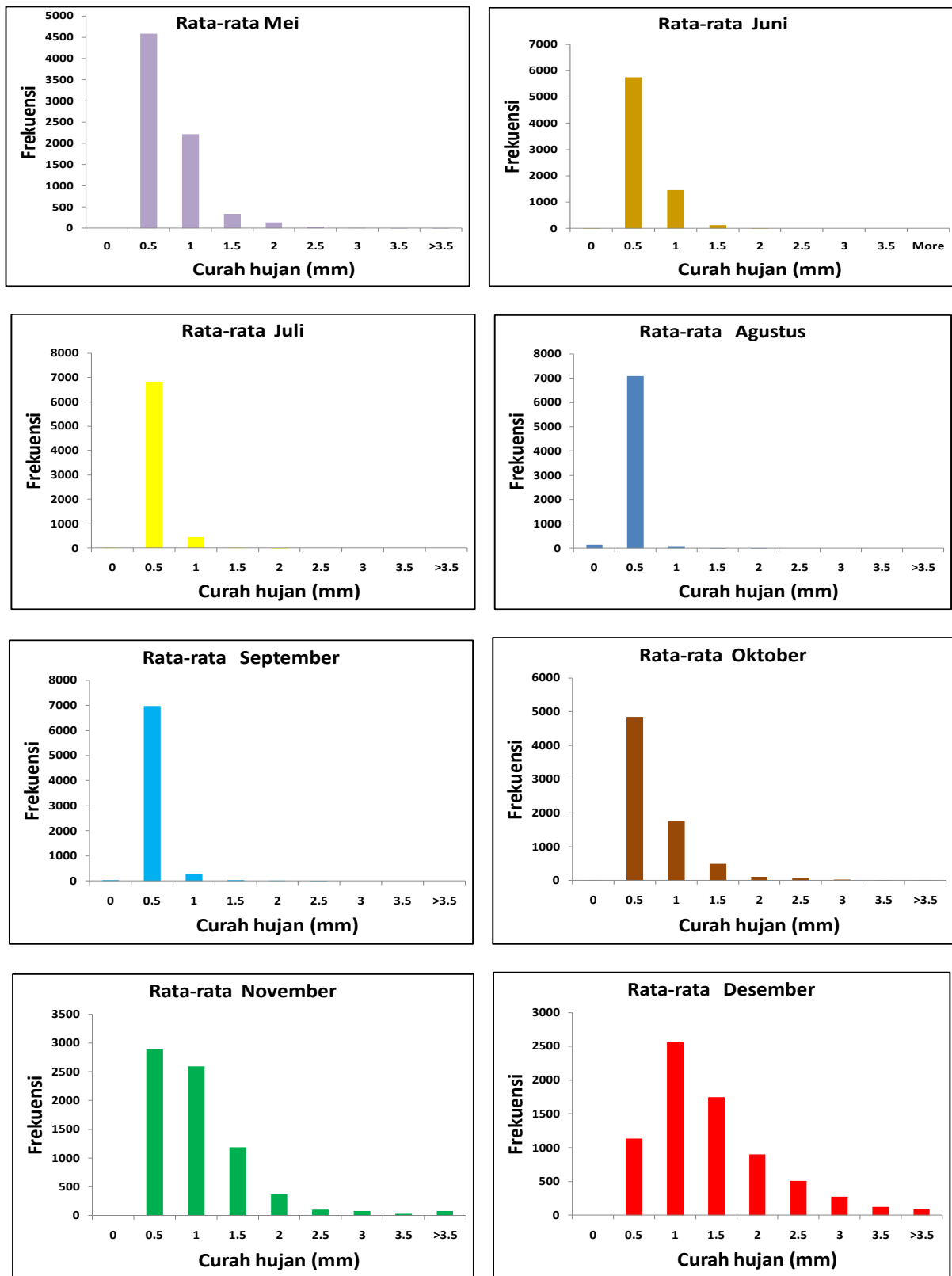
Nilai *outlier* sangat tergantung terhadap waktu dan lokasi. Nilai yang normal pada suatu saat atau suatu lokasi bisa menjadi nilai *outlier* untuk waktu dan lokasi lain. Sebagai contoh diperlihatkan contoh nilai *outlier* untuk bulan Januari pada Tabel 4-2. Pada tabel

tersebut nampak bahwa curah hujan sebesar 2,782 menjadi *outlier* di lokasi 6,875 LS; 109,88 BT akan tetapi menjadi nilai yang normal (misalnya) di 7,125 LS; 109,13 BT karena di lokasi ini nilai *outliernya* adalah 4,716.

