

PENERAPAN *CLUSTERING TIME SERIES* UNTUK MENGGEROMBOLKAN PROVINSI DI INDONESIA BERDASARKAN NILAI PRODUKSI PADI

Andrew Donda Munthe

Badan Pusat Statistik Kota Kupang – Provinsi Nusa Tenggara Timur
dondaandrew@gmail.com

Abstract. *In fulfilling national food needs, rice plants are an important main commodity. The characteristics of various regions that differ from one another, greatly affect the production of rice produced. The purpose of this research is to cluster provinces in Indonesia by applying clustering time series analysis. The application of clustering time series analysis is based on rice production data for each province in Indonesia in the period 1968 – 2015. Clustering time series analysis with hierarchical and non-hierarchical methods results in the distribution of the same cluster members in the 3 optimal groups. The 1st cluster consists of 3 provinces, 2nd cluster consists of 15 provinces, while 3rd cluster consists of 8 provinces. On average, from the 3 clusters formed, the silhouette coefficient value is 0.64, so the clustering is categorized in Good Classification.*

Keyword: *cluster, rice, silhouette, time series.*

Abstraksi. *Dalam memenuhi kebutuhan pangan nasional, tanaman padi merupakan komoditas utama yang diupayakan setiap tahun agar produksinya terus meningkat. Karakteristik wilayah yang berbeda-beda sangat mempengaruhi produksi padi yang dihasilkan. Tujuan yang ingin dicapai dalam penelitian ini adalah untuk mengelompokkan provinsi-provinsi di Indonesia dengan melakukan penerapan analisis gerombol deret waktu (clustering time series analysis). Penerapan analisis gerombol deret waktu dilakukan berdasarkan perkembangan data produksi padi setiap provinsi di Indonesia pada periode tahun 1968 – 2015. Clustering time series analysis dengan metode hierarki dan tidak berhierarki menghasilkan distribusi anggota gerombol yang sama pada 3 gerombol optimal. Gerombol 1 terdiri dari 3 provinsi, gerombol 2 terdiri dari 15 provinsi, sedangkan gerombol 3 terdiri dari 8 provinsi. Secara rata-rata dari 3 gerombol yang terbentuk menghasilkan nilai koefisien silhouette sebesar 0.64 atau terkategori Good Classification.*

Kata kunci: *deret waktu, gerombol, padi, silhouette*

PENDAHULUAN

Sektor pertanian merupakan salah satu sektor strategis dan prioritas bagi nadi perekonomian di hampir seluruh wilayah Indonesia. Definisi sektor pertanian secara luas mencakup beberapa subsektor yaitu tanaman pangan, hortikultura, perkebunan, peternakan, perikanan, dan juga kehutanan. Sektor pertanian,

khususnya usaha pertanian lahan sawah, memiliki nilai multifungsi yang besar dalam rangka meningkatkan ketahanan pangan nasional, kesejahteraan petani, dan juga menjaga kelestarian lingkungan hidup (Sudaryanto dan Rusastra, 2006).

Dalam memenuhi kebutuhan pangan nasional, subsektor tanaman pangan (terutama padi) adalah komoditas yang

difokuskan untuk dapat ditingkatkan produksinya terus menerus. Hal ini karena tanaman padi merupakan bahan makanan pokok bagi mayoritas masyarakat Indonesia. Produksi padi dari tahun ke tahun (1993-2015) secara umum mengalami peningkatan yang signifikan. Beberapa faktor yang menjadi penyebabnya adalah peningkatan luas areal panen serta meningkatnya tingkat produktivitas lahan.

Produksi padi di dalam negeri yang terus meningkat ternyata belum mampu mengimbangi pesatnya laju pertumbuhan penduduk serta tingginya konsumsi beras di Indonesia. Kebutuhan beras nasional memang dapat dipenuhi dari produksi dalam negeri dan impor. Akan tetapi, ketergantungan akan pangan impor menyebabkan ketahanan pangan menjadi rentan. Hal ini tentunya akan menimbulkan dampak negatif terhadap berbagai aspek kehidupan masyarakat yaitu aspek sosial, ekonomi, dan bahkan politik (Sembiring, 2008).

Karakteristik wilayah yang berbeda-beda sangat mempengaruhi nilai produksi padi yang dihasilkan di setiap provinsi. Karakteristik wilayah yang berbeda-beda tersebut dapat diidentifikasi dengan melakukan analisis statistik yang disebut dengan analisis gerombol (*cluster analysis*). Analisis ini perlu dilakukan agar program kebijakan yang dilakukan oleh pemerintah dapat lebih fokus dan terarah sesuai dengan karakteristik dari gerombol yang terbentuk.

