Pengelompokkan Data Runtun Waktu menggunakan Analisis *Cluster* (Studi Kasus: Nilai Ekspor Komoditi Migas dan Nonmigas Provinsi Kalimantan Timur Periode Januari 2000-Desember 2016)

Grouping of Time Series Data using Cluster Analysis (*Case Study: Export Value of Oil and Non-oilCommodities in East Kalimantan Provinces Period January 2000-December 2016*)

Andrea Tri Rian Dani[1], Sri Wahyuningsih[2], dan Nanda Arista Rizki[3]
[1]Laboratorium Statistika Ekonomi dan Bisnis FMIPA Universitas Mulawarman
[2]Laboratorium Statistika Terapan FMIPA Universitas Mulawarman
[3]Laboratorium Statistika Komputasi FMIPA Universitas Mulawarman
[1]E-mail: andrikadoko@gmail.com

Abstract

The export value of East Kalimantan Province has big data conditions with time series and multivariable data types. Cluster analysis can be applied to time series data, where there are different procedures and grouping algorithms compared to grouping cross section data. Algorithms and procedures in the cluster formation process are done differently, because time series data is a series of observational data that occur based on a time index in sequence with a fixed time interval. The purpose of this research is to obtain the best similarity measurement using the cophenetic correlation coefficient and get the optimal c-value using the silhouete coefficient. In this study, the grouping algorithm used is a single linkage with four measurements of similarity, namely the Pearson correlation distance, euclidean, dynamic time warping and autocorrelation based distance. The sample in this study is the data on the export value of oil and non-oil commodities in East Kalimantan Province from January 2000 to December 2016 consisting of 10 variables. Based on the results of the analysis, the distance of the best similarity measurement in clustering the export value of oil and non-oil commodities in East Kalimantan Province is the dynamic time warping distance with the optimal c-value of 3 clusters.

Keywords*: Cluster,copheneticcorrelation coefficient, silhouette coefficient, time series*

Pendahuluan

Dengan kemajuan teknologi informasi dewasa ini, kebutuhan akan informasi yang akurat sangat dibutuhkan dalam kehidupan sehari-hari, sehingga informasi akan menjadi elemen penting dalam perkembangan masyarakat sekarang ini dan waktu mendatang. Namun kenyataannya kebutuhan informasi yang tinggi terkadang tidak diimbangi dengan penyajian informasi yang memadai, sering kali informasi tersebut hanya menjadi bongkahan informasi yang terus menerus menumpuk dan jumlahnya sangat besar. Menurut Haryati, dkk. (2015), pertumbuhan yang sangat pesat dari akumulasi data telah menciptakan kondisi kaya akan data tetapi minim informasi. *Data mining* muncul didasarkan pada kenyataan bahwa jumlah data yang tersimpan dalam basis data semakin besar, sehingga mendorong penerapan teknik pengolahan data dari berbagai bidang pengetahuan seperti statistika.

Prasetyo (2012) mendefinisikan *data mining* sebagai proses untuk mendapatkan informasi yang berguna dari gudang basis data yang besar. *Data mining* juga dapat diartikan sebagai proses ekstraksi informasi baru yang diambil dari data berskala besar dalam membantu proses pengambilan keputusan. Istilah *data mining* terkadang disebut juga *knowledge discovery in database* (KDD). Pekerjaan yang berkaitan dengan *data mining* dapat dibagi menjadi empat bagian, yaitu teknik pemodelan (*predictive modelling*), analisis *cluster* (*cluster analysis*),analisis asosiasi (*association analysis*) dan deteksi anomali (*anomaly detection*).

Analisis *cluster* adalah salah satu alat yang penting dalam pengolahan data statistik untuk melakukan analisis data. Analisis *cluster* merupakan seperangkat metode yang secara otomatis mengelompokkan objek ke dalam sebuah *cluster* berdasarkan kemiripannya. *Cluster* yang baik adalah *cluster* yang mempunyai homogenitas yang tinggi antar anggota dalam satu *cluster* dan heterogenitas yang tinggi antar *cluster* yang satu dengan *cluster* yang lain (Prasetyo, 2012).

Analisis *cluster* dapat diterapkan pada data runtun waktu, dimana terdapat prosedur dan algoritma pengelompokkan yang berbeda dibandingkan dengan pengelompokkan data *cross section*. Algoritma dan prosedur dalam proses pembentukan *cluster* dilakukan berbeda, karena data runtun waktu merupakan serangkaian data pengamatan yang terjadi berdasarkan indeks waktu secara berurutan dengan interval waktu yang tetap. Selama berkembangnya proses pengelompokkan pada data runtun waktu, banyak teknik yang dikembangkan di antaranya adalah

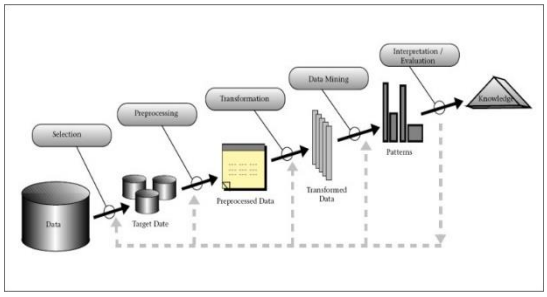
penggunaan jarak pengukuran kemiripan yang sesuai dengan karakteristik data runtun waktu. Jarak yang digunakan dalam mengukur kemiripan dua data runtun waktu pada penelitian ini adalah korelasi Pearson, *euclidean*, *dynamic time warping*dan *autocorrelation based distance*. Secara umum, penelitian terkait analisis *cluster*dapat digunakan pada sektor ekonomi di antaranya data nilai ekspor, nilai impor dan lain sebagainya.

Menurut Undang-Undang No. 17 tahun 2006 menjelaskan bahwa ekspor adalah kegiatan mengeluarkan barang dari daerah pabean. Menurut Sukirno (2010), ekspor suatu negara terjadi karena adanya manfaat yang diperoleh akibat transaksi perdagangan luar negeri. Perdagangan luar negeri dapat memperbesar kapasitas konsumsi suatu negara serta membantu berbagai usaha untuk melakukan pembangunan, meningkatkan peranan sektor yang mempunyai keunggulan kompetitif karena efisiensi dalam faktor produksi.

