SPEKTRUM AKADEMIKA
Volume. 1, Nomor. 2, 2026

PERBANDINGAN K-MEANS DAN AGGLOMERATIVE CLUSTERING UNTUK PENGELOMPOKAN KINERJA REALISASI PAD OPD SURABAYA 2022–2024

Putri Amanda Khairunnisa[1,*] , Aviolla Terza Damaliana, S.Si., M.Stat.[1]
[1] Fakultas Ilmu Komputer, Program Studi Sains Data, Universitas Pembangunan Nasional "Veteran" Jawa Timur, Surabaya, Indonesia
Email: [1,*]putriamandakhairunnisa16@gmail.com, [1]aviolla.terza.sada@upnjatim.ac.id
(* Email Corresponding Author: putriamandakhairunnisa16@gmail.com)

Abstrak

Pendapatan Asli Daerah (PAD) Kota Surabaya dihimpun oleh beberapa Organisasi Perangkat Daerah (OPD) dengan kinerja realisasi yang beragam, tetapi pengelompokan kinerja tersebut secara kuantitatif dan lintas tahun belum banyak dilakukan. Penelitian ini membandingkan K-Means dan Agglomerative Clustering untuk mengelompokkan kinerja realisasi PAD OPD Surabaya tahun 2022–2024 dengan unit analisis OPD per tahun. Data diekstraksi dari tiga dokumen rincian realisasi APBD per organisasi, kemudian diaudit dan diagregasi pada akun 4.1.01–4.1.04. Audit menemukan 22 baris pendapatan ganda pada dokumen 2024 yang bernilai Rp1.303,06 miliar untuk akun PAD; baris tersebut dihapus sebelum analisis. Dari 47 rekaman OPD-tahun yang memiliki akun PAD, 44 observasi dianalisis karena tiga rekaman tidak memiliki anggaran PAD. Variabel utama adalah persentase realisasi PAD dan kontribusi PAD; pertumbuhan realisasi hanya diuji pada skenario 2023–2024 karena data 2021 tidak tersedia. Data distandardisasi dengan StandardScaler, jumlah cluster ditentukan melalui metode Elbow, Silhouette Score, Davies-Bouldin Index (DBI), Calinski-Harabasz Index (CHI), dan dendrogram. Pada k = 5, K-Means memperoleh Silhouette 0,615, DBI 0,376, dan CHI 205,94, sedangkan Agglomerative Ward memperoleh Silhouette 0,603, DBI 0,365, dan CHI 189,52. K-Means unggul pada dua dari tiga indeks, dan kesepakatan kedua algoritma sangat tinggi (Adjusted Rand Index 0,939; hanya satu observasi berbeda). Cluster yang terbentuk meliputi penghimpun dominan, realisasi jauh melampaui target, memenuhi target, di bawah target, dan realisasi rendah. Sebanyak 7 dari 14 OPD yang lengkap tiga tahun berpindah cluster, dengan kecenderungan bergeser ke cluster memenuhi target. Hasil ini menunjukkan K-Means sedikit lebih baik pada skenario utama, tetapi selisihnya kecil, sementara Dinas Perhubungan konsisten berada pada cluster realisasi rendah.

Kata Kunci: Agglomerative Clustering, Davies-Bouldin Index, K-Means, Pendapatan Asli Daerah, Silhouette Score, Organisasi Perangkat Daerah

Abstract

Local own-source revenue (Pendapatan Asli Daerah, PAD) of Surabaya City is collected by several regional apparatus organizations (OPD) whose realization performance varies, yet quantitative, multi-year grouping of that performance remains scarce. This study compares K-Means and Agglomerative Clustering for grouping the PAD realization performance of Surabaya OPDs in 2022–2024, using the OPD-year as the unit of analysis. Data were extracted from three budget realization reports itemized by organization, then audited and aggregated over accounts 4.1.01–4.1.04. The audit detected 22 duplicated revenue rows in the 2024 report worth IDR 1,303.06 billion in PAD accounts, which were removed. Of 47 OPD-year records holding PAD accounts, 44 observations were analyzed because three records had no PAD budget. The main variables are the PAD realization percentage and PAD contribution; realization growth was tested only in a 2023–2024 scenario because 2021 data were unavailable. Data were standardized with StandardScaler, and the number of clusters

Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

was selected using the Elbow method, Silhouette Score, Davies-Bouldin Index (DBI), Calinski-Harabasz Index (CHI), and a dendrogram. At k = 5, K-Means achieved a Silhouette of 0.615, DBI of 0.376, and CHI of 205.94, while Ward Agglomerative achieved 0.603, 0.365, and 189.52. K-Means performed better on two of three indices, and agreement between the algorithms was very high (Adjusted Rand Index 0.939; only one observation differed). The resulting clusters are a dominant collector, far above target, meeting target, below target, and low realization. Seven of the 14 OPDs observed in all three years changed clusters, mostly shifting toward the meeting-target cluster.

