Penerapan Metode *Hierarchical Clustering Multiscale Bootstrap* untuk Pengelompokan Indikator Indeks Pembangunan Manusia Tahun 2021 di Jawa Barat

Sophia Annisa Faisal*, Nur Azizah Komara Rifai

Prodi Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Islam Bandung, Indonesia.

*sophiaafaisal14@gmail.com, muslimahstatistician@gmail.com

Abstract. Cluster analysis is a technique for grouping objects that have the same characteristics into one group and between different groups. In general there are two methods, namely hierarchical and non-hierarchical. The average linkage method is one of the methods in the hierarchical cluster analysis method that can be used to group data, one of which is the Human Development Index (HDI) data. This study uses HDI indicator data in West Java in 2021. The average linkage method only provides solutions based on a measure of proximity, so this study uses the multiscale bootstrap method to obtain the validity of the groups formed. There are four clusters formed by the average linkage method. Of the four groups formed, there is one valid cluster, namely the fourth cluster which consists of the group with the highest average HDI score, namely Bandung City, Bekasi City, and Depok City.

Keywords: *Average Linkage, Hierarchical Cluster Analysis, Human Development Index, Multiscale Bootstrap.*

Abstrak. Analisis *cluster* adalah teknik untuk mengelompokkan objek-objek yang memiliki karakteristik sama ke dalam satu kelompok dan antar kelompok berbeda. Secara umum terdapat dua metode yaitu hierarki dan non-hierarki. Metode *average linkage* merupakan salah satu metode pada analisis *cluster* metode hierarki yang dapat digunakan untuk mengelompokkan data, salah satunya yaitu data Indeks Pembangunan Manusia (IPM). Penelitian ini menggunakan data indikator IPM di Jawa Barat Tahun 2021. Metode *average linkage* hanya memberikan solusi berdasarkan ukuran kedekatan jarak, sehingga pada penelitian ini menggunakan metode *multiscale bootstrap* untuk memperoleh validitas dari kelompok yang terbentuk. Terdapat empat *cluster* yang terbentuk dengan metode *average linkage*. Dari keempat kelompok yang terbentuk, terdapat satu *cluster* yang valid yaitu *cluster* keempat yang terdiri dari kelompok dengan nilai rata-rata IPM tertinggi yaitu Kota Bandung, Kota Bekasi, dan Kota Depok.

Kata Kunci: *Analisis Cluster Hierarki, Average Linkage, Indeks Pembangunan Manusia, Multiscale Bootstrap*.

Corresponding Author
Email: muslimahstatistician@gmail.com

A. Pendahuluan

Pengelompokan objek-objek yang memiliki karakteristik sama ke dalam satu kelompok dan antar kelompok berbeda satu sama lain adalah analisis cluster. Secara umum terdapat dua metode pengelompokan data dalam analisis cluster, yaitu hierarki dan non-hierarki. Beberapa metode dalam analisis cluster hierarki yaitu metode single linkage, complete linkage, centroid linkage, average linkage dan ward. Sedangkan metode non-hierarki diantaranya adalah k-means, k-median, dan fuzzy. Kelebihan metode hierarki adalah proses pengolahan data yang singkat dan hasil yang berbentuk tingkatan sehingga menghemat waktu.

Penggunaan metode hierarki dengan pengukuran jarak hanya didasarkan pada ukuran kemiripan yang digunakan, sehingga validitas dalam pengelompokannya tidak ditemukan. Oleh karena itu, diperlukan metode lain untuk mengatasi permasalahan ini, yaitu salah satunya dengan menggunakan multiscale bootstrap. Menurut Efron & Tibshirani (1998), metode multiscale bootstrap bekerja dengan pendekatan bootstrap resampling untuk setiap kelompok yang terbentuk. Dalam penerapan analisis cluster dapat menggunakan berbagai macam data yang memiliki skala metrik maupun non-metrik, salah satu contoh data skala metrik yaitu data Indeks Pembangunan Manusia (IPM).