Analisis gerombol merupakan salah satu teknik analisis peubah ganda yang digunakan untuk mengidentifikasi dan mengelompokkan individu/objek yang memiliki kemiripan ke dalam kelompok atau gerombol tertentu. Karakteristik objek-objek dalam suatu gerombol

tertentu akan memiliki tingkat kemiripan yang tinggi sedangkan karakteristik antar objek pada satu gerombol dengan gerombol yang lain memiliki tingkat kemiripan yang rendah. Dengan kata lain, keragaman dalam suatu gerombol menjadi minimum sedangkan keragaman antar gerombol yang terbentuk akibat penerapan analisis ini menjadi maksimum (Mattjik dan Sumertajaya, 2011).

Penerapan analisis gerombol semakin berkembang dengan penggunaan data deret waktu. Kumpulan data deret waktu tersebut dapat dikelompokkan berdasarkan karakteristik pada masing-masing data deret waktu. Penggunaan analisis gerombol deret waktu (*clustering time series analysis*) dilakukan dengan cara mengelompokkan objek berdasarkan pola deret waktunya. Akan tetapi, pemilihan jarak dan metode penggerombolan yang digunakan harus dapat memadai struktur data deret waktu yang sifatnya sangat dinamis (Liao, 2005).

Berdasarkan uraian di atas, penelitian ini bertujuan untuk menggerombolkan provinsi-provinsi di Indonesia dengan melakukan penerapan analisis gerombol deret waktu (*clustering time series analysis*). Penerapan analisis gerombol deret waktu tersebut dilakukan berdasarkan data produksi padi di 26 provinsi di Indonesia pada periode tahun 1968–2015.

METODE PENELITIAN

1. Sumber Data

Penelitian ini menggunakan data sekunder dari Badan Pusat Statistik (BPS) dengan cakupan 26 provinsi di Indonesia. Data yang digunakan adalah data produksi padi (ton) dalam bentuk Gabah Kering Giling (GKG). Periode waktu

produksi padi tersebut sejak tahun 1968 – 2015. Sebagai catatan, analisis gerombol deret waktu tidak dapat digunakan apabila terdapat data yang kosong atau hilang (*missing data*). Oleh karena itu, penelitian ini hanya menerapkan analisis gerombol pada provinsi-provinsi yang memiliki data yang lengkap. Terdapat 26 provinsi dari total 34 provinsi di Indonesia yang datanya lengkap periode tahun 1968–2015.

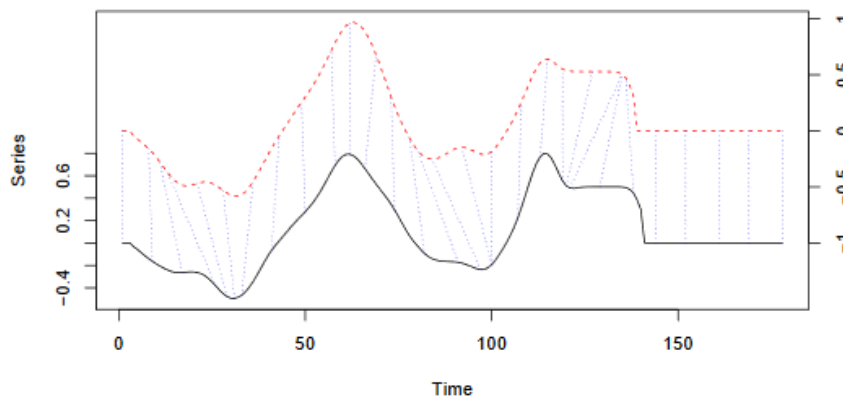
2. Metode Analisis

Data produksi padi dari 26 provinsi di Indonesia periode tahun 1968 - 2015 diolah menggunakan Ms. Exel dan *software* statistik R 3.4.4 dengan *Package library* yang digunakan adalah “TSclust”.

Langkah-langkah analisis data yang dilakukan dalam penelitian ini adalah sebagai berikut:

a. Melakukan eksplorasi data

b. Melakukan penggerombolan metode berhierarki analisis gerombol deret waktu dengan menggunakan ukuran jarak *Dynamic Time Warping* (DTW). Jarak *dynamic time warping* (DTW) adalah sebuah algoritma yang membandingkan dua seri data dan melakukan perhitungan untuk menemukan lintasan optimum diantara dua seri data tersebut. Jarak DTW merupakan generalisasi dari algoritma klasik yang membandingkan urutan diskret dengan urutan nilai kontinu. Apabila terdapat dua deret waktu yaitu Q dan R , dimana $Q = q_1, q_2, \dots, q_i, \dots, q_n$ dan $R = r_1, r_2, \dots, r_j, \dots, r_m$ maka penggunaan jarak DTW menyelaraskan data pada dua deret waktu tersebut sehingga perbedaan jarak keduanya minimum. Ilustrasi penerapan jarak DTW seperti ditunjukkan pada Gambar 1.



