Jurnal Statistika dan Aplikasinya Volume 7 Issue 2, December 2023 e-ISSN: 2620-8369

Received: 27 October 2023
Revised: 27 December 2023
Accepted: 30 December 2023
Published: 31 December 2023

Evaluasi Perbandingan Kinerja Algoritma *Cheng and Church Biclustering* Terhadap Algoritma *Clustering* Klasik *K-Means* untuk Mengidentifikasi Pola Distribusi Barang Ekspor Indonesia

Seta Baehera[1, a)], Utami Dyah Syafitri[1, b)], Agus Mohamad Soleh[1, c)]

[1]*Program Studi Statistika dan Sains Data, Institut Pertanian Bogor.*

Email: [a)]setabaehera@apps.ipb.ac.id, [b)]utamids@apps.ipb.ac.id, [c)]agusms@apps.ipb.ac.id

Abstract

Clustering is a process of grouping data into several groups (clusters) so that data in one cluster has a homogeneous level of similarity and data between clusters has heterogeneous similarity. A common example of a clustering algorithm is K-Means Clustering. Compared with classical clustering algorithms, the biclustering algorithm is a two-dimensional data grouping process. The biclustering algorithm functions to find data submatrices, namely row subgroups and column subgroups that have high correlation. One example of a biclustering algorithm is Cheng and Church Biclustering (CC Biclustering). The aim of this research is to evaluate the performance of the biclustering algorithm against classical clustering algorithms. Analysis applied to CC Biclustering and K-Means Clustering to identify distribution patterns of Indonesian export goods in the period 2013 to 2022. Based on research results, the optimal scenario for the K-Means algorithm is scenario 2, that is the application of the 7 cluster K-Means algorithm with pre- processing data scaling. Meanwhile, the optimal scenario for the CC Biclustering algorithm is scenario 1, that is the application of the CC Biclustering algorithm with a tolerance value of 0.10 with data scaling pre-processing. However, from these two scenarios, based on the MSR/Volume value, it was concluded that the best scenario is scenario 1 in the application of the CC Biclustering algorithm which has an MSR/Volume value of 0.077.

Keywords: Clustering Analysis, K-Means Clustering, Cheng and Church Biclustering

Abstrak

Clustering merupakan suatu proses pengelompokan data menjadi beberapa kelompok (*cluster*) hingga data dalam satu *cluster* mempunyai tingkat kemiripan yang homogen dan data antar *cluster* mempunyai kemiripan yang heterogen. Contoh umum dari algoritma *clustering* adalah *K-Means Clustering*. Dibandingkan dengan algoritma *clustering* klasik, algoritma *biclustering* merupakan proses pengelompokan data dua dimensi. Algoritma *biclustering* berfungsi mencari submatriks data yaitu subgrup baris dan subgrup kolom yang mempunyai korelasi tinggi. Salah satu contoh algoritma *biclustering* adalah *Cheng and Church Biclustering* (*CC Biclustering*).

Tujuan dari penelitian ini adalah untuk mengevaluasi kinerja algoritma *biclustering* terhadap algoritma *clustering* klasik. Analisis yang diterapkan pada *CC Biclustering* dan *K-Means Clustering* untuk mengidentifikasi pola distribusi barang ekspor Indonesia pada periode 2013 hingga 2022. Berdasarkan hasil penelitian, skenario optimal pada algoritma *K-Means* adalah skenario 2, yaitu penerapan algoritma *K-Means* 7 *cluster* dengan pra-prosesing *data scaling.* Sedangkan skenario optimal pada algoritma *CC Biclustering* adalah skenario 1, yaitu penerapan algoritma *CC Biclustering* dengan nilai toleransi 0,10 dengan pra-prosesing *data scaling.* Namun dari dua skenario tersebut berdasarkan nilai MSR/Volume didapatkan kesimpulan bahwa skenario terbaik adalah skenario 1 pada penerapan algoritma *CC Biclustering* yang memiliki nilai MSR/Volume sebesar 0,077.

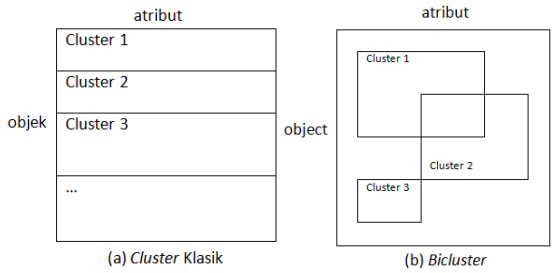
Kata-kata kunci: Analisis *Cluster, K-Means Clustering, Cheng and Church Biclustering*

PENDAHULUAN

Clustering adalah salah satu algoritma *unsupervised learning*, yaitu merupakan jenis *learning* yang hanya mempunyai variabel input tapi tidak mempunyai variabel output yang berhubungan. Menurut Tan (2006), *clustering* adalah sebuah proses untuk mengelompokan data ke dalam beberapa kelompok (*cluster*) sehingga data dalam satu *cluster* memiliki tingkat kemiripan yang homogen dan data antar *cluster* memiliki kemiripan yang heterogen. Salah satu algoritma *clustering* yang umum dipakai sekarang ini adalah *K-Means Clustering.*

Sedikit berbeda dengan analisis *clustering,* analisis *biclustering* merupakan pengelompokan untuk data dua arah atau dimensi. *Biclustering* pada umumnya juga dikenal dengan istilah *simultaneous clustering*, *two-mode clustering*, *co-clustering*, *two-way clustering* atau *block clustering* (Madeira & Oliveira, 2004). *Biclustering* memiliki kemampuan untuk menemukan himpunan bagian yang bermakna dari subjek dan variabel secara bersamaan, yang mungkin tidak terdeteksi oleh algoritma *clustering* klasik. Salah satu metode analisis yang termasuk ke dalam *biclustering* yaitu algoritma Cheng dan Church (CC) atau yang biasa disebut δ-*biclustering* (Cheng & Church, 2000). Algoritma δ-*biclustering* adalah sebuah *iterative greedy search algorithm* yang mencoba untuk menemukan *bicluster* maksimal dengan kesamaan yang tinggi (Nurmawiya dan Kurniawan, 2020).

Algoritma *cluster* klasik menemukan kelompok objek homogen atau kelompok atribut yang juga homogen, seperti pada GAMBAR 1(a). Tetapi, algoritma *bicluster* adalah algoritma untuk menemukan grup yang berisi sebagian objek dan atribut secara bersamaan, seperti pada GAMBAR 1(b). Sementara algoritma *cluster* klasik mengasumsikan bahwa objek dalam *cluster* yang sama adalah serupa di seluruh atribut, algoritma *bicluster* menganggap objek yang serupa hanya dalam subset atribut sebagai anggota *bicluster* meskipun mereka berbeda dalam atribut lainnya.