Pembangunan yang dilakukan oleh pemerintah bertujuan untuk mensejahterakan rakyat sehingga untuk mengukur keberhasilan pembangunan digunakan salah satu indikator yaitu Indeks Pembangunan Manusia (IPM). Bagi Indonesia, IPM adalah ukuran kinerja pemerintah juga dipakai sebagai alokator penentuan Dana Alokasi Umum (DAU). Tinggi rendahnya nilai IPM menandakan bahwa suatu daerah berhasil atau tidak dalam pembangunan. Pembangunan yang dilakukan perlu didukung oleh strategi perencanaan yang baik. IPM Provinsi Jawa Barat di tahun 2021 berada pada peringkat 10 besar dan mengalami peningkatan dari 66,15 pada tahun 2010 menjadi 72,45 pada tahun 2021. Meskipun demikian, nilai IPM di tiap Kabupaten/Kota di Jawa Barat berbeda-beda yang menandakan bahwa strategi perencanaan tiap daerah tidak sama. Misalnya, Kabupaten Cianjur memiliki nilai IPM terendah yaitu 65,56 maka dari itu pemerintah Kabupaten setempat menyiapkan beberapa program salah satunya membangun Program Kegiatan Belajar Masyarakat (PKBM). Pengelompokan Kabupaten/Kota di Provinsi Jawa Barat berdasarkan indikator IPM perlu dilakukan untuk mengetahui Kabupaten/Kota yang mirip.

Berdasarkan latar belakang yang telah diuraikan, maka perumusan masalah dalam penelitian ini sebagai berikut: "Bagaimana hasil pembentukan *cluster* dengan menggunakan metode *average linkage?*" dan "Bagaimana validitas dari hasil *cluster* yang telah terbentuk?". Selanjutnya, tujuan dalam penelitian ini sebagai berikut: "Mendapatkan hasil pembentukan *cluster* dengan menggunakan metode *average linkage*." dan "Mengetahui validitas dari hasil *cluster* yang telah terbentuk.".

B. Metodologi Penelitian

Standarisasi Data

Standarisasi data dilakukan untuk menghindari masalah yang akan dihasilkan dari penggunaan nilai skala yang berbeda antar objek. Standarisasi data yang paling umum adalah konversi setiap nilai objek terhadap nilai standar atau *z-score* dengan melakukan substraksi nilai tengah dan membaginya dengan standar deviasi tiap objek. Rumus standarisasi untuk setiap objek (Walpole & Myers, 1995):

$$Z = \frac{x_i - \bar{x}}{s}$$

dimana:
Z = standarisasi data
x_i = data ke-i
$\bar{x}$ = rata-rata keseluruhan data setiap objek
s = standar deviasi

Ukuran Jarak

Pada analisis *cluster*, ada tiga ukuran yang dapat digunakan untuk mengukur kesamaan antar objek salah satunya ukuran kedekatan. Ukuran kedekatan adalah ukuran yang paling sering

diaplikasikan untuk data berskala metrik (interval atau rasio). Sebenarnya ukuran ini merupakan ukuran ketidakmiripan, dimana jarak yang besar mengindikasikan sedikit kesamaan, dan jarak yang pendek mengindikasikan banyak kesamaan. Ukuran ini memiliki fokus terhadap besarnya nilai atau objek sehingga memiliki kesamaan nilai walaupun polanya berbeda. Salah satu ukuran kedekatan ini adalah jarak *Euclidean*.

Jarak *Euclidean* adalah jarak yang diukur lurus antara sentral fasilitas satu dengan sentral fasilitas lainnya. Metode pengukuran ini sering digunakan. Contoh aplikasi dari jarak *euclidean* misalnya pada analisis *cluster* metode hierarki yaitu *single linkage, complete linkage, average linkage,* dan *centroid linkage.* Menurut Tiskadewi (2017), jarak *euclidean* memiliki kelebihan yaitu tingkat penentuan kesamaannya lebih tinggi jika dibandingkan dengan metode lainnya. Untuk menyimpulkan jarak *Euclidean* sentral fasilitas satu dengan lainnya menggunakan rumus sebagai berikut:

$$\boldsymbol{d_{ij}} = \left[\left(\boldsymbol{x_i} - \boldsymbol{x_j}\right)^{\mathbf{2}} + \left(\boldsymbol{y_i} - \boldsymbol{y_j}\right)^{\mathbf{2}}\right]^{\frac{\mathbf{1}}{\mathbf{2}}}$$

dimana:
$\boldsymbol{x_i}$ = koordinat *x* pada pusat fasilitas *i*
$\boldsymbol{y_i}$ = koordinat *y* pada pusat fasilitas *i*
$\boldsymbol{x_j}$ = koordinat *x* pada pusat fasilitas *j*
$\boldsymbol{y_j}$ = koordinat *y* pada pusat fasilitas *j*
$\boldsymbol{d_{ij}}$ = jarak antara pusat fasilitas *i* dan *j*

