I PENDAHULUAN

1.1 Latar Belakang

Letak astronomis berpengaruh terhadap iklim di Indonesia, berada di garis lintang 6° LU – 11° LS dan garis bujur 95° BT – 141° BT menjadikan Indonesia sebagai negara beriklim tropis. Iklim tropis dicirikan dengan hutan hujan tropis yang luas, kelembapan udara yang tinggi dan curah hujan yang cukup tinggi. Perubahan kondisi cuaca yang sekarang ini dirasakan dapat mengakibatkan peningkatan intensitas curah hujan. Hal ini dapat memberikan manfaat tetapi juga berpotensi menimbulkan bencana seperti banjir dan longsor.

Indonesia merupakan negara agraris artinya pertanian memegang peranan penting dari perekonomian nasional. Curah hujan tentu akan sangat berdampak pada bidang pertanian yang produktivitasnya dipengaruhi cuaca dan ketersediaan air. Intensitas curah hujan yang sangat tinggi berpeluang untuk terjadinya gagal panen dan rendahnya intensitas hujan juga akan berpotensi terjadinya kekeringan. Oleh karena itu, informasi mengenai dugaan curah hujan sangatlah diperlukan untuk mengantisipasi atau mengurangi dampak bencana dan menunjang keberhasilan produksi pertanian. Namun kondisi geografis, daratan, laut, atmosfer yang kompleks mempersulit pendugaan curah hujan di walayah Indonesia.

Metode *Statistical downscaling* (SD) merupakan salah satu metode yang dapat menangani permasalahan pendugaan curah hujan harian. Metode ini membuat Informasi data berskala global GCM diproyeksikan terhadap informasi skala lokal di stasiun klimatologi (Zorita dan Von Storch 1999). GCM adalah model berbasis komputer yang menggambarkan sejumlah subsistem dari iklim di bumi, proses di atmosfer, lautan dan daratan, dan menstimulasi kondisi iklim skala besar (Wigena 2006). Data luaran GCM adalah data berskala global dan menjadi peubah prediktor. Data curah hujan pada stasiun cuaca yang dikeluarkan oleh BMKG (Badan Meteorologi dan Geofisika) adalah data berskala lokal dan menjadi peubah respon. Peubah prediktor dalam SD harus lebih dari satu agar bisa melingkupi peubah lokal (Benestad *et al*. 2008).

Data luaran GCM mengandung multikolinieritas yang tinggi (Wigena 2006). Multikolinieritas akan menyebabkan model tidak stabil, dan salah satu solusi untuk mengatasinya adalah dengan mereduksi peubah prediktor menggunakan Analisis Komponen Utama (AKU). AKU mentransformasi perubah awal ke peubah yang saling ortogonal atau melakukan proyeksi peubah-peubah prediktor awal berdimensi besar menjadi sejumlah peubah baru berdimensi kecil yang disebut komponen utama (Wigena 2006). Data curah hujan pada stasiun hujan lokal memiliki ragam yang besar sehingga regresi gerombol menjadi salah satu solusi untuk meminimalkan nilai ragam. Regresi gerombol adalah prosedur statistik multivariat yang mengelompokkan objek dengan tujuan meminimalkan tingkat kesalahan dan ragam pada model regresi dalam gerombol (Brusco *et al*. 2008).

Curah hujan adalah kejadian alami yang hanya memiliki nilai positif sehingga dianggap sebagai peubah acak dengan rentang nilai lebih dari 0 dan salah satu sebaran yang dapat digunakan adalah sebaran Gamma dengan 2-parameter (Soleh 2015). Bentuk sebaran Gamma dengan ekor kanan dan non-simetrik sesuai dengan karakteristik curah hujan. Pengembangan metode SD telah banyak dilakukan untuk pendugaan curah hujan. Penelitian SD dengan sebaran Gamma menggunakan metode LASSO dan Regresi Komponen Utama (RKU) dan metode LASSO memberikan memberikan nilai RMSEP terkecil dibanding dengan regresi komponen utama (RKU) (Soleh *et al*. 2015). Penelitian lain dengan menggunakan model SD sebaran Gamma dilakukan oleh Permatasari *et al*. (2017) dan pendekatan data curah hujan menyebar Gamma cukup baik untuk menduga curah hujan lokal. Peubah *dummy* ditambahkan dan mampu memperbaiki model pendugaan curah hujan dengan menghasilkan nilai statistik RMSEP terkecil.

Nadya (2018) menggunakan regresi linear gerombol dan pemodelan dua tahap, metode regresi gerombol menghasilkan nilai galat pendugaan yang lebih kecil dibandingkan dengan penambahan peubah *dummy* menggunakan metode *Classification and Regression Trees* (CART), Regresi Komponen Utama (RKU) dan RKU dengan peubah *dummy*. Butar-butar *et al*. (2019) melakukan Pendugaan curah hujan menggunakan regresi gerombol dan menyimpulkan regresi gerombol lebih baik dibandingkan RKU dan RKTP. Syafruddin *et al*. (2020) melakukan penelitian SD dengan analisis gerombol dengan sebaran Gamma dan Normal, namun metode tersebut belum mampu membuat pendugaan yang diukur dengan nilai RMSEP. Oleh karena itu, pada penelitian ini akan dilakukan pengembangan algoritma pendugaan curah hujan harian menggunakan regresi gerombol sebaran Gamma dengan analisis komponen utama untuk mengatasi multikolinearitas.

1.2 Tujuan Penelitian

Tujuan penulisan penelitian ini adalah:

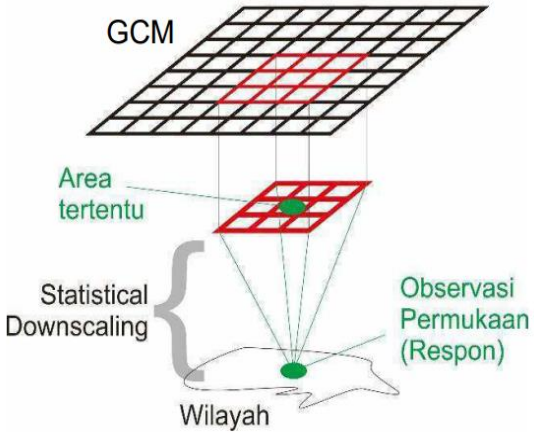
1. Mengembangkan simulasi regresi gerombol dengan sebaran campuran terdiri dari sebaran normal dan sebaran Gamma yang terbagi kedalam simulasi dua kelompok dan tiga kelompok.
2. Menerapkan regresi gerombol pada data curah hujan dan melakukan pendugaan curah hujan di stasiun Bandung, Bogor, Citeko dan Jatiwangi.

II TINJAUAN PUSTAKA

2.1 *Statistical Downscaling*

Pendekatan *Statistical downscaling* (SD) disusun berdasarkan adanya hubungan antara peubah prediktor (grid skala besar) dengan peubah respon (grid skala lokal) yang dinyatakan dengan model statistik (Zorita dan Von Storch, 1999). SD diperlukan untuk menjembatani antara skala besar GCM dengan skala di wilayah studi iklim dilaksanakan. GCM merupakan sumber informasi primer dan dapat digunakan untuk memprediksi iklim dan cuaca secara numerik. Data global GCM tidak dapat langsung digunakan sebagai peubah prediktor, karena data GCM memiliki kemungkinan untuk terjadi korelasi spasial antar data pada grid yang beda dalam satu domain dan multikolinearitas.

Domain menentukan keakuratan pendugaan model sehingga penentuan domain merupakan langkah penting dalam melakukan SD. Peubah prediktor adalah domain yang berdimensi banyak dan besar. Hal tersebut memungkinan terjadi korelasi spasial antar grid dalam domain dan multikolinearitas antar peubah prediktor. Semakin besar domain dan semakin banyak peubah yang digunakan akan mengakibatkan model semakin kompleks. Pembuatan model SD yang baik harus memmperhatikan tiga hal, yaitu keeratan hubungan antara peubah respon dan peubah prediktor, peubah prediktor disimulasikan dengan baik oleh peubah GCM, dan hubungan antara peubah respon dan prediktor tidak berubah dengan adanya perubahan waktu dan tetap sama meskipun terdapat perubahan iklim (Busuioc *et al*. 2001). Ilustrasi *statistical downscaling* dijelaskan oleh Gambar 1.



Gambar 1 Ilustrasi *statistical downscaling* (Wicaksono 2019)

Bentuk model SD secara umum adalah:

$$y_{(t\times l)} = f\left(X_{(t\times p)}\right) \quad (1)$$

Notasi $y_{(t\times l)}$ adalah vektor peubah iklim lokal (curah hujan) atau peubah respon dan $X_{(t\times p)}$ adalah matriks peubah luaran GCM (presipitasi) atau peubah

prediktor. *t* adalah amatan dalam satuan waktu dan *p* adalah banyaknya grid. Semakin banyak peubah *y* dan peubah *x* yang terlibat dalam *statistical downscaling*, maka model akan semakin kompleks.

2.2 Regresi Gerombol

Regresi gerombol (*clusterwise regression*) pertama kali dikenalkan dengan menggunakan algoritma pertukaran (Spath 1979). Regresi gerombol adalah kombinasi yang berasal dari dua teknik yaitu penggerombolan dan analisis regresi. Regresi gerombol dapat menduga jumlah gerombol dan parameter regresi di setiap gerombol secara bersama-sama sehingga efisien digunakan pada data set yang membutuhkan dua atau lebih fungsi regresi untuk mengetahui struktur data (Bagirov *et al.* 2017). Regresi gerombol sangat efisien dalam dataset deret waktu di mana penggunaan dua atau lebih fungsi regresi diperlukan untuk meringkas struktur data yang paling baik. Model regresi gerombol secara umum (Grun dan Leisch 2007) adalah:

$$y_i = \sum_{j=1}^{j} a_{ij} E(y|x) \qquad (2)$$

y_i = amatan ke-*i* dari peubah respon
a_{ij} = bernilai 1 jika amatan ke-i masuk gerombol ke-*j* atau bernilai 0 jika amatan ke-*i* tidak masuk ke gerombol ke-*j*
$E(y|x)$ = model regresi dari amatan ke-*i*

Langkah penerapan regresi gerombol dengan algoritma pertukaran (*exchange algorithm*) dimulai dengan menginisialisasi gerombol kemudian melakukan pendugaan model regresi pada setiap gerombol. Pada setiap gerombol parameter regresi dilakukan dengan menggunakan model linear terampat atau *generalized linear model* (GLM). Setiap amatan bisa berpindah pada gerombol lain ataupun tetap dalam gerombol awal sesuai dengan dimana nilai error terkecilnya berada. Hal ini dilakukan hingga amatan terakhir dan terbentuklah gerombol baru

2.3 Model Linear Terampat

Model Linier Terampat atau *Generalized Linear Model* (GLM) merupakan rampatan dari model model linier, dalam hal ini peubah respons berasal dari keluarga sebaran eskponensial dan adanya fungsi hubung yang menghubungkan antara nilai harapan dengan komponen sistematik dari model linier (Soleh 2015). Pendugaan parameter dalam regresi gerombol menggunakan Model linier terampat (MLT). Komponen-komponen dari Model Linier Terampat (MLT) adalah sebagai berikut (McCullagh dan Nelder 1989):

- Komponen acak
  Merupakan komponen yang menentukan sebaran bersyarat dari peubah respon. Peubah respons (Y) berasal dari keluarga sebaran eksponensial dengan $E(Y) = \mu$.
- Komponen sistematik
  Merupakan peubah prediktor x yang dikombinasikan dalam model sebagai fungsi linier dari parameter-parameter. Komponen ini memiliki penduga linier: $\eta = \beta_0 + \sum_i^p x_j\beta_j$.
- Fungsi hubung (*link*)
  Yaitu fungsi yang menghubungkan kombinasi linier dari peubah prediktor, $\eta = \sum_i^p x_j\beta_j$ dengan taksiran μ yang disesuaikan dengan distribusi dari peubah respon y dengan lambang $g(\mu)$. yaitu $g(\mu) = \eta$.

Bentuk umum dari fungsi kepekatan peubah acak dari keluarga sebaran eksponensial dengan menambahkan parameter skala (konstan) yaitu φ (Lindsey 1997) sebagai berikut:

$$f(y_i; \theta_i, \varphi) = \exp\left[\frac{y_i\theta_i - b(\theta_i)}{a_i(\varphi)} + c(y_i, \varphi)\right] \quad (3)$$

Dalam hal ini $a(\cdot)$, $b(\cdot)$ dan $c(\cdot)$ adalah fungsi tertentu, θ adalah parameter kanonik yang merupakan parameter yang bentuknya menyesuaikan dengan fungsi penghubung dari peubah respon y, dan φ adalah parameter disperse yang bersifat konstan. Fungsi kepekatan sebaran Gamma 2-parameter (v, ξ) adalah sebagai berikut:

$$f(y; v, \xi) = \frac{v^{\xi}}{\Gamma(\xi)} y^{\xi-1}\exp(-vy) \quad (4)$$

Dalam hal ini v adalah parameter laju (*rate*) dan ξ adalah parameter bentuk (*shape*). Pemodelan linier terampat perlu dilakukan parameterisasi ulang parameter v dengan $v = \xi / \mu$ dengan hubungan μ dalam sebaran Gamma adalah $\mu = \xi / v$ dan parameter ξ (*shape*) bernilai konstan (Soleh 2015). Parameter β_j dalam komponen sistematik digunakan untuk menduga nilai parameter μ sesuai dengan fungsi hubungnya. Fungsi hubung kanonik untuk sebaran Gamma adalah *reciprocal* yaitu:

$$\eta_i = g(\mu_i) = \frac{1}{\mu_i} \quad (5)$$

Karena fungsi hubung menggunakan:

$$\log \rightarrow g(\mu_i) = log(\mu_i)$$

maka:

$$\mu = \frac{1}{\log \sum_{j=1}^p x_{ij}\beta_j}$$

$$\mu = \exp \sum_{j=1}^{p} x_{ij}\beta_j \quad (6)$$

Pendugaan parameter β_j menggunakan metode kemungkinan maksimum yang diperoleh dengan prosedur numerik *Iterated Re-Weighted Least Squares* (IRLS). Fungsi kepekatan sebaran normal adalah sebagai berikut:

$$f(y;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \quad (7)$$

Dimana σ^2 adalah ragam dan μ adalah nilai harapan dari peubah acak *Y*. Fungsi hubung kanonik untuk sebaran normal adalah *identity*, yaitu:

$$n_i = g(\mu_i) = \mu_i$$

maka:

$$\mu = \sum_{j=1}^{p} x_{ij}\beta_j \quad (8)$$

III METODE

3.1 Data

Peubah prediktor adalah luaran GCM (*Generalized Circulation Model*) yang digunakan adalah data yang dikeluarkan oleh *National Centers for Environmental Prediction* (NCEP) yang dinamakan data *Climate Forecast System Reanalysis* (CFSRv2) yang menyediakan parameter iklim yang diukur setiap 6 jam dalam sehari. CFSR merupakan model yang menggambarkan interaksi global antara daratan, lautan dan udara yang ada di bumi (Saha *et al.* 2010). Parameter yang digunakan dari data CFS adalah *precipitation rate* atau curah hujan. Data yang digunakan adalah data harian. Data CFSR dapat diunduh dari situs http://rda.ucar.edu/ .