GAMBAR 1. Algoritma *Clustering* Klasik dan *Biclustering*

Penelitian ini merujuk pada beberapa penelitian yang telah dilakukan sebelumnya. Seperti penelitian yang telah dilakukan oleh R.Novidianto dan R.Irfani dengan judul "*Bicluster CC Algorithm Analysis to Identify Pattern of Food Insecurity in Indonesia*" pada tahun 2020. Penelitiannya bertujuan untuk mengetahui pola kerawanan pangan di setiap Provinsi di Indonesia secara spasial dengan menggunakan algoritma *bicluster*. Di tahun yang sama Nurmawiya dan Robert Kurniawan melakukan penelitian untuk mengelompokan wilayah kabupaten/kota di Indonesia dengan menggunakan variabel

indikator kesiapan masyarakat yang terdapat dalam *Networked Readiness Index* (NRI) untuk mengetahui tingkat kesiapan masyarakat Indonesia dalam menghadapi Revolusi Industri 4.0. Penelitian ini dipublikasi pada seminar nasional dengan judul "Pengelompokan Wilayah Indonesia Dalam Menghadapi Revolusi Industri 4.0 dengan Metode *Biclustering*". Pada tahun 2021, dilakukan penelitian oleh C.A. Putri, R. Irfani dan B. Sartono dengan judul "*Recognizing Poverty Pattern in Central Java using Biclustering Analysis*". Tujuan penelitiannya adalah untuk mengetahui pola kemiskinan di Provinsi Jawa Tengah menurut wilayah dan variabel dimensi kemiskinan secara simultan menggunakan analisis *biclustering*. Pada tahun 2022 Wiwik Andriyani L.N., I.M.Sumertajaya dan A.Saefuddin juga melakukan penelitian dengan menggunakan Algoritma *Biclustering* untuk mendapatkan gambaran tentang pola spasial dan karakteristik kerentanan ekonomi dan pandemi COVID-19 di Indonesia. Penelitiannya diberi judul "*Biclustering Application in Indonesian Economic and Pandemic Vulnerability*".

Tujuan dilakukannya penelitian ini adalah untuk melakukan evaluasi terhadap sebuah pemodelan *bicluster*, yaitu algoritma Cheng dan Church (CC) atau biasa disebut juga δ-*biclustering* serta *K-Means clustering* sebagai perbandingan dari algoritma *clustering* klasik. Seluruh algoritma pengelompokan tersebut akan diterapkan pada data ekspor Indonesia pada kurun waktu 2013 hingga 2022. Berkaitan dengan karakteristik data ekspor Indonesia dimana nilai ekspor memiliki variasi antar negara sehingga perlu dilakukan pra-prosesing data sebelum dilakukan analisis *cluster*. Pra-prosesing data tersebut dilakukan dengan menggunakan metode *data scaling* dan juga *Principal Component Analysis* (PCA). Proses evaluasi terhadap kualitas *clusters* yang telah dibangun diukur dengan menggunakan Rasio Rata-rata *Bicluster* (Chakraborty & Maka, 2005). Sedangkan proses evaluasi terhadap kualitas *bicluster* diukur dengan menggunakan Rataan MSR per volume dan indeks Liu dan Wang (Liu dan Wang, 2007).

METODOLOGI

Sumber Data dan Peubah Penelitian

Data ekspor yang digunakan pada penelitian ini berisi data kode golongan barang (kode HS) beserta negara tujuan ekspornya selama kurun waktu tahun 2013 hingga tahun 2022. Negara tujuan ekspor akan menjadi unit pengamatan, sedangkan kode HS akan menjadi peubah dalam penelitian ini.

Unit pengamatan penelitian dalam hal ini adalah negara tujuan ekspor, yaitu negara-negara yang melakukan kegiatan perdagangan dengan Indonesia. Diantara 260 negara di dunia, tercatat dalam data ekspor bahwa selama kurun waktu tahun 2013 hingga tahun 2022 hanya sebanyak 238 negara didunia yang melakukan kerjasama perdagangan dengan Indonesia. Data negara tujuan ini berisi kode negara tujuan ekspor dengan kode 2 dijit huruf.

Data kode HS yang menjadi peubah dalam penelitian merupakan kode golongan barang dengan 2 dijit angka. Kode HS 2 dijit memiliki nilai kategorik sebanyak 98 kode HS. Kode HS berisi nilai kode berurutan dari 01 sampai dengan kode 99, kecuali kode HS 77 yang merupakan kode golongan barang yang dicadangkan untuk kemungkinan penggunaan di masa depan (*reserved for possible future use*).

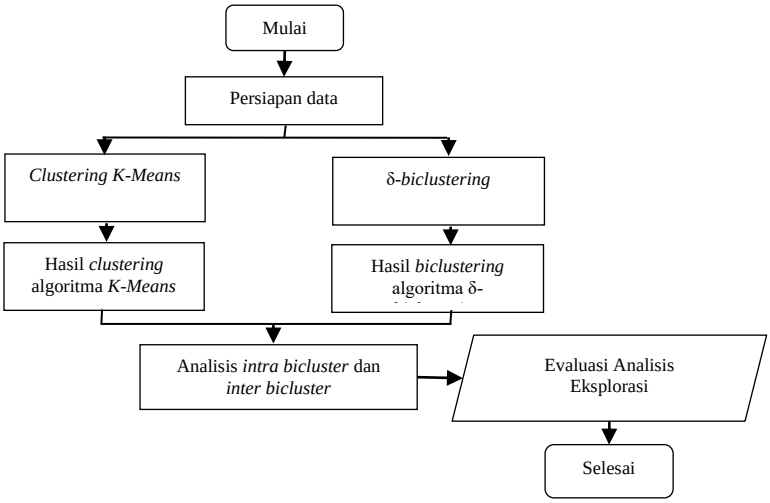
Data ekspor tersebut akan dibentuk sedemikian rupa hingga membentuk sebuah matriks data. Nilai data yang berada di dalam matriks merupakan nilai komoditas barang yang diekspor ke negara tujuan atau disebut juga dengan nilai FOB (*Freight on Board*). Nilai FOB tersebut dalam satuan US dollar (USD).

Metode Penelitian

Langkah Analisis

Dalam melakukan penelitian ini penulis menggunakan aplikasi R *for Windows 64 bit*. Berikut tiga tahapan umum yang akan dijalankan yaitu Persiapan, *Clustering* dan *Bicluster*, serta Evaluasi serta pemilihan skenario terbaik (GAMBAR 2).

Dalam tahap persiapan data hal pertama yang akan dilakukan adalah melakukan praproses data berupa normalisasi data. Normalisasi data dilakukan dengan metode *data scaling* dan PCA (*Principal Component Analysis*). Selanjutnya membentuk dataset menjadi matriks data untuk dilakukan proses *clustering* dan *biclustering.*



GAMBAR 2. Diagram Alir Langkah Analisis

Data ekspor yang telah diubah menjadi bentuk matriks data selanjutnya akan dilakukan normalisasi data. Tujuan normalisasi data adalah untuk memastikan setiap unit pengamatan pada matriks data tetap konsisten. Pada penelitian ini proses normalisasi data menggunakan metode *data scaling* dan juga *Principal Component Analysis* (PCA).

Tahapan normalisasi ini terbagi menjadi 3 jenis, yaitu:

1. Matriks data ekspor yang telah dilakukan proses *data scaling*.
2. Matriks data ekspor yang telah dilakukan proses PCA.
3. Matriks data ekspor yang telah dilakukan proses PCA dengan reduksi banyaknya komponen utama sebanyak 10% (PCA 90%).

