Jurnal Matematika UNAND
Vol. **VIII** No. **1** Hal. 323 – 331
Edisi Mei 2019
ISSN : 2303–291X

PENERAPAN ANALISIS *CLUSTER ENSEMBLE* UNTUK MENGELOMPOKKAN PROVINSI DI INDONESIA BERDASARKAN INDIKATOR KESEHATAN LINGKUNGAN

YULIZA DIANA PUTRI, IZZATI RAHMI HG, HAZMIRA YOZZA

Program Studi S1 Matematika,
Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Andalas,
Kampus UNAND Limau Manis Padang, Indonesia,
email: yulizadianaputrii@gmail.com

Diterima 9 Maret 2019 Direvisi 7 April 2019 Dipublikasikan 7 Mei 2019

Abstrak. Kesehatan lingkungan merupakan bagian dari pada kesehatan masyarakat pada umumnya. Setiap daerah memiliki keadaan kesehatan lingkungan yang berbeda-beda jika dikaitkan dengan indikator kesehatan lingkungan tersebut. Oleh karena itu prioritas program penyehatan lingkungan pun berbeda pada setiap daerah. Suatu hal yang menarik untuk diketahui adalah bagaimana kesamaan/kemiripan dari masing-masing daerah tersebut berdasarkan indikator kesehatan lingkungan. Kemiripan tersebut selanjutnya dapat dijadikan dasar untuk melakukan pengelompokan daerah-daerah tersebut, sehingga daerah yang memiliki kondisi kesehatan lingkungan yang hampir sama akan berada pada satu kelompok dan sebaliknya, daerah-daerah dengan kondisi kesehatan lingkungan yang tidak sama akan berada pada kelompok yang berbeda. Dengan adanya pengelompokan tersebut akan mempermudah pemerintah untuk menentukan prioritas bagi pembangunan kesehatan lingkungan di daerah-daerah tersebut. Dalam penelitian ini metode *cluster ensemble* akan diterapkan untuk mengelompokkan provinsi di Indonesia berdasarkan 8 indikator kesehatan lingkungan. Penelitian ini menghasilkan solusi pengklasteran terbaik yaitu solusi dengan 2 *cluster*, dimana anggota dari *cluster* 1 merupakan provinsi dengan lingkungan sehat yang lebih baik dibandingkan anggota dari *cluster* 2.

Kata Kunci: Cluster Ensemble, Cluster Hierarki, k-Means Cluster

1. Pendahuluan

Lingkungan memiliki peranan yang sangat penting dalam mewujudkan derajat kesehatan masyarakat yang optimal, disamping faktor-faktor lain seperti kualitas pelayanan kesehatan dan perilaku masyarakat. Kesehatan lingkungan merupakan bagian dari pada kesehatan masyarakat pada umumnya. Menyadari pentingnya kesehatan lingkungan diperlukan program-program penyehatan lingkungan yang tujuannya adalah untuk membina dan mempercepat terwujudnya derajat kesehatan masyarakat yang optimal, baik fisik, mental, maupun sosial.

Setiap daerah memiliki keadaan kesehatan lingkungan yang berbeda-beda jika dikaitkan dengan indikator kesehatan lingkungan tersebut. Oleh karena itu priori-

tas program penyehatan lingkungan pun berbeda pada setiap daerah. Suatu hal yang menarik untuk diketahui adalah bagaimana kesamaan/kemiripan dari masing-masing daerah tersebut berdasarkan indikator kesehatan lingkungan. Kemiripan tersebut selanjutnya dapat dijadikan dasar untuk melakukan pengelompokan daerah daerah tersebut, sehingga daerah yang memiliki kondisi kesehatan lingkungan yang hampir sama akan berada pada satu kelompok dan sebaliknya, daerah-daerah dengan kondisi kesehatan lingkungan yang tidak sama akan berada pada kelompok yang berbeda. Dengan adanya pengelompokan tersebut akan mempermudah pemerintah untuk menentukan prioritas bagi pembangunan kesehatan lingkungan di daerah-daerah tersebut.