Gambar 1. Ilustrasi pengukuran jarak *Dynamic Time Warping* (DTW) (Sumber: Sardá-Espinosa, 2017)

Penerapan ukuran jarak DTW menggunakan perhitungan melalui sebuah matriks $n \times m$ dimana elemen (i, j) dari matriks tersebut terdapat jarak (q_i, r_j) . Jarak *Euclidean* digunakan untuk menghitung lintasan *warping* dengan $W = w_1, w_2, \dots, w_k$,

\dots, w_k dimana $\max(m, n) \leq K \leq m + n - 1$ merupakan himpunan elemen matriks yang memenuhi tiga kendala yaitu kondisi batas, kontinuitas, dan monotonisitas. Kendala kondisi batas membutuhkan lintasan *warping* untuk memulai dan menyelesaikan

perhitungan secara diagonal berlawanan sudut sel dari matriks, dengan $W_1 = (1,1)$ dan $W_K = (m, n)$. Jalur *warping* yang memiliki jarak minimum antara dua deret waktu rumusnya adalah sebagai berikut:

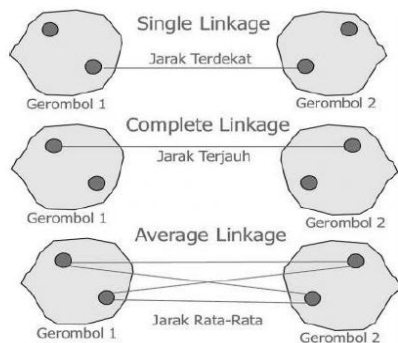
$$d_{DTW} = \min \frac{\sum_{k=1}^K w_k}{K}$$

dengan :

d_{DTW} = Jarak *Dynamic Time Warping*

w_k = lintasan *warping* ke- k

Perhitungan jarak DTW dilakukan dengan 3 metode keterpautan yaitu “*single linkage*”, “*complete linkage*” dan “*average linkage*”. Ilustrasi ketiga keterpautan seperti ditunjukkan pada Gambar 2.



Gambar 2. Ilustrasi keterpautan

- c. Melakukan perhitungan korelasi *Cophenetic*.

Koefisien korelasi *Cophenetic* merupakan koefisien korelasi antara elemen-elemen asli matriks ketidakmiripan (matriks jarak *Euclidean*) dan elemen-elemen yang dihasilkan oleh dendogram (matriks *Cophenetic* berdasarkan ukuran jarak dan metode keterhubungan yang digunakan). Formula untuk menghitung koefisien korelasi *Cophenetic* adalah sebagai berikut:

$$r_{Coph} = \frac{\sum_{i < k} (d_{ik} - \bar{d})(d_{Cik} - \bar{d}_C)}{\sqrt{[\sum_{i < k} (d_{ik} - \bar{d})^2][\sum_{i < k} (d_{Cik} - \bar{d}_C)^2]}}$$

dengan:

r_{Coph} = koefisien korelasi *Cophenetic*

d_{ik} = jarak *Euclidean* objek ke- i dan ke- k

\bar{d} = rata-rata d_{ik}

d_{Cik} = jarak *Cophenetic* objek ke- i dan ke- k

\bar{d}_C = rata-rata d_{Cik}

Nilai dari r_{Coph} berkisar antara -1 dan 1. Nilai yang semakin mendekati 1 berarti solusi yang dihasilkan dari proses gerombol semakin baik. Ukuran jarak DTW dengan metode keterhubungan yang menghasilkan nilai *cophenetic* terbesar merupakan solusi terbaik pada metode penggerombolan berhierarki.

- d. Melakukan penggerombolan tak berhierarki dengan menggunakan algoritma *k-means*.
- e. Melakukan perhitungan koefisien *silhouette* untuk melihat ketepatan penggerombolan algoritma *k-means*. Rumus perhitungan koefisien *silhouette* adalah sebagai berikut:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

dimana:

$a(i)$ = rata-rata jarak antar anggota dalam *cluster*

$b(i)$ = jarak terkecil antara anggota *cluster* dengan anggota *cluster* terdekatnya (*nearest neighbor*)

Kriteria ketepatan dan kualitas hasil penggerombolan berdasarkan nilai koefisien *silhouette* adalah sebagai berikut:

Tabel 1 Kategori penggerombolan berdasarkan koefisien *silhouette*

Nilai koefisien <i>silhouette</i> (SC)	Kategori Penggerombolan
0.71 – 1.00	<i>Strong Classification</i>
0.51 – 0.70	<i>Good Classification</i>
0.26 – 0.50	<i>Weak Classification</i>
0 – 0.25	<i>Bad Classification</i>

Sumber: Kauffman dan Rousseuw (1990)

HASIL DAN PEMBAHASAN

Deskripsi Data

Data produksi padi tersedia mulai periode tahun 1961–2015 pada 34 provinsi di Indonesia (Gambar 3). Dari 34 provinsi tersebut, terdapat 8 provinsi yang baru terbentuk hasil pemekaran dari provinsi induk dan datanya tidak lengkap. Provinsi-provinsi hasil pemekaran tersebut adalah Provinsi Kepulauan Riau, Provinsi Bangka Belitung, Provinsi Banten, Provinsi Kalimantan Utara, Provinsi Gorontalo, Provinsi Sulawesi Barat, Provinsi Maluku Utara, Provinsi Papua Barat.