Average Linkage
Metode ini menghitung jarak rata-rata antara dua *cluster* dengan semua pasangan objek dimana salah satu anggota dari pasangan dimiliki oleh setiap *cluster*. Langkah-langkah metode ini (Johnson & Wichern, 1992):

1. Menemukan jarak terkecil dalam $\boldsymbol{D} = \{\boldsymbol{d_{ik}}\}$.
2. Menggabungkan objek yang sesuai, misalkan *U* dan *V* untuk mendapatkan *cluster* (*UV*).
3. Menghitung jarak antar *cluster* (UV) dan *cluster W* lainnya dengan cara:

$$\boldsymbol{d_{(UV)W}} = \frac{\sum_i \sum_k \boldsymbol{d_{ik}}}{\boldsymbol{N_{(UV)} N_W}}$$

dimana:
$\boldsymbol{d_{ik}}$= jarak antar objek *i* dalam *cluster* (*UV*) dan objek *k* dalam *cluster W*
$\boldsymbol{N_{(UV)}}$ = jumlah objek pada *cluster* (*UV*)
$\boldsymbol{N_W}$ = jumlah objek pada *cluster W*

Profiling Cluster
Menurut Sugiyono (2007), statistika deskriptif adalah statistik yang digunakan untuk menganalisis data dengan menjelaskan data yang diteliti tanpa membuat kesimpulan yang berlaku untuk umum. Bentuk penyajian hasil analisis deskriptif ini bergantung dari jenis atau skala dari objek yang sedang dianalisis.

Menurut Supranto (2000), salah satu bagian dari statistika deskriptif yaitu ukuran pemusatan. Ukuran pemusatan adalah suatu ukuran yang dapat menggambarkan data secara menyeluruh. Salah satu ukuran pemusatan ini adalah rata-rata. Nilai rata-rata adalah nilai yang mewakili kelompok data. Langkah ini dilakukan dengan menghitung nilai rata-rata dari setiap variabel pada *cluster*.

Berikut merupakan cara menghitung rata-rata:

$$\overline{\boldsymbol{x}} = \frac{\mathbf{1}}{\boldsymbol{n}} \sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{n}} \boldsymbol{x_i}$$

dimana:
$\overline{\boldsymbol{x}}$ = rata-rata
$\boldsymbol{n}$ = jumlah pengamatan ke-*n*
$\boldsymbol{x_i}$ = nilai objek pengamatan ke-*i*

Metode *Bootstrap* dan Multiscale Bootstrap

Menurut Efron (1979), metode *bootstrap* adalah suatu teknik yang berdasarkan *resampling* data sampel dengan syarat pengembalian pada datanya dalam menyelesaikan statistik ukuran suatu sampel dengan harapan sampel tersebut dapat mewakili data populasinya. Menurut Efron & Tibshirani (1993), pada metode ini dilakukan pengembalian sampel yang disebut *resampling bootstrap*.

Untuk melihat validitas dari hasil analisis *cluster* hirarki dapat menggunakan *multiscale bootstrap*. Menurut Anuraga (2015), metode ini bergerak dengan pendekatan *resampling bootstrap* untuk setiap *cluster* hingga didapatkan dua nilai *p-values* yaitu *probability bootstrap* (BP) *value* dan *approximately unbiassed* (AU) *p-values*. Perbedaan antara metode *bootstrap* dan metode *multiscale bootstrap* adalah panjang urutan replikasi *bootstrap*. Berikut langkah-langkah menghitung nilai-*p* dari uji AU (Shimodaira, 2002):