Proses *biclustering* pada penelitian ini menggunakan algoritma δ-*biclustering* (CC *algorithm*). Sedangkan sebagai pembanding dari algoritma *clustering* klasik akan digunakan *K-Means clustering.* Pada tahap evaluasi hasil pembentukan *cluster* dan *bicluster* dilakukan dengan metode *intra-bicluster* dan *inter-bicluster*.

Principal Component Analysis (PCA)

PCA merupakan singkatan dari *Principal Component Analysis* atau dikenal juga sebagai Analisis Komponen Utama. PCA adalah salah satu metode statistika multivariabel yang digunakan untuk menganalisis dan mengurangi dimensi data. Tujuan utama dari PCA adalah untuk mengidentifikasi pola dalam data dengan mengurangi jumlah dimensi peubah yang ada, sehingga memudahkan pemahaman dan interpretasi data.

PCA mencari kombinasi linear dari peubah dalam data yang disebut sebagai komponen utama. Komponen utama ini diurutkan berdasarkan sejauh mana mereka menjelaskan ragam dalam data. Komponen pertama adalah yang paling penting karena menjelaskan ragam paling besar, sedangkan komponen berikutnya menjelaskan ragam yang kurang penting.

K-Means Clustering

Algoritma *K-Means* pertama kali diusulkan oleh MacQueen (1967) dan dikembangkan oleh Hartigan dan Wong (1979). Algoritma ini merupakan salah satu algoritma yang bersifat *unsupervised learning. K-Means* dapat menerima data tanpa adanya label kategori.

Algoritma *K-Means* bekerja mengelompokan kumpulan data ke dalam beberapa *cluster* yang sebelumnya jumlahnya telah ditentukan, dimana setiap *data point* hanya dimiliki oleh satu *cluster* saja.Secara umum langkah-langkah dalam proses *K-Means clustering*, antara lain:

1. Tentukan jumlah (*k*) *cluster* yang akan dibentuk.
2. Tentukan *initial data point* (*initial centroid*) sebanyak *k.*

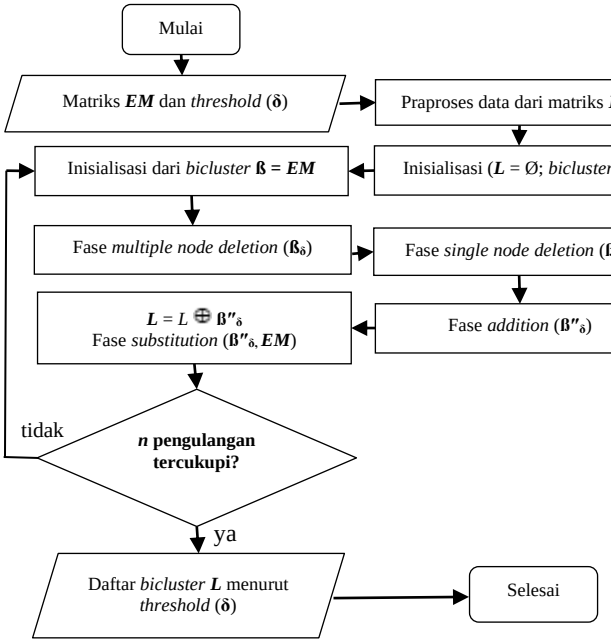
3. Label semua data berdasarkan titik *data point* (*centroid*) terdekat.
4. Tentukan titik *data point* (*centroid*) baru berdasarkan *cluster* yang terbentuk.
5. Label ulang data berdasarkan jarak terdekat terhadap *data point* (*centroid*) baru
6. Ulangi langkah 4 s.d 5 sampai tidak ada lagi perpindahan *data point* (*centroid*) di setiap cluster.

δ-Biclustering (CC Biclustering)

δ-*biclustering* diusulkan pertama kali oleh Cheng dan Church pada tahun 2000. Menurut Nurmawiya dan R Kurniawan (2020), δ-*biclustering* merupakan sebuah algoritma serakah (*greedy algorithm*) yang berusaha mencari *bicluster* maksimum dengan kemiripan tinggi. Tujuan algoritma ini adalah untuk menemukan *bicluster* dengan nilai *Mean Squared Residual* (MSR) lebih kecil dari suatu nilai toleransi (δ) yang telah ditentukan (Kaban et al. 2019).

Algoritma δ-*biclustering* membangun pola *bicluster* yang terbentuk dari beberapa nilai δ (*threshold*) dan jumlah *bicluster* (nilai *k*) serta menentukan nilai δ (*threshold*) dan jumlah *bicluster* terbaik (Cheng dan Church, 2000). *Biclustering* mengambil subgroup objek yang serupa dalam satu subgroup variable dan berbeda dalam variable lainnya.

Pada tahap ini penelitian akan mencoba beberapa jenis δ (*threshold*) dan akan mengevaluasi serta memilih δ (*threshold*) yang memberikan hasil *bicluster* yang optimal. Menghitung nilai MSR/V dari setiap *cluster* yang terbentuk pada *bicluster* hasil algoritma δ-*biclustering.* Menentukan *bicluster* berdasarkan algoritma CC dengan melakukan *tuning parameter* δ.



Deskripsi notasi:

J	: kolom (variabel)
I	: baris (variable)
EM	: matriks dengan ukuran *I x J*
δ	: *threshold*
$b_{Ij} = \frac{\sum_{i \in I} b_{ij}}{\lvert I \rvert}$	: rataan baris
$b_{iJ} = \frac{\sum_{j \in J} b_{ij}}{\lvert J \rvert}$	: rataan kolom
$b_{IJ} = \frac{\sum_{i \in I, j \in J} b_{ij}}{\lvert I \rvert \lvert J \rvert}$	: rataan di seluruh *bicluster*
$RS_{IJ}(i,j) = b_{ij} - b_{Ij} - b_{iJ} + b_{IJ}$	: residual dari elemen matriks
$H(I,J) = \sum_{i \in I, j \in J} \frac{RS_{ij}^2}{\lvert I \rvert \lvert J \rvert}$	: *mean squared residual* (MSR)
$d(i) = \frac{1}{\lvert J \rvert} \sum_{j \in J} RS_{ij}^2$	: nilai sisa kuadrat dari baris
$d(j) = \frac{1}{\lvert I \rvert} \sum_{i \in I} RS_{ij}^2$	: nilai sisa kuadrat dari kolom

GAMBAR 3. Diagram Alir Algoritma δ-*biclustering*

Ilustrasi δ-*biclustering* seperti yang terlihat pada diagram alir pada GAMBAR 3. Secara umum proses δ-*biclustering* terdapat fase penghapusan, penambahan dan substitusi (Pontes B, 2015). Fase penghapusan dan penambahan bersifat iteratif, memastikan bahwa MSR lebih besar dari nilai δ (*threshold*) yang telah ditentukan sebelumnya. Sementara itu, tahap substitusi dilakukan untuk mencegah terjadinya *overlapping* antar bicluster yang dihasilkan.