Secara visual berdasarkan plot data, terlihat tiga pola yang terjadi pada perkembangan produksi padi di Indonesia periode tahun 1961 – 2015. Pola pertama terjadi pada provinsi-provinsi yang mengalami kecenderungan peningkatan produksi dengan tingkat fluktuasi yang rendah. Beberapa contoh provinsi dengan pola seperti ini adalah Provinsi Aceh,

Provinsi Sumatera Utara, Provinsi Sumatera Barat.

Pola kedua terjadi pada provinsi-provinsi yang mengalami kecenderungan peningkatan produksi dengan tingkat fluktuasi yang tinggi. Beberapa contoh provinsi dengan pola seperti ini adalah Provinsi Kalimantan Timur, Provinsi Sulawesi Utara, Provinsi Maluku.

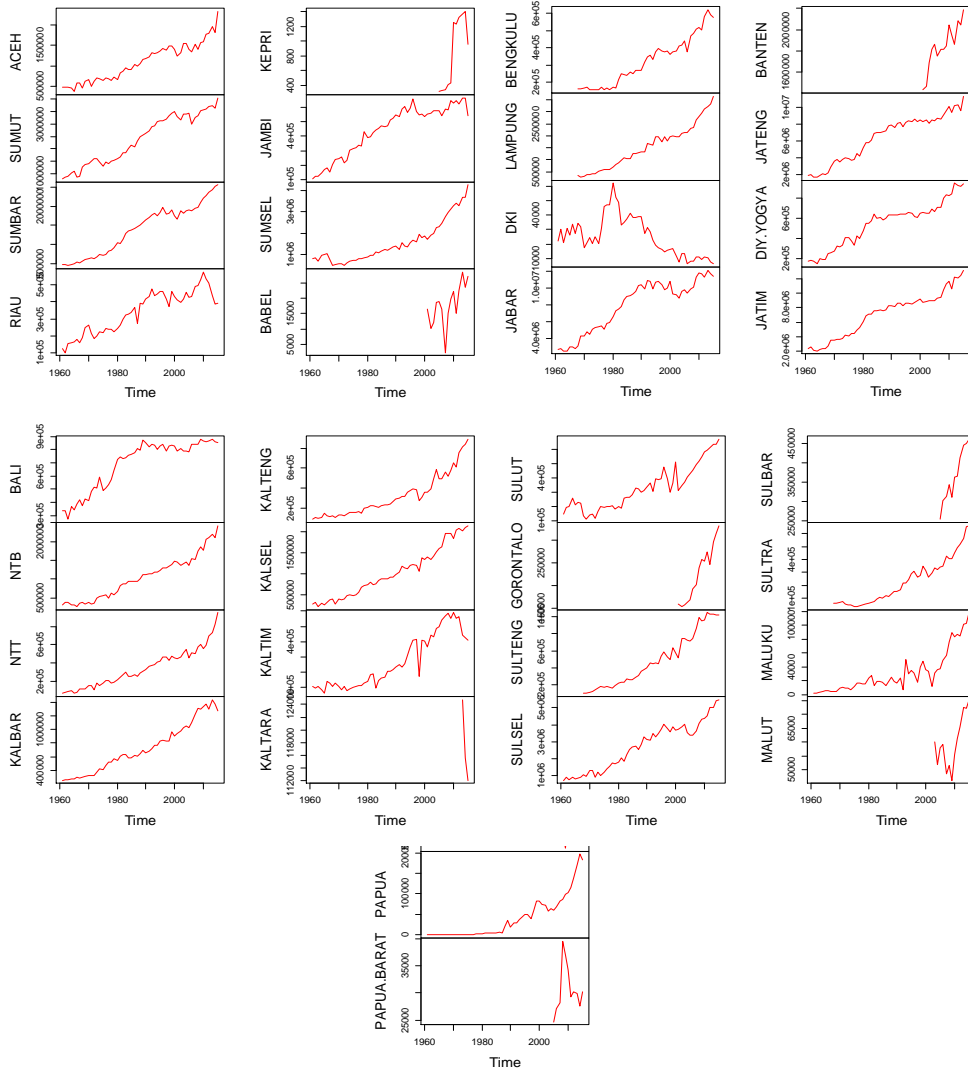
Pola ketiga adalah provinsi yang mengalami kecenderungan penurunan produksi yaitu Provinsi DKI Jakarta.