Analisis statistika untuk mengelompokkan objek-objek adalah analisis *cluster*. Analisis *cluster* saat ini semakin berkembang pesat seiring dengan kemajuan teknologi dan informasi yang melahirkan data besar (*big data*). Banyak metode *cluster* yang telah dikembangkan oleh para pakar dan telah banyak pula diterapkan pada berbagai bidang. Selain metode konvensional, yaitu analisis *cluster* berhirarki dan tak berhirarki, salah satu metode yang hingga saat ini banyak dikembangkan adalah metode *cluster ensemble*. Metode ini diperkenalkan oleh Strehl dan Gosh pada tahun 2002 [9]. Ide dasar dari cluster ensemble adalah mengombinasikan sekumpulan hasil *cluster* yang dibentuk berdasarkan metode-metode yang biasa dilakukan. Menurut [9], *cluster ensemble* dapat memberikan hasil pengklasteran yang lebih berkualitas.

Berdasarkan uraian tersebut, penelitian ini bertujuan menerapkan metode *cluster ensemble* untuk mengelompokkan provinsi di Indonesia berdasarkan indikator kesehatan lingkungan tahun 2016.

2. Landasan Teori

2.1. *Analisis Cluster*

Analisis *cluster* merupakan metode dengan analisis peubah ganda untuk mengelompokkan n objek ke dalam m cluster ($m < n$) berdasarkan karakteristiknya. Pengelompokan dilakukan berdasarkan pada sifat kemiripan atau sifat ketidakmiripan antar objek. Objek yang berada dalam kelompok yang sama akan lebih mirip dibandingkan dengan objek pada kelompok yang berbeda.

2.1.1. *Konsep Jarak*

Dalam analisis *cluster*, objek dikelompokkan berdasarkan kemiripan atau ketidakmiripan antar objek. Salah satu ukuran ketidakmiripan yang paling sering digunakan adalah ukuran jarak dan ukuran jarak yang sering digunakan sebagai ukuran ketidakmiripan antar objek adalah jarak *euclid*. Penggunaan jarak *euclid* dilakukan jika tidak ada kerelasi antar peubah. Jarak ini didefinisikan sebagai berikut:

$$d_{ij} = [(\mathbf{x_i} - \mathbf{x_j})^T(\mathbf{x_i} - \mathbf{x_j})]^{1/2} \tag{2.1}$$

Jika satuan pengukuran data tidak sama, maka data perlu ditransformasi ke bentuk baku (Z) sebelum dilakukan perhitungan jarak *euclid*. Jarak *mahalanobis* juga dapat digunakan untuk mengatasi korelasi antar peubah. Jarak *mahalanobis* dihitung

dengan rumus berikut:

$$d_{ij} = [(\mathbf{x_i} - \mathbf{x_j})^T \mathbf{S^{-1}} (\mathbf{x_i} - \mathbf{x_j})]^{1/2} \tag{2.2}$$

2.1.2. *Metode Berhirarki*

Metode berhirarki biasanya digunakan jika peneliti belum mengetahui banyaknya cluster yang akan dibentuk dan ukuran contoh relatif kecil. Terdapat dua prosedur pada metode berhirarki ini yaitu *agglomerative hieararchical clustering* dan *divisive hierarchical clustering*. Dalam metode *agglomerative* terdapat lima metode perbaikan jarak yang dapat digunakan.

(a) Pautan Tunggal.
Jarak dua *cluster* diukur dengan jarak terdekat antara sebuah objek dalam *cluster* yang satu dengan sebuah objek dalam cluster yang lain.

$$d_{(uv)w} = \min(d_{uw}, d_{vw}). \tag{2.3}$$

(b) Pautan Lengkap.
Jarak dua *cluster* diukur dengan jarak terjauh antara sebuah objek dalam *cluster* yang satu dengan sebuah objek dalam cluster yang lain.

$$d_{(uv)w} = \max(d_{uw}, d_{vw}). \tag{2.4}$$

(c) Pautan *Centroid*.
Jarak antara dua buah *cluster* diukur sebagai jarak *Euclidian* antara kedua rataan *(centroid) cluster*.