Gambar 2 Peta sebaran 4 stasiun hujan terpilih di Jawa Barat (peta tidak berskala)

Data CFSR yang digunakan menyesuaikan domain yang berada pada sekitar stasiun amatan. Domain pada data CFSR yang digunakan dalam penelitian ini yaitu berukuran 6×6 dengan jarak antar grid adalah 0,5º x 0,5º pada masing masing stasiun amatan hujan. Pengolahan data dilakukan dengan menggunakan *software* R. Data respon menggunakan data cutah hujan harian yang dipublikasikan oleh Badan Meteorologi dan Geofisika (BMKG), dapat diunduh dari situs http://dataonline.bmkg.go.id/data_iklim/ . Data yang akan digunakan adalah stasiun hujan yang ada di provinsi Jawa Barat yaitu stasiun hujan Bandung, Bogor, Citeko dan Jatiwangi. Letak stasiun terpilih dijelaskan dalam Gambar 2 dengan titik Lintang dan Bujur sebagai berikut:

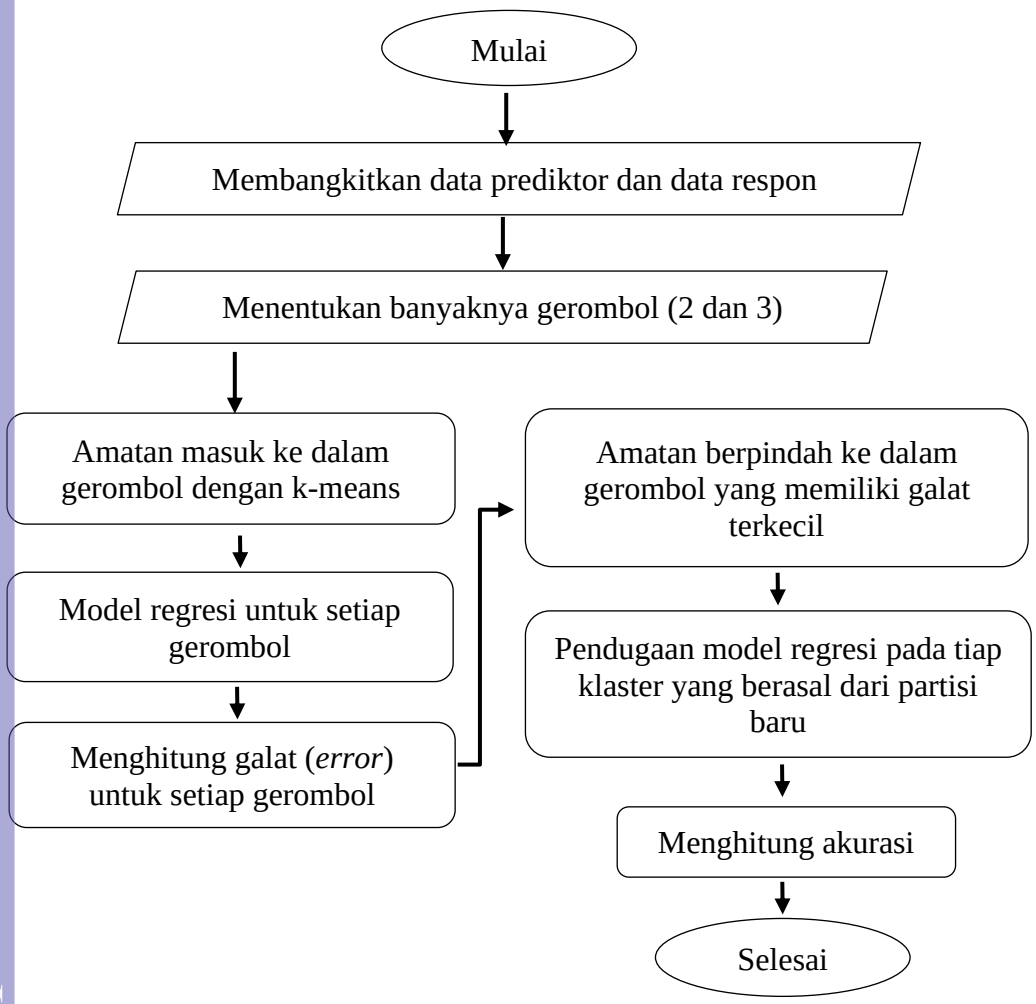
- Bogor (-6.550, 106,750) → lat = [-8, -5,5] lon = [105,5, 108]
- Citeko (-6,700, 106,850) → lat = [-8, -5,5] lon = [105,5, 108]
- Jatiwangi (-6,734, 108,263) → lat = [-8, -5,5] lon = [105, 107,5]
- Bandung (-6,883, 107,597) → lat = [-8, -5,5] lon = [106,5, 109]

3.2 Prosedur Analisis Data

Langkah-langkah analisis yang dilakukan pada penelitian ini dibagi ke dalam dua tahapan utama, yaitu tahap simulasi regresi gerombol dan penerapan regresi gerombol pada empat stasiun hujan terpilih yaitu stasiun Bandung, Bogor, Citeko dan Jatiwangi. Skenario simulasi terbagi menjadi simulasi dua kelompok dan tiga kelompok.

3.2.1 Simulasi Sebaran Gamma, Normal dan Campuran

Tahap simulasi ini bertujuan untuk melihat kemampuan regresi gerombol dalam mengelompokkan data sesuai dengan sebaran sebenarnya. Simulasi ini terbagi ke dalam simulasi dua kelompok dan simulasi tiga kelompok. Sebaran yang digunakan terdiri dari tiga sebaran yaitu Gamma, Normal dan campuran Gamma-Normal. Dimulai dengan pengembangan regresi gerombol, tahapan-tahapan regresi gerombol dijelaskan dalam Gambar 3 sebagai berikut:



Gambar 3 Diagram alur regresi gerombol

Tahapan–tahapan simulasi dengan regresi gerombol sebagai berikut:

a. Membangkitkan data prediktor yang mengikuti sebaran Seragam dengan rentang nilai 0 hingga 10.
b. Data respon dibangkitkan dari sebaran yang berasal dari sebaran normal, Gamma dan campuran Gamma-Normal dengan proporsi masing masing sebaran sama.
   - Data respon sebaran normal dibangkitkan dengan menggunakan persamaan $y = x\beta + \varepsilon$. Dalam hal ini, dibangkitkan galat mengikuti sebaran normal $\varepsilon \sim Normal(0, \sigma^2)$ dengan nilai $\sigma = 1$ dan 3. Tenentukan parameter β kemudian hitung peubah prediktor dengan masuk dalam persamaan normal.
   - Data respons sebaran Gamma dibangkitkan dengan cara menentukan parameter bentuk (ξ) dengan nilai ξ = 0,5, 0,85 dan 15. Hitung nilai $\mu = 1/X\beta$. Selanjutnya adalah menghitung parameter $v = \xi/\mu$, kemudian bangkitkan data respon Gamma $y \sim Gamma\ (\xi, v)$.
c. Data respon dibangkitan dengan proporsi yang sama bagi masing masing sebaran. Simulasi dua kelompok, terdiri dari 5 model simulasi dijelaskan pada Tabel 1.

Tabel 1 Simulasi Model Regresi dua kelompok

Simulasi respon	Peubah prediktor	Peubah Respon		
Gamma **(GG1)**	X ~ Seragam (0,10)	Y ~ Gamma (ξ , v) Y ~ Gamma (ξ , v)	$y_1 = (-2) + 0{,}25.x_1$ $y_2 = 0 + 0{,}4.x_2$	; ξ = 0,5 ; ξ = 15
Gamma **(GG2)**	X ~ Seragam (0,10)	Y ~ Gamma (ξ , v) Y ~ Gamma (ξ , v)	$y_1 = 0{,}25 + 0{,}25.x_1$ $y_2 = 0{,}55 + 0{,}45.x_2$	; ξ = 0,5 ; ξ = 15
Normal **(NN)**	X ~ Seragam (0,10)	Y ~ Normal (μ, σ^2) ε ~ Normal (0,1) Y ~ Normal (μ, σ^2) ε ~ Normal (0,1)	$y_1 = 5 + 4.x_1 + \varepsilon$ $y_2 = 10 + 10.x_2 + \varepsilon$	
Gamma Normal **(GN1)**	X ~ Seragam (0,10)	Y ~ Gamma (ξ , v) Y ~ Normal (μ, σ^2) ε ~ Normal (0,1)	$y_1 = 0 + 0{,}5.x_1$ $y_2 = 1 + 4.x_2 + \varepsilon$	; ξ = 15
Gamma Normal **(GN2)**	X ~ Seragam (0,10)	Y ~ Gamma (ξ , v) Y ~ Normal (μ, σ^2) ε ~ Normal (0,1)	$y_1 = 3 + 0{,}5.x_1$ $y_2 = 1000 + (-50).x_2 + \varepsilon$	; ξ = 0,85

- model Gamma-Gamma (GG) sebanyak 2 model: GG1 dan GG2. Sebaran GG1 terdiri dari 2 sebaran Gamma berbeda parameter yang dibuat hampir berdekatan antara sebaran satu dan sebaran dua. Sebaran GG2 dibuat terpisah antara sebaran satu dan sebaran dua.
- model Normal-Normal (NN) sebanyak 1 model. Sebaran NN hanya terdiri dari 1 skenario dengan 2 sebaran normal berbeda parameter yang dibuat terpisah.
- model Gamma-Normal (GN) sebanyak 2 model: GN1 dan GN2. sebaran GN1 terdiri dari 1 sebaran Gamma dan 1 sebaran Normal yang dibuat

berdekatan antara sebaran satu dan sebaran dua. Sebaran GN2 yang dibuat terpisah antara sebaran satu dan sebaran dua. Skenario simulasi tiga kelompok dijelaskan dalam Tabel 2.

Tabel 2 Simulasi model regresi tiga kelompok

Simulasi respon	Peubah prediktor	Peubah Respon		
Gamma **(GGG)**	X ~ Seragam (0,10)	Y ~ Gamma (ξ, v) Y ~ Gamma (ξ, v) Y ~ Gamma (ξ, v)	$y_1 = (-2) + 0{,}25.x_1$ $y_2 = 0{,}5 + 0{,}48.x_2$ $y_3 = 2 + 0{,}4.x_3$	$;\xi = 0{,}5$ $;\xi = 15$ $;\xi = 15$
Normal **(NNN)**	X ~ Seragam (0,10)	Y ~ Normal (μ, σ^2) ε ~ Normal (0.1) Y ~ Normal (μ, σ^2) ε ~ Normal (0,1) Y ~ Normal (μ, σ^2) ε ~ Normal (0,3)	$y_1 = 0 + 4.x_1 + \varepsilon$ $y_2 = 1 + 10.x_2 + \varepsilon$ $y_3 = 3 + 15.x_3 + \varepsilon$	
Gamma Normal Normal **(GNN1)**	X ~ Seragam (0,10)	Y ~ Gamma (ξ, v) Y ~ Normal (μ, σ^2) ε ~ Normal (0,1) Y ~ Normal (μ, σ^2) ε ~ Normal (0,1)	$y_1 = 0 + 0{,}6.x_1$ $y_2 = 130 + 25.x_2 + \varepsilon$ $y_3 = 80 + 6.x_3 + \varepsilon$	$;\xi = 1$
Gamma Normal Normal **(GNN2)**	X ~ Seragam (0,10)	Y ~ Gamma (ξ, v) Y ~ Normal (μ, σ^2) ε ~ Normal (0,1) Y ~ Normal (μ, σ^2) ε ~ Normal (0,1)	$y_1 = 0 + 0{,}6.x_1$ $y_2 = 85 + 5.x_2 + \varepsilon$ $y_3 = 120 + 40.x_3 + \varepsilon$	$;\xi = 15$
Gamma Gamma Normal **(GGN1)**	X ~ Seragam (0,10)	Y ~ Gamma (ξ, v) Y ~ Gamma (ξ, v) Y ~ Normal (μ, σ^2) ε ~ Normal (0,1)	$y_1 = (-2) + 0{,}25.x_1$ $y_2 = 0 + 0{,}48.x_2$ $y_3 = 50 + (-3.x_3) + \varepsilon$	$;\xi = 0.5$ $;\xi = 15$
Gamma Gamma Normal **(GGN2)**	X ~ Seragam (0,10)	Y ~ Gamma (ξ, v) Y ~ Gamma (ξ, v) Y ~ Normal (μ, σ^2) ε ~ Normal (0,1)	$y_1 = 0 + 0{,}25.x_1$ $y_2 = 0{,}2 + 0{,}5.x_2$ $y_3 = 40 + 4.x_3 + \varepsilon$	$;\xi = 0{,}5$ $;\xi = 15$

Simulasi tiga kelompok teridiri dari dari 6 model simulasi yaitu:

- model Gamma-Gamma-Gamma (GGG) sebanyak 1 model: GGG. Sebaran GGG terdiri dari 3 sebaran Gamma dengan parameter berbeda dengan bentuk sebaran terpisah antara satu dan lain.
- model Normal-Normal-Normal (NNN) sebanyak 1 model: NNN. Sebaran NNN terdiri dari 3 sebaran Normal dengan parameter berbeda dengan bentuk sebaran terpisah antara satu dan lain
- model Gamma-Normal-Normal (GNN) sebanyak 2 model: GNN1 dan GNN2. Terdiri dari 1 sebaran Gamma dan 2 sebaran normal dengan parameter berbeda, sebaran GNN1 dan GNN2 dibuat terpisah satu sama lain.
- model Gamma-Gamma-Normal (GGN) sebanyak 2 model GGN1 dan GGN2. Terdiri dari 2 sebaran Gamma dan 1 sebaran normal dengan parameter berbeda, sebaran GNN1 dan GNN2 dibuat terpisah satu sama lain.

d. Untuk melakukan evaluasi hasil simulasi digunakan akurasi atau ketepatan klasifikasi yang menggambarkan keakuratan model regresi gerombol dalam mengklasifikasi amatan sesuai dengan sebaran sebenarnya. Ketepatan klasifikasi tersaji dalam Tabel 3.

Tabel 3 Perhitungan akurasi atau ketepatan klasifikasi

Kelas sebenarnya	Kelas prediksi	
	1	0
1	*True Positive* (TP)	*False Negative* (FN)
0	*False Positive* (FP)	*True Negative* (TN)

$$\text{ketepatan klasifikasi} = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

3.2.2 Penerapan Algoritma Regresi Gerombol pada Data Curah Hujan

Algoritma regresu gerombol diaplikasikan pada data curah hujan dengan tahapan sebagai berikut:

1. Data global CFSR diubah formatnya dari format kg/m^{2}s menjadi mm/hari.
2. Eksplorasi data dengan statistik deskriptif kemudian membuat diagram pencar untuk data curah hujan lokal untuk melihat karakteristik dan pola data.
3. Analisis komponen utama (AKU) digunakan untuk mereduksi perubah prediktor yaitu data CFSR.

   Analisis komponen utama membentuk peubah baru yang merupakan kombinasi linear dari seluruh peubah asal, yang disebut komponen utama. Meskipun dibutuhkan p komponen untuk menunjukkan keseluruhan variasi data, seringkali variasi ini dapat diwakili oleh k komponen utama, dengan $k \ll p$ (Jolliffe 2002). Komponen utama yang bebas dari multikolinearitas diperoleh, maka komponen utama tersebut menjadi peubah bebas baru yang dapat diregresikan atau dianalisa pengaruhnya terhadap peubah tak bebas (Y). Misalkan suatu data terdiri dari n pengamatan dan p peubah dapat dinyatakan dalam matriks $\boldsymbol{X}$ pada persamaan (10).

$$\boldsymbol{X}_{n\times p} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix} \tag{10}$$

Persamaan (10) dapat dinyatakan dalam bentuk vektor yaitu $\boldsymbol{X} = [\boldsymbol{x_1'}, \boldsymbol{x_2'}, \dots \boldsymbol{x_n'}]^T$ dan $\mathbf{x_i'} = \left[x_{i1}, x_{i2}, \dots, x_{ip}\right]$ $i = 1,2, \dots n$ dengan vektor peubah *acak* $\boldsymbol{X}$ yang berisi p peubah *acak* dapat ditulis, $\boldsymbol{X} = [X_1, X_2, \dots X_P]'$. Komponen utama dapat dibentuk menggunakan matriks kovarian $(\boldsymbol{\Sigma})$, tetapi jika peubah-peubah bebas yang diamati mempunyai perbedaan satuan pengukuran sangat besar digunakan matriks korelasi $(\boldsymbol{\rho})$.