Analisis gerombol deret waktu tidak dapat digunakan apabila terdapat data yang hilang. Untuk itu, dalam penelitian ini penggerombolan dan analisis lebih lanjut hanya akan dilakukan pada 26 provinsi yang memiliki data yang lengkap yaitu pada periode tahun 1968–2015.

Pengerombolan Deret Waktu

Analisis gerombol deret waktu dalam penelitian ini bertujuan untuk menggerombolkan provinsi di Indonesia berdasarkan nilai produksi padi (ton GKG). Langkah pertama yang dilakukan dalam analisis ini adalah membuat matrik ukuran jarak *Dynamic Time Warping* (DTW) seperti ditunjukkan pada Tabel 2.

Jarak terdekat berdasarkan jarak *Dynamic Time Warping* (DTW) adalah antara Provinsi Papua dan Provinsi Maluku dengan nilai 606 728. Oleh karena itu, penggerombolan pertama kali penerapan jarak DTW dilakukan dengan mengelompokkan Provinsi Papua dan Provinsi Maluku.



Gambar 3. Plot produksi padi (ton GKG) dirinci per provinsi periode tahun 1961 - 2015

Tabel 2 Matrik ukuran jarak Jarak *Dynamic time warping* (DTW)

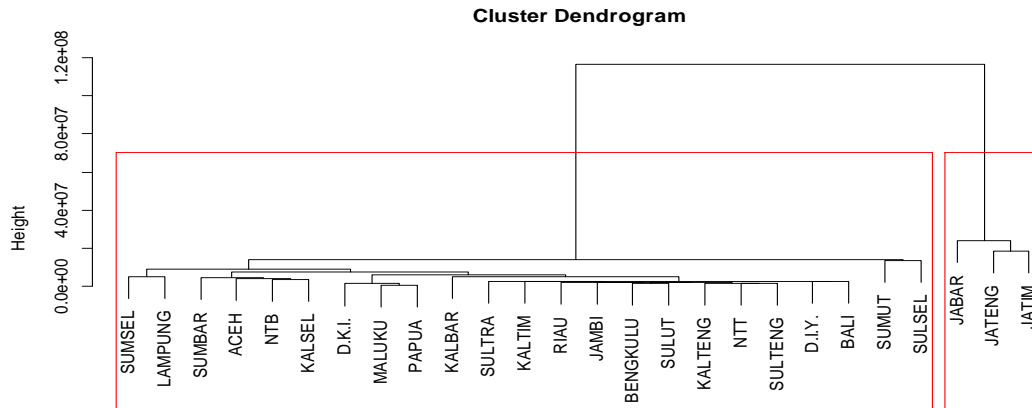
Provinsi	ACEH	SULSEL	SULTRA	MALUKU	PAPUA
ACEH	0					
.....	0				
SULSEL	52 947 261	0			
SULTRA	36 992 222	1.46 e+08	0		
MALUKU	70 443 742	1.79 e+08	7 359 233	0	
PAPUA	66 913 491	1.76 e+08	5 890 033	606 728	0

Sumber : Hasil olah data

Pengerombolan Berhierarki

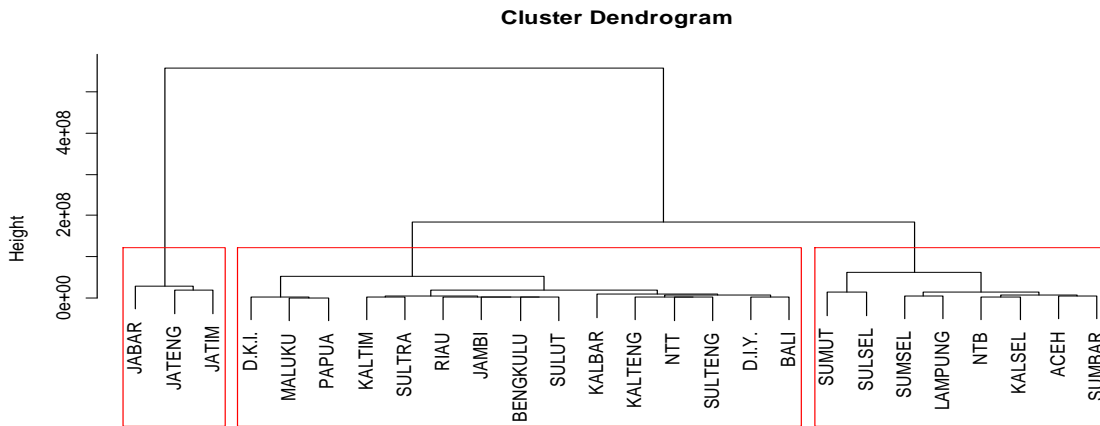
Langkah selanjutnya adalah menggambarkan dendrogram analisis gerombol deret waktu berhierarki dengan jarak DTW dan penerapan berbagai

metode keterpautannya (“single linkage”, “complete linkage”, dan “average linkage”). Hasil penggambaran dan pemotongan dendrogramnya adalah sebagai berikut:



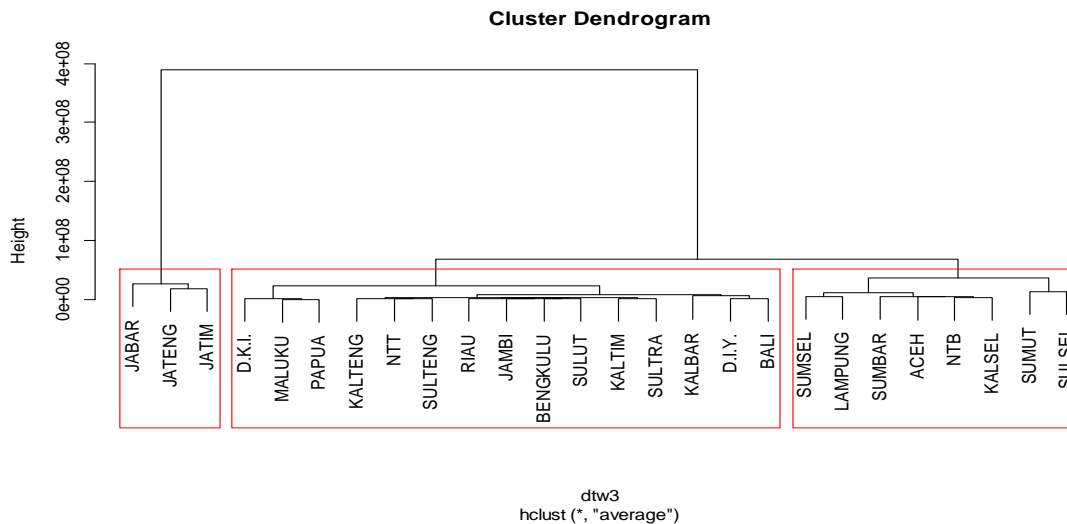
dtw1
hclust (*, "single")

Gambar 4. Dendrogram analisis gerombol dengan ukuran jarak DTW metode keterpautan “single linkage”



dtw2
hclust (*, "complete")

Gambar 5. Dendrogram analisis gerombol dengan ukuran jarak DTW metode keterpautan “complete linkage”



Gambar 6. Dendrogram analisis gerombol dengan ukuran jarak DTW metode keterpautan “average linkage”

Berdasarkan Pemotongan dendrogram analisis gerombol dengan ukuran jarak DTW dan keterpautan “single linkage” maka hasil gerombol yang terbentuk adalah sebanyak 2 gerombol (Gambar 4). Sedangkan untuk metode keterpautan “complete linkage” dan “average linkage” menghasilkan 3 gerombol (Gambar 5 dan Gambar 6).

Langkah selanjutnya yang dilakukan dalam penelitian ini adalah menghitung nilai korelasi *cophenetic* seperti yang ditunjukkan pada Tabel 3.

Tabel 3 Nilai korelasi *cophenetic* ukuran jarak DTW berdasarkan metode keterpautannya

No.	Metode Keterpautan	Korelasi <i>Cophenetic</i>
1.	<i>Single Linkage</i>	0.9652
2.	<i>Complete Linkage</i>	0.9520
3.	<i>Average Linkage</i>	0.9687

Sumber: Hasil olah data

Dari Tabel 3, nilai korelasi *Cophenetic* tertinggi (0.9687) adalah dengan menggunakan ukuran jarak DTW dan metode keterpautan *average linkage*. Dengan demikian maka berdasarkan

analisis gerombol deret waktu (*clustering time series*) metode berhierarki gerombol optimal yang terbentuk adalah sebanyak 3 gerombol. Distribusi provinsi-provinsi pada 3 gerombol optimal metode berhierarki adalah:

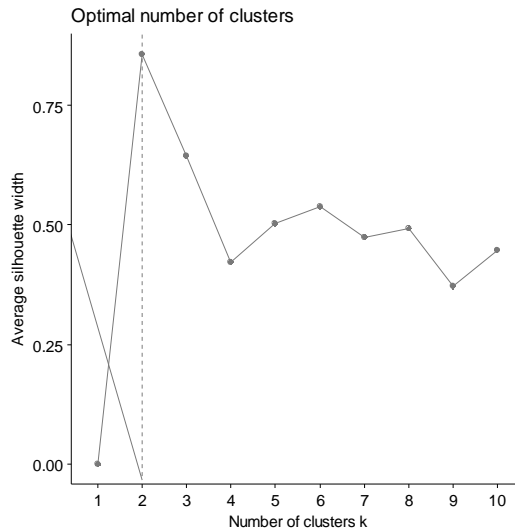
1. Gerombol 1 terdiri dari 3 provinsi yaitu Provinsi Jawa Barat, Jawa Tengah, dan Jawa Timur.
2. Gerombol 2 terdiri dari 15 provinsi yaitu Provinsi DKI Jakarta, Maluku, Papua, Kalimantan Tengah, Nusa Tenggara Timur, Sulawesi Tengah, Riau, Jambi, Bengkulu, Sulawesi Utara, Kalimantan Timur, Sulawesi Tenggara, Kalimantan Barat, DI Yogyakarta, dan Bali.
3. Gerombol 3 terdiri dari 8 provinsi yaitu Provinsi Sumatera Selatan, Lampung, Sumatera Barat, Aceh, Nusa Tenggara Barat, Kalimantan Selatan, Sumatera Utara, dan Sulawesi Selatan.

Pengerombolan Tak Berhierarki

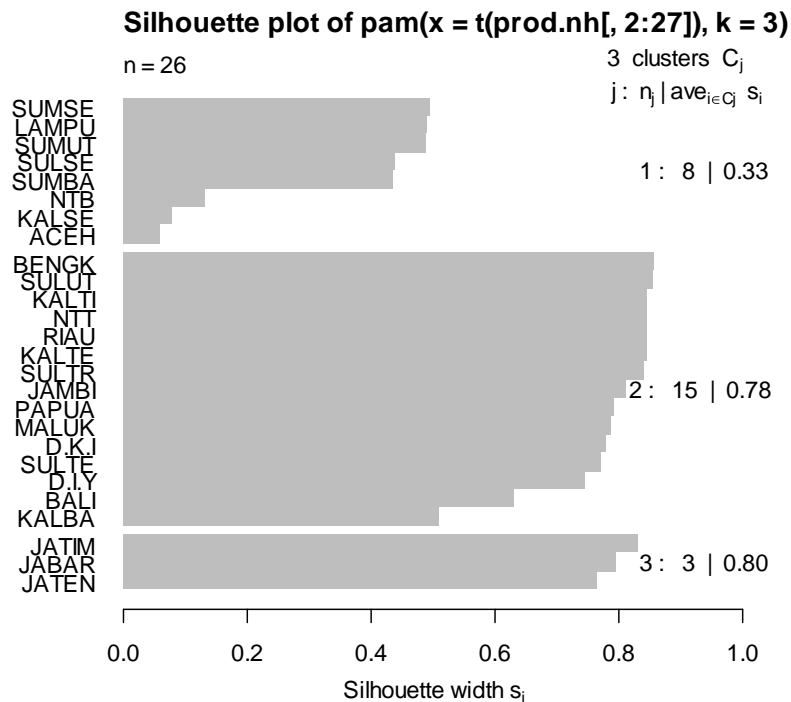
Salah satu cara untuk menduga kualitas dari gerombol yang terbentuk hasil algoritma *k-means* maka dapat dilihat dari pendekatan rata-rata nilai

silhouette. Berdasarkan Gambar 7 maka banyak gerombol optimal yang terbentuk adalah pada $k=2$, sedangkan banyak gerombol optimal yang kedua adalah pada $k=3$. Hasil penggerombolan metode

tidak berhirarki ini untuk dapat dibandingkan dengan metode hierarki maka yang digunakan dan dianalisis lebih lanjut adalah pada jumlah 3 gerombol ($k=3$).



Gambar 7. Gerombol optimal yang dihasilkan oleh algoritma *k-means*



Gambar 8. Plot *Silhouette* data produksi padi Indonesia periode 1968 - 2015

Perhitungan menggunakan algoritma *k-means* dan jumlah gerombol optimal sebanyak 3 menghasilkan distribusi gerombol yang sama dengan penggerombolan berhierarki. Distribusi penggerombolan dan hasil perhitungan koefisien *silhouette* untuk melihat ketepatan penggerombolan seperti ditunjukkan pada Gambar 8. Berdasarkan kriteria penggerombolan dengan nilai koefisien *silhouette*, gerombol 1 dengan nilai 0.33 terkategori dalam *Weak Classification*. Gerombol 2 dengan nilai 0.78 dan Gerombol 3 dengan nilai 0.80 terkategori dalam *Strong classification*. Secara rata-rata dari 3 gerombol yang terbentuk, nilai koefisien *silhouette* yang dihasilkan adalah sebesar 0.64. Hal ini berarti penggerombolan yang dilakukan sudah terkategori dalam *Good Classification*.