$$d_{(uv)w} = \frac{n_u d_{uw} + n_v d_{vw}}{n_u + n_v} - \frac{n_u n_v d_{uv}}{(n_u + n_v)^2}. \tag{2.5}$$

(d) Pautan Rataan.
Jarak antara dua *cluster* diukur dengan jarak rataan antara sebuah objek dalam *cluster* yang satu dengan sebuah objek dalam *cluster* yang lain.

$$d_{(uv)w} = \frac{n_u d_{uw} + n_v d_{vw}}{n_u + n_v}. \tag{2.6}$$

(e) Pautan *Ward*.
Jarak antara dua buah *cluster* sebagai jarak antar median, dan *cluster-cluster* dengan jarak terkecil akan digabungkan.

$$d_{(uv)w} = \frac{1}{2} d_{uw} + \frac{1}{2} d_{vw} - \frac{1}{4} d_{uv}. \tag{2.7}$$

2.1.3. *Metode Tak Berhirarki*

Metode tak berhirarki digunakan untuk pengelompokan objek dimana banyaknya *cluster* yang akan dibentuk dapat ditentukan terlebih dahulu sebagai bagian dari prosedur pengklasteran. Metode ini dapat diterapkan pada data yang lebih besar dibandingkan metode hirarki. Metode tak berhirarki yang umum digunakan adalah k-rataan (*k-means*).

Secara ringkas, langkah-langkah pengklasteran menggunakan metode tak berhirarki adalah sebagai berikut [2]:

(1) Bagi objek-objek tersebut ke dalam K *cluster* awal.
(2) Masukkan tiap objek ke suatu *cluster* berdasarkan rataan terdekat. Jarak biasanya ditentukan dengan menggunakan *Euclidean*. Hitung kembali rataan untuk *cluster* yang mendapat objek dan yang kehilangan objek.
(3) Ulangi langkah 2 sampai tidak ada lagi pemindahan objek antar *cluster*.

2.2. *Cluster Ensemble*

Cluster Ensemble diperkenalkan oleh Strehl dan Gosh (2002), yaitu sebuah metode yang digunakan untuk mengombinasikan sekumpulan solusi *cluster*. Metode ini memiliki keunggulan dibanding metode pengklasteran lainnya, yakni mampu meningkatkan kualitas dan kekekaran solusi *cluster* [9]. Secara umum, pengklasteran objek dengan metode *cluster ensemble* dilakukan dalam dua tahap menurut [1], yaitu:

(1) Membentuk anggota ensemble yang anggotanya adalah solusi dari berbagai metode pengklasteran yang berbeda.
(2) Mengombinasikan seluruh anggota ensemble untuk memperoleh satu solusi akhir yang dinamakan solusi *Consensus*.

Fungsi *Consensus* didefinisikan sebagai fungsi yang memetakan sekumpulan solusi *cluster* menjadi solusi gabungan hingga diperoleh satu hasil pengklasteran akhir yang disebut solusi *consensus*. Fungsi ini memiliki beragam algoritma, salah satunya algoritma *Meta-Clustering*.

Berikut ini adalah algoritma *meta-clustering* yang dikembangkan oleh [5]:

(1) Mentransformasi sekumpulan solusi *cluster* atau anggota *ensemble* menjadi sebuah matriks indikator. Matriks indikator adalah matriks yang kolom-kolomnya menggambarkan *cluster* dari setiap solusi sedangkan baris-baris matriks indikator menggambarkan objek pengamatan. Matriks ini terdiri dari angka biner 1 dan 0. Objek bernilai 1 pada kolom tertentu jika merupakan angka *cluster* yang bersesuaian dengan kolom tersebut dan bernilai 0 jika sebaliknya.
(2) Mengelompokkan kembali objek pengamatan dengan menganggap kolom-kolom matriks indikator sebagai peubah baru yang digunakan dalam analisis *cluster*. Tahap ini disebut dengan *clustering on cluster (CC)*. Tahap *clustering on cluster* (CC) menggunakan metode *k-means* sebagai analisis cluster. Pengklasteran pada tahap ini dilakukan secara berulang dengan inisialisasi acak. Pengulangan ini dilakukan untuk mengatasi adanya pengaruh inisialisasi yang berbeda terhadap solusi yang dihasilkan. Kedua langkah di atas dilakukan berulang kali hingga tidak terjadi perubahan keanggotaan *cluster* yang dihasilkan.