Tabel 4-2: NILAI OUTLIER UNTUK DATA RATA-RATA BULAN JANUARI

Nilai outlier	Lokasi outlier		Nilai outlier	Lokasi outlier	
	Lintang	Bujur		Lintang	Bujur
2,922	105,63	- 5,375	5,163; 2,845	108,63	-6,625
2,832	106,88	-5,875	5,554; 5,484	108,88	-7,125
5,318	107,88	-7,125	5,707	108,88	-6,875
5,727	108,13	-7,125	4,716	109,13	-7,125
5,672; 4,816	108,13	-6,875	2,782	109,88	-6,875
4,844	108,13	-6,625	3,142	110,13	-6,875
6,016; 4,965	108,38	-7,125	5,072	111,88	-7,625
6,018	108,38	-6,875			
5,661	108,38	-6,625	5,378	111,88	-7,375
5,465	108,63	-7,375	4,984	111,88	-7,375
6,305; 5,169	108,63	-7,125	5,538	112,13	-7,875
6,42	108,63	-6,875	6,094	112,13	-7,625
2,911	110,38	-6,875	6,379	112,13	-7,375
3,009	110,63	-6,875	5,702	112,38	-7,875
2,969	110,63	-6,625	6,444	112,38	-7,625
5,322	111,13	-7,625	6,222	112,38	-7,375
4,766	111,13	-7,375	4,650	112,38	-7,375
5,204	111,38	-7,625	5,676	112,63	-7,875
5,065	111,38	-7,375	5,954	112,63	-7,625
5,024	111,63	-7,625	5,573	112,63	-7,375
4,995	111,63	-7,375	4,806	112,63	-7,125
5,405	111,88	-7,875	2,969	113,63	-5,375
5,668	111,88	-7,625	3,202	114,13	-8,125





Gambar 4-2: Histogram curah hujan setiap 3 jam TRMM rata-rata bulan Februari sampai dengan Desember

4.2 Pembahasan

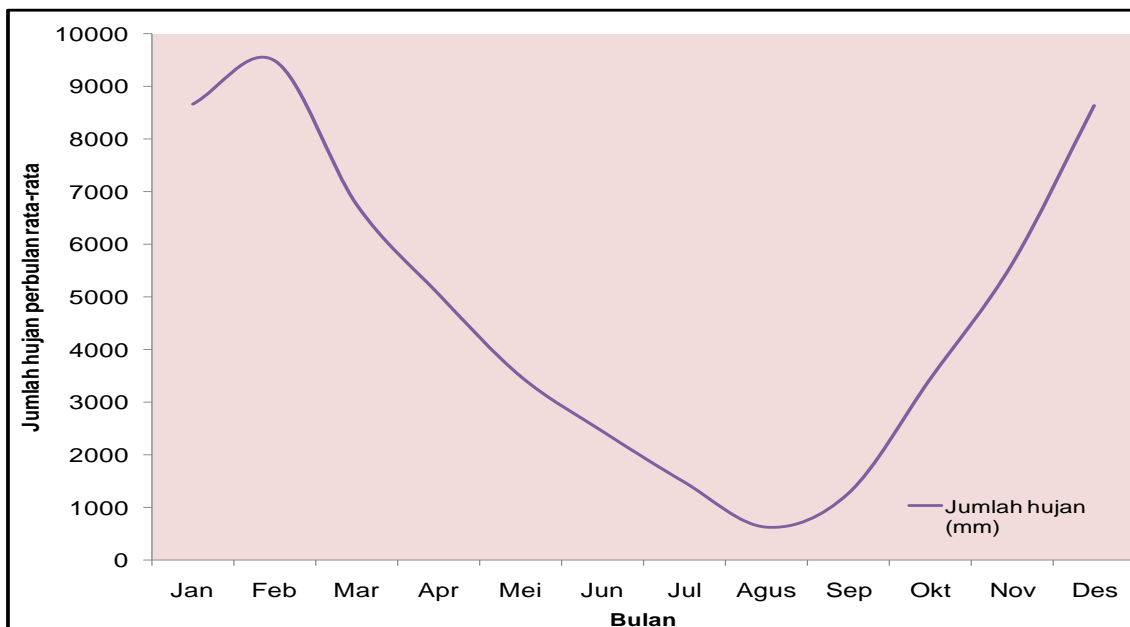
Berdasarkan hasil yang diperlihatkan pada Tabel 4-1, *outlier* banyak muncul pada kejadian hujan siang sampai malam hari (catatan: data TRMM dalam satuan

waktu sinoptik, jadi untuk konversi ke waktu setempat harus ditambah 7 untuk waktu Indonesia barat/WIB, ditambah 8 untuk waktu Indonesia tengah/WITA dan ditambah 9 untuk waktu Indonesia

timur/WIT). Hal ini bersesuaian dengan intensitas maksimum curah hujan yang juga terjadi pada saat tersebut (Juaeni, 2011). Dengan demikian, *outlier* berkaitan dengan frekuensi curah hujan. Semakin sering terjadi hujan, maka semakin berpotensi terjadinya *outlier* atau dengan kata lain probabilitas ditemukan *outlier* semakin tinggi jika frekuensi hujan semakin tinggi. Penjelasan lain adalah bahwa *outlier* berkaitan dengan proses pembentukan hujannya, karena waktu kejadian hujan secara tidak langsung menunjukkan proses pembentukannya. Tabel 4-1 juga menunjukkan bahwa *outlier* semakin sering terjadi pada bulan-bulan peralihan (Juli, Agustus dan September, Gambar 4-3) (Juaeni *et al.*, 2011). Pada bulan-bulan peralihan tersebut distribusi spasial curah hujan sangat bersifat lokal. Hal ini disebabkan hujan sudah mulai turun di wilayah Pulau Jawa. Sementara di wilayah lain (Bali dan Lombok) belum turun hujan. Variasi spasial inilah yang kemudian mendorong terdeteksinya *outlier* yang lebih banyak. Pada musim hujan (November, Desember Januari sampai Februari,

Gambar 4-3) (Juaeni *et al.*, 2011), jumlah *outlier* mengikuti pola kejadiannya yaitu maksimum pada sore sampai malam hari. Pada waktu tersebut distribusi curah hujan hampir merata di seluruh wilayah penelitian, sehingga jumlah *outlier* hanya berkaitan dengan proses pembentukan hujan.