Misalkan vektor *random* $\boldsymbol{X} = [X_1, X_2, \dots, X_p]^T$ dengan pasangan akar ciri dan vektor ciri yaitu $(\lambda_1, \boldsymbol{e_1}), (\lambda_2, \boldsymbol{e_2}), \dots, (\lambda_p, \boldsymbol{e_p})$ dengan $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$. Komponen utama ke-$i$ didefinisikan pada Persamaan (11) sebagai berikut.

$$\text{KU}_i = \mathbf{e_i'x} = e_{1i}X_1 + e_{2i}X_2 + \cdots + e_{pi}X_p \quad , i = 1,2, \dots, p \tag{11}$$

4. Menggabungkan data curah hujan lokal sebagai peubah respon dan data global CFSR sebagai peubah prediktor menjadi satu data *frame*.
5. Melakukan pembagian data menjadi dua yaitu data *training* dan data *testing*. Dengan komposisi data *training* 80% dan data *testing* 20%. Perulangan dilakukan sebanyak 50 kali.
6. Melakukan pemodelan curah hujan menggunakan regresi gerombol.

a. Tentukan banyaknya gerombol *j* yang akan terbentuk dan dalam penelitian ini ditentukan sebanyak dua dan tiga gerombol.
b. Tiap amatan digerombolkan ke dalam satu dari *j* buah gerombol dengan teknik pengegrombolan k-means.
c. Pemodelan awal regresi disetiap gerombol dengan data yang sudah terbentuk pada penggerombolan k-means.
d. Menghitung galat untuk setiap model regresi. Tiap amatan akan mempunyai nilai galat sebanyak model regresi.
e. Amatan akan berpindah ke dalam gerombol dengan galat terkecil. Jika galat terkecil berada pada gerombolnya maka amatan tidak berpindah, tetapi jika galat terkecil berada di gerombol lain, maka amatan berpindah. Langkah ini akan membentuk partisi yang baru dan dilakukan hingga amatan terakhir.
f. Pendugaan pada parameter regresi untuk setiap gerombol baru yang terbentuk dari partisi baru.

7. Selanjutnya adalah prediksi data *testing* yang dilakukan dengan 2 jenis metode, sebagai berikut:

a. Cara 1

Memprediksi peubah respon pada data testing dengan mencari nilai jarak antara peubah respon dengan nilai *centroid* tiap gerombol yang terbentuk dari proses penggerombolan sebelumnya. Nilai *centroid* adalah rata-rata pada setiap gerombol. Jarak yang digunakan adalah jarak *Euclidean* yang dapat digunakan apabila semua peubah yang digunakan adalah kontinu (Sumertajaya *et al*. 2007).

$$d_{(i,j)} = \sqrt{\sum_{j=1}^{p} \left(x_{ij} - \bar{x}_{ij}\right)^2} \tag{12}$$

x_{ij} = amatan pada data *testing* peubah ke-j pada gerombol ke-i
$\bar{x}_{ij}$ = rata-rata pada data *training* peubah ke-j gerombol ke-i.

Amatan data testing dihitung jaraknya pada tiap model dan amatan masuk ke gerombol dengan jarak terdekat dan dilakukan prediksi data.

b. Cara 2

Memprediksi peubah respon pada data testing dengan mencari nilai jarak antara peubah respon dengan nilai rataan jarak tiap amatan pada gerombol yang terbentuk dari proses penggerombolan sebelumnya. Amatan data testing dihitung jaraknya pada tiap model dan amatan data testing akan memiliki nilai jarak pada masing masing gerombol kemudian dan amatan masuk ke gerombol dengan jarak terdekat dan dilakukan prediksi data.

8. Kinerja prediksi pada model dievaluasi dengan membandingkan curah hujan yang diamati dan diprediksi yang dihitung dari set data uji. Model terbaik ditentukan melalui nilai penduga galat yang terkecil. *Root Means Square Error* (RMSE) atau *Root Means Square Error Prediction* (RMSEP) digunakan untuk tujuan tersebut. RMSEP mengukur perbedaan antara nilai prediksi dengan nilai aktual yang didefinisikan sebagai berikut:

$$\text{RMSE / RMSEP} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \qquad (13)$$

Simbol untuk y_i adalah amatan ke-i, dengan n data, sedangankan $\hat{y}_i$ adalah dugaan amatan ke-i.

IV HASIL DAN PEMBAHASAN

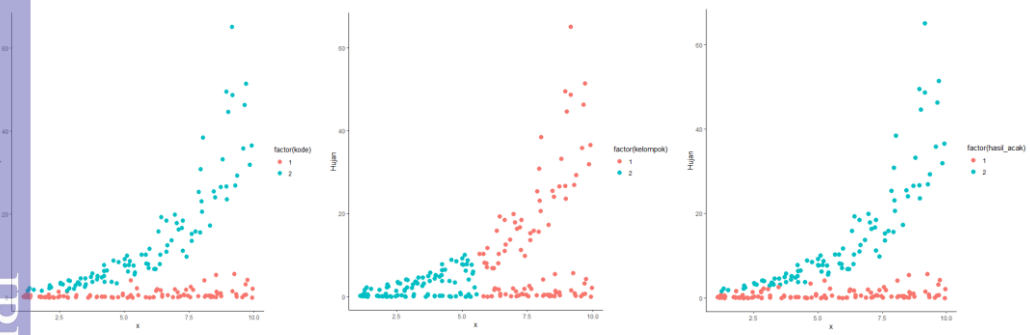
4.1 Simulasi Model 2 Gerombol Sebaran Gamma-Gamma

Tahap pertama dalam penelitian adalah simulasi regresi regresi gerombol dengan tiga model dengan sebaran Gamma, sebaran normal dan sebaran campuran Gamma-normal (GN). Hasil regresi gerombol dijelaskan menggunakan diagram pencar dengan indeks warna berbeda tiap sebaran. Tiga gambar diagram pencar digunakan untuk menjelaskan proses penggerombolan. Peubah respon mengikuti sebaran Gamma dua parameter dengan persamaan $y = 1 / (\beta_0 + \beta_1.x)$. Peubah prediktor mengikuti sebaran seragam dengan rentang 0 hingga 10. Regresi gerombol dua kelompok Gamma disajikan pada Gambar 4.

Simulasi sebaran Gamma dua kelompok, dicobakan dua skenario dengan parameter ξ (*shape*) dan β yang berbeda. Skenario pertama, dibangkitan dua kelompok data, sebaran pertama dengan koefisien β_0 = -2, β_1 = 0,25 dan ξ = 0,5 dan sebaran kedua dengan koefisien β_0 = 0, β_1 = 0,5 dan ξ = 15. Skenario kedua, dibangkitan dua kelompok data, sebaran pertama dengan koefisien β_0 = 0,25, β_1 = 0,25 dan ξ = 0,5 dan sebaran kedua dengan koefisien β_0 = 0,55, β_1 = 0,45 dan ξ = 15. Terlihat jelas perbedaan letak sebaran pertama dan sebaran kedua pada simulasi GG1. Berbeda dengan simulasi GG1, pada simulasi GG2 sebaran Gamma dibuat berdekatan. Hal ini bertujuan untuk melihat kemampuan regresi gerombol dalam memisahkan kelompok data yang berdekatan ataupun berjauhan. Gambar 4 dan Gambar 5 menjelaskan proses simulasi regresi gerombol Gamma dua kelompok. Gambar (a) merupakan diagram sebaran data bangkitan sesungguhnya, Gambar (b) merupakan diagram sebaran data hasil pengelompokkan data bangkitan dengan k-means dan Gambar (c) adalah diagram sebaran data hasil pengelompokkan menggunakan regresi gerombol.

a) Simulasi Gamma 2 kelompok (GG1)

1) $y_1 = (-2) + 0{,}25.x_1 \quad ; \xi = 0{,}5$
2) $y_2 = 0 + 0{,}5.x_2 \quad ; \xi = 15$



(a) (b) (c)

Gambar 4 Diagram sebaran data simulasi Gamma 2 kelompok

b) Simulasi Gamma 2 kelompok (GG2)

1) $y_1 = 0{,}25 + 0{,}25.x_1 \quad ; \xi = 0{,}5$
2) $y_2 = 0{,}55 + 0{,}45.x_2 \quad ; \xi = 15$

(a) (b) (c)

Gambar 5 Diagram sebaran data simulasi Gamma 2 kelompok (2)

Tahap awal data dibangkitkan dengan sebaran Gamma berbeda parameter, kemudian data dikelompokkan dengan k-means. Masing-masing kelompok akan dimodelkan dengan regresi Gamma dan setiap amatan akan mempunyai nilai galat terhadap setiap model Gamma. Amatan bisa tetap pada kelompoknya atau berpindah pada kelompok lain, hal ini bergantung pada model mana nilai galat terkecilnya berada. Hasil simulasi GG1 dan GG2 terbukti regresi gerombol dapat memisahkan data dengan baik pada kedua simulasi. Pada simulasi GG1, regresi gerombol terbukti dapat memisahkan data dengan baik walaupun berdekatan letak antara sebaran satu dan sebaran dua, menghasilkan rataan akurasi tertinggi pada model Gamma sebesar 96,41%, akurasi model normal sebesar 87,03% dan akurasi model GN sebesar 85,93%. Simulasi GG2 menghasilkan rataan akurasi tertinggi pada model Gamma sebesar 91,38%, akurasi model normal sebesar 77,30% dan akurasi model GN sebesar 48,06%.

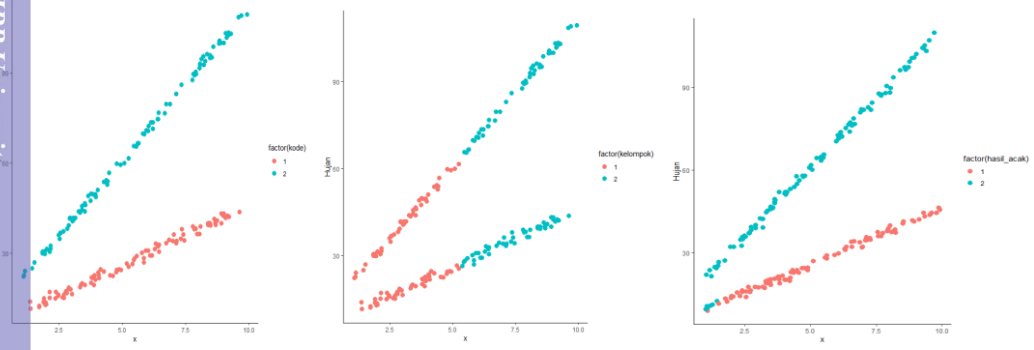
4.2 Simulasi Model 2 Gerombol Sebaran Normal-Normal

Simulasi regresi gerombol dengan persamaan $y = \beta_0 + \beta_1.x + \varepsilon$ dan galat menyebar normal dengan $\varepsilon \sim \mathrm{N}(\mu, \sigma^2)$. Terdapat satu skenario dengan galat peubah respon mengikuti sebaran normal yang dibangkitkan dengan parameter β dan galat yang berbeda. Peubah prediktor mengikuti sebaran seragam dengan rentang 0 hingga 10. Dibangkitan dua kelompok data, sebaran pertama dengan koefisien β_0 = 5, β_1 = 4 dan galat $\varepsilon \sim \mathrm{N}(0,1)$. Sebaran kedua dengan koefisien β_0 = 10, β_1 = 10 dan galat $\varepsilon \sim \mathrm{N}(0,1)$. Peubah respon dan peubah prediktor dibangkitkan, kemudian data akan dikelompokkan dengan k-means. Setiap kelompok akan dimodelkan dengan regresi normal dan setiap amatan akan mempunyai nilai galat terhadap setiap model normal. Hasil simulasi normal dua kelompok ini, regresi gerombol terbukti dapat memisahkan data dengan sangat baik pada kedua gerombol. Simulasi NN menghasilkan rataan akurasi tertinggi pada model normal sebesar 98,40%, akurasi model Gamma sebesar 56,00% dan akurasi model GN sebesar

69,39%. Regresi gerombol mampu mengelompokkan sebaran sesuai dengan sebaran sebenarnya, terlihat dari Gambar 6. Gambar (a) diagram sebaran data bangkitan sesungguhnya, Gambar (b) diagram sebaran data hasil pengelompokkan data bangkitan dengan k-means dan Gambar (c) diagram sebaran data hasil pengelompokkan menggunakan regresi gerombol.

Simulasi Normal 2 kelompok (NN)

1) $y_1 = 5 + 4.x_1 + \varepsilon$; $\varepsilon \sim N(0,1)$
2) $y_2 = 10 + 10.x_2 + \varepsilon$; $\varepsilon \sim N(0,1)$



(a) (b) (c)

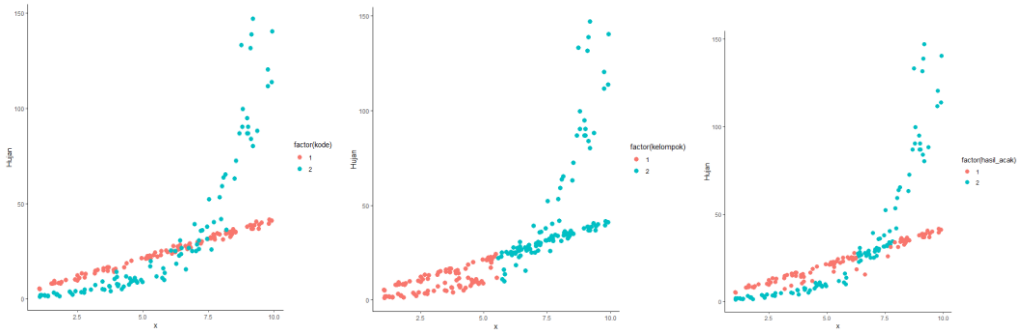
Gambar 6 Diagram sebaran data simulasi normal 2 kelompok

4.3 Simulasi Model 2 Gerombol Sebaran Gamma-Normal

Model terakhir dalam simulasi dua kelompok adalah model regresi campuran Gamma-normal (GN). Terbagi kedalam dua model yaitu GN1 dan GN2. Tahap awal adalah membangkitkan peubah respon yang berasal dari dua kelompok Gamma dan Normal. Peubah prediktor berdistribusi seragam dengan rentang 0 hingga 10. Peubah respon dibangkitkan dengan proporsi data yang sama yaitu 50%, kemudian kedua data digabungkan dan diacak untuk mejadi satu peubah respon. Simulasi sebaran GN, dicobakan dua skenario dengan parameter ξ (*shape*) dan β yang berbeda. Skenario pertama GN1 dibangkitan sebaran Gamma dengan koefisien $\beta_0 = 0$, β_1 = 0,5 dan ξ = 15 dan sebaran normal dengan koefisien $\beta_0 = 1$, β_1 = 4 dan galat $\varepsilon \sim N(0,1)$. Skenario kedua GN2 dibangkitan sebaran Gamma dengan koefisien $\beta_0 = 3$, β_1 = 0,5 dan ξ = 15, sebaran normal dengan koefisien $\beta_0 = 1000$, β_1 = -50 dan galat $\varepsilon \sim N(0,1)$. Data dibangkitan sebaran Gamma-Normal, kemudian data akan dikelompokkan dengan k-means. Setiap kelompok akan dimodelkan dengan regresi Gamma dan regresi normal dan setiap amatan akan mempunyai nilai galat terhadap setiap model regresi.