Algoritma *Meta-clustering* dilakukan secara terpisah untuk kumpulan solusi yang menghasilkan 2, 3, dan 4 *cluster*.

2.2.1. *Nilai Reproducibility*

Nilai *reproducibility* untuk membandingkan solusi *cluster* yang menghasilkan jumlah cluster yang sama dapat diperoleh dengan cara berikut:

(1) Membuat tabulasi silang solusi *cluster* ke-i dengan solusi *cluster* ke-j untuk $i \neq j$. Misalkan terdapat dua solusi yang sama-sama menghasilkan 3 *cluster*. Tabulasi silang dapat dilihat sebagai berikut:

Tabel 1. Tabulasi Silang Solusi *Cluster*

Solusi 1,2	*cluster* 1	*cluster* 2	*cluster* 3	Total
cluster 1	0	3	0	3
cluster 2	0	0	2	2
cluster 3	2	0	0	2
Total	2	3	2	7

Misal solusi *cluster* 1 dan solusi *cluster* 2 yang menghasilkan solusi yang sama namun tampak berbeda. Perbedaan tersebut menimbulkan masalah ketika dilakukan tabulasi silang antar solusi sehingga mempengaruhi nilai *reproducibility*. Untuk mengatasi hal tersebut perlu dilakukan penukaran posisi kolom ke posisi yang bersesuaian dengan posisis baris yang mempunyai nilai terbesar pada kolom tersebut sehingga dapat memaksimalkan jumlah diagonal. Setelah dilakukan penukaran posisi kolom maka tabulasi silang antara solusi *cluster* 1 dan solusi *cluster* 2 berubah seperti Tabel 2.

Tabel 2. Tabulasi Silang Solusi *Cluster* Setelah Penukaran Posisi Kolom

Solusi 1,2	*cluster* 1	*cluster* 2	*cluster* 3	Total
cluster 1	3	0	0	3
cluster 2	0	2	0	2
cluster 3	0	0	0	2
Total	3	2	2	7

(2) Menghitung nilai

$$R_{ij} = \text{ jumlah diagonal/jumlah objek.}$$

Sebagai contoh jumlah objek yang konsisten berada pada cluster tertentu pada solusi cluster 1 dan solusi cluster 2 adalah (3+2+2=7) dan nilai *reproducibility* berpasangan solusi cluster 1 dan solusi cluster 2 adalah 1.

(3) Menetukan nilai *reproducibility* total solusi *cluster* ke-i dengan cara sebagai berikut:

$$R_i = \frac{(R_{i1} + R_{i2} + \cdots + R_{ij})}{j}.$$

Sebagai contoh untuk mengetahui nilai *reproducibility* total solusi cluster 1 maka perlu dihitung nilai *reproducibility* solusi cluster 1 dan solusi cluster 2 sehingga diperoleh nilai *reproducibility* total solusi cluster 1 sebesar:

$$R_1 = \frac{(R_{12} + R_{13})}{2}.$$

Nilai *reproducibility* dihitung untuk membandingkan solusi dengan jumlah *cluster* yang sama. Untuk membandingkan solusi dengan jumlah yang berbeda, dilakukan penyesuaian terhadap *reproducibility* ini, yang dirumuskan sebagai:

$$RA = \frac{(kR) - 1}{k - 1}.$$

3. Metodologi

3.1. *Data*

Data yang digunakan pada penelitian ini adalah data indikator kesehatan lingkungan dari seluruh provinsi Indonesia tahun 2016. Data tersebut diperoleh dari Kementrian Kesehatan Republik Indonesia, yang meliputi 34 provinsi dan 8 peubah berskala numerik. Daftar peubah ditampilkan pada Tabel 3.