Analisis Intra-Bicluster dan Inter-Bicluster

Fungsi evaluasi *intra-bicluster* adalah fungsi yang mengukur kualitas suatu *bicluster* menggunakan tingkat koherensi dalam suatu *bicluster* (Ben Saber dan Elloumi, 2014). Sedangkan fungsi evaluasi *inter-bicluster* merupakan fungsi evaluasi yang mengukur kualitas kelompok *bicluster* dengan menilai

keakuratan suatu algoritma untuk memperoleh *bicluster* yang sebenarnya dalam suatu matriks data (Ben Saber dan Elloumi, 2014).

Ukuran fungsi evaluasi *intra-bicluster* yang digunakan adalah *mean squared residue* (MSR). Berdasarkan penelitian Cheng dan Church (2000), *residue* dari elemen a_{ij} dalam *bicluster* yang ditujukan oleh subset *I* dan *J* adalah sebagai berikut:

$$\varepsilon_{ij} = a_{ij} - a_{iJ} - a_{Ij} + a_{IJ} \quad (1)$$

dimana a_{iJ} adalah rata-rata dari baris ke-i dalam *bicluster*, a_{Ij} adalah rata-rata dari kolom ke-j dalam *bicluster*, dan a_{IJ} adalah rata-rata dari seluruh elemen dalam *bicluster*. Sementara itu, *mean squared residual* adalah ragam dari himpunan seluruh elemen dalam *bicluster* ditambah ragam baris rata-rata dan ragam kolom rata-rata. Dalam hal ini *bicluster* yang memiliki *mean squared residual* rendah dengan dimensi yang besar adalah *bicluster* yang baik. Berikut adalah penghitungan dari *mean squared residual:*

$$MSR(I,J) = \frac{1}{|I||J|} \sum_{i \in I, j \in J} (a_{ij} - a_{iJ} - a_{Ij} + a_{IJ})^2 \quad (2)$$

dimana,

$$a_{iJ} = \frac{1}{|J|} \sum_{j \in J} a_{ij}, \qquad a_{Ij} = \frac{1}{|I|} \sum_{i \in I} a_{ij} \quad (3)$$

$$a_{IJ} = \frac{1}{|I||J|} \sum_{i \in I, j \in J} a_{ij} \quad (4)$$

Kualitas dari kelompok *bicluster* yang berbasis MSR selanjutnya diukur dengan menghitung rataan dari MSR dibagi dengan volume dan didefinisikan oleh persamaan 5

$$rataan\ MSR\ per\ volume = \frac{1}{b} \sum_{i=1}^{b} \frac{MSR_i}{Volume_i} \quad (5)$$

dengan *b* merupakan banyak *bicluster* yang dihasilkan metode tertentu.

Sedangkan ukuran fungsi inter-bicluster yang digunakan adalah *Liu and Wang index*. Indeks Liu dan Wang adalah metrik yang digunakan dalam analisis *bicluster* untuk mengukur kualitas dan signifikansi *bicluster*. Indeks Liu dan Wang digunakan untuk membandingkan dua solusi (hasil clustering) dengan mempertimbangkan baris dan kolom suatu bicluster. Indeks Liu dan Wang didefinisikan oleh persamaan 6 (Liu dan Wang, 2007)

$$I_{Liu\&Wang}(M_{opt}, M) = \frac{1}{K_{opt}} \sum_{i=1}^{K_{opt}} max\left(\frac{|G_i \cap G_j| + |C_i \cap C_j|}{|G_i \cup G_j| + |C_i \cup C_j|}\right) \quad (6)$$

dengan M_{opt} adalah kelompok *cluster* yang memiliki nilai rataan MSR per volume terkecil dan *M* adalah kelompok *cluster* yang lain. K_{opt} merupakan banyaknya *cluster* yang ada pada M_{opt}, $|G_i \cap G_j|$ merupakan banyaknya baris *G* pada M_{opt} yang beririsan dengan baris *M*, dan $|C_i \cap C_j|$ merupakan banyaknya kolom *C* pada M_{opt} yang beririsan dengan kolom pada *M*. $|G_i \cup G_j|$ merupakan banyaknya gabungan baris dari M_{opt} dan *M*, serta $|C_i \cup C_j|$ merupakan banyaknya gabungan kolom dari M_{opt} dan *M*. Nilai indeks Liu dan Wang menunjukan seberapa baik suatu kelompok *bicluster* optimal (M_{opt}) akan memiliki kesamaan dengan kelompok *bicluster* yang lain(*M*). Ketika $M_{opt} = M$ maka nilai indeks Liu dan Wang akan bernilai 1 (Prelic et al. 2006).

Evaluasi Analisis Eksplorasi

Tahap evaluasi analisis eksplorasi dari hasil *cluster/bicluster* ini dilakukan dengan 3 tahapan, yaitu: interpretasi *cluster/bicluster*, validasi hasil *cluster/bicluster* dan profiling *cluster/bicluster*. Interpretasi *cluster/bicluster* adalah memberi nama spesifik untuk menggambarkan isi *cluster/bicluster* tersebut. Proses validasi bertujuan menjamin bahwa hasil *cluster/bicluster* adalah representatif terhadap populasi secara umum, dan dengan demikian dapat digeneralisasi untuk objek serupa lainnya dan stabil untuk waktu tertentu. Tahap profiling meliputi penggambaran karakteristik masing-masing *cluster/bicluster* untuk menjelaskan bagaimana kelompok-kelompok tersebut bisa berbeda secara

relevan pada tiap dimensi. Tujuan proses ini adalah untuk menjelaskan karakteristik tiap *cluster/bicluster* berdasarkan profil tertentu. Proses *profiling* ini menitikberatkan pada karakteristik yang secara signifikan berbeda antar tiap *cluster* dan memprediksi anggota dalam *cluster* tersebut.

HASIL DAN PEMBAHASAN

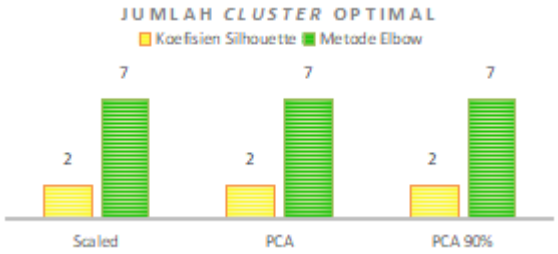
Proses *K-Means Clustering*

Algoritma *K-Means* adalah salah satu algoritma *clustering* non-hierarki dimana jumlah *cluster* harus diketahui dari awal. Pemilihan jumlah *cluster* yang optimal dilakukan dengan menggunakan metode koefisien *silhouette* dan metode *elbow.*

Nilai koefisien *silhouette* adalah ukuran seberapa mirip suatu obyek dengan *cluster*-nya sendiri (kohesi) dibandingkan dengan *cluster* yang lainnya (separasi). *Cluster* dengan nilai koefisien *silhouette* terbesar merupakan jumlah *cluster* yang optimal. Sedangkan metode *elbow* menentukan jumlah *cluster* yang optimal berdasarkan penurunan nilai *Sum of Square Error* (SSE). Pada metode *elbow*, penentuan jumlah *cluster* optimal berdasarkan penurunan nilai SSE yang signifikan.

Kedua metode penentuan jumlah *cluster* optimal ini diterapkan pada ketiga jenis matriks yang telah dipersiapkan pada tahap sebelumnya, yaitu: matriks data ekspor dengan *scaling*, matriks data ekspor dengan PCA, serta matriks data ekspor dengan PCA reduksi komponen utama sebanyak 10% (PCA 90%).