SIMPULAN

Analisis gerombol deret waktu dalam penelitian ini bertujuan untuk menggerombolkan provinsi di Indonesia berdasarkan nilai produksi padi (ton

GKG). Metode yang digunakan adalah hierarki dan tidak berhierarki. Nilai korelasi *cophenetic* yang tertinggi adalah pada ukuran jarak *Dynamic Time Warping* (DTW) keterpautan *average linkage* dengan nilai sebesar 0.9687. Dendrogram dengan menggunakan metode hierarki DTW *average linkage* menunjukkan terbentuknya 3 gerombol optimal.

Penggerombolan tak berhierarki dengan 3 gerombol menghasilkan distribusi anggota-anggota provinsi yang sama dengan anggota gerombol pada penggerombolan berhierarki. Gerombol 3 (3 provinsi) merupakan daerah dengan kategori produksi padi “tinggi”. Gerombol 1 (8 provinsi) merupakan daerah dengan kategori produksi padi “sedang”, sedangkan gerombol 2 (15 provinsi) merupakan daerah-daerah dengan kategori produksi padi “rendah”. Hasil 3 gerombol yang terbentuk secara rata-rata diperoleh nilai koefisien *silhouette* sebesar 0.64 yang berarti penggerombolan sudah terkategori dalam *Good Classification*.

DAFTAR PUSTAKA

- Aghabozorgi S., Shirkorshidi A.S., Wah T.Y. (2015). Time-series clustering – a decade review. *Information Systems*, 53,16-38.
- Badan Pusat Statistik. <https://www.bps.go.id/linkTableDinamis/view/id/865> [Internet] [diakses tanggal 1 Desember 2018].
- Badan Pusat Statistik. (2015). *Pedoman Pengumpulan Data Statistik Pertanian (SP) Tanaman Pangan 2015*. Jakarta: Badan Pusat Statistik.
- Badan Pusat Statistik. (2015). *Pedoman Pengumpulan Data Survei Ubinan Tanaman Pangan 2015*. Jakarta: Badan Pusat Statistik.
- Galeano, P., Peña, D. P. (2000). Multivariate analysis in vector time series. *Resenhas do Instituto de Matemática e Estatística da Universidade de São Paulo*, 4(4), 383-403.

- Gan, G., Ma, C., Wu, J. (2007). *Data clustering: theory, algorithms, and applications* (Vol. 20). Philadelphia: American Statistical Association – Society for Industrial and Applied Mathematics (ASA – SIAM).
- Giorgino, T. (2009). Computing and visualizing dynamic time warping alignments in R: the dtw package. *Journal of statistical Software*, 31(7), 1-24.
- Kaufman, L., Rousseuw, P.J. (1990). *Finding Group in Data : An Introduction to Cluster Analysis*. New Jersey: John Wiley & Sons, Inc.
- Liao, T. W. (2005). Clustering of time series data—a survey. *Pattern recognition*, 38(11), 1857-1874.
- Mattjik, A.A., Sumertajaya I.M. (2011). *Sidik Peubah Ganda Dengan Menggunakan SAS*. Bogor: IPB Press.
- Petitjean, F., Ketterlin, A., & Gançarski, P. (2011). A global averaging method for dynamic time warping, with applications to clustering. *Pattern Recognition*, 44(3), 678-693.
- Rahayu. 2016. Analisis Pencapaian Program Swasembada Beras Pada Tahun 2017 di Indonesia [Tesis]. Bogor: Institut Pertanian Bogor.
- Rani, S., Sikka, G. (2012). Recent techniques of clustering of time series data: a survey. *International Journal of Computer Applications*, 52(15).
- Sardá-Espinosa, A. (2017). Comparing time-series clustering algorithms in r using the dtwclust package. *R package vignette*, 12.
- Sembiring, H. (2008). Kebijakan penelitian dan rangkuman hasil penelitian BB Padi dalam mendukung peningkatan produksi beras nasional. *Prosiding Seminar Apresiasi Hasil Penelitian Padi Menunjang P2BN*. Balai Besar Penelitian Tanaman Padi (pp. 39-59).
- Sharma S. (1996). *Applied Multivariate Techniques*. New York (US): John Willey & Sons Inc.
- Sudaryanto, T., Rusastra, I.W. (2006). Kebijakan Strategis Usaha Pertanian Dalam Rangka Peningkatan Produksi Dan Pengentasan Kemiskinan. *Litbang Pertanian* 25(4): 115–22.
- Wang, X., Smith, K., & Hyndman, R. (2006). Characteristic-based clustering for time series data. *Data mining and knowledge Discovery*, 13(3), 335-364.