Berdasarkan uraian tersebut, penulis tertarik untuk membahas mengenai analisis *cluster* pada proses pengelompokkan data runtun waktu nilai ekspor komoditi migas dan nonmigas Provinsi Kalimantan Timur.

Data Mining

Data mining merupakan salah satu bidang yang berkembang pesat karena besarnya kebutuhan akan nilai tambah dari data berskala besar (*big data*). Prasetyo (2012) mendefinisikan *data mining* sebagai proses untuk mendapatkan informasi yang berguna dari gudang basis data yang besar. *Data mining* juga dapat diartikan proses ekstraksi informasi baru yang diambil dari data berskala besar yang membantu dalampengambilan keputusan. Istilah *data mining* terkadang disebut juga *knowledge discovery in database* (KDD). Menurut Mabrus dan Lubis (2012), tahapan dalam proses *data mining* ditampilkan pada Gambar 1 sebagai berikut:



Gambar 1. Tahapan *data mining*

Analisis *Cluster*

Analisis *cluster* merupakan seperangkat metode yang secara otomatis digunakan untuk mengelompokkan objek atau data ke dalam sebuah *cluster* berdasarkan kemiripan dan kedekatannya (Prasetyo, 2012). Menurut Supranto (2010), analisis *cluster* merupakan suatu teknik yang digunakan untuk mengelompokkan objek ke dalam *cluster* atau kelompok yang relatif homogen. Tujuan analisis *cluster* adalah tidak untuk menghubungkan ataupun membedakan objek yang satu dengan objek lainnya, melainkan untuk mengidentikasi sekelompok objek yang mempunyai kemiripan dan karakteristik tertentu yang dapat dipisahkan dengan kelompok lainnya. Objek yang berada dalam kelompok yang sama relatif lebih homogen daripada objek yang berada dalam kelompok yang berbeda.

Normalisasi Data

Tujuan dari analisis *cluster* adalah mengelompokkan objek-objek yang mirip dalam *cluster* yang sama. Objek dengan jarak yang lebih dekat akan lebih mirip satu sama lain dibandingkan jarak yang lebih jauh. Jika rentang nilai antar objek memiliki perbedaan skala yang cukup besar yang dapat menyebabkan bias dalam analisis *cluster*, maka data asli perlu dilakukan normalisasi. Normalisasi dapat menyingkirkan atau menghilangkan pengaruh dari unit pengukuran dan dapat memperkecil perbedaan antara kelompok atau *cluster* (Supranto, 2010).

Normalisasi data dapat dilakukan dengan cara semua dimensi atau sub-variabel penyusun ditransformasi ke dalam data standar (nilai rata-rata sama dengan nol, variansi sama dengan satu). Menurut Sartono, dkk. (2003), cara menentukan nilai normalisasi adalah dengan menghitung nilai rata-rata dan deviasi standar yaitu:

$$\overline{Z} = \left(\frac{1}{n}\right)\sum_{t=1}^{n} Z(t) \qquad (1)$$

dan

$$S_z = \sqrt{\frac{1}{n-1}\sum_{t=1}^{n}(Z(t)-\overline{Z})^2} \qquad (2)$$

Kemudian menghitung data hasil normalisasi dengan menggunakan Persamaan (3) sebagai berikut:

$$\tilde{Z}(t) = \frac{Z(t)-\overline{Z}}{S_z}, \qquad (3)$$

dengan:

$Z(t)$: data Z pada waktu ke-t

n : banyaknya data

$\overline{Z}$: rata-rata dari $Z(t)$

S_z : deviasi standar dari $Z(t)$

$\tilde{Z}(t)$: normalisasi data Z pada waktu ke-t.

Pengukuran Kemiripan

Pada dasarnya proses pembentukan*cluster* yaitu mencari dan mengelompokkan objek-objek

berdasarkan kemiripan dan kedekatan antar objek yang satu dengan objek lainnya. Langkah pertama adalah mengukur seberapa dekat kemiripan dan kedekatan antar objek tersebut. Adapun pengukuran kemiripan yang digunakan untuk mengelompokkan data runtun waktu pada penelitian ini adalah sebagai berikut:

1. Jarak korelasi Pearson

Jarak korelasi Pearson merupakan salah satu ukuran korelasi yang digunakan mengukur kekuatan dan arah hubungan linear dua data runtun waktu. Jarak korelasi Pearson memperhitungkan hubungan linear antara dua data runtun waktu yang didefinisikan sebagai berikut:

$$r_{Z,Y} = \frac{\text{cov}(Z,Y)}{S_Z S_Y} \qquad (4)$$

Sehingga diperoleh nilai korelasi Pearson sebagai berikut:

$$r_{Z,Y} = \frac{n\sum_{t=1}^{n} Z(t)Y(t) - \sum_{t=1}^{n} Z(t)\sum_{t=1}^{n} Y(t)}{\sqrt{n\sum_{t=1}^{n} Z(t)^2 - (\sum_{t=1}^{n} Z(t))^2}\sqrt{n\sum_{t=1}^{n} Y(t)^2 - (\sum_{t=1}^{n} Y(t))^2}} \qquad (5)$$

dengan:

$r_{Z,Y}$: koefisien korelasi Pearson
$Z(t)$: data Z pada waktu ke-t
$Y(t)$: data Y pada waktu ke-t
n : banyaknya data.

Ukuran jarak yang memperhitungkan korelasi antara dua data runtun waktu akan menghasilkan nilai jarak terendah untuk dua data runtun waktu yang berkorelasi positif, karena dua data runtun waktu ini memiliki kemiripan terdekat. Jarak korelasi Pearson didefinisikan sebagai berikut:

$$d_{cor}(Z,Y) = 1 - r_{Z,Y} \qquad (6)$$

Alternatif pengukuran kemiripan dengan menggunakan jarak korelasi Pearson adalah koefisien korelasi *Rank Spearman* dan koefisien korelasi *Tau Kendall* (Pereira dan Mello, 2013).