Tabel 3. Daftar peubah Penelitian

Peubah	Keterangan
X1	Persentase jumlah desa/kelurahan yang melaksanakan sanitasi total berbasis masyarakat
X2	Persentase kabupaten/kota yang menyelenggarakan tatanan kawasan sehat
X3	Persentase rumah tangga menurut sumber air minum layak
X4	Persentase rumah tangga yang memliki akses terhadap sanitasi layak menurut provinsi
X5	Persentase tempat-tempat umum (TTU) yang memenuhi syarat kesehatan
X6	Persentase tempat pengelolaan makanan (TPM) yang memenuhi syarat kesehatan
X7	Persentase rumah tangga yang menempati rumah layak huni menurut provinsi
X8	Persentase rumah tangga kumuh menurut provinsi

3.2. *Analisis Data*

Langkah-langkah analisis data pada penelitian ini adalah sebagai berikut:

(1) Melakukan eksplorasi data untuk melihat gambaran umum mengenai kesehatan lingkungan di Indonesia.
(2) Membentuk anggota *ensemble* dengan cara mengelompokkan data dengan menggunakan berbagai metode pengklasteran (metode hirarki dan tak berhirarki), dengan berbagai metode perbaikan jarak dan pentuan pusat *cluster* awal serta dengan berbagai jumlah *cluster*. Beberapa metode yang digunakan adalah:
   (a) Metode *cluster* berhirarki dengan metode pautan rataan sebagai metode perbaikan jarak.
   (b) Metode *cluster* berhirarki dengan metode pautan lengkap sebagai metode perbaikan jarak.
   (c) Metode *cluster* tak berhirarki dengan metode *Distance Based Starting Point* untuk menentukan k titik pusat awal.
   (d) Metode *cluster* tak berhirarki dengan metode *Density Based Starting Point* untuk menentukan k titik pusat awal.

(e) Metode *cluster* tak berhirarki dengan metode *Hierarchical Starting Point* untuk menentukan k titik pusat awal.

(3) Untuk kumpulan solusi yang mengandung dua *cluster* lakukan:

(a) Dari berbagai solusi yang diperoleh, dihitung nilai *reproducibility* dan ditentukan solusi terbaik sementara dengan 2, 3, dan 4 *cluster*. Solusi terbaik inilah yang akan menjadi acuan pemberhentian metode pengklasteran pada tahap sebelumnya.

(b) Mentransformasikan anggota ensemble ke dalam bentuk matriks indikator.

(c) *Cluster on clustering*, yakni mengelompokkan kembali objek berdasarkan kolom-kolom pada matriks indikator. Pengklasteran dilakukan dengan menggunakan metode k-*means* dengan berbagai metode pemilihan titik pusat *cluster* awal dengan berbagai jumlah cluster.

(d) Hitung kembali nilai *reproducibility* dari setiap solusi yang diperoleh, solusi dengan nilai textitreproducibility tertinggi adalah solusi terbaik sementara yang dapat diambil.

(e) Bila solusi yang diperoleh pada tahap d sama dengan solusi terbaik sementara yang diperoleh pada tahap sebelumnya, maka metode pengklasteran *cluster ensemble* dihentikan. Jika tidak, lakukan kembali langkah b dan c hingga tidak terjadi perubahan keanggotaan *cluster* dan solusi sebelumnya dijadikan solusi akhir *cluster*.

(4) Lakukan langkah 3 untuk kumpulan solusi dengan 3 dan 4 *cluster*

(5) Lakukan perhitungan *reproducibility adjusted* untuk memilih solusi *cluster* mana yang paling baik sebagai solusi akhir *consensus*.

4. Hasil dan Pembahasan

4.1. *Hasil Pengelompokan dengan Cluster Ensemble*

(a) Solusi 2 *cluster*
Solusi dua *cluster* yang akan dibentuk dari 34 objek pengamatan terhadap 8 peubah, dimana 34 objek tersebut akan menghasilkan 3 *cluster*, 22 provinsi pada *cluster* 1, dan 12 provinsi pada cluster 2. Nilai *reproducibility* pada solusi 2 *cluster* adalah 0,97, hal ini menunjukkan bahwa terdapat 97% dari seluruh provinsi di Indonesia yang konsisten berada dalam *cluster* tertentu untuk seluruh solusi.