a) Simulasi Gamma-Normal 2 kelompok (GN1)

1) $y_1 = 0 + 0{,}5.x_1$; $\xi = 15$
2) $y_2 = 1 + 4.x_2 + \varepsilon$; $\varepsilon \sim N(0,1)$



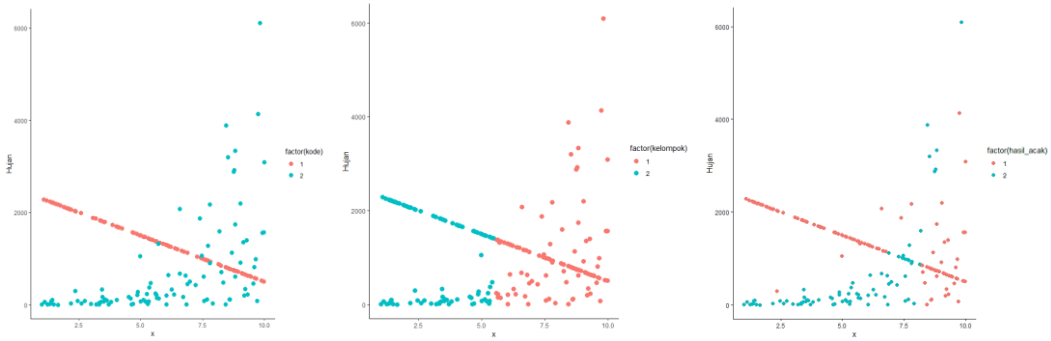
(a) (b) (c)

Gambar 7 Diagram sebaran data simulasi GN 2 kelompok

Amatan bisa tetap pada kelompoknya atau berpindah pada kelompok lain, hal ini bergantung pada model mana nilai galat terkecilnya berada. Hal ini dilakukan hingga amatan terakhir. Hasil simulasi GN1, regresi gerombol terbukti dapat memisahkan data dengan baik pada kedua gerombol. Gambar 7 terlihat bahwa sebaran Gamma dan normal berdekatan dan regresi gerombol mampu memisahkan dengan baik. Hasil rataan akurasi simulasi GN1 tertinggi pada model GN sebesar 84.90%, akurasi model Gamma sebesar 70,14% dan akurasi model normal sebesar 61,15%. Hasil simulasi GN2, regresi gerombol terbukti dapat memisahkan data dengan baik pada kedua gerombol Gambar 8 terlihat sebaran Gamma dan normal berjauhan dan regresi gerombol mampu memisahkan dengan baik. Hasil simulasi GN2 menghasilkan rataan akurasi tertinggi pada model GN sebesar 80,86%, akurasi model Gamma sebesar 57,20% dan akurasi model normal sebesar 73,55%.

b) Simulasi Gamma-Normal 2 kelompok (GN2)

1) $y_1 = 3 + 0{,}5.x_1$; $\xi = 0{,}85$
2) $y_2 = 2500 + (-200).x_2 + \varepsilon$; $\varepsilon \sim N(0,1)$



(a) (b) (c)

Gambar 8 Diagram sebaran data simulasi GN 2 kelompok (2)

Simulasi regresi gerombol dua kelompok memiliki tiga model simulasi yaitu model Gamma (GG), model normal (NN) dan model Gamma-normal (GN1 dan GN2). Hasil pada semua model menunjukkan bahwa regresi gerombol mampu memisakan data dengan baik sesuai dengan sebaran sebenarnya. Simulasi sebaran Gamma, akurasi tertinggi berada pada model gerombol Gamma. Simulasi sebaran

normal, akurasi tertinggi berada pada model gerombol normal. Simulasi sebaran GN, akurasi tertinggi berada pada model gerombol GN. Sehingga dapat disimpulkan algoritma yang dibentuk mampu mengelompokkan data sesuai dengan karakteristik data sebenarnya. Rangkuman hasil regresi gerombol dua kelompok disajikan dalam Tabel 4 berikut.

Tabel 4 Rangkuman akurasi model dua kelompok pada data simulasi

no	Skenario	Akurasi Gamma	Akurasi Normal	Akurasi GN
1	Model GG 1	**96,41 %**	87,03 %	85,93 %
2	Model GG 2	**91,38 %**	77,30 %	48,06 %
3	Model NN	56,00 %	**98,40 %**	69,39 %
4	Model GN 1	70,14 %	61,15 %	**84,90 %**
5	Model GN 2	57,20 %	73,55 %	**80,86 %**

Simulasi dua kelompok Gamma memiliki dua skenario dengan parameter berbeda, model GG1 dan model GG2. Skenario pertama dan skenario kedua menghasilkan nilai rataan akurasi yang tinggi sebesar 96,41% dan 91,38%. Regresi gerombol dapat memisahkan data sesuai dengan data aktual dengan sangat baik pada simulasi Gamma dua kelompok. Simulasi dua kelompok normal memiliki satu skenario yaitu model NN dan menghasilkan nilai rataan akurasi yang nyaris sempurna pada model normal sebesar 98,40%. Regresi gerombol memisahkan data dengan sangat baik pada simulasi normal dua kelompok. Simulasi dua kelompok campuran GN memiliki dua skenario, model GN1 dan model GN2. Skenario pertama menghasilkan nilai rataan akurasi tinggi pada model Gamma-Normal sebesar 84,90%. Skenario kedua menghasilkan nilai akurasi cukup tinggi pada model GN sebesar 80,86%. Tabel 5 menampilkan rataan akurasi tertinggi pada pada setiap model dan proporsi untuk gerombol terbaik.

Tabel 5 Akurasi dan proporsi gerombol terbaik pada model simulasi dua kelompok

No	Regresi gerombol	Akurasi	Gerombol	Proporsi
1	Model GG 1	96,41 %	1	52,50 %
			2	47,50 %
2	Model GG 2	91,38 %	1	53,00 %
			2	47,00 %
3	Model NN	98,40 %	1	48,00 %
			2	52,00 %
4	Model GN 1	84,90 %	1	45,50 %
			2	54,50 %
5	Model GN 2	80,86 %	1	63,00 %
			2	37,00 %

Data aktual mempunyai proporsi yang sama yaitu 50% pada setiap gerombol. Model GG1 dan GG2 memiliki rataan akurasi yang tinggi. Gerombol satu dan

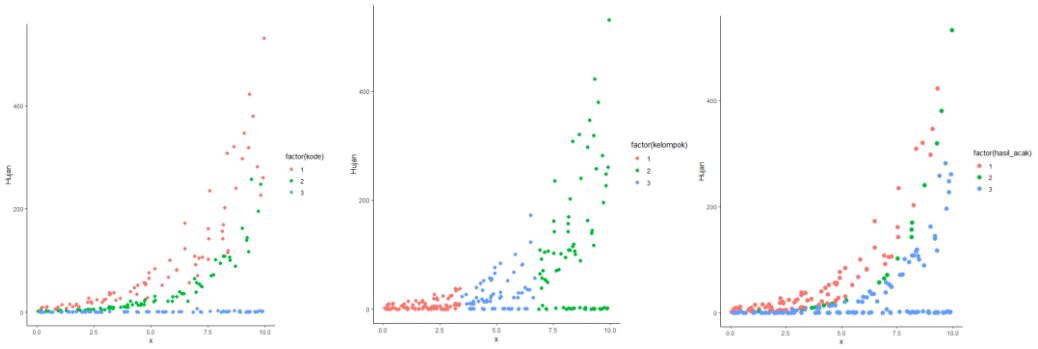
gerombol dua pada model GG1 dan GG2 memiliki proporsi yang hampir sama dan proprosi tertinggi ada pada gerombol satu. Model NN memiliki rataan akurasi yang sangat tinggi dengan memiliki proporsi yang hampir seimbang pada tiap gerombol. Model GN1 dan GN2 dengan rataan akurasi yang cukup tinggi, dapat dilihat bahwa proporsi GN1 memiliki proporsi lebih seimbang dibandingkan dengan GN2 dengan masing-masing proporsi tertinggi berada pada gerombol satu. Kemampuan regresi gerombol dalam memisahkan data dengan sebaran yang lebih kompleks dilakukan dengan simulasi data tiga kelompok yang berasal dari sebaran Gamma, normal dan campuran Gamma-normal.

4.4 Simulasi Model 3 Gerombol Sebaran Gamma-Gamma-Gamma

Tahap kedua dalam simulasi adalah simulasi regresi gerombol dengan tiga kelompok dengan model sebaran Gamma (GGG), model sebaran normal (NNN), model sebaran Gamma-normal-normal (GNN) sebanyak dua dan model sebaran Gamma-Gamma-normal (GGN) sebanyak dua. Diagram pencar digunakan untuk memudahkan melihat hasil regresi gerombol. Simulasi GGG memiliki peubah respon mengikuti sebaran Gamma dua parameter dengan persamaan $y = 1/(\beta_0 + \beta_1.x)$. Peubah prediktor mengikuti sebaran seragam dengan rentang nilai 0 hingga 10. Simulasi sebaran Gamma tiga kelompok memiliki satu skenario dengan parameter ξ (*shape*) dan β yang berbeda.

Simulasi Gamma 3 sebaran (GGG)

1) $y_1 = (-2) + 0{,}25.x_1 \quad ; \xi = 0{,}5$
2) $y_2 = 0{,}5 + 0{,}48.x_2 \quad ; \xi = 15$
3) $y_3 = 2 + 0{,}40.x_3 \quad ; \xi = 15$



(a) (b) (c)

Gambar 9 Diagram sebaran data simulasi Gamma 3 kelompok

Regresi gerombol tiga kelompok Gamma disajikan pada Gambar 9. Gambar (a) diagram sebaran data bangkitan sesungguhnya, Gambar (b) diagram sebaran data hasil pengelompokkan data bangkitan dengan k-means dan Gambar (c) diagram sebaran data hasil pengelompokkan menggunakan regresi gerombol. Model GGG dibangkitan tiga kelompok data, sebaran pertama dengan koefisien β_0 = -2, β_1 = 0,25 dan ξ =0,5. Sebaran kedua dengan koefisien β_0 = 0,5, β_1 = 0,48

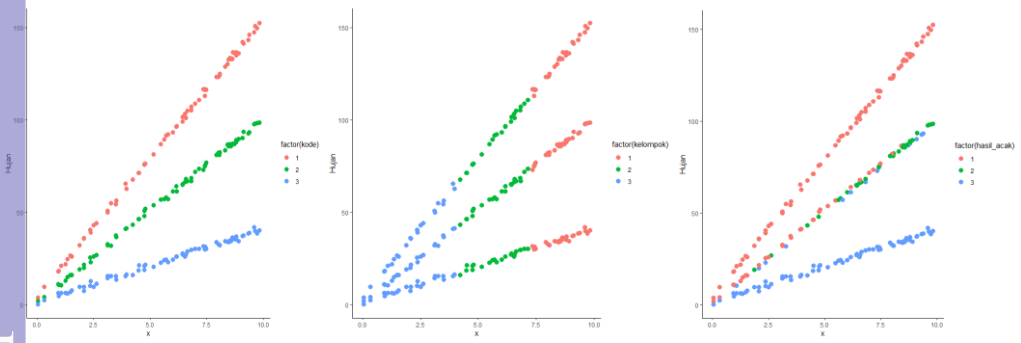
dan ξ = 15. Sebaran ketiga dengan koefisien β_0 =2, β_1 = 0,4 dan ξ = 15. Hasil simulasi Gamma tiga kelompok terbukti regresi gerombol dapat memisahkan data dengan cukup baik. Simulasi GGG menghasilkan rataan akurasi tertinggi pada model Gamma sebesar 68,05%, akurasi model normal sebesar 53,31%, akurasi model GNN sebesar 45,07% dan akurasi model GGN sebesar 64,52%. Berdasarkan diagram pencar terlihat bahwa regresi gerombol kurang mampu memisahkan beberapa amatan dalam gerombol dua dan gerombol tiga. Salah klasifikasi amatan gerombol dua menjadi anggota gerombol tiga, menyebabkan nilai akurasi menjadi kurang baik. Hal ini dikarenakan tiga kelompok Gamma yang sedikit berdekatan satu dengan yang lain.

4.5 Simulasi Model 3 Gerombol Sebaran Normal-Normal-Normal

Simulasi sebaran dengan persamaan $y = \beta_0 + \beta_1.x + \varepsilon$ dengan galat menyebar Normal $\varepsilon \sim N(\mu, \sigma^2)$ memiliki satu skenario. Tahap pertama pada peubah respon adalah membangkitkan galat beridistribusi normal, kemudian menghitung nilai peubah respon dengan memasukan galat pada persamaan dengan nilai β_0 dan β_1 yang telah ditentukan. Sebaran pertama dengan koefisien β_1 = 4 dan galat $\varepsilon \sim N(0,1)$. Sebaran kedua dengan koefisien β_0 = 1, β_1 = 10 dan galat $\varepsilon \sim N(0,1)$, sebaran ketiga dengan koefisien β_0 =3, β_1 = 15 dan galat $\varepsilon \sim N(0,3)$. Hasil simulasi normal tiga kelompok ini regresi gerombol dapat memisahkan data dengan sangat baik pada ketiga kelompok. Proses gerombol tiga kelompok normal dijelaskan dalam Gambar 10. Gambar (a) diagram sebaran data bangkitan sesungguhnya, Gambar (b) diagram sebaran data hasil pengelompokkan data bangkitan dengan k-means dan Gambar (c) diagram sebaran data hasil pengelompokkan menggunakan regresi gerombol.

Simulasi Normal 3 sebaran (NNN)

1) $y_1 = 0 + 4.x_1 + \varepsilon \quad ; \varepsilon \sim N(0,1)$
2) $y_2 = 1 + 10.x_2 + \varepsilon \quad ; \varepsilon \sim N(0,1)$
3) $y_3 = 3 + 15.x_3 + \varepsilon \quad ; \varepsilon \sim N(0,3)$



(a) (b) (c)

Gambar 10 Diagram sebaran data simulasi normal 3 kelompok

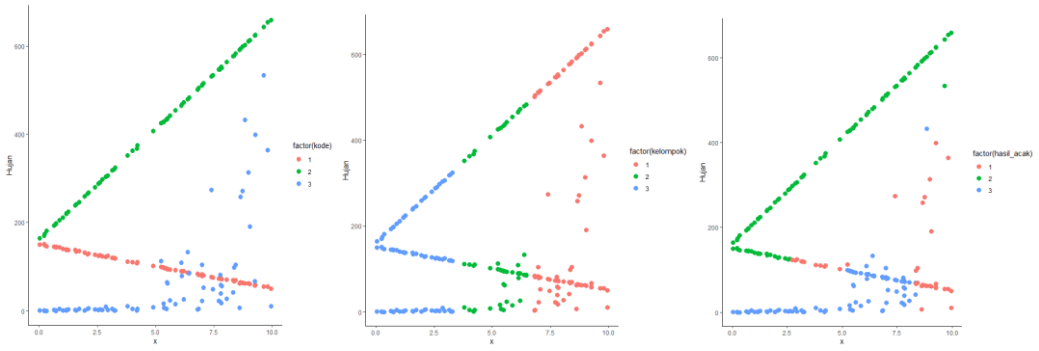
Berdasarkan diagram pencar terlihat bahwa regresi gerombol mampu memisahkan data dengan sangat baik. Anggota dalam gerombol satu, gerombol dua dan gerombol tiga terpisah dengan tepat sesuai dengan sebaran sebenarnya. Hasil simulasi NNN menghasilkan rataan akurasi tertinggi berada pada model normal sebesar 95,93%, akurasi model Gamma sebesar 41,93%, akurasi model GNN sebesar 65% dan akurasi model GGN sebesar 44,85%.