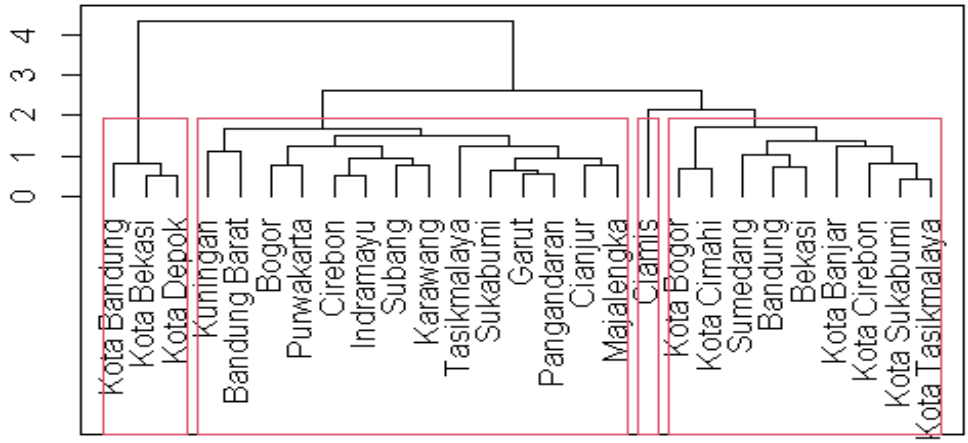
1. Tentukan konstanta penskalaan $\boldsymbol{r_1, \dots, r_K}$ dan jumlah replikasi $\boldsymbol{B_1, \dots, B_K}$ untuk $K \geq 2$ kumpulan replikasi *bootstrap.*
2. Buat replikasi *bootstrap* $\boldsymbol{B_k}$ dengan panjang urutan $\boldsymbol{N' = r_k N}$ untuk $k = 1, \dots, K$
$$\boldsymbol{X^{*1}(r_k), \dots, X^{*B_k}(r_k)}$$
Dan hitung nilai BP
$$\boldsymbol{BP_1 = \#\frac{\left\{X^{*b}(r_k) = X; b = 1, \dots, B_k\right\}}{B_k}}$$
Pertama, hasilkan $\boldsymbol{X^{*b}, b = 1, \dots, B_k}$. Kemudian hitung replikasi *bootstrap* menggunakan rumus:
$$\boldsymbol{X^{*b}(r_k) = \bar{X} + \sqrt{r_0/r_k}\left(X^{*b} - \bar{X}\right), b = 1, \dots, B_k}$$
dimana $\boldsymbol{r_0 = 1}$ dan $\boldsymbol{\bar{X} = \sum_{b=1}^{B_k} Y^{*b}/B_k}$
3. Perkirakan *v* dan *c* dengan metode kuadrat terkecil tertimbang (WLS) yaitu dengan meminimalkan jumlah sisa kuadrat (RSS):
$$\boldsymbol{RSS(v, c) = \sum_{k=1}^{K} V_k^{-1}\left\{v\sqrt{r_k} + \frac{c}{\sqrt{r_k}} - \Phi^{-1}[1 - BP_1(r_k)]\right\}^2}$$
Dimana bobot untuk setiap *k* adalah kebalikan dari varians yang diberikan:
$$\boldsymbol{V_k = \frac{BP_1(r_k)[1 - BP_1(r_k)]}{\phi\{\Phi^{-1}[BP_1(r_k)]\}^2 B_k}}$$
$\boldsymbol{\Phi^{-1}(.)}$ adalah fungsi kuantil dan untuk memperolehnya menggunakan kalkulator distribusi khusus. $\boldsymbol{\phi(.)}$ adalah fungsi kerapatan.
4. Hitung approximately unbiassed (AU) p-values dengan rumus:
$$\boldsymbol{AU_1 = 1 - \Phi(v - c)}$$
dimana:
X = sampel data asli
X^* = sampel bootstrap dari *X*
N = jumlah sampel data asli
N' = jumlah sampel *bootstrap*
r = panjang urutan relatif dari ulangan *bootstrap*
K = jumlah simulasi *bootstrap*
v, c = estimasi parameter dari setiap *cluster*
Φ = fungsi distribusi normal standar

C. Hasil Penelitian dan Pembahasan

Average Linkage

Pada Gambar 1 diketahui bahwa terdapat empat cluster yang terbentuk. Cluster pertama terdiri dari Kabupaten Kuningan, Kabupaten Bandung Barat, Kabupaten Bogor, Kabupaten Purwakarta, Kabupaten Cirebon, Kabupaten Indramayu, Kabupaten Subang, Kabupaten Karawang, Kabupaten Tasikmalaya, Kabupaten Sukabumi, Kabupaten Garut, Kabupaten