Pemilihan *cluster* dibatasi dari 2 *cluster* hingga 10 *cluster*, hal ini dilakukan untuk mempermudah interpretasi karakteristik dari *cluster* yang terbentuk. Berdasarkan metode koefisien *silhouette* pada ketiga jenis data matriks tersebut, diperoleh jumlah *cluster* yang optimal sebanyak 2 *cluster* (GAMBAR 4). Sedangkan pada metode *elbow*, terlihat penurunan nilai SSE yang cukup signifikan terdapat pada jumlah *cluster* 7. Sehingga berdasarkan metode *elbow*, jumlah *cluster* optimal adalah 7 (GAMBAR 4).



GAMBAR 4. Nilai Koefisien *Silhouette* dan Metode *Elbow*

GAMBAR 4 memperlihatkan bahwa proses normalisasi data yang telah digunakan pada pra-proses data, yaitu: *data scaling* dan PCA, tidak memiliki pengaruh yang signifikan pada penentuan jumlah *cluster* yang optimal. Penggunaan metode koefisien *silhouette* pada ketiga data matriks tersebut sama-sama menghasilkan jumlah *cluster* optimal sebanyak 2 *cluster*. Pada penerapan metode *elbow*, jumlah *cluster* optimal untuk semua jenis pra-proses menghasilkan nilai 7 *cluster.*

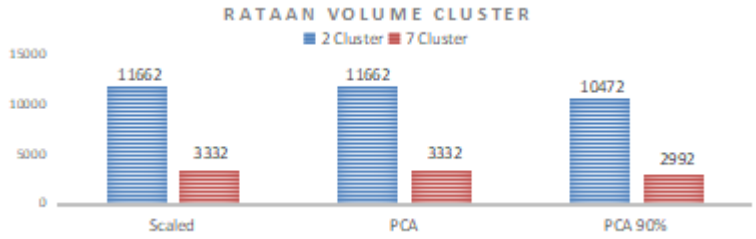
Berdasarkan kedua metode tersebut penerapan algoritma *K-Means* pada penelitian ini menggunakan jumlah *cluster* optimal sebanyak 2 *cluster* dan 7 *cluster*. Berdasarkan jumlah *cluster* optimal yang dihasilkan dari kedua metode tersebut dapat dibuat beberapa skenario penelitian seperti tertera pada TABEL 1.

TABEL 1. Skenario percobaan pada penerapan algoritma *K-Means*

Skenario 1	Skenario 2	Skenario 3
Algoritma: *K-Means* Jumlah *cluster* : 2 Normalisasi: *data scaling*	Algoritma: *K-Means* Jumlah *cluster* : 7 Normalisasi: *data scaling*	Algoritma: *K-Means* Jumlah *cluster* : 2 Normalisasi: *PCA*
Skenario 4	**Skenario 5**	**Skenario 6**
Algoritma: *K-Means* Jumlah *cluster* : 7 Normalisasi: *PCA*	Algoritma: *K-Means* Jumlah *cluster* : 2 Normalisasi: *PCA 90%*	Algoritma: *K-Means* Jumlah *cluster* : 7 Normalisasi: *PCA 90%*

Berdasarkan metode koefisien *silhouette* yang menghasilkan jumlah *cluster* optimal sebanyak 2, maka dibentuklah *cluster* dengan menggunakan algoritma *K-Means*. Hasil *cluster* dengan menggunakan algoritma *K-Means* baik pada matriks data dengan *data scaling*, matriks data dengan PCA maupun matriks data dengan PCA 90% sama-sama memiliki keanggotaan *cluster* yang serupa. *Cluster* 1 merupakan *cluster* dengan jumlah anggota terbanyak yaitu 228 negara. *Cluster* 2 memiliki anggota sebanyak 10 negara.

Seperti halnya pada algoritma *K-Means* 2 *cluster*, pada algoritma *K-Means* 7 *cluster* yang dibentuk berdasarkan metode *elbow* ini pun, baik pada matriks data dengan *data scaling*, matriks data dengan PCA maupun matriks data dengan PCA 90% sama-sama memiliki keanggotaan *cluster* yang serupa. *Cluster* 1, 2, 4 dan 5 merupakan *cluster* dengan hanya 1 anggota negara. Berturut-turut negara tersebut antara lain: Tiongkok, Amerika Serikat, Singapura dan Jepang. *Cluster* 3 berisi 5 negara, yaitu: India, Malaysia, Filipina, Thailand dan Vietnam. *Cluster* 6 merupakan *cluster* dengan jumlah anggota terbanyak. *Cluster* 7 berisi 12 negara.



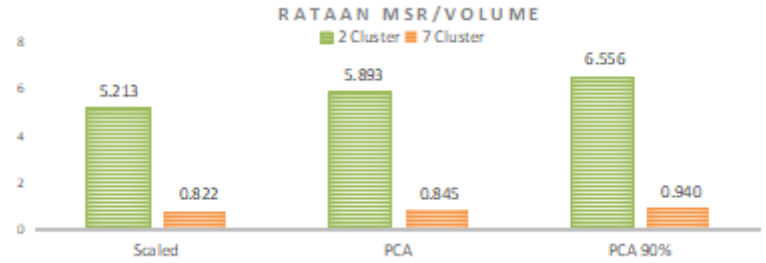
GAMBAR 5. Nilai rataan volume *cluster* dari setiap skenario

Ukuran rataan volume *cluster* yang dibentuk pada algoritma *K-Means* 2 *cluster* berukuran lebih besar dibandingkan rataan volume *cluster* yang dibentuk pada algoritma *K-Means* 7 *cluster* (GAMBAR 5). Ukuran rataan volume pada algoritma *K-Means* 2 *cluster* pada matriks data dengan *data scaling* dan PCA sama-sama bernilai 11.662. Sedangkan pada matriks data dengan PCA 90% berukuran lebih kecil yaitu 10.472 karena terdapat reduksi peubah sebesar 10%.

Pada algoritma *K-Means* 7 *cluster* ukuran rataan volume *cluster* jauh lebih kecil. Pada matriks data dengan *data scaling* dan PCA berukuran 3.332, sedangkan pada matriks data dengan PCA 90% hanya bernilai 2.992.

Hasil perhitungan Intra-Cluster pada K-Means Clustering

Metode *intra-cluster* berfungsi untuk mengukur kualitas suatu *cluster* menggunakan tingkat koherensi dalam suatu *cluster* (Ben Saber dan Elloumi, 2014). Ukuran fungsi evaluasi dalam metode *intra-cluster* yang digunakan adalah menghitung nilai *mean square residue* dari setiap *cluster* yang terbentuk dan kemudian membuat rataannya dari seluruh *cluster* yang ada. Semakin kecil nilai rataan MSR/volume semakin baik tingkat koherensi dalam suatu *cluster*, maka semakin bagus juga kualitas *cluster* yang telah terbentuk.