Nilai *outlier* seperti yang ditunjukkan pada Tabel 4-2, bersifat sangat khas. Ini menunjukkan ketergantungannya terhadap waktu dan tempat, untuk selanjutnya menunjukkan ketergantungan terhadap proses pembentukan hujan. Dengan demikian, analisis data tahap berikutnya harus memperhatikan faktor-faktor lokal. Misalnya, jika pendeteksian *outlier* akan digunakan untuk mendefinisikan curah hujan ekstrim, maka harus dilakukan untuk skala lokal dengan memperhatikan faktor-faktor yang berpengaruh terhadap proses pembentukan hujan secara mikro. Karena curah hujan ekstrim juga berkaitan dengan dampak, maka selain kaidah matematis, faktor daya dukung permukaan harus dilibatkan untuk mendefinisikan nilai curah hujan ekstrim.



Gambar 4-3: Jumlah curah hujan rata-rata perbulan berbasis TRMM (Juaeni *et al.*, 2011)

5 KESIMPULAN

Outlier dengan batasan simpangan yang lebih besar dari 3 kali deviasi standar terjadi dengan frekuensi rata-rata 1,16% dari seluruh kejadian hujan (Tabel 4-1) pada wilayah dengan cakupan lintang sekitar 2° ($-5,375^{\circ}$ LS sampai $-8,125^{\circ}$ LS) dan cakupan bujur sekitar $8,5^{\circ}$ ($105,63^{\circ}$ BT - $114,13^{\circ}$ BT). Penentuan outlier dengan data curah hujan setiap 3 jam dari TRMM memberikan gambaran bahwa frekuensi kejadian outlier berkaitan dengan waktu kejadian hujan dan variasi spasial curah hujan. *Outlier* semakin tinggi frekuensinya ketika musim hujan belum merata diseluruh wilayah penelitian. Nilai outlier itu sendiri sangat khas, dan secara tidak langsung menunjukkan proses pembentukan hujannya.

Ucapan Terima Kasih

Tulisan ini merupakan bagian dari laporan penelitian IPKPP-2011. Penulis mengucapkan terima kasih kepada RISTEK dan LAPAN atas dukungan dana penelitian, juga kepada seluruh anggota penelitian atas kerjasamanya.

DAFTAR RUJUKAN

- Barnett, V. and Lewis, T., 1994. *Outliers in Statistical Data*. John Wiley & Sons., 3rd edition.
- Grubbs, F. E., 1969. *Procedures for Detecting Outlying Observations in Samples*. *Technometrics* 11, 1–21.
- Hans-Peter Kriegel, Peer Kröger, Arthur Zimek, 2009. *Outlier Detection Techniques (Tutorial)*. 13th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2009) (Bangkok, Thailand). Retrieved 2010-06-05.
- http://disc2.nascom.nasa.gov/Giovanni/tovas/TRMM_V6.3B42.2.shtml.
- http://www.mathworks.com/help/techdoc/data_analysis/.
- Juaeni, I.; Bambang Siswanto; Teguh Harjana, Nurzaman; Martono; Farid Lasmono, 2011. *Pemetaan Wilayah Konveksi Untuk Mendukung Pariwisata*, Laporan Penelitian IPKPP-RISTEK 2011.
- Juaeni, I.; Dewi Yuliani, Risana Ayahbi, Noersomadi, Teguh Harjana, Nurzaman, 2010. *Pengelompokan Wilayah Curah Hujan Kalimantan Barat Berbasis Metoda Ward dan Fuzzy Clustering*, *Jurnal Sains Dirgantara*, 7, 2, LAPAN.
- Moore, D. S. and McCabe, G. P., 1999. *Introduction to the Practice of Statistics*, 3rd ed. New York: W. H. Freeman, 1999.
- Renze, John, 2011. *Outlier*, From *Math World*--A Wolfram Web Resource, created by Eric W, <http://mathworld.wolfram.com/Outlier.html>.
- Rousseeuw, P. and Leroy, A., 1996. *Robust Regression and Outlier Detection*. John Wiley & Sons., 3rd edition.
- Varun Chandola, Arindam Banerjee, and Vipin Kumar, 2009. *Anomaly Detection: A Survey*, *ACM Computing Surveys*, Vol. 41(3), Article 15, July 2009.
- Weisstein, Eric, 2011. *Interquartile Range*, From *MathWorld* - A Wolfram Web Resource.