2. Jarak *euclidean*

Menurut Johnson dan Wichern (2002), jarak *euclidean* merupakan tipe pengukuran jarak dalam analisis *cluster* yang paling umum digunakan untuk mengukur jarak dari objek data ke pusat *cluster*. Jarak *euclidean* merupakan jarak geometris antar dua objek data. Semakin dekat jarak, maka semakin mirip suatu objek data tersebut dengan objek lainnya. Jarak *euclidean* dapat diperoleh dengan menggunakan Persamaan (7) sebagai berikut:

$$d_{euclid}(Z,Y) = \sqrt{\sum_{t=1}^{n} (Z(t) - Y(t))^2}, \qquad (7)$$

dengan:

$Z(t)$: data Z pada waktu ke-t
$\overline{Z}$: rata-rata dari $Z(t)$
$Y(t)$: data Y pada waktu ke-t
$\overline{Y}$: rata-rata dari $Y(t)$.

3. Jarak *dynamic time warping* (DTW)

Dynamic time warping diperkenalkan pertama kali oleh Sakoe dan Chiba pada tahun 1978. *Dynamic time warping* adalah algoritma untuk menghitung *warping path* yang optimal antara dua data runtun waktu sehingga *output*-nya adalah nilai-nilai *warping path* dan jarak di antara kedua data runtun waktu tersebut. Algoritma *dynamic time warping* dapat digunakan untuk mengukur kedekatan dua data runtun waktu dengan jumlah data yang berbeda. Misalkan terdapat dua data runtun waktu dengan panjang yang berbeda yaitu

$$Z(t) = Z(1), Z(2), Z(3), ..., Z(i), ..., Z(m)$$

dan

$$Y(t) = Y(1), Y(2), Y(3), ..., Y(j), ..., Y(n)$$

Langkah pertama adalah membuat matriks **C** yang berukuran $n \times m$. Elemen ke-(i,j) dalam matriks **C** didefinisikan sebagai selisih antara $Z(i)$ dengan $Y(j)$, kemudian ditambah dengan nilai minimum tiga elemen yang berdekatan $\{c_{(i-1)(j-1)}, c_{(i-1),j}, c_{(i)(j-1)}\}$ dengan $0 < i \le m$ dan $0 < j \le n$. Elemen ke-(i,j) dalam matriks **C** dapat ditulis menjadi

$$c_{ij} = w_{ij} + \min\{c_{(i-1)(j-1)}, c_{(i-1),j}, c_{(i)(j-1)}\} \qquad (8)$$

Dalam hal ini nilai w_{ij} merupakan selisih antara $Z(i)$ terhadap $Y(j)$ dengan perhitungan dapat dituliskan pada Persamaan (9) sebagai berikut:

$$w_{ij} = |Z(i) - Y(j)| \qquad (9)$$

Berdasarkan Persamaan (8) dan (9), maka jarak *dynamic time warping* antara dua data runtun waktu $Z(t)$ terhadap $Y(t)$ dapat didefinisikan:

$$d_{DTW}(Z,Y) = \min_{\forall w \in P}\left\{\sqrt{\sum_{i,j=1}^{K} c_{ij}}\right\}, \qquad (10)$$

di mana P adalah sekumpulan dari semua *warping path* yang mungkin, c_{ij} adalah elemen (i,j) pada *warping path* serta K adalah panjang dari *warping path* (Montero dan Vilar, 2014).

4. Jarak *autocorrelation based distance* (ABD)

Galeano dan Pena (2000) melakukan penelitian mengenai hubungan dua data runtun waktudengan menggunakan pendekatan fungsi otokorelasi (FOK). Ilustrasi untuk perhitungan jarak FOK adalah sebagai berikut, misalkan diberikan dua data runtun waktudengan ukuran n yaitu:

$$Z(t) = Z(1), Z(2), Z(3), ..., Z(n)$$

dan

$$Y(t) = Y(1), Y(2), Y(3), ..., Y(n).$$

Sehingga dapat dicari

$$\hat{\boldsymbol{\rho}}_{Z} = (\hat{\rho}_1, \hat{\rho}_2, \hat{\rho}_3 ..., \hat{\rho}_n)'$$

dan

$$\hat{\boldsymbol{\rho}}_{Y} = (\hat{\rho}_1, \hat{\rho}_2, \hat{\rho}_3 ..., \hat{\rho}_n)'$$

adalah vektor-vektor otokorelasi hasil pendugaan dari data runtun waktu $Z(t)$ dan $Y(t)$. Jarak *autocorrelation based distance*dapat dituliskan sebagai berikut:

$$d_{FOK}(Z,Y) = \sqrt{(\hat{\boldsymbol{\rho}}_{Z} - \hat{\boldsymbol{\rho}}_{Y})' \boldsymbol{\Omega} (\hat{\boldsymbol{\rho}}_{Z} - \hat{\boldsymbol{\rho}}_{Y})}\ , \qquad (11)$$

dengan $d_{FOK}(Z,Y)$ adalah jarak otokorelasi antara dua data runtun waktu $Z(t)$ terhadap $Y(t)$ sedangkan $\boldsymbol{\Omega}$ adalah matriks identitas (Riyadi, dkk. 2016).

Metode dalam Analisis *Cluster*

Metode pengklasteran merupakan prosedur yang relatif sederhana yang tidak didukung dengan suatu penalaran statistik yang ekstensif. Terdapat dua metode yang dapat digunakan untuk melakukan analisis *cluster*, yaitu metode pengelompokkan hierarki dan non-hierarki.

Metode pengelompokkan hierarki merupakan metode pengelompokkan yang berusaha untuk membangun sebuah hierarki kelompok. Strategi untuk pengelompokkan hierarki pada umumnya dibagi menjadi dua jenis yaitu *agglomerative* (pemusatan) dan *divisive* (penyebaran). Algoritma pengelompokkan yang digunakan pada penelitian ini adalah *single linkage*(pautan tunggal).

Algoritma *single linkage* (pautan tunggal) merupakan prosedur pengelompokkan yang didasarkan pada jarak minimum atau jarak terdekat antar objek. Algoritma pengelompokkan *single linkage* diawali dengan memilih jarak terkecil dalam matriks **D**, kemudian menggabungkan objek-objek yang bersesuaian misalnya U dan V untuk mendapatkan *cluster* (UV). Langkah selanjutnya adalah mencari nilai jarak antara (UV) dengan *cluster* lainnya, misalnya W sehingga dapat dituliskan sebagai berikut:

$$d_{(UV)W} = \min(d_{UW}, d_{VW})\ , \qquad (12)$$

dengan d_{UW} adalah jarak tetangga terdekat dari *cluster* U dan W serta d_{VW} adalah jarak tetangga terdekat dari *cluster* V dan W.