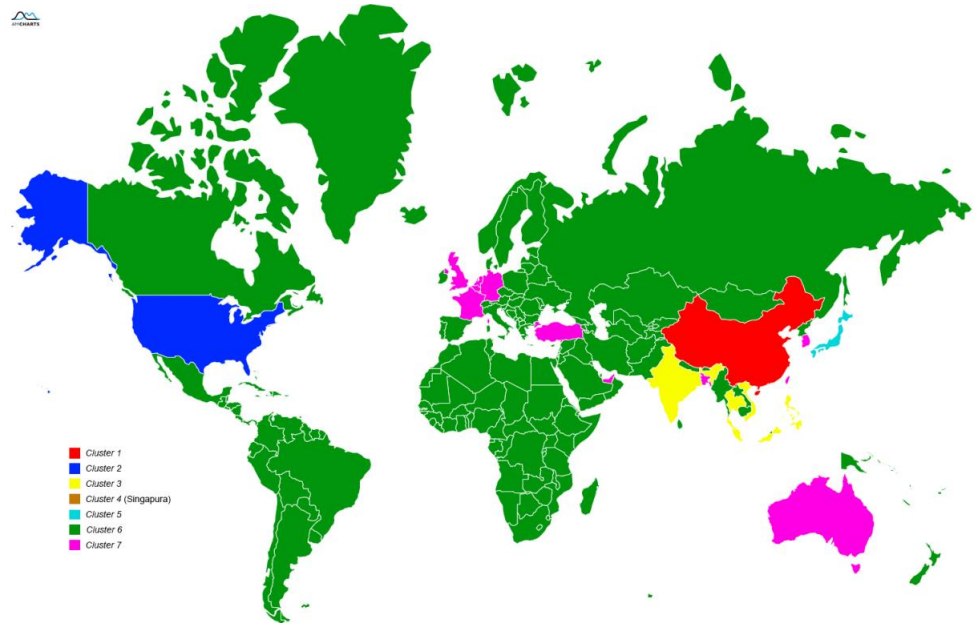


GAMBAR 6. Nilai Rataan MSR/volume pada algoritma *K-Means*

Pada hasil *clustering* dengan menggunakan algoritma *K-Means* ini dihitung nilai rataan MSR/volume dari setiap matriks data yang tersedia. Dari ketiga jenis matriks data yang digunakan,

matriks data dengan pra-proses *data scaling* memiliki nilai rataan MSR/volume yang cenderung lebih kecil dibandingkan dengan matriks data dengan PCA maupun matriks data dengan PCA 90%.

Pada GAMBAR 6, dapat diperhatikan bahwa nilai rataan MSR/volume dengan nilai 0,822 pada skenario 2, yaitu: penerapan *K-Means* 7 *cluster* dengan normalisasi data menggunakan *data scaling* merupakan yang terkecil dibandingkan dengan nilai rataan MSR/volume pada kombinasi yang lainnya. Dengan demikian skenario 2 merupakan skenario terbaik yang diperoleh ketika penerapan algoritma *K-Means* pada data ekspor Indonesia kurun waktu tahun 2013-2022. Hasil *K-Means* pada skenario 2 dapat dilihat pada GAMBAR 7.



GAMBAR 7. Anggota *Cluster* pada algoritma *K-Means Clustering* skenario 2

Pada GAMBAR 7, *cluster* 1, *cluster* 2, *cluster* 4 dan *cluster* 5 adalah *cluster* dengan jumlah anggota 1. Berturut-turut anggota *cluster*-nya yaitu, Tiongkok, Amerika Serikat, Singapura dan Jepang. *Cluster* 3 mayoritas berisi negara-negara Asia Selatan dan Asia Tenggara. *Cluster* 6 merupakan *cluster* dengan jumlah anggota terbanyak. Sedangkan *cluster* 7 beranggotakan beberapa negara Eropa Barat, Uni Emirat Arab, Turki, Bangladesh, Taiwan, Korea Selatan dan juga Australia.

Proses δ-*Biclustering* (CC *Biclustering*)

Algoritma CC *biclustering* merupakan algoritma *biclustering* yang menggunakan nilai *mean square residual* (MSR) sebagai nilai toleransi (δ) dalam membentuk sebuah *bicluster*. Kumpulan subkelompok baris dan subkelompok kolom disebut *bicluster* jika memiliki nilai MSR dibawah *level/threshold*(δ). Oleh karena itu penentuan nilai toleransi (δ) amatlah penting dalam algoritma CC *biclustering.*

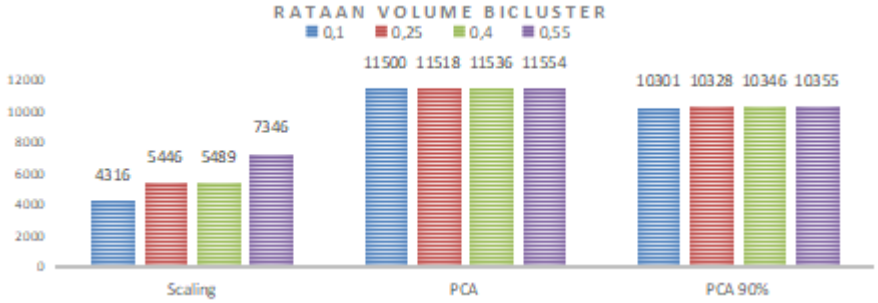
Dalam penelitian ini nilai toleransi (δ) yang digunakan adalah 0,10; 0,25; 0,40 dan 0,55. Setiap tingkatan nilai toleransi (δ) ini akan diterapkan pada ketiga jenis matriks yang telah dipersiapkan pada tahap sebelumnya, yaitu: matriks data ekspor dengan *scaling*, matriks data ekspor dengan PCA serta matriks data ekspor dengan reduksi komponen utama sebanyak 10% (PCA 90%). Berdasarkan beberapa jenis pra-prosesing data dan tingkat nilai toleransi (δ) sehingga didapatkan skenario penelitian yang dapat dilihat pada TABEL 2.

TABEL 2. Skenario percobaan pada penerapan algoritma CC *Biclustering*

Skenario 1	Skenario 2	Skenario 3	Skenario 4
Algoritma : *CC Biclustering* Nilai Toleransi (δ) : 0,10 Normalisasi : *data scaling*	Algoritma : *CC Biclustering* Nilai Toleransi (δ) : 0,25 Normalisasi : *data scaling*	Algoritma : *CC Biclustering* Nilai Toleransi (δ) : 0,40 Normalisasi : *data scaling*	Algoritma : *CC Biclustering* Nilai Toleransi (δ) : 0,55 Normalisasi : *data scaling*
Skenario 5	Skenario 6	Skenario 7	Skenario 8

Algoritma : *CC Biclustering* Nilai Toleransi (δ) : 0,10 Normalisasi : PCA	Algoritma : *CC Biclustering* Nilai Toleransi (δ) : 0,25 Normalisasi : PCA	Algoritma : *CC Biclustering* Nilai Toleransi (δ) : 0,40 Normalisasi : PCA	Algoritma : *CC Biclustering* Nilai Toleransi (δ) : 0,55 Normalisasi : PCA
Skenario 9	Skenario 10	Skenario 11	Skenario 12
Algoritma : *CC Biclustering* Nilai Toleransi (δ) : 0,10 Normalisasi : PCA 90%	Algoritma : *CC Biclustering* Nilai Toleransi (δ) : 0,25 Normalisasi : PCA 90%	Algoritma : *CC Biclustering* Nilai Toleransi (δ) : 0,40 Normalisasi : PCA 90%	Algoritma : *CC Biclustering* Nilai Toleransi (δ) : 0,55 Normalisasi : PCA 90%

Dari hasil pembentukan *biclustering* dari seluruh skenario yang ada diketahui bahwa besaran nilai toleransi (δ) dapat mempengaruhi jumlah *bicluster* yang terbentuk dan juga ukuran volume *bicluster* tersebut. Ukuran rataan volume hasil proses CC *biclustering* dapat dilihat pada GAMBAR 8.