4.6 Simulasi Model 3 Gerombol Sebaran Gamma-Normal-Normal

Simulasi sebaran Gamma-normal-normal (GNN) memiliki dua skenario dibuat dengan kondisi sebaran yang berbeda. Terdiri dari model GNN1 dan GNN2 yang bertujuan untuk melihat kemampuan regresi gerombol dalam kondisi sebaran campuran yang berbeda bentuk sebarannya. Peubah respon dibuat dengan membangkitkan galat beridistribusi Gamma dan Normal. Simulasi sebaran Normal dengan persamaan $y = \beta_0 + \beta_1.x + \varepsilon$ dan galat $\varepsilon \sim N(\mu, \sigma^2)$. Sebaran Gamma dengan persamaan $y = 1/(\beta_0 + \beta_1.x)$, kedua kelompok dibangkitkan dengan parameter β yang berbeda. Simulasi GNN1, sebaran pertama dengan koefisien β_1 = 0,6 dan ξ = 1, sebaran kedua dengan koefisien β_0 = 150, β_1 = -10, dan galat $\varepsilon \sim N(0,1)$. Sebaran ketiga dengan koefisien β_0 = 160, β_1 = 50 dan galat $\varepsilon \sim N(0,1)$. Simulasi GNN2, sebaran pertama dengan koefisien β_1 = 0,6 dan ξ = 15. Sebaran kedua dengan koefisien β_0 = 85 dan β_1 = 5 dan galat $\varepsilon \sim N(0,1)$. Sebaran ketiga dengan koefisien β_0 = 120, β_1 = 40 dan galat $\varepsilon \sim N(0,1)$. Respon sebaran Normal dengan membangkitkan galat dan masuk pada persamaan dengan nilai β_0 dan β_1 yang telah ditentukan. Respon sebaran Gamma dengan parameter β dan ξ dan yang telah ditentukan dan kedua data digabungkan menjadi satu peubah respon dengan proporsi masing masing sebaran sama.

a) Simulasi Gamma-Normal-Normal 3 sebaran (GNN1)

1) $y_1 = 0 + 0{,}6.x_1 \quad ; \xi = 1$
2) $y_2 = 150 + (-10).x_2 + \varepsilon \quad ; \varepsilon \sim N(0,1)$
3) $y_3 = 160 + 50.x_3 + \varepsilon \quad ; \varepsilon \sim N(0,1)$



(a) (b) (c)

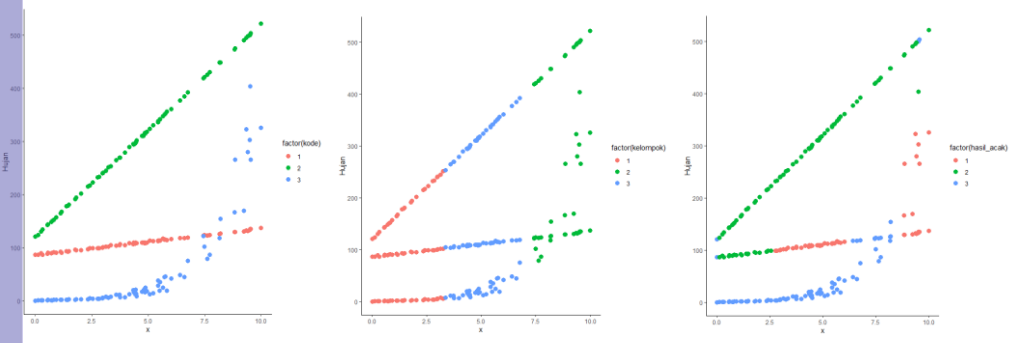
Gambar 11 Diagram sebaran data simulasi GNN 3 kelompok

Proses gerombol GNN1 dan GNN2 dijelaskan dalam Gambar 11 dan Gambar 12. Gambar (a) diagram sebaran data bangkitan sesungguhnya, Gambar (b) diagram sebaran data hasil pengelompokkan data bangkitan dengan k-means dan Gambar (c) diagram sebaran data hasil pengelompokkan menggunakan regresi gerombol. Hasil simulasi GNN1, regresi gerombol dapat memisahkan data dengan sangat baik pada ketiga gerombol. Simulasi GNN1 menghasilkan rataan akurasi tertinggi berada pada model GNN sebesar 81,37%, akurasi pada model Gamma sebesar 53,35%, akurasi pada model normal sebesar 80,06% dan akurasi pada model GGN sebesar 62,96%. Berdasarkan diagram pencar Gambar 11, terlihat regresi gerombol mampu mengelompokkan gerombol dua dengan baik (warna hijau), tetapi ada beberapa salah klasifikasi pada gerombol satu menjadi gerombol tiga.

Simulasi GNN2 menghasilkan rataan akurasi tertinggi pada model GNN sebesar 81,23%, akurasi pada model Gamma sebesar 51,27%, akurasi pada model normal sebesar 79,45% dan akurasi model GGN sebesar 61,68%. Berdasarkan diagram pencar terlihat bahwa regresi gerombol cukup mampu memisahkan anggota dalam gerombol satu, gerombol dua dan gerombol tiga. Berdasarkan diagram pencar Gambar 12, regresi gerombol mampu mengelompokkan gerombol dengan baik. Beberapa kesalahan terletak pada gerombol satu yang terklasifikasi ke dalam gerombol dua.

b) Simulasi Gamma-Normal-Normal 3 sebaran (GNN2)

1) $y_1 = 0 + 0{,}6.x_1 \quad ; \xi = 15$
2) $y_2 = 85 + 5.x_2 + \varepsilon \quad ; \varepsilon \sim N(0,1)$
3) $y_3 = 20 + 40.x_3 + \varepsilon \quad ; \varepsilon \sim N(0,1)$



(a) (b) (c)

Gambar 12 Diagram sebaran data simulasi GNN 3 kelompok (2)

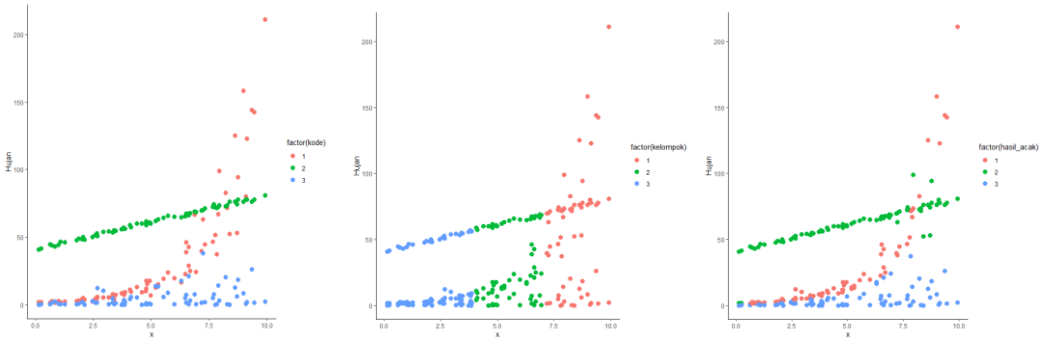
4.7 Simulasi Model 3 Gerombol Sebaran Gamma-Gamma-Normal

Tahap terakhir untuk simulasi adalah simulasi sebaran Gamma-Gamma-Normal (GGN). Terdapat dua skenario dibuat dengan kondisi sebaran yang berbeda yaitu GGN1 dan GGN2. Hal ini bertujuan untuk melihat kemampuan regresi gerombol dalam memisahkan data. Sebaran Normal memiliki persamaan $y = \beta_0 + \beta_1.x + \varepsilon$ dengan galat menyebar normal $\varepsilon \sim N(0,1)$ dan sebaran Gamma

dengan persamaan $y = 1/(\beta_0 + \beta_1.x)$. Kedua kelompok dibangkitkan dengan parameter β dan ξ yang berbeda. Simulasi GGN1, sebaran pertama dengan koefisien β_1 = 0,25 dan ξ = 0,5. Sebaran kedua dengan koefisien β_0 = 0,2, β_1 = 0,5 dan ξ = 15, sebaran ketiga dengan koefisien β_0 = 40, β_1 = 4 dan galat ε~N(0,1). Simulasi GGN2, sebaran pertama dengan koefisien β_0 = -2, β_1 = 0,25 dan ξ = 0,5, sebaran kedua dengan koefisien β_1 = 0,48 dan ξ = 15, sebaran ketiga dengan koefisien β_0 = 50, β_1 = -3 dan galat ε~N(0,1). Hasil simulasi dijelaskan dengan diagram pencar. Gambar 13 dan Gambar 14 menjelaskan proses pengerombolan GGN1 dan GGN2.

a) Simulasi Gamma-Gamma-Normal 3 sebaran (GGN1)

1) $y_1 = 0 + 0{,}25.x_1$; $\xi = 0{,}5$
2) $y_2 = 0{,}2 + 0{,}5.x_2$; $\xi = 15$
4) $y_3 = 40 + 4.x_3 + \varepsilon$; ε~N(0,1)



(a) (b) (c)

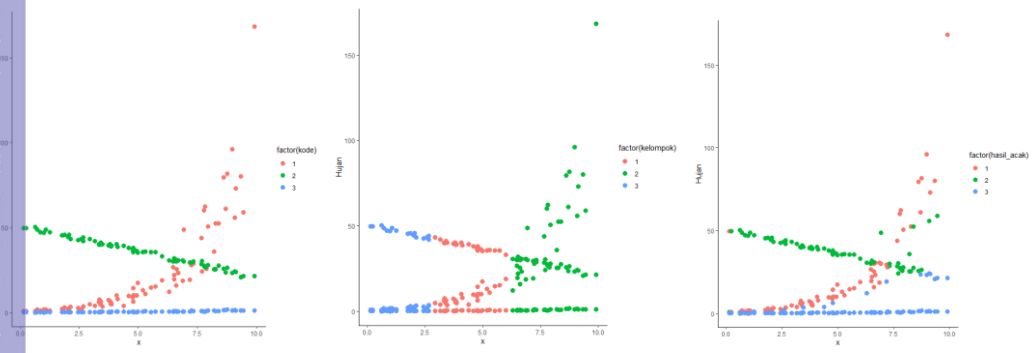
Gambar 13 Diagram sebaran data simulasi GGN 3 kelompok

Gambar (a) diagram sebaran data bangkitan sesungguhnya, Gambar (b) diagram sebaran data hasil pengelompokkan data bangkitan dengan k-means. Gambar (c) adalah diagram sebaran data hasil pengelompokkan menggunakan regresi gerombol. Diagram pencar menunjuukan regresi gerombol dapat memisahkan data sesuai dengan sebaran asli. Kesalahan kalsifikasi pada gerombol satu yang masuk ke gerombol dua terlihat pada Gambar 13 (c).

Simulasi GGN1 menghasilkan rataan akurasi tertinggi pada model GGN sebesar 84,41%, akurasi model Gamma sebesar 73,89%, akurasi model normal sebesar 60,25% dan akurasi model GNN sebesar 76,15%. Simulasi GGN2 menghasilkan rataan akurasi tertinggi pada model GGN sebesar 92,65%, akurasi model Gamma sebesar 77,96%, akurasi model normal sebesar 56,88% dan akurasi model GNN sebesar 77,6%. Diagram pencar pada Gambar 14 menunjukkan regresi gerombol mampu memisahkan data dengan baik pada gerombol satu, dua dan tiga. Kedua model simulasi menghasilkan rataan akurasi tertinggi pada model GNN, sesuai dengan sebaran sebenarnya.

b) Simulasi Gamma-Gamma-Normal 3 sebaran (GGN2)

1) $y_1 = (-2) + 0.25.x_1 \quad ; \xi = 0.5$
2) $y_2 = 0 + 0.48.x_2 \quad ; \xi = 15$
3) $y_3 = 50 + (-3).x_3 + \varepsilon \quad ; \varepsilon \sim N(0,1)$



(a) (b) (c)

Gambar 14 Diagram sebaran data simulasi GGN 3 kelompok (2)

Simulasi regresi gerombol tiga kelompok terdiri dari empat model. Terdiri dari model GGG, model NNN, model GNN1, model GNN2, model GGN1 dan model GGN2. Hasil pada semua model menunjukkan bahwa regresi gerombol mampu memisakan data dengan baik sesuai dengan sebaran sebenarnya. Hasil menunjukkan simulasi sebaran Gamma, akurasi tertinggi berada pada model gerombol Gamma. Simulasi sebaran normal, akurasi tertinggi berada pada model gerombol normal. Simulasi sebaran GNN, akurasi tertinggi berada pada model gerombol GNN, dan terakhir pada simulasi sebaran GGN akurasi tertinggi berada pada amodel gerombol GGN. Simulasi dua gerombol dan simulasi tiga gerombol sebanyak memiliki 11 model simulasi, dan dapat disimpulkan bahwa algoritma yang dibentuk mampu mengelompokkan data sesuai dengan karakteristik data sebenarnya atau data bangkitannya. Rangkuman hasil regresi gerombol tiga kelompok disajikan dalam Tabel 6 berikut.

Tabel 6 Rangkuman akurasi model tiga kelompok pada data simulasi

no	Skenario	Akurasi Gamma	Akurasi Normal	Akurasi GNN	Akurasi GGN
1	Model GGG	**68,05 %**	53,31 %	45,07 %	64,52 %
2	Model NNN	41,93 %	**95,94** %	65,00 %	44,85 %
3	Model GNN 1	53,35 %	80,06 %	**81,37** %	62,96 %
4	Model GNN 2	52,17 %	79,45 %	**81,23** %	61,68 %
5	Model GGN 1	73,89 %	60,25 %	76,15 %	**84,41** %
6	Model GGN 2	77,96 %	56,88 %	77,60 %	**92,65** %

Simulasi tiga kelompok Gamma menghasilkan nilai rataan akurasi yang cukup baik pada model Gamma sebesar 68,15%. Simulasi tiga kelompok normal memiliki

satu skenario dan menghasilkan nilai rataan akurasi yang sangat baik pada model normal sebesar 95,94%. Simulasi tiga kelompok campuran GNN memiliki dua skenario. Skenario pertama dan kedua menghasilkan nilai rataan akurasi cukup tinggi pada model GNN sebesar 81,37% dan 81,23%. Simulasi tiga kelompok campuran GGN memiliki dua skenario. Skenario pertama dan kedua menghasilkan nilai rataan akurasi cukup tinggi pada model GGN sebesar 84,41% dan 92,65%. Simulasi dengan tiga kelompok menghasilkan akurasi yang selaras dengan simulasi dengan dua kelompok sebelumnya.

Simulasi data sebaran campuran GN, GNN1, GNN2, GGN1 dan GGN2 dibuat dengan skenario lebih dari 1 dengan parameter berbeda beda. Hal ini untuk memastikan regresi gerombol dapat memisahkan data dengan sebaran yang berbeda. Simulasi sebaran campuran menghasilkan rataan akurasi tertinggi pada model sesuai dengan sebaran sebenarnya. Berdasarkan simulasi dua kelompok dan simulasi tiga kelompok, dapat disimpulkan bahwa algoritma regresi gerombol yang dibentuk mampu mengelompokkan data sesuai dengan karakteristik asli atau sebaran asli data.