Uji Validitas

Adapun uji validitas yang digunakan dalam penelitian ini adalah:

1. Validitas jarak

Uji validitas jarak diperlukan untuk melihat kebaikan (*goodness*) dan kualitas (*quality*) dari hasil analisis *cluster*. Ukuran yang digunakan untuk menguji validitas jarak pengukuran kemiripan pada penelitian ini adalah koefisien korelasi *cophenetic*. Koefisien korelasi *cophenetic* merupakan koefisien korelasi antara elemen-elemen asli matriks ketidakmiripan (*dissimilarity distance*) dan elemen-elemen yang dihasilkan oleh dendogram (matriks *cophenetic*). Formulasi yang digunakan untuk menghitung koefisien korelasi *cophenetic* sebagai berikut:

$$r_{coph} = \frac{\sum_{i<j}^{n} (d_{ij} - \bar{d})(d_{coph\sim ij} - \bar{d}_{coph})}{\sqrt{\left[\sum_{i<j}^{n} (d_{ij} - \bar{d})^2\right]\left[\sum_{i<j}^{n} (d_{coph\sim ij} - \bar{d}_{coph})^2\right]}} \qquad (13)$$

dengan:

r_{coph} : koefisien korelasi *cophenetic*

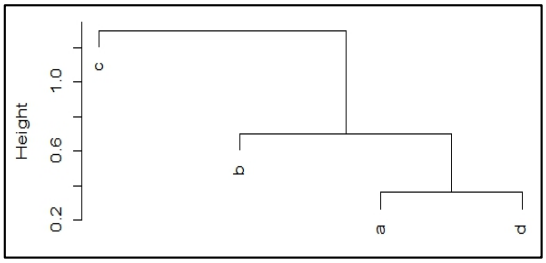
d_{ij} : jarak asli antara objek ke-*i* dan ke-*j*

$\bar{d}$: rata-rata d_{ij}

$d_{coph\sim ij}$: jarak *cophenetic* objek ke-*i* dan ke-*j*

$\bar{d}_{coph}$: rata-rata $d_{coph\sim ij}$.

Nilai koefisien korelasi *cophenetic* berkisar antara -1 hingga 1, nilai koefisien korelasi *cophenetic* yang mendekati 1 berarti jarak yang digunakan dalam proses pembentukan *cluster* cukup baik. Ilustrasi dari jarak *cophenetic* ditampilkan pada Gambar 2 sebagai berikut:



Gambar 2. Ilustrasi jarak *cophenetic*

Berdasarkan Gambar 2, untuk mencari jarak *cophenetic* dari objek a dengan d maka bisa dilihat dari tinggi dendogram di mana kedua objek pertama kali bergabung yaitu 0,36. Jarak *cophenetic* diperoleh berdasarkan tinggi dari dendogram, ketika dua objek tersebut pertama kali bergabung (Saracli, dkk., 2013).

2. Validitas *cluster*

Menurut Kaufman dan Rousseeuw (1990), salah satu metode evaluasi yang dapat digunakan untuk melihat kualitas dan kekuatan *cluster* adalah metode koefisien *silhouette*. Tahapan perhitungan mencari koefisien *silhouette* adalah:

a. Untuk setiap objek *i*, hitung rata-rata jarak dari suatu objek ke-*i* dengan semua objek pada satu *cluster* yang sama.
b. Kemudian untuk setiap objek *i*, hitung rata-rata jarak dari suatu objek ke-*i* dengan semua data yang berada pada *cluster* yang berbeda, kemudian ambil nilai yang paling kecil.
c. Selanjutnya menghitung nilai koefisien *silhouette* dengan Persamaan (14)

$$S_i = \frac{(b_i - a_i)}{\max(a_i, b_i)}, \qquad (14)$$

dengan:

S_i : nilai koefisien *silhouette*

b_i : rata-rata jarak objek ke-*i* dengan semua objek pada satu *cluster* yang berbeda

a_i : rata-rata jarak objek ke-*i* dengan semua objek pada satu *cluster* yang sama.

Hasil perhitungan nilai koefisien *silhouette* dapat bervariasi antara -1 hingga 1. Hasil *cluster* dikatakan baik jika nilai koefisien *silhouette* mendekati 1, yang berarti objek ke-*i* sudah berada dalam *cluster* yang tepat.

Ekspor

Perdagangan internasional didefinisikan sebagai perdagangan yang dilakukan suatu negara dengan negara lain atas dasar saling percaya dan saling menguntungkan. Undang-Undang Republik Indonesia Nomor 17 tahun 2006 menjelaskan bahwa ekspor adalah kegiatan mengeluarkan barang dari daerah pabean. Menurut Sukirno (2010), ekspor suatu negara terjadi karena adanya manfaat yang diperoleh akibat transaksi perdagangan luar negeri. Perdagangan dapat memperbesar kapasitas konsumsi suatu negara serta membantu berbagai usaha untuk melakukan pembangunan, meningkatkan peranan sektor yang mempunyai keunggulan komperatif karena efisiensi dalam faktor produksi.

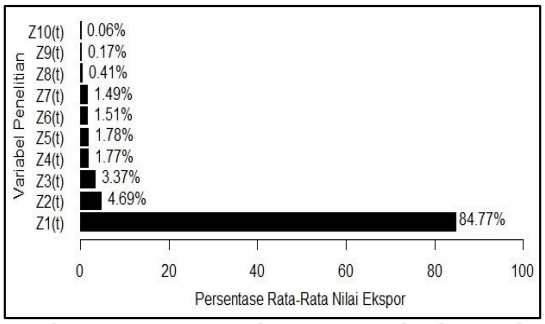
Hasil Penelitian dan Pembahasan

Variabel yang digunakan dalam penelitian ini adalah nilai ekspor komoditi migas dan nonmigas Provinsi Kalimantan Timur yang direkapitulasi oleh KPw Bank Indonesia Provinsi Kalimantan Timur, terdiri dari 10 variabel yang dinotasikan $Z_i(t)$ dengan i=1,2,…,10.

Sampel yang digunakan dalam penelitian ini adalah data nilai ekspor komoditi migas dan nonmigas Provinsi Kalimantan Timur dari bulan Januari tahun 2000 sampai dengan bulan Desember tahun 2016 sebanyak 204 data untuk setiap variabel penelitian.