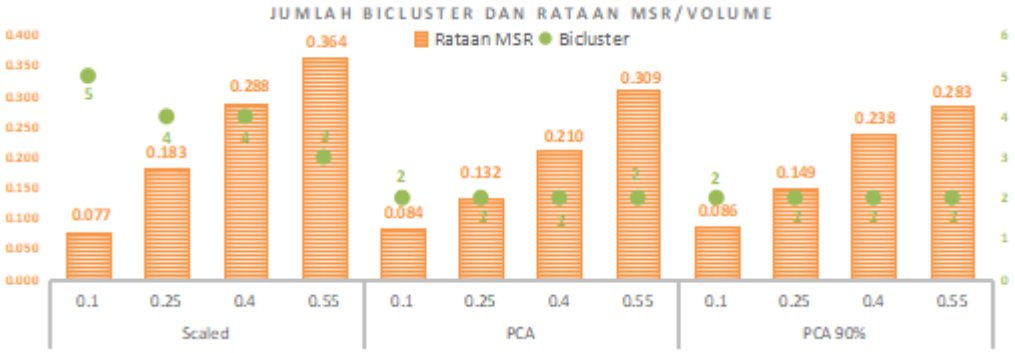


GAMBAR 8. Nilai rataan MSR/volume pada algoritma *CC Biclustering*

Pada skenario 1, yaitu penerapan *CC Biclustering* dengan nilai toleransi (δ) 0,10 dan metode normalisasi dengan *data scaling* hasil *bicluster* yang terbentuk adalah sebanyak 5 *bicluster*. Skenario 2 dan skenario 3 menghasilkan 4 *bicluster*, Sedangkan skenario 4 menghasilkan 3 *bicluster*. Skenario 1 memiliki nilai rataan volume *bicluster* terkecil, yaitu 4.316.

Pada skenario 5 sampai skenario 12 yang menggunakan proses normalisasi data dengan metode PCA dan PCA 90% sama-sama menghasilkan jumlah *bicluster* yang terbentuk sebanyak 2 *bicluster*. Meskipun menghasilkan jumlah *bicluster* yang sama yaitu 2, ukuran rataan volume dari setiap *bicluster* yang terbentuk relatif berbeda. Rataan volume *bicluster* pada matriks data dengan PCA memiliki nilai berkisar 11.500, sedangkan pada matriks data dengan PCA 90% memiliki nilai berkisar 10.300.

Hasil perhitungan Intra-Cluster pada δ-Biclustering



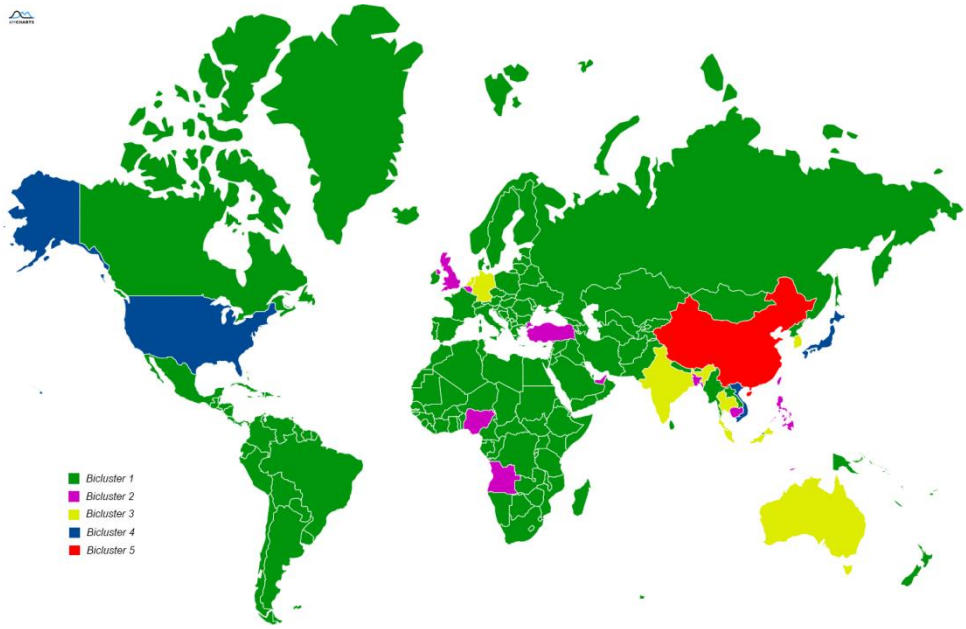
GAMBAR 9. Jumlah *bicluster* dan nilai rataan MSR/volume pada algoritma *K-Means*

Dapat dilihat pada GAMBAR 9, nilai MSR/Volume terkecil dari setiap jenis matriks data terjadi ketika nilai toleransi (δ) bernilai 0,10, yaitu pada skenario 1, skenario 5 dan skenario 9. Berturut-turut nilai tersebut adalah 0,077; 0,084 dan 0,086.

Nilai rataan MSR/Volume terkecil diperoleh pada algoritma *CC Biclustering* menggunakan matriks data dengan *data scaling* pada nilai toleransi (δ) 0,10 (skenario 1), dengan nilai 0,077. Skenario 1 merupakan *CC bicluster* dengan jumlah *bicluster* sebanyak 5. Dengan demikian skenario 1 merupakan skenario terbaik yang diperoleh ketika menggunakan algoritma *CC Biclustering* yang diterapkan pada

data ekspor Indonesia kurun waktu tahun 2013-2022. Hasil *CC Biclustering* pada skenario 1 dapat dilihat pada GAMBAR 10.