Pangandaran, Kabupaten Cianjur, dan Kabupaten Majalengka. Cluster kedua terdiri dari Kota Bogor, Kota Cimahi, Kabupaten Sumedang, Kabupaten Bandung, Kabupaten Bekasi, Kota Banjar, Kota Cirebon, Kota Sukabumi, dan Kota Tasikmalaya. Cluster ketiga terdiri dari Kabupaten Ciamis. Dan cluster keempat terdiri dari Kota Bandung, Kota Bekasi, dan Kota Depok.



Gambar 1. Dendrogram Average Linkage

Profiling Cluster

Setelah cluster terbentuk, langkah selanjutnya adalah memberi ciri spesifik untuk masing-masing cluster yaitu dengan melihat nilai rata-rata objek tiap variabel yang terdapat dalam cluster.

Tabel 1. Profiling Cluster

Klaster	IPM			
	AHH	**HLS**	**RLS**	**PpK**
1	71.62	12.16	7.53	9567
2	72.97	13.27	9.64	11010
3	72.02	14.20	7.90	9259
4	74.76	14.08	11.25	16106

dimana:

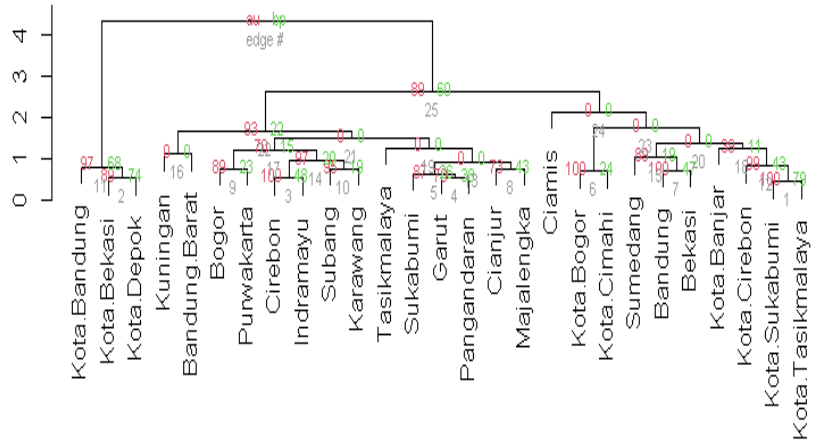
= sangat tinggi = sedang

= tinggi = rendah

Pada Tabel 1 diketahui bahwa cluster pertama, memiliki tingkat AHH, HLS, dan RLS yang “rendah” dengan PpK yang “sedang”. Cluster kedua, memiliki tingkat AHH, RLS, dan PpK yang “tinggi” dengan HLS yang “sedang”. Cluster ketiga, memiliki tingkat AHH dan RLS yang “sedang”, HLS yang “sangat tinggi” dan PpK yang “rendah”. Dan Cluster keempat, memiliki tingkat AHH, RLS, dan PpK yang “sangat tinggi” dengan HLS yang “tinggi”.

Hierarchical Clustering Multiscale Bootstrap

Pada Gambar 2 disajikan hasil analisis hierarchical clustering multiscale bootstrap. Dari hasil analisis ini dapat kita ketahui validitas dari cluster.



Gambar 2. Dendrogram Multiscale Bootstrap

Gambar 2 menunjukkan dendrogram dengan nilai Approximately Unbiassed (AU) p-values dan Probability Bootstrap (BP) value untuk 27 Kabupaten/Kota di Provinsi Jawa Barat tahun 2021. Cluster yang memiliki nilai AU ≥ 0,95 dikatakan sebagai cluster yang valid.

Pada cluster pertama, nilai AU diperoleh sebesar 0,93 maka dapat disimpulkan bahwa cluster pertama tidak valid. Kemudian pada cluster kedua, nilai AU diperoleh sebesar 0,00 maka dapat disimpulkan bahwa cluster kedua tidak valid. Kemudian pada cluster ketiga, nilai AU diperoleh sebesar 0,00 maka dapat disimpulkan bahwa cluster ketiga tidak valid. Dan pada cluster keempat, nilai AU diperoleh sebesar 0,97 maka dapat disimpulkan bahwa cluster keempat valid.