Tabel 7 Akurasi dan proporsi gerombol terbaik pada model simulasi tiga kelompok

No	Regresi gerombol	Akurasi	Gerombol	Proporsi
1	Model GGG	68,05 %	1	39,11 %
			2	14,67 %
			3	46,22 %
2	Model NNN	92,66 %	1	35,11 %
			2	32,44 %
			3	32,44 %
3	Model GNN 1	81,37 %	1	16,89%
			2	39,56%
			3	43,55%
4	Model GNN 2	81,23 %	1	24,44 %
			2	42,67 %
			3	32,89 %
5	Model GGN 1	84,41 %	1	28,00 %
			2	38,67 %
			3	33,33 %
6	Model GGN 2	92,65 %	1	28,44 %
			2	36,00 %
			3	35,56 %

Proporsi gerombol terbaik dijelaskan didalam Tabel 7. Data dibangkitkan dengan proporsi yang sama pada setiap gerombol. Model GGG menunjukkan rataan akurasi yang cukup baik sebesar 68,15%. Proporsi menunjukan gerombol satu dapat terpisahkan dengan baik tetapi proporsi gerombol dua dan gerombol tiga

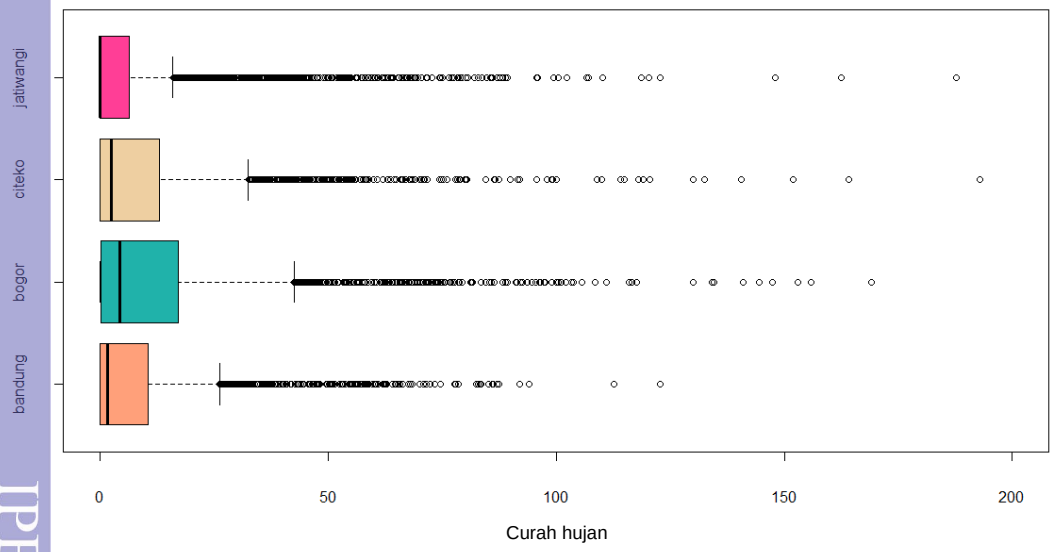
memiliki selisih nilai proporsi yang cukup tinggi. Gerombol terbesar GGG berada pada gerombol tiga dengan proporsi sebesar 46,22%. Model NNN menghasikan rataan akurasi yang sangat tinggi dan nilai proporsi yang hampir seimbang pada tiap gerombolnya. Model GNN1 dan model GNN2 menghasilkan rataan akurasi yang baik sebesar 81,37% dan 81,23% dengan proporsi yang hampir seimbang pada tiap gerombol. Proporsi tertinggi model GNN1 berada pada gerombol 3 dengan 43,55%, proporsi tertinggi model GNN2 berada pada gerombol 2 dengan 42,67%. Model GGN1 didapatkan akurasi lebih rendah dibandingkan model GGN2. Proporsi model GGN1 dan GG2 nilainya hampir seimbang dengan proporsi tertinggi pada gerombol 2.

4.8 Aplikasi Regresi Gerombol dengan Sebaran Gamma

4.8.1 Eksplorasi Data

Regresi gerombol diaplikasikan pada data curah hujan harian. Curah hujan harian adalah jumlah air yang jatuh di permukaan tanah datar selama satu hari yang diukur dengan satuan tinggi milimeter (mm/hari) di atas permukaan horizontal. Data respon yang digunakan adalah data curah hujan harian yang bersumber dari BMKG (Badan Meteorologi Klimatologi dan Geofisika). Periode waktu 2010 hingga 2019 pada keempat stasiun hujan terpilih yang berada di provinsi Jawa Barat. Diantaranya stasiun hujan Bandung, Bogor, Citeko dan Jatiwangi.

Pola sebaran curah hujan masing-masing stasiun hujan digambarkan dalam kotak dan garis. Berdasarkan Gambar 15, semua stasiun hujan memiliki data pencilan yang menjulur kanan menandakan banyaknya curah hujan yang tinggi atau curah hujan ekstrem. Letak diagram kotak dan garis cenderung di kiri artinya data cenderung berkumpul dibawah, yaitu data curah hujan yang berada dalam jangkauan 0 hingga 20 mm. Hal ini menunjukkan bahwa tidak hujan ataupun hujan dengan intensitas kecil sering terjadi.



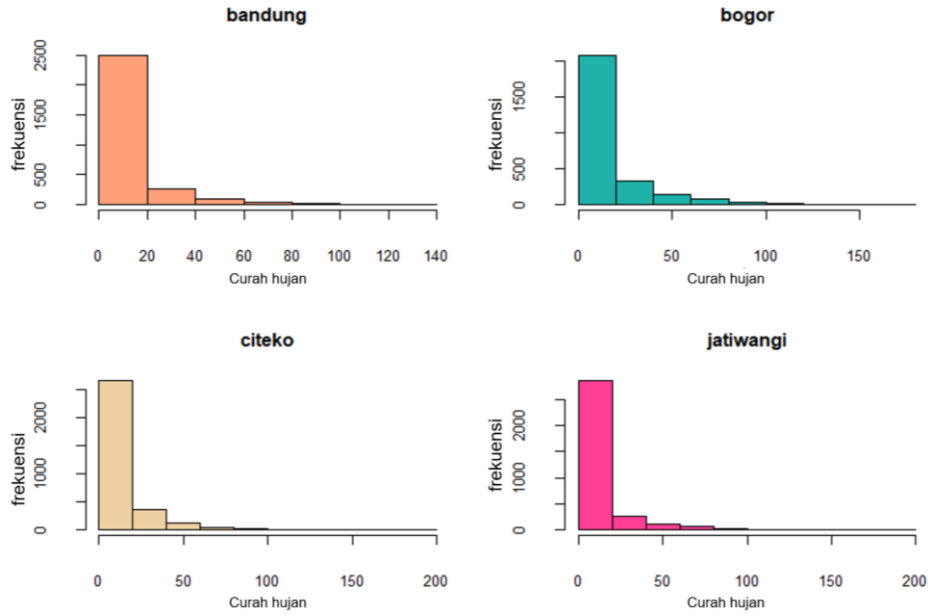
Gambar 15 Diagram kotak-garis curah hujan pada 4 stasiun hujan

Berdasarkan stastistik deskriptif pada Tabel 8, terlihat bahwa rata-rata hujan pada semua stasiun berada dalam rentang 8-13 mm/hari, tetapi nilai maksimum sangat besar berada di rentang ratusan. Perhatikan bahwa rata-rata lebih besar dari median, sehingga data menjulur ke kanan (kemiringan positif). Stasiun hujan dengan keragaman tertinggi berada pada stasiun hujan Bogor dengan standar deviasi sebesar 20,60. Keragaman terendah berada pada stasiun hujan Bandung dengan standar deviasi 14,33. Daerah dengan rata-rata hujan tertinggi yaitu Bogor sebesar 8,3 mm/hari. Curah hujan maksimum paling tinggi selama 10 tahun terakhir berada pada stasiun Citeko sebesar 192,8 mm/hari sedangkan curah hujan maksimum paling rendah selama 10 tahun terakhir berada di stasiun Bandung sebesar 122,9 mm/hari.

Tabel 8 Statistik deskriptif curah hujan pada 4 stasiun hujan

Stasiun	Minimum	Median	Rataan	Maksimum	Std. Deviasi
Bandung	0	1,6	8,30	122,9	14,33
Bogor	0	4,3	13,17	169,1	20,60
Citeko	0	2,5	10,07	192,8	17,12
Jatiwangi	0	0,0	7,86	187,8	17,28

Berdasarkan histogram pada Gambar 16 untuk semua stasiun hujan, data cenderung berkumpul di curah hujan 0-20 mm/hari. Semakin tinggi curah hujan, frekuensi cenderung menurun, artinya curah hujan yang tinggi jarang sekali terjadi. Terutama pada stasiun hujan Bandung, frekuensi curah hujan rendah 0-20 mm/hari mencapai 2500 hari. Curah hujan ini lebih tinggi jika dibanding stasiun hujan lainnya menunjukkan bahwa hujan lebih jarang terjadi di daerah Bandung. Pada stasiun Bogor, frekuensi curah hujan rendah 0-20mm/hari kurang dari 2000 menunjukkan bahwa hujan lebih sering terjadi di daerah Bogor.



Gambar 16 Histogram curah hujan pada 4 stasiun hujan

4.8.2 Analisis Komponen Utama

Peubah prediktor yang digunakan adalah data global GCM dengan domain 6 × 6 grid dan ukuran grid 0,5° × 0,5°, maka dalam analisis ini terdapat 36 peubah prediktor bagi masing masing stasiun hujan. Peubah prediktor yang cukup banyak dapat mengakibatkan multikolinearitas. Hal ini dapat menyebabkan bias dalam pembentukan model, dan perlu penanganan multikolinearitas tanpa menghilangkan peubah prediktor. Solusi untuk masalah ini adalah Analisis komponen utama yang menghasilkan komponen-komponen baru tanpa menghilangkan informasi dari data asli. Kriteria Pemilihan Komponen Utama yang terbentuk dapat ditentukan dengan menggunakan 2 kriteria yaitu diagram *scree*, dan proporsi kumulatif keragaman.

Diagram *scree* merupakan kriteria pemilihan jumlah Komponen Utama yang lebih bersifat grafis atau visual. Diagram *Scree* merupakan diagram antara akar ciri (*eigenvalue*) dan Komponen Utama yang terbentuk. Untuk menentukan jumlah Komponen Utama yang terbentuk perlu diperhatikan letak dimana terjadi patahan siku dari diagram *scree* atau dipilih suatu titik terakhir dimana setelah titik tersebut diagram cenderung meluruh (Rencher 1998). Komponen utama yang dipilih adalah Komponen Utama yang mempunyai akar ciri bernilai lebih besar dari 1.

Kriteria ini digunakan terutama untuk Komponen Utama yang diperoleh dari matriks korelasi. Alasan untuk membandingkannya dengan λ_i lebih besar dari 1 adalah ketika Komponen Utama yang diperoleh dari matriks korelasi (standarisasi data), variansi dari masing-masing peubahnya sama dengan satu. Jika suatu Komponen Utama tidak dapat menerangkan variansi melebihi suatu peubah dapat menerangkan dirinya sendiri, maka komponen utama tersebut tidak signifikan atau dengan kata lain komponen utama yang mempunyai akar ciri kurang dari 1 dapat diabaikan. Proporsi kumulatif keragaman data asal yang dijelaskan oleh *k* komponen utama minimal 75%. Persentase keragaman total yang dijelaskan oleh *k* Komponen Utama dihitung menggunakan Persamaan (14).

$$\frac{\sum_{i=1}^{k}\lambda_i}{\sum_{i=1}^{p}\lambda_i} \times 100\% \ , \quad k \leq p \tag{14}$$

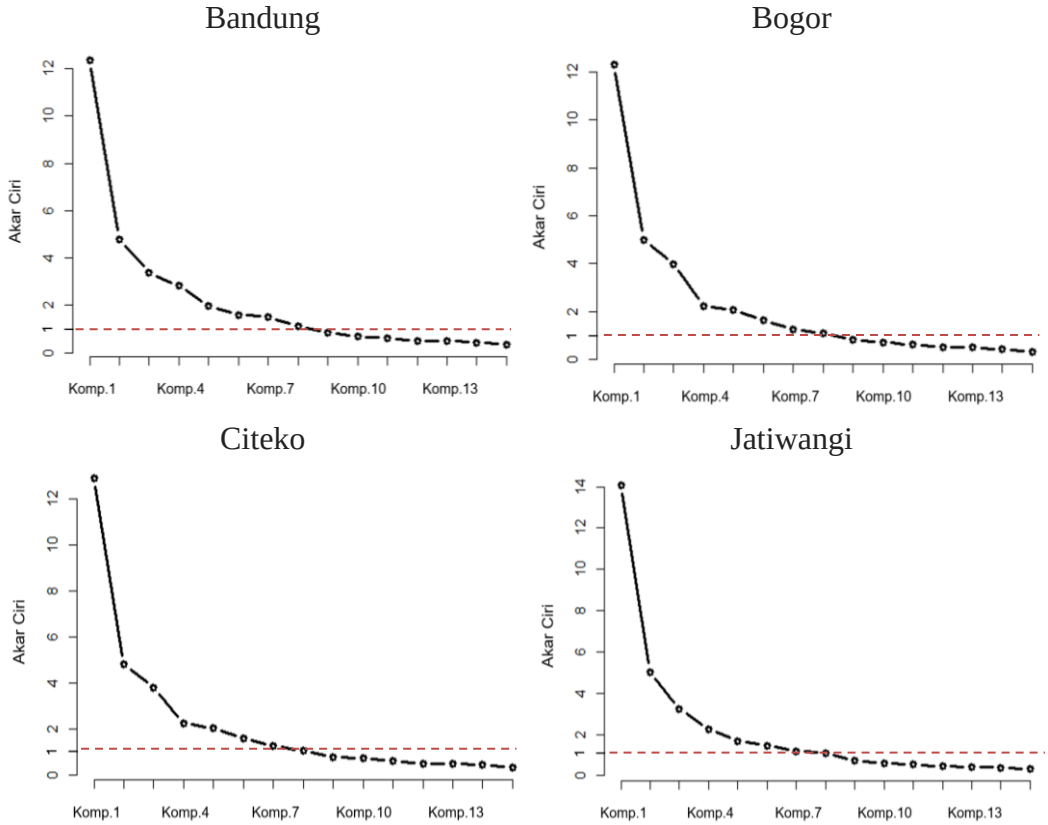
Tabel 9 Proporsi kumulatif keragaman komponen utama pada 4 stasiun hujan

Stasiun	K1	K2	K3	K4	K5	K6	K7	K8
Bandung	0,35	0,48	0,57	0,65	0,70	0,75	**0,79**	0,82
Bogor	0,37	0,49	0,60	0,66	0,71	**0,76**	0,79	0,82
Citeko	0,37	0,49	0,60	0,66	0,71	**0,76**	0,79	0,82
Jatiwangi	0,39	0,53	0,62	0,68	0,72	**0,77**	0,80	0,83

Berdasarkan Tabel 9, dipilih komponen yang dengan proporsi kumulatif keragaman lebih dari 75% pada tiap stasiun hujan. Pada stasiun Bandung terpilih 7 komponen dengan persentase keragaman sebesar 79%. Stasiun Bogor dan Citeko

memiliki nilai komponen utama yang sama karena berada pada cakupan grid yang sama dan terpilih 6 komponen dengan persentase keragaman sebesar 76%. Stasiun Jatiwangi terpilih 6 komponen dengan persentase sebesar 77%. Selain proporsi kumulatif keragaman, menentukan komponen utama dilihat dari akar ciri (*eigenvalue*), dengan syarat pada komponen terpilih adalah komponen dengan nilai akar ciri lebih dari 1.