(b) Solusi 3 *cluster*
Solusi tiga *cluster* yang akan dibentuk dari 34 objek pengamatan terhadap 8 peubah, dimana 34 objek tersebut akan menghasilkan *cluster* 1, *cluster* 2, dan *cluster* 3. Solusi *consensus* dengan 3 *cluster* ini menghasilkan 16 provinsi pada *cluster* 1, 10 provinsi pada *cluster* 2, dan 8 provinsi pada *cluster* 3. Nilai *reproducibility* pada solusi 3 *cluster* adalah 0,88, hal ini menunjukkan bahwa terdapat 88,1% dari seluruh provinsi di Indonesia yang konsisten berada dalam *cluster* tertentu untuk seluruh solusi.

(c) Solusi 4 *cluster*
Solusi empat *cluster* yang akan dibentuk dari 34 objek pengamatan terhadap

8 peubah, dimana 34 objek tersebut akan menghasilkan *cluster* 1, *cluster* 2, *cluster* 3, dan *cluster* 4. *cluster* 1 terdiri dari 18 anggota *cluster*, *cluster* 2 terdiri dari 9 anggota *cluster*, *cluster* 3 terdiri dari 4 anggota *cluster*, dan *cluster* 4 terdiri dari 3 anggota cluster. Nilai *reproducibility* pada solusi 4 *cluster* adalah 0,77, hal ini menunjukkan bahwa terdapat 77,0% dari seluruh provinsi di Indonesia yang konsisten berada dalam *cluster* tertentu untuk seluruh solusi.

Nilai *reproducibility adjusted* yang diperoleh solusi pengklasteran terbaik adalah solusi dengan dua cluster, karena memiliki nilai *reproducibility adjusted* yang paling tinggi.

5. Kesimpulan

Penerapan analisis *cluster ensemble* untuk mengelompokkan provinsi di Indonesia dilakukan berdasarkan delapan peubah indikator kesehatan lingkungan. Dalam proses pengklasteran objek ke dalam 2, 3, dan 4 *cluster*, hasil analisis diperoleh nilai *reproducibility adjusted* sebesar 0,94 untuk solusi 2 cluster, 0,82 untuk solusi 3 *cluster*, dan 0,69 untuk solusi 4 *cluster*. Dapat disimpulkan bahwa dari nilai *reproducibility adjusted* yang diperoleh solusi pengklasteran terbaik adalah solusi dengan 2 *cluster*, dimana anggota dari *cluster* 1 merupakan provinsi dengan lingkungan sehat yang lebih baik dibandingkan anggota dari *cluster* 2.

Daftar Pustaka

[1] Iam-on N, Garret S. 2010. LinkCluE: A MATLAB Package for Link- Based *Cluster* Ensemble *Journal of Statistical Software*. **36**(9): 1 – 3

[2] Johson R.A, and D. W Winchern. 2007. *Applied Multivariate Statistical Analysis*. New Jersey, Prentice Hall

[3] Kementrian Kesehatan Republik Indonesia.2016. *Profil Kesehatan Indonesia Tahun 2016*. Kementrian Kesehatan Republik Indonesia, Jakarta

[4] Mattjik A.A, Sumertajaya IM. 2011. *Sidik Peubah Ganda dengan Menggunakan SAS*. Wibawa GNA, Hadi AF, editor. Bogor (ID):IPB Press

[5] Orme,B. dan Johsons,R.2008. Improving *K-Means Cluster Analysis : Ensemble Analysis Instead of Highest Reproducibility Replicates*. Sawtooth software

[6] Pokok-Pokok Hasil RISKESDAS 2013. http://terbitan.litbang.depkes.go.id. (24 Februari 2016).

[7] Rachmatin D. 2014. Aplikasi metode-metode *agglomerative* dalam analisis klaster pada data tingkat polusi udara. *Jurnal Ilmiah Program Studi Matematika STKIP Siliwangi Bandung*. **3**(2): 133 – 149

[8] Report of WHO Technical Consultation. WHO/CDS/RBM/2001.35. Geneva, WHO 2001

[9] Strehl A, Gosh J. 2012. A Knowledge Reuse Framework for Combining Partitionings.*The Journal of Machine learning Research*. **3**(1): 583 - 586