D. Kesimpulan

Berdasarkan pembahasan dalam penelitian ini, peneliti menyimpulkan beberapa hasil penelitian sebagai berikut:

1. Terdapat empat kelompok yang terbentuk dengan menggunakan metode average linkage. Berdasarkan hasil pengelompokan, cluster 1 memiliki status IPM yang "rendah" terdiri dari Kuningan, Bandung Barat, Bogor, Purwakarta, Cirebon, Indramayu, Subang, Karawang, Tasikmalaya, Sukabumi, Garut, Pangandaran, Cianjur, dan Majalengka. Cluster 2 memiliki status IPM yang "tinggi" terdiri dari Kota Bogor, Kota Cimahi, Sumedang, Bandung, Bekasi, Kota Banjar, Kota Cirebon, Kota Sukabumi, dan Kota Tasikmalaya. Cluster 3 memiliki status IPM yang "sedang" terdiri dari Kabupaten Ciamis. Dan cluster 4 memiliki status IPM yang "sangat tinggi" terdiri dari Kota Bandung, Kota Bekasi, dan Kota Depok.
2. Dari keempat cluster yang terbentuk, terdapat satu cluster yang valid yaitu cluster 4 dengan nilai AU sebesar 0,97.

Acknowledge

Penelitian ini dapat terlaksana dengan baik atas bantuan, bimbingan, saran dan dukungan dari berbagai pihak. Oleh karena itu, peneliti mengucapkan terima kasih kepada Yang Maha Kuasa Allah SWT atas Rahmat dan Hidayah-Nya, kedua orang tua serta keluarga yang selalu mendoakan dan memberi semangat, Ibu Nur Azizah Komara Rifai, S.Si., M.Si., yang telah memberikan sumbangan pikiran, pengetahuan, ilmu, serta motivasi, para dosen Statistika Unisba yang telah memberikan ilmu pengetahuannya, dan teman-teman yang telah memberikan kritikan, bantuan, dan dorongannya.

Daftar Pustaka

[1] Anuraga, G. (2015, Mei). *Hierarchical Clustering Multiscale Bootstrap Untuk Pengelompokan Kemiskinan Di Jawa Timur*. Statistica, 1(3), 27-33. Retrieved February 2022.

[2] Efron, B. (1979). *Bootstrap Methods: Another Look At The Jacknife*. Annals of Statistics, 7(1), 1-26.

[3] Efron, B., & Tibshirani, R. (1993). *Book An Introduction to the Bootstrap, Monographs On Statistics and Applied Probability 57*. New York: Chapman and Hall (CRC).

[4] Efron, B., & Tibshirani, R. J. (1998). An Introduction to the Bootstrap. *New York: Chapman Hall*.

[5] Johnson, R. A., & Wichern, D. W. (1992). *Applied Multivariate Statistical Analysis, Fifth Edition*. New Jersey: Prentice Hall.

[6] Shimodaira, H. (2002, July). *An Approximately Unbiased Test of Phylogenetic Tree Selection*. Systematic Biology, 51(3), 492-508. doi:10.1080/10635150290069913

[7] Sugiyono. (2007). M*etode Penelitian Kuantitatif, Kualitatif dan R&D. Bandung: Alfabeta.* Retrieved January 5, 2023

[8] Supranto, J. (2000). *Teknik Sampling Untuk Survei Dan Eksperimen.* Bandung: PT. Rineka Cipta.

[9] Tiskadewi, N. (2017). *IDENTIFIKASI CITRA IRIS MATA DENGAN METODE KNN (K-Nearest Neighbor).* YOGYAKARTA: STMIK AKAKOM. Retrieved March 30, 2022, from https://eprints.utdi.ac.id/4974/2/2_135410017_BAB_I.pdf

[10] Walpole, R. E., & Myers, R. H. (1995). Ilmu Peluang dan Statistika untuk Insinyur dan Ilmuwan. Bandung: ITB.

[11] Wildan, Karyana, Yayat. (2021). Evaluasi Kesalahan Proyeksi Penduduk Tahun 2020 untuk Memproyeksikan Penduduk Tahun 2025 Provinsi Jawa Barat. Jurnal Riset Statistika 1(2). 92-98.