Gambar 17 merupakan diagram *scree* untuk seluruh stasiun hujan. Berdasarkan diagram *scree*, pada stasiun hujan Bandung terpilih 7 komponen utama dan keragaman sebesar 79% pada 7 komponen. Oleh karena itu, terpilih 7 komponen utama untuk stasiun hujan Bandung. Pada stasiun hujan Bogor terpilih 6 komponen utama berdasarkan diagram *scree* dan nilai persentase keragaman sebesar 76% pada 6 komponen, maka dari itu terpilih 6 komponen utama untuk stasiun hujan Bogor. Stasiun hujan Citeko terpilih 6 komponen utama berdasarkan scree diagram dan nilai persentase keragaman sebesar 76% pada 6 komponen, maka dari itu terpilih 6 komponen utama untuk stasiun hujan Citeko. Stasiun hujan Jatiwangi, terpilih 6 komponen utama berdasarkan scree diagram dan nilai persentase keragaman sebesar 77% pada 6 komponen, maka dari itu terpilih 6 komponen utama untuk stasiun hujan Jatiwangi. Pemilihan komponen utama pada setiap stasiun hujan kemudian menjadi peubah prediktor. Selanjutnya dilakukan pemodelan regresi gerombol dengan peubah prediktor yang berasal dari komponen utama.



Gambar 17 Diagram *scree* pada 4 stasiun hujan

4.8.3 Pemodelan Regresi Gerombol

Berdasarkan hasil simulasi sebelumnya, regresi gerombol terbukti mampu menggerombolkan data berdasarkan kesesuaian model dan sebaran data. Regresi gerombol tersebut diterapkan pada *statistical downscaling* yang memproyeksikan data curah hujan lokal dengan data presipitasi global (GCM). Analisis Komponen Utama digunakan terlebih dahulu untuk pereduksian variabel prediktor. Proses reduksi peubah ini menghasilkan enam atau tujuh komponen pada setiap stasiun hujan. Variabel laten / komponen utama ini digabungkan dengan data hujan lokal harian pada stasiun Bandung, Bogor, Citeko dan Jatiwangi sebagai data respon.

Model tanpa penggerombolan dibentuk dari regresi sebaran Gamma. Model dua gerombol yaitu model campuran Gamma-Normal (GN). Model tiga gerombol sebanyak dua model, dengan model pertama dibentuk dari dua kelompok regresi Gamma dan satu sebaran regresi normal (GGN). Model kedua dibentuk dari dua kelompok regesi normal dan satu sebaran regresi Gamma (GNN). Kinerja model dievaluasi dengan membandingkan curah hujan yang diamati dan yang diprediksi menggunakan nilai RMSE (*Root Mean Squared Error*) dan RMSEP (*Root Mean Squared Error Prediction*). RMSE disebut juga akar rata deviasi kuadrat, adalah ukuran perbedaan antara nilai prediksi oleh model dan nilai observasi yang dimodelkan. RMSE bertujuan untuk menilai sebarapa baik model dapat menggerombolkan, RMSEP bertujuan untuk menilai seberapa baik model dapat memprediksi curah hujan.

Analisis komponen utama menghasilkan komponen-komponen utama untuk setiap stasiun hujan. Komponen utama tersebut dibagi menjadi data *training* sebesar 80% dan data *testing* sebesar 20%. Data *training* akan masuk ke dalam proses regresi gerombol untuk mendapatkan gerombol baru dan kebaikan gerombol baru diukur menggunakan RMSE. Gerombol baru masuk ke dalam proses prediksi dan diukur menggunakan RMSEP. Tabel 10 menyajikan nilai RMSE pada regresi gerombol untuk setiap stasiun hujan.

Tabel 10 Nilai RMSE regresi gerombol dari 4 stasiun hujan

Stasiun hujan	RMSE		
	GN (2)	GNN (3)	GGN (3)
Bandung	11,87	**9,93**	10,43
Bogor	17,93	**13,76**	14,91
Citeko	13,63	12,76	**11,70**
Jatiwangi	13,66	12,62	**12,37**

Secara umum, nilai RMSE pada tiga gerombol lebih baik dari nilai RMSE dua gerombol. Nilai RMSE pada Stasiun Bandung, untuk model GN sebesar 11,87 dan nilai RMSE terbaiknya pada model GNN sebesar 9,93. Nilai RMSE pada stasiun Bogor, tertinggi dibanding stasiun lainnya. Nilai RMSE model GN sebesar 17,93 dan RMSE terbaik stasiun Bogor pada model GNN sebesar 13,76. Stasiun Citeko

memiliki nilai RMSE GN 13,63 dan RMSE terbaik pada model GGN sebesar 11,70. Stasiun Jatiwangi memiliki nilai RMSE GN 13,66 dan nilai RMSE terbaik pada model GGN sebesar 12,37. Hal ini menunjukkan bahwa model tiga gerombol lebih baik dalam menggerombolkan data pada stasiun Bandung, Bogor, Citeko dan Jatiwangi. Gerombol baru yang didapatkan melalui regresi gerombol, kemudian dilakukan prediksi pada data *testing* dengan dua acara. Prediksi dengan pendekatan jarak amatan pada titik *centroid* tiap gerombol (cara A) dan dengan pendekatan jarak amatan pada setiap anggota dalam gerombol (cara B). Tabel 11 menyajikan nilai RMSEP untuk stasiun hujan Bandung, Bogor, Citeko dan Jatiwangi.

Tabel 11 Nilai RMSEP regresi gerombol dari 4 stasiun hujan

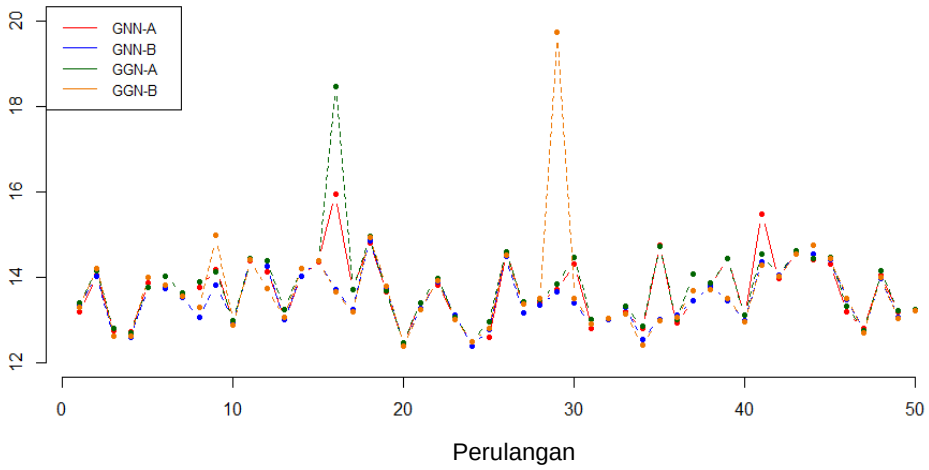
Stasiun hujan	Tanpa gerombol	Penggerombolan					
		2 (GN)		3 (GNN)		3 (GGN)	
		A	B	A	B	A	B
Bandung	20,43	13,51	13,44	13,76	**13,40**	13,79	13,66
Bogor	21,35	18,90	18,87	**18,59**	18,71	18,67	18,85
Citeko	37,15	15,33	15,45	15,20	**15,12**	15,29	15,20
Jatiwangi	38,80	15,40	15,35	**15,34**	15,66	15,73	15,78

Berdasarkan tabel 11, kolom tanpa gerombol yang artinya hanya dilakukan analisis denga regresi Gamma dan menghasilkan nilai RMSEP tinggi pada semua stasiun hujan. Model dua gerombol yaitu model Gamma-Normal (GN) dan model tiga gerombol Gamma-Normal-Normal (GNN) dan Gamma-Gamma-Normal (GGN). Nilai RMSEP tanpa gerombol tertinggi ada pada stasiun Jatiwangi sebesar 38,80 dan terkecil pada stasiun hujan Bandung sebesar 20,43. Dua atau tiga gerombol menghasil nilai RMSEP yang lebih rendah dibandingkan dengan tanpa dilakukan penggerombolan. Tidak ada perbedaan signifikan nilai RMSEP antara dua gerombol ataupun tiga gerombol pada semua stasiun hujan.

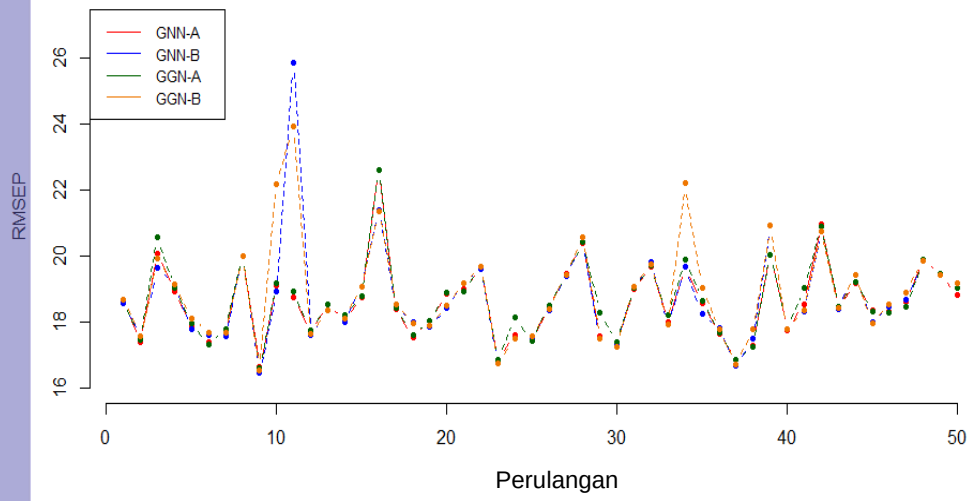
Model tanpa penggerombolan menghasilkan rataan RMSEP sebesar 20,43 pada stasiun hujan Bandung. Model terbaik untuk stasiun hujan Bandung adalah model GNN prediksi B dengan nilai rataan RMSEP sebesar 13.40. Nilai rataan RMSEP tanpa penggerombolan menghasilkan sebesar 21,35 pada stasiun hujan Bogor. Model terbaik untuk stasiun hujan Bogor adalah model GNN prediksi A dengan nilai rataan RMSEP sebesar 18,59. Stasiun hujan Citeko dimodelkan tanpa penggerombolan menghasilkan rataan RMSE sebesar 37,15. Model terbaik untuk stasiun hujan Citeko berada pada model GNN prediksi B dengan nilai rataan RMSEP sebesar 15,12. Stasiun hujan Jatiwangi dimodelkan tanpa penggerombolan menghasilkan rataan RMSE sebesar 38,8. Model terbaik untuk stasiun hujan Jatiwangi berada pada model GNN prediksi A dengan nilai rataan RMSEP sebesar 15,34.

Perulangan untuk setiap stasiun hujan dilakukan sebanyak 50 kali bagi masing masing model gerombol. Gambar 18, Gambar 19, Gambar 20 dan Gambar 21,

merupakan sebaran nilai RMSEP bagi setiap stasiun hujan Bandung, Bogor, Citeko dan Jatiwangi. Model terbaik untuk seluruh stasiun hujan berada pada model 3 gerombol, maka pada plot sebaran akan tersaji nilai RMSEP 3 gerombol. Model tersebut adalah model GNN prediksi A, GNN prediksi B, GGN prediksi A dan GGN prediksi B.



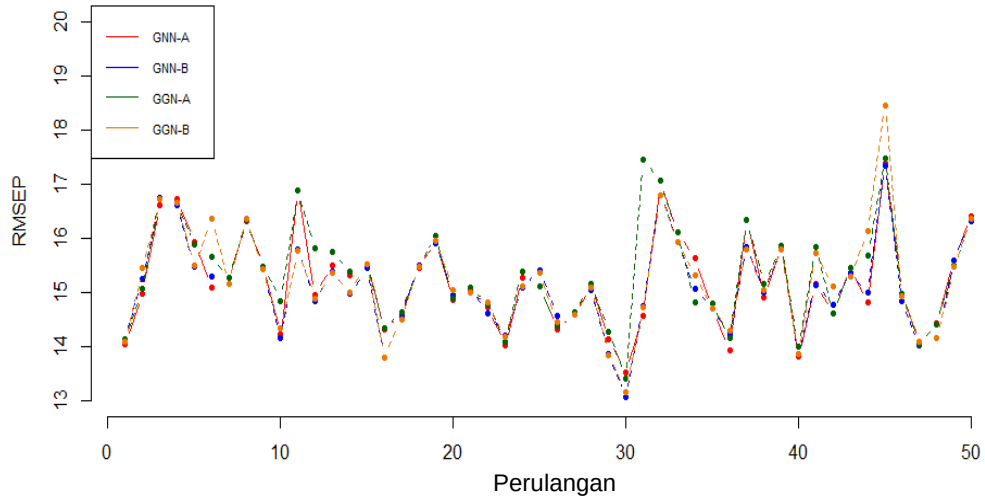
Gambar 18 Nilai RMSEP pada 50 ulangan untuk tiga gerombol stasiun Bandung



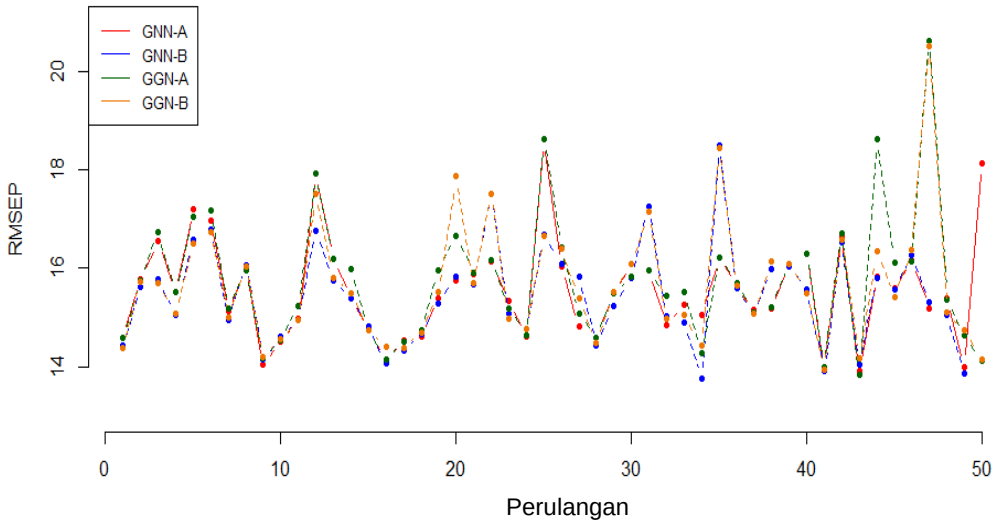
Gambar 19 Nilai RMSEP pada 50 ulangan untuk tiga gerombol stasiun Bogor

Stasiun Bandung dengan nilai RMSEP menyebar pada rentang 12,5 hingga 20 mm/hari. RMSEP terbaik pada model GNN dengan prediksi B yang ditunjukkan dengan garis biru pada Gambar 18. Model GNN mampu menghasilkan nilai RMSEP yang lebih konstan dibandingkan dengan model GGN. Nilai RMSEP stasiun Bogor menyebar pada rentang 16 hingga 26 mm/hari, dimana nilai ini lebih tinggi jika dibandingkan dengan stasiun lainnya. RMSEP terbaik stasiun Bogor

berada pada model GNN dengan prediksi A yang ditunjukkan dengan garis merah pada Gambar 19.



Gambar 20 Nilai RMSEP pada 50 ulangan untuk tiga gerombol stasiun Citeko



Gambar 21 Nilai RMSEP pada 50 ulangan untuk tiga gerombol stasiun Jatiwangi

RMSEP terbaik stasiun hujan Citeko berada pada model GNN dengan prediksi B yang ditunjukkan dengan garis biru pada Gambar 20. Nilai RMSEP menyebar dalam rentang 14 hingga 21 mm/hari. Model GNN menghasilkan RMSEP yang lebih konstan jika dibandingkan dengan model GGN. Stasiun hujan Jatiwangi mendapatkan RMSEP terbaik pada model GNN dengan prediksi A yang ditunjukkan dengan garis merah pada Gambar 21. Terlihat bahwa model GNN menghasilkan nilai RMSEP yang lebih konstan dibanding model GGN.