1. Statistika deskriptif

Pembahasan akan diawali dengan menampilkan statistika deskriptif berupa diagram batang.



Gambar 3. Diagram batang untuk data nilai ekspor

Berdasarkan Gambar 3, dapat diketahui persentasebterbesar untuk kegiatan ekspor Provinsi Kalimantan Timur adalah komoditi migas yaitu bahan bakar dan mineral sebesar 84,77%. Komoditi-komoditi lainnya selain migas (nonmigas) menyumbang dengan persentase yang relatif lebih kecil.

2. Normalisasi data

Setiap variabel penelitian memiliki nilai ekspor yang berbeda-beda dengan selisih yang cukup berjauhan satu sama lain, sehingga perlu dilakukan normalisasi data. Normalisasi data disini bertujuan untuk membuat semua variabel penelitian berada dalam jangkauan yang sama dan memperkecil perbedaan antar variabel. Dalam normalisasi data digunakan Persamaan (1), (2) dan (3) dan diperoleh hasil sebagai berikut:

$$\mathbf{O} = \begin{array}{c} \begin{matrix} \tilde{Z}_1(t) & \tilde{Z}_2(t) & \cdots & \tilde{Z}_{10}(t) \end{matrix} \\ \begin{bmatrix} -1,16 & 1,16 & \cdots & -0,10 \\ -1,16 & 1,98 & \cdots & -0,15 \\ \vdots & \vdots & \ddots & \vdots \\ 0,69 & -1,11 & \cdots & -0,22 \end{bmatrix} \end{array}$$

Matriks **O** adalah matriks yang berisikan hasil normalisasi data nilai ekspor komoditi migas dan nonmigas Provinsi Kalimantan Timur.

3. Algoritma *single linkage*

Algoritma *single linkage* merupakan salah satu algortima yang didasarkan pada jarak minimum atau jarak terdekat antar objek. Pengukuran kemiripan yang digunakan pada penelitian ini dalam proses pengelompokkan menggunakan algoritma *single linkage* sebagai berikut:

a. Jarak korelasi Pearson

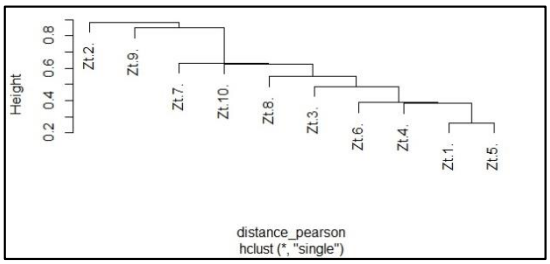
Setelah melakukan normalisasi data nilai ekspor komoditi migas dan nonmigas Provinsi Kalimantan Timur yang hasil normalisasinya disebut dengan matriks **O**. Langkah selanjutnya menghitung nilai korelasi Pearson menggunakan Persamaan (5) terhadap setiap variabel penelitian, kemudian membentuk matriks korelasi **R**.

$$\mathbf{R} = \begin{array}{c} \\ \tilde{Z}_1(t) \\ \tilde{Z}_1(t) \\ \vdots \\ \tilde{Z}_{10}(t) \end{array} \begin{array}{c} \begin{array}{cccc} \tilde{Z}_1(t) & \tilde{Z}_2(t) & \cdots & \tilde{Z}_{10}(t) \end{array} \\ \begin{bmatrix} 1,00 & 1,16 & \cdots & 0,18 \\ -0,08 & 1,00 & \cdots & 0,01 \\ \vdots & \vdots & \ddots & \vdots \\ 0,18 & 0,01 & \cdots & 1,00 \end{bmatrix} \end{array}$$

Berdasarkan matriks **R** di atas yang berisikan perhitungan nilai korelasi Pearson untuk masing-masing variabel penelitian, kemudian langkah selanjutnya menghitung jarak korelasi Pearson menggunakan Persamaan (6) dan membentuk matriks jarak **D** sebagai berikut:

$$\mathbf{D} = \begin{array}{c} \\ \tilde{Z}_1(t) \\ \tilde{Z}_1(t) \\ \vdots \\ \tilde{Z}_{10}(t) \end{array} \begin{array}{c} \begin{array}{cccc} \tilde{Z}_1(t) & \tilde{Z}_2(t) & \cdots & \tilde{Z}_{10}(t) \end{array} \\ \begin{bmatrix} 0,00 & 1,08 & \cdots & 0,81 \\ 1,08 & 0,00 & \cdots & 0,97 \\ \vdots & \vdots & \ddots & \vdots \\ 0,81 & 0,97 & \cdots & 0,00 \end{bmatrix} \end{array}$$

Langkah selanjutnya setelah melakukan perhitungan jarak korelasi Pearson adalah melakukan proses pengelompokkan menggunakan algoritma *single linkage*. Setiap variabel penelitian dimulai sebagai *cluster*, sehingga awalnya pada penelitian ini terdapat 10 *cluster*yang terbentuk. Algoritma pengelompokkan *single linkage* dimulai dengan memilih jarak terkecil dalam matriks **D**. Jarak variabel $\tilde{Z}_1(t)$ dengan variabel $\tilde{Z}_5(t)$ adalah sebesar 0,26, artinya bahwa kedua variabel tersebut yang pertama kali bergabung membentuk satu *cluster*. Langkah selanjutnya adalah menggabungkan variabel $\tilde{Z}_1(t)$ dengan variabel $\tilde{Z}_5(t)$ ke dalam satu *cluster*, kemudian mencari nilai jarak minimum terhadap setiap variabel lainya berdasarkan Persamaan (12). Proses pengelompokkan akan terus berlanjut sampai dengan hanya tersisa 2 *cluster* dan diperoleh dendogram sebagai berikut:



Gambar 4. Dendogram jarak korelasi Pearson

Berdasarkan Gambar 4, diketahui bahwa variabel yang pertama kali bergabung adalah $\tilde{Z}_1(t)$ yaitu bahan bakar dan mineral dengan variabel $\tilde{Z}_5(t)$ yaitu lemak hewani dan nabati, kemudian variabel selanjutnya yang bergabung adalah $\tilde{Z}_4(t)$ yaitu pupuk.Proses pengelompokkan terus berlanjut sampai dengan variabel terakhir yang bergabung adalah $\tilde{Z}_2(t)$ yaitu kayu dan kerajinan kayu.