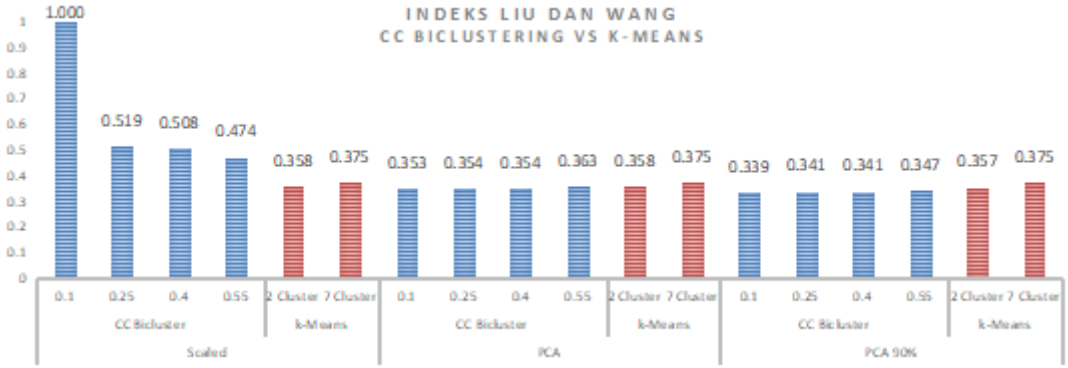
Pada GAMBAR 10, *bicluster* 1 adalah *bicluster* dengan jumlah anggota terbanyak dan juga berisikan seluruh peubah (kode HS). *Bicluster* 2 berisikan 12 negara beberapa diantaranya adalah Nigeria, Angola, Bangladesh, Turki, Uni Emirat Arab, Inggris dan beberapa negara Asia Tenggara. *Bicluster* 2 ini juga berisi sebanyak 37 peubah. *Bicluster* 3 berisi negara Jerman, Australia, Korea Selatan, beberapa negara Asia Selatan dan Asia Tenggara. Sebanyak 13 peubah termasuk kedalam *bicluster* 3. *Bicluster* 4 berisi negara Amerika Serikat, Jepang dan Vietnam. Peubah yang termasuk ke dalam *bicluster* 4 hanya sebanyak 5. *Bicluster* 4 merupakan *bicluster* dengan ukuran volume paling kecil diantara yang lain. *Bicluster* 5 berisi Tiongkok dan Singapura dengan jumlah peubah sebanyak 29 peubah.



GAMBAR 10. Anggota *Bicluster* pada algoritma *CC Biclustering* skenario 1

Hasil perhitungan Inter-Cluster pada K-Means dan δ-Biclustering

Indeks Liu dan Wang umumnya digunakan untuk mengukur sejauh mana suatu *cluster* mewakili pola yang signifikan dalam data. Tujuan utamanya adalah untuk menghasilkan *cluster* yang memiliki nilai indeks Liu dan Wang yang tinggi, yang menunjukan bahwa *cluster* tersebut memiliki pola yang signifikan dalam data.



GAMBAR 11. Nilai indeks Liu dan Wang pada algoritma *CC Biclustering* dan algoritma *K-Means*

Berdasarkan nilai rataan MSR/Volume yang terkecil dari seluruh skenario baik dari hasil penerapan algoritma *CC Biclustering* dan *K-Means* diperoleh kesimpulan bahwa skenario 1 dari penerapan algoritma *CC Biclustering* merupakan skenario paling optimal yang dapat diterapkan pada data ekspor Indonesia kurun waktu tahun 2013-2022.

Pada GAMBAR 11, skenario 1 yang memiliki nilai indeks Liu dan Wang sebesar 1 menunjukan bahwa kelompok *bicluster* pada skenario tersebut adalah solusi pengelompokan yang optimal. Nilai indeks Liu dan Wang 0,519 - 0,474, didapatkan pada skenario 2, skenario 3 dan skenario 4 pada algoritma *CC Biclustering.*

KESIMPULAN DAN SARAN

Berdasarkan nilai MSR/Volume dan nilai indeks Liu dan Wang yang dihasilkan dari setiap skenario baik dari penerapan algoritma *CC Biclustering* dan *K-Means Clustering*, diperoleh kesimpulan bahwa skenario 1 dari penerapan algoritma *CC Biclustering* dengan nilai toleransi (δ) 0,10 dan proses normalisasi data menggunakan *data scaling* merupakan skenario optimal pada kasus pengelompokan data ekspor Indonesia kurun waktu tahun 2013 hingga tahun 2022.

Penelitian ini terbatas pada algoritma *K-Means* dan *CC Biclustering. CC Biclustering* merupakan algoritma *biclustering* tanpa adanya kemampuan membentuk *overlapping bicluster.* Oleh karena itu penelitian ini masih dapat dikembangkan dengan mencoba membandingkannya kembali menggunakan algoritma *biclustering* yang memiliki kemampuan *overlapping bicluster*, seperti *Plaid Model Biclustering* ataupun ISA (*Iterative Signature Algorithm*).

REFERENSI

Ben Saber H, Elloumi M. 2014. A Comparative Study of Clustering and Biclustering of Microarray Data. Int J Comput Sci Inf Technol. 6(6):93–111.

C. A. Putri, R. Irfani, and B. Sartono, 2021. "Recognizing poverty pattern in Central Java using Biclustering Analysis," J. Phys. Conf. Ser., vol. 1863.

Chakraborty, A., & Maka, H. 2005. Biclustering of gene expression data using genetic algorithm. Di dalam : 2005 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (pp. 1-8). IEEE.

Cheng, Y., & G. M. Church. 2000. Biclustering of expression data. ISMB-00 Proceedings (pp. 93 - 103). International Conference on Intelligent Systems for Molecular Biology.

Frongillo, E.A., Nguyen, H.T., Smith, M.D. and Coleman-Jensen, A., 2019. Food Insecurity is More Strongly Associated with Poor Subjective Well-Being in More-Developed Countries than in Less-Developed Countries. The Journal of Nutrition, 149(2), pp.330-335.

Hartigan, J.A.; Wong, M. A. 1979. "Algorithm AS 136: A K-Means Clustering Algorithm". Journal of the Royal Statistical Society, Series C. 28 (1): 100–108. JSTOR 2346830

Kaban PA, Kurniawan R, Caraka RE, Pardamean B, Yuniarto B, Sukim. 2019. Biclustering method to capture the spatial pattern and to identify the causes of social vulnerability in Indonesia: A new recommendation for disaster mitigation policy. Procedia Comput Sci. 157:31–37.

Liu X, Wang L. 2007. Computing the maximum similarity bi-clusters of gene expression data. Bioinformatics. 23(1):50–56.

MacQueen, J. B. 1967. Some Methods for classification and Analysis of Multivariate Observations. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. Vol. 1. University of California Press.

Madeira, S., & Oliveira, A. 2004. Biclustering algorithms for biological data analysis: a survey. EEE/ACM Trans Comput Biol Bioinform (TCBB), 1(1), 24–45.

Nurmawiya dan R. Kurniawan, 2020. Pengelompokan Wilayah Indonesia Dalam Menghadapi Revolusi Industri 4.0 Dengan Metode Biclustering, pp. 790–797.

Ningsih, Wiwik Andriyani Lestari & Sumertajaya, I Made & Saefuddin, Asep. 2022. Biclustering Application in Indonesian Economic and Pandemic Vulnerability. Barekeng: Jurnal Ilmu Matematika dan Terapan. 16. 1453-1464.

Pontes, B., Girldez, R., & Aguilar-Ruiz, J. S. 2015. Quality measures for gene expression biclusters. PloS one, 10(3).

Prelić A, Bleuler S, Zimmermann P, Wille A, Bühlmann P, Gruissem W, Hennig L, Thiele L, Zitzler E. 2006. A systematic comparison and evaluation of biclustering methods for gene expression data. Bioinformatics. 22(9):1122– 1129.

R. Novidianto and R. Irfani, 2020. "Bicluster CC Algoritm Analysis to Identify Patterns of Food Insecurity in Indonesia," J. Mat. Stat. dan Komputasi, vol. 17, no. 2, pp. 325–338.

Tan, P.N., Steinbach, M., Kumar, V. 2006. Introduction to Data Mining. Boston:Pearson Education.