Tabel 12 menunjukkan karakteristik gerombol terbaik untuk masing – masing stasiun hujan. Gerombol yang digunakan untuk stasiun Bandung adalah tiga gerombol. Proporsi tertinggi stasiun Bandung berada pada sebaran Gamma dengan

proporsi sebesar 72,3%. Gerombol satu dengan sebaran Gamma terdiri dari 416 data dengan rataan sebesar 9,27 dan median 7,23. Gerombol dua dan tiga dengan sebaran normal terdiri dari 10 dan 149 data dengan rataan yang cukup berbeda signifikan yaitu sebesar 3,52 dan 12,78. Gerombol yang digunakan untuk stasiun Bogor adalah tiga gerombol. Proporsi tertinggi stasiun Bogor berada pada sebaran Gamma dengan proporsi sebesar 62,9%. Gerombol satu dengan sebaran Gamma sebanyak 332 data dengan rataan sebesar 10,88. Gerombol dua dan gerombol tiga dengan sebaran normal sebanyak 52 dan 144 data dengan rataan sebesar 18,42 dan 15,93.

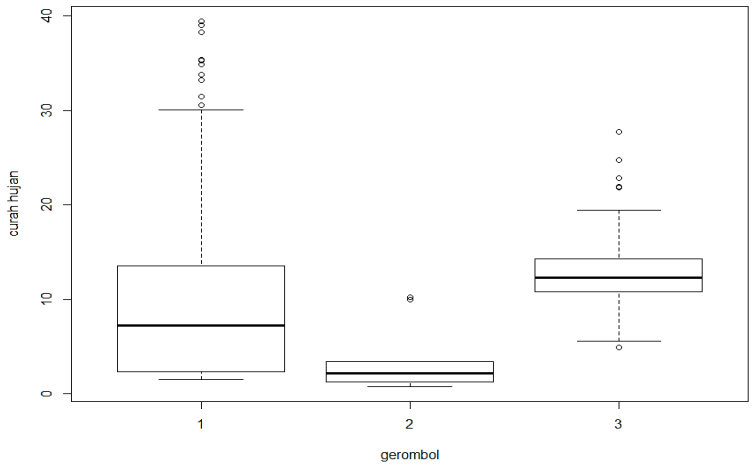
Tabel 12 Proporsi gerombol terbaik pada 4 stasiun hujan

Stasiun hujan	Gerombol	Sebaran	Proporsi
Bandung	1	Gamma	62,9 %
	2	Normal	9,8 %
	3	Normal	27,3 %
Bogor	1	Gamma	62,9 %
	2	Normal	9,8 %
	3	Normal	27,3 %
Citeko	1	Gamma	13,3 %
	2	Normal	8,5 %
	3	Normal	78,2 %
Jatiwangi	1	Gamma	26,6 %
	2	Normal	6,5 %
	3	Normal	66,9 %

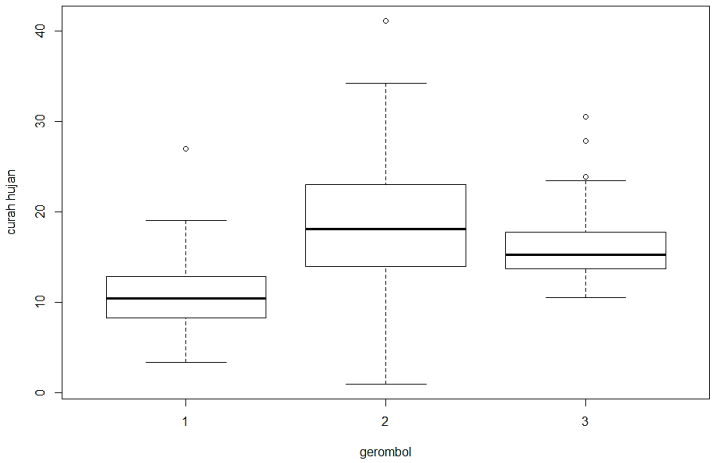
Gerombol yang digunakan untuk stasiun Citeko adalah tiga gerombol. Proporsi tertinggi stasiun Citeko berada pada sebaran normal dengan proporsi sebesar 78,2%. Gerombol satu dengan sebaran Gamma sebanyak 85 data dengan nilai rataan sebesar 18,44, gerombol dua dan gerombol tiga dengan sebaran normal sebanyak 54 dan 499 data dengan nilai rataan sebesar 12,64 dan 6,52. Stasiun terakhir, yaitu stasiun Jatiwangi dengan gerombol sebanyak tiga gerombol. Proporsi tertinggi stasiun Jatiwangi berada pada sebaran Normal dengan proporsi sebesar 66,9%. Frekuensi pada gerombol satu dengan sebaran Gamma sebanyak 176 data dengan nilai rataan 13,47, untuk gerombol dua dan tiga dengan sebaran normal sebanyak 43 dan 442 data dengan nilai rataan sebesar 10,34 dan 6,33.

Diagram kotak dan garis digunakan untuk menjelaskan kondisi gerombol terbaik pada masing-masing stasiun hujan. Gambar 22 memperlihatkan data curah hujan di setiap gerombol pada stasiun Bandung. Prediksi data curah hujan stasiun Bandung dengan model GNN prediksi B sebagai model terbaik. Setiap gerombol memiliki nilai pencilan dengan nilai maksimum di gerombol satu sebesar 39,4 dengan rataan 9,27. Gerombol dua dengan nilai maksimum sebesar 10 dengan rataan 3,52 dan gerombol tiga dengan nilai maksimum sebesar 27,75 dengan rataan 12,78. Berdasarkan Gambar 23 menunjukan data curah hujan untuk stasiun Bogor

di setiap gerombol dengan model terbaik adalah model GNN prediksi A. Nilai maksimum pada gerombol satu sebesar 26,98 dengan rataan 10,88. Gerombol dua dengan nilai maksimum sebesar 41,09 dengan rataan 18,42 dan gerombol tiga dengan nilai maksimum sebesar 30,47 dengan rataan 15,93.



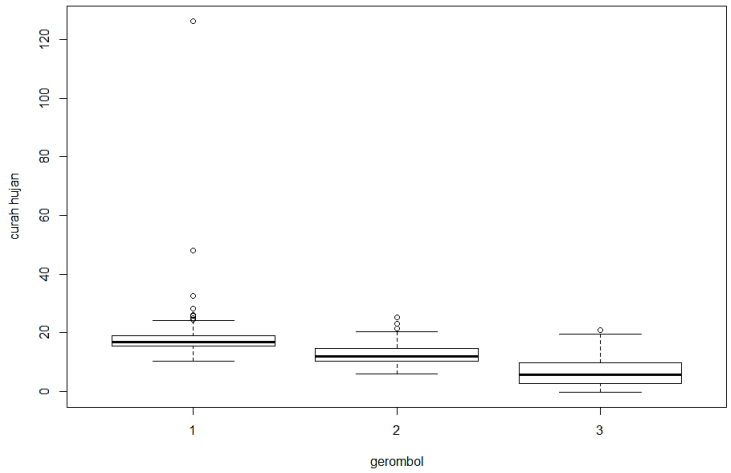
Gambar 22 Diagram kotak-garis dugaan curah hujan pada gerombol terbaik stasiun Bandung



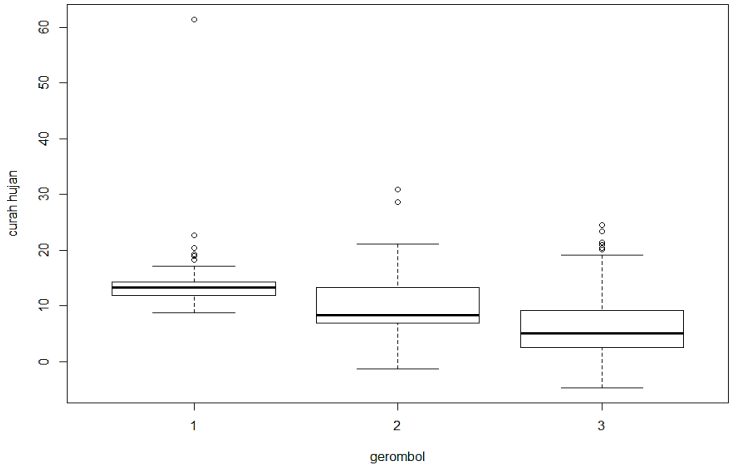
Gambar 23 Diagram kotak-garis dugaan curah hujan pada gerombol terbaik stasiun Bogor

Gambar 24 memperlihatkan data curah hujan di setiap gerombol pada stasiun Citeko. Model terbaik untuk stasiun Citeko adalah model GNN dengan prediksi B. Berdasarkan Gambar 23, terlihat pencilan jauh pada gerombol satu dengan nilai maksimum sebesar 126,34 dan rataan 18,44. Gerombol dua memiliki nilai maksimum sebesar 25,21 dengan rataan 12,64. Gerombol tiga memiliki nilai maksimum sebesar 20,91 dengan rataan 6,52. Berdasarkan Gambar 25, pada stasiun Jatiwangi, model terbaik adalah model GNN dengan prediksi A. Setiap gerombol memiliki pencilan, pada gerombol satu memiliki nilai maksimum sebesar 61,39

dengan rataan 13,47. Gerombol dua memiliki nilai maksimum 30,91 dengan rataan 10,34 dan gerombol tiga memiliki nilai maksimum 24,45 dengan rataan 6,33.



. Gambar 24 Diagram kotak-garis dugaan curah hujan pada gerombol terbaik stasiun Citeko



Gambar 25 Diagram kotak-garis dugaan curah hujan pada gerombol terbaik stasiun Jatiwangi

V SIMPULAN DAN SARAN

5.1 Simpulan

Pengembangan regresi gerombol dengan sebaran Gamma, normal dan campuran melalui simulasi terbukti dapat mengelompokkan data sesuai dengan sebaran sebenarnya. Nilai ketepatan klasifikasi lebih besar dari 90% untuk model dengan sebaran homogen. Sebaran campuran menghasilkan nilai ketepatan klasifikasi kurang dari 90% yaitu pada model Gamma-Normal (GN) sebesar 84,90%, dan model Gamma-Normal-Normal (GNN) sebesar 81,37%. Nilai ketepatan klasifikasi yang kurang dari 90% pada sebaran campuran dipengaruhi oleh bentuk sebaran Gamma dan Normal yang beririsan menyebabkan salah klasifikasi pada beberapa amatan. Penerapan regresi gerombol menghasilkan model Gamma-Normal-Normal (GNN) adalah model paling sesuai untuk memodelkan curah hujan di seluruh stasiun hujan dengan RMSEP terkecil dibanding model Gamma-Normal (GN) dan Gamma-Gamma-Normal (GGN). Nilai RMSEP stasiun Bandung sebesar 13,4, stasiun Bogor sebesar 18,59, stasiun Citeko sebesar 15,122 dan pada stasiun Jatiwangi sebesar 15,345. Oleh karena itu, model campuran dengan sebaran Gamma dan normal dapat dikatakan model terbaik untuk menduga curah hujan harian ketika model Gamma saja belum mampu menghasilkan model terbaik.

DAFTAR PUSTAKA

Bagirov AM, Mahmood A, Barton A. 2017. Prediction of Monthly Rainfall in Victoria, Australia: Clusterwise Linear Regression. J Atmosres. 20-29.

Benestad RE, Chen D, Hanssen-Bauer I. 2008. *Empirical-statistical downscaling*. Singapore: World Scientific Publishing Co.

Brusco MJ, Cradit JD, Steinley D, Fox GL. 2008. Cautionary Remarks on the Use of Clusterwise Regression, *Multivariate Behavioral Research*, 43:1, 29-49.

Busuioc A, Chen D, Hellstrom C. 2001. Performance of Statistical downscaling Models in GCM Validation and Regional Climate Change Estimates: Application for Swedish Precipitaion. *International Journal of Climatology* 21: 557–578.

Butar-butar VP, Soleh AM, Wigena AH. 2019. Pemodelan Clusterwise Regression Pada Statistical downscaling Untuk Pendugaan Curah Hujan Bulanan. *Indonesian Journal of Statistics and Its Applications*, *3*(3), 236–246. https://doi.org/10.29244/ijsa.v3i3.310

Grun B, Leisch F. 2007. FlexMix: An R package for finite mixture modelling. R News, 7(1): 8–13.

Jolliffe IT. 2002. *Principal Component Analysis. 2nd Edition*. New York: Springer.

Lindsey JK. 1997. *Applying Generalized Linear Models*. New York: Springer.

McCullagh P, Nelder JA. 1989. *Generalized Linear Models*. 2nd Edition. London: Chapman and Hall.

Nadya AR. 2018. Pemodelan Statistical downscaling untuk Menduga Curah Hujan degan Regresi Linear Gerombol dan Pemodelan Dua Tahap [Tesis]. Bogor (ID): Institut Pertanian Bogor.

Permatasari S, Djuraidah A, Soleh AM. 2016. Statistical downscaling with Gamma Distribution and Elastic Net Regularization. In *The 2nd International Conference on Applied Statistics (ICAS 2016), Departement of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Padjadjaran*. pp. 121–129.

Rencher AC. 1998. *Multivariate Statistical Inference and Applications*. New York: John Wiley & Sons, Inc.

Saha S, Moorthi S, Pan HL, Wu X, Wang J, Nadiga S, Tripp P, Kistler R, Woollen J, Behringer D, others. 2010. The NCEP climate forecast system reanalysis. *Bull. Am. Meteorol. Soc*. 91(8) 1015–1058.

Soleh AM, Wigena AH, Djuraidah A, Saefudin A. 2015. Statistical downscaling to predict monthly rainfall using linear regression with L1 regularization (LASSO). *Applied Mathematics Sciences*. 9(108):5361-5369.

Soleh AM. 2015. Pemodelan Linear Sebaran Gamma dan Pareto Terampat dengan Regularisasi L1 pada Statistical Downscaling untuk Pendugaan Curah Hujan Bulanan [disertasi]. Bogor (ID): Institut Pertanian Bogor.

Spath H. 1979. Algorithm 39: Clusterwise Linear Regression. Computing 22. 367-373.

Sumertajaya IM, Erfiani, Putri WDY. 2007. Analisis gerombol menggunakan metode *two step cluster* (Studi kasus: data Potensi Desa Sensus Ekonomi 2003 wilayah Jawa Barat). *Forum Statistika dan Komputasi* p: 18-23 Vol 12 No.1

Syafruddin R, Soleh AM, Wigena AH. 2019. Clusterwise Regression Model Development with Gamma Distribution. In *1st International Conference on Statistics and Analytics* (ICSA).

Wigena AH. 2006. Pemodelan Statistical downscaling Dengan Regresi Projection Pursuit Untuk Peramalan Curah Hujan Bulanan (Kasus Curah hujan bulanan di Indramayu) [Disertasi] Bogor (ID): Institut Pertanian Bogor.

Wicaksono AS. 2019. Pendugaan Curah Hujan Harian Melalui Pemodelan Dua Tahap Dengan Klasifikasi Bagging dan Boosting Pada Statistical downscaling [tesis]. Bogor (ID): Institut Pertanian Bogor.

Zorita E, Von Storch H. 1999. The analog method as a Simple statistical downscaling technique. *Journal of Climate and Applied Meteorology* 12: 2474-2489.