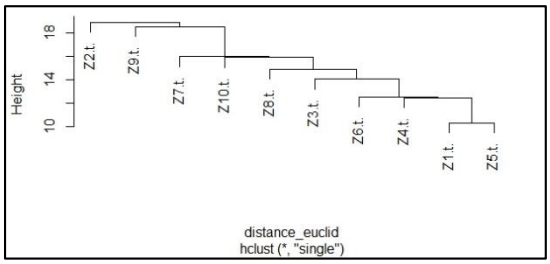
b. Jarak *euclidean*

Jarak pengukuran kemiripan selanjutnya yang digunakan adalah jarak *euclidean* dengan perhitungan menggunakan Persamaan (7) dan dilakukan terhadap setiap variabel penelitian, kemudian membentuk matriks jarak **D**sebagai berikut:

$$\mathbf{D} = \begin{array}{c} \\ \tilde{Z}_1(t) \\ \tilde{Z}_1(t) \\ \vdots \\ \tilde{Z}_{10}(t) \end{array} \begin{array}{c} \begin{array}{cccc} \tilde{Z}_1(t) & \tilde{Z}_2(t) & \cdots & \tilde{Z}_{10}(t) \end{array} \\ \begin{bmatrix} 0,00 & 20,90 & \cdots & 18,20 \\ 20,90 & 0,00 & \cdots & 20,03 \\ \vdots & \vdots & \ddots & \vdots \\ 18,20 & 20,03 & \cdots & 0,00 \end{bmatrix} \end{array}$$

Langkah selanjutnya setelah melakukan perhitungan jarak *euclidean* adalah melakukan proses pengelompokkan menggunakan algoritma *single linkage*. Setiap variabel penelitian dimulai sebagai *cluster*, sehingga awalnya pada penelitian ini terdapat 10 *cluster*yang terbentuk. Algoritma pengelompokkan *single linkage* dimulai dengan memilih jarak terkecil dalam matriks **D**. Jarak variabel $\tilde{Z}_1(t)$ dengan variabel $\tilde{Z}_5(t)$ adalah sebesar 10,34, artinya bahwa kedua variabel tersebut yang pertama kali bergabung membentuk satu *cluster*.Langkah selanjutnya adalah menggabungkan variabel $\tilde{Z}_1(t)$ dengan variabel $\tilde{Z}_5(t)$ ke dalam satu *cluster*, kemudian mencari nilai jarak minimum terhadap setiap variabel lainya berdasarkan Persamaan (12). Proses

pengelompokkan akan terus berlanjut sampai dengan hanya tersisa 2 *cluster* dan diperoleh dendogram sebagai berikut:



Gambar 5. Dendogram jarak *euclidean*

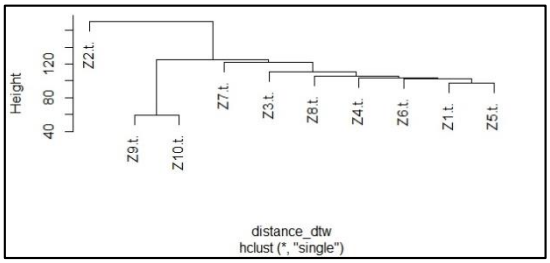
Berdasarkan Gambar 5, diketahui bahwa variabel yang pertama kali bergabung adalah $\tilde{Z}_1(t)$ yaitu bahan bakar dan mineral dengan variabel $\tilde{Z}_5(t)$ yaitu lemak hewani dan nabati, kemudian variabel selanjutnya yang bergabung adalah $\tilde{Z}_4(t)$ yaitu pupuk. Proses pengelompokkan terus berlanjut sampai dengan variabel terakhir yang bergabung adalah $\tilde{Z}_2(t)$ yaitu kayu dan kerajinan kayu. Dendogram yang dihasilkan menggunakan jarak *euclidean* mempunyai hasil pengelompokkan yang sama dengan dendogram yang dihasilkan menggunakan jarak korelasi Pearson.

c. Jarak *dynamic time warping* (DTW)

Jarak pengukuran kemiripan selanjutnya yang digunakan adalah jarak *dynamic time warping* (DTW) dengan perhitungan menggunakan Persamaan (8), (9) dan (10) dan dilakukan terhadap setiap variabel penelitian, kemudian membentuk matriks jarak **D** sebagai berikut:

$$\mathbf{D} = \begin{array}{c} \\ \tilde{Z}_1(t) \\ \tilde{Z}_1(t) \\ \vdots \\ \tilde{Z}_{10}(t) \end{array} \begin{array}{c} \begin{array}{cccc} \tilde{Z}_1(t) & \tilde{Z}_2(t) & \cdots & \tilde{Z}_{10}(t) \end{array} \\ \begin{bmatrix} 0,00 & 195,13 & \cdots & 147,42 \\ 195,13 & 0,00 & \cdots & 170,59 \\ \vdots & \vdots & \ddots & \vdots \\ 147,72 & 170,59 & \cdots & 0,00 \end{bmatrix} \end{array}$$

Langkah selanjutnya setelah melakukan perhitungan jarak DTW adalah melakukan proses pengelompokkan menggunakan algoritma *single linkage*. Setiap variabel penelitian dimulai sebagai *cluster*, sehingga awalnya pada penelitian ini terdapat 10 *cluster*yang terbentuk. Algoritma pengelompokkan *single linkage* dimulai dengan memilih jarak terkecil dalam matriks **D**. Jarak variabel $\tilde{Z}_9(t)$ dengan variabel $\tilde{Z}_{10}(t)$ adalah sebesar 59,38, artinya bahwa kedua variabel tersebut yang pertama kali bergabung membentuk satu *cluster*. Langkah selanjutnya adalah menggabungkan variabel $\tilde{Z}_9(t)$ dengan variabel $\tilde{Z}_{10}(t)$ ke dalam satu *cluster*, kemudian mencari nilai jarak minimum terhadap setiap variabel lainya berdasarkan Persamaan (12). Proses pengelompokkan akan terus berlanjut sampai dengan hanya tersisa 2 *cluster* dan diperoleh dendogram sebagai berikut:



Gambar 6. Dendogram jarak DTW

Berdasarkan Gambar 6, diketahui bahwa variabel yang pertama kali bergabung adalah $\tilde{Z}_9(t)$ yaitu kendaraan selain kereta api dengan variabel $\tilde{Z}_{10}(t)$ yaitu barang dari besi dan baja, kemudian variabel selanjutnya yang bergabung adalah $\tilde{Z}_1(t)$ yaitu bahan bakar dan mineral dengan $\tilde{Z}_5(t)$ yaitu lemak hewani dan nabati. Proses pengelompokkan terus berlanjut sampai dengan variabel terakhir yang bergabung adalah $\tilde{Z}_2(t)$ yaitu kayu dan kerajinan kayu.

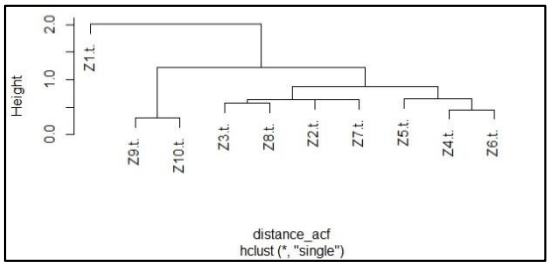
d. Jarak *autocorrelation based distance* (ABD)

Jarak pengukuran kemiripan yang terakhir digunakan pada penelitian ini adalah jarak *autocorrelation based distance* (ABD) dengan perhitungan menggunakan Persamaan (11) dan dilakukan terhadap setiap variabel penelitian, kemudian membentuk matriks jarak **D** sebagai berikut:

$$\mathbf{D} = \begin{array}{c} \\ \tilde{Z}_1(t) \\ \tilde{Z}_1(t) \\ \vdots \\ \tilde{Z}_{10}(t) \end{array} \begin{array}{c} \begin{array}{cccc} \tilde{Z}_1(t) & \tilde{Z}_2(t) & \cdots & \tilde{Z}_{10}(t) \end{array} \\ \begin{bmatrix} 0,00 & 3,61 & \cdots & 4,73 \\ 3,61 & 0,00 & \cdots & 1,43 \\ \vdots & \vdots & \ddots & \vdots \\ 4,73 & 1,43 & \cdots & 0,00 \end{bmatrix} \end{array}$$

Langkah selanjutnya setelah melakukan perhitungan jarak ABD adalah melakukan proses pengelompokkan menggunakan algoritma *single linkage*. Setiap variabel penelitian dimulai sebagai *cluster*, sehingga awalnya pada penelitian ini terdapat 10 *cluster*yang terbentuk. Algoritma pengelompokkan *single linkage* dimulai dengan memilih jarak terkecil dalam matriks **D**. Jarak variabel $\tilde{Z}_9(t)$ dengan variabel $\tilde{Z}_{10}(t)$ adalah sebesar 0,31, artinya bahwa kedua variabel tersebut yang pertama kali bergabung membentuk satu *cluster*.Langkah selanjutnya adalah

menggabungkan variabel $\tilde{Z}_9(t)$ dengan variabel $\tilde{Z}_{10}(t)$ ke dalam satu *cluster*, kemudian mencari nilai jarak minimum terhadap setiap variabel lainya berdasarkan Persamaan (12).Proses pengelompokkan akan terus berlanjut sampai dengan hanya tersisa 2 *cluster* dan diperoleh dendogram sebagai berikut:



Gambar 7. Dendogram jarak ABD

Berdasarkan Gambar 7, diketahui bahwa variabel yang pertama kali bergabung adalah $\tilde{Z}_9(t)$ yaitu kendaraan selain kereta api dengan variabel $\tilde{Z}_{10}(t)$ yaitu barang dari besi dan baja, kemudian variabel selanjutnya yang bergabung adalah $\tilde{Z}_4(t)$ yaitu pupuk dengan $\tilde{Z}_6(t)$ yaitu perikanan. Proses pengelompokkan terus berlanjut sampai dengan variabel terakhir yang bergabung adalah $\tilde{Z}_1(t)$ yaitu bahan bakar dan mineral.

4. Uji Validitas

Pengujian validitas pada penelitian ini bertujuan untuk menghasilkan proses *clustering* yang optimal, artinya proses pembentukan *cluster* dengan algoritma *single linkage* didasarkan padajarak pengukuran kemiripan serta jumlah *cluster* yang optimal.

a. Validitas jarak

Setelah melakukan pengukuran kemiripan untuk masing-masing variabel penelitian, tahapan selanjutnya adalah melakukan pemilihan jarak pengukuran kemiripan terbaik pada data nilai ekspor komoditi migas dan nonmigas Provinsi Kalimantan Timur. Uji validitas jarak yang digunakan dalam penelitian ini adalah koefisien korelasi *cophenetic* berdasarkan Persamaan (13).

Tabel 1. Nilai Koefisien Korelasi *Cophenetic*

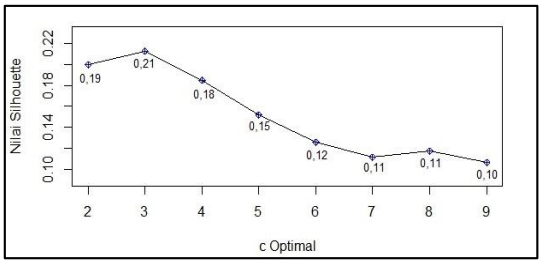
Jarak Pengukuran	Koefisien Korelasi *Cophenetic*
Korelasi Pearson	0,83
Euclidean	0,85
Dynamic Time Warping	0,92
Autocorrelation Based Distance	0,87

Nilai dari koefisien korelasi *cophenetic* berkisar antara -1 sampai dengan 1, yang artinya ketika nilai koefisien korelasi mendekati 1 berarti jarak yang digunakan dalam proses *clustering* cukup baik. Berdasarkan Tabel 1, dapat diketahui bahwa jarak pengukuran kemiripan terbaik dalam proses pengelompokkan nilai ekspor komoditi migas dan nonmigas Provinsi Kalimantan Timur adalah jarak DTW dengan nilai koefisien korelasi *cophenetic* terbesar yaitu 0,92. Jarak DTWini nantinya akan digunakan dalam proses analisis selanjutnya untuk menentukan nilai *c*-optimal dalam proses *clustering* algoritma *single linkage*.

b. Validitas *cluster*

Setelah mendapatkan jarak pengukuran kemiripan terbaik yaitu jarak DTW, langkah selanjutnya adalah menentukan nilai *c*-optimal dalam proses *clustering* algoritma *single linkage*. Uji validitas *cluster* yang digunakan pada penelitian ini adalah metode koefisien *silhouette* berdasarkan Persamaan (14).

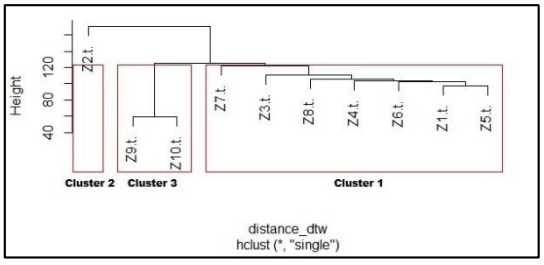
Nilai koefisien *silhouette* dapat bervariasi antara -1 hingga 1. Jumlah *cluster* dikatakan optimal jika nilai koefisien *silhouette* mendekati 1. Berdasarkan Gambar 8, dapat diketahui bahwa nilai *c*-optimal dalam mengelompokkan nilai ekspor komoditi migas dan nonmigas Provinsi Kalimantan Timur adalah 3 *cluster* dengan nilai koefisien *silhouette* terbesar yaitu 0,21.



Gambar 8. Diagram garis nilai koefisien *silhouette*

5. Profilisasi dan Interpretasi Hasil *Cluster*

Setelah mendapatkan proses *clustering* yang optimal. Langkah selanjutnya adalah melakukan profilisasi dan interpretasi hasil *cluster*. Pada proses pengelompokkan nilai ekspor komoditi migas dan nonmigas Provinsi Kalimantan Timur menggunakan analisis *cluster*, jarak pengukuran kemiripan yang digunakan adalah *dynamic time warping*(DTW)dengan nilai *c*-optimal yaitu 3 *cluster* sehingga diperoleh dendogram hasil pengelompokkan yang ditampilkan pada Gambar 9.



Gambar 9. Dendogram jarak DTW dengan 3 *cluster*

Berdasarkan Gambar 9, dapat diketahui hasil pengelompokkan nilai ekspor komoditi migas dan nonmigas Provinsi Kalimantan Timur. Pada *cluster* 1 terdapat 7 variabel penelitian yang bergabung diantaranya $\tilde{Z}_1(t)$ yaitu bahan bakar dan mineral, $\tilde{Z}_5(t)$ yaitu lemak hewani dan nabati, $\tilde{Z}_6(t)$ yaitu perikanan, $\tilde{Z}_4(t)$ yaitu pupuk, $\tilde{Z}_8(t)$ yaitu bahan reaksi nuklir, $\tilde{Z}_3(t)$ yaitu bahan-bahan kimia anorganik serta $\tilde{Z}_7(t)$ yaitu bahan-bahan kimia organik. Pada *cluster* 2 hanya terdapat 1 variabel penelitian yang bergabung yaitu $\tilde{Z}_2(t)$ atau kayu dan kerajinan kayu, sedangkan pada *cluster* 3 terdapat 2 variabel penelitian yang bergabung diantaranya $\tilde{Z}_9(t)$ yaitu kendaraan selain kereta api serta $\tilde{Z}_{10}(t)$ yaitu barang dari besi dan baja.

Kesimpulan

Berdasarkan hasil penelitian dan pembahasan, maka kesimpulan yang diperoleh adalah sebagai berikut:

1. Pengukuran kemiripan terbaik dalam proses pembentukan *cluster* data nilai ekspor komoditi migas dan nonmigas Provinsi Kalimantan Timur menggunakan algoritma *single linkage* adalah jarak *dynamic time warping* (DTW) dengan nilai koefisien korelasi *cophenetic* sebesar 0,92.
2. Nilai *c*-optimal dalam proses pembentukan *cluster* data nilai ekspor komoditi migas dan nonmigas Provinsi Kalimantan Timur menggunakan algoritma *single linkage* dengan jarak *dynamic time warping* adalah 3 *cluster*. Nilai koefisien *silhouette* yang diperoleh sebesar 0,21.

Daftar Pustaka

Haryati, S., Sudarsono, A., dan Suryana, E. (2015). Implementasi Data Mining untuk Memprediksi Masa Studi Menggunakan Algoritma C4.5. *Jurnal Media Infotama*. 11 (2), 130-138.

Johnson, R. A. dan Wichern, D. W. (2002). *Applied Multivariate StatisticalAnalysis, Fifth Edition*. New Jersey: Pearson Prentice Inc.

Kaufman, L. dan Rousseeuw, P. J. (1990). *Finding Groups in Data An Introduction to Cluster Analysis*. New Jersey: John Wiley & Sons Inc Publication.

Mabrur, A. G. dan Lubis, R. (2012). Penerapan Data Mining untuk Memprediksi Kriteria Nasabah Kredit. *Jurnal Komputer dan Informatika*. 1 (1), 53-57.

Montero, P. dan Vilar, J. A. (2014). TSclust: An R Package for Time Series Clustering. *Journal of Statistical Software*. 62 (1), 01-43.

Pereira, C. M. M dan Mello, R. F. (2013). Common Dissimilarity Measures are Inappropriate for Time Series Clustering. *Revista de Informatica Teorica e Aplicada* (*RITA*). 20 (1), 25-48.

Prasetyo, E. (2012). *Data Mining:Konsep dan Aplikasi Menggunakan MATLAB*. Yogyakarta: Penerbit Andi.

Riyadi, M. A. A., Fithriasari, K. dan Dwiatmono. (2016). Data Mining Peramalan Konsumsi Listrik dengan Pendekatan Cluster Time Series sebagai Preprocessing. *Jurnal Sains dan Seni ITS*. 5 (1), 121-126.

Saracli, S., Dogan, N. dan Dogan, I. (2013). Comparison of Hierarchical Cluster Analysis Methods by Cophenetic Correlation. *Journal of Inequalities and Applications*. doi: 10.1186/1029-242X-2013-203.

Sartono, B., Affendi, F. M., Sumertajaya, I. M. dan Angraeni, Y. (2003). *Analisis Peubah Ganda*. Bogor: Fakultas Matematika dan Ilmu Pengetahuan Alam IPB.

Sukirno, S. (2010). *Makroekonomi: Teori Pengantar*. Jakarta: Rajawali Pers.

Supranto, J. (2010). *Statistik: Teori dan Aplikasi Edisi 8*. Jakarta: Erlangga.