Keywords: *Agglomerative Clustering, Davies-Bouldin Index, K-Means, Local Own-Source Revenue, Silhouette Score, Regional Apparatus Organization*

1. PENDAHULUAN

Pendapatan Asli Daerah (PAD) merupakan indikator penting kemandirian fiskal pemerintah daerah karena menunjukkan kemampuan daerah membiayai urusan pemerintahan dari sumber yang digali sendiri. Undang-Undang Nomor 1 Tahun 2022 menegaskan bahwa PAD terdiri atas pajak daerah, retribusi daerah, hasil pengelolaan kekayaan daerah yang dipisahkan, dan lain-lain PAD yang sah [1]. Klasifikasi tersebut dioperasionalkan dalam struktur akun pendapatan 4.1.01 sampai 4.1.04 sebagaimana diatur dalam pedoman teknis pengelolaan keuangan daerah [2].

Kota Surabaya merupakan salah satu daerah dengan basis PAD terbesar di Indonesia. Berdasarkan dokumen rincian realisasi APBD yang dianalisis dalam penelitian ini, realisasi PAD seluruh Organisasi Perangkat Daerah (OPD) mencapai Rp5,32 triliun pada 2022, Rp5,77 triliun pada 2023, dan Rp6,11 triliun pada 2024. Meskipun demikian, rasio realisasi terhadap anggaran PAD tingkat kota masih berada di bawah 100%, yaitu 84,20%, 87,55%, dan 92,69% pada ketiga tahun tersebut. Kajian rasio keuangan sebelumnya juga mencatat rata-rata efektivitas PAD Surabaya periode 2019–2023 sebesar 87,97% dan mengategorikannya kurang efektif [3]. Angka agregat tersebut belum menjelaskan OPD mana yang menjadi penyumbang kesenjangan, karena kinerja realisasi di tingkat OPD sangat beragam: sebagian OPD melampaui target, sementara OPD lain hanya merealisasikan kurang dari 60% anggarannya [4].

Permasalahan utama yang diangkat adalah belum tersedianya pengelompokan kinerja realisasi PAD OPD yang objektif, berbasis data, dan dapat diikuti lintas tahun. Evaluasi yang lazim dilakukan berupa perbandingan rasio realisasi per OPD secara satu per satu, sehingga sulit menangkap pola gabungan antara tingkat capaian target dan besarnya peran OPD dalam struktur PAD kota. Selain itu, dokumen sumber yang tersedia dalam format PDF berukuran ratusan halaman memerlukan proses ekstraksi dan audit agar tidak terjadi kesalahan penghitungan. Solusi yang ditawarkan adalah penerapan teknik *clustering* sebagai metode *unsupervised learning* yang mengelompokkan objek berdasarkan kemiripan karakteristik tanpa label awal [5], [6]. Hasil pengelompokan dapat dijadikan dasar pemetaan prioritas pembinaan, penetapan target yang lebih realistis, dan pemantauan pergeseran kinerja OPD dari tahun ke tahun.

Sejumlah penelitian terkait telah menerapkan *clustering* pada data keuangan daerah. Aji dkk. menerapkan K-Means pada data realisasi PAD dan menunjukkan bahwa algoritma tersebut mampu memisahkan kelompok realisasi tinggi dan rendah [7]. Rahmadianissa mengelompokkan provinsi di Indonesia berdasarkan APBD 2021 menggunakan K-Means dan DBSCAN, dan menemukan bahwa K-Means menghasilkan dua cluster dengan DKI Jakarta sebagai cluster tersendiri, sedangkan DBSCAN memperlakukan provinsi ekstrem sebagai *noise* [8]. Wulaningrum dkk. memakai K-Means untuk mengelompokkan pemerintah provinsi berdasarkan kepemilikan aset tetap dan pertumbuhan ekonomi menjadi empat cluster [9]. Pada sisi metodologi, Abdalla membandingkan K-Means dan *agglomerative hierarchical clustering* pada dataset berukuran kecil dan menyimpulkan bahwa keunggulan algoritma sangat bergantung pada ukuran kemiripan yang digunakan [10]. Purnama dkk. membandingkan K-Means, *hierarchical clustering*, dan K-Medoids untuk segmentasi pasar dan memperoleh selisih Silhouette yang sangat kecil antaralgoritma [11], sedangkan Sholeh dan Aeni menunjukkan pentingnya kombinasi metode Elbow, Silhouette, dan Davies-Bouldin dalam menentukan jumlah cluster [12].

Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

Dari kajian tersebut terdapat beberapa celah penelitian. Pertama, studi *clustering* keuangan daerah umumnya menggunakan unit analisis wilayah (provinsi atau kabupaten/kota), bukan OPD sebagai pelaksana pemungutan PAD di dalam satu pemerintah daerah. Kedua, sebagian besar studi hanya menerapkan satu algoritma, atau membandingkan algoritma tanpa memastikan skenario data dan parameter yang identik. Ketiga, studi terdahulu jarang melaporkan audit kualitas data sumber, padahal dokumen anggaran berformat PDF berpotensi memuat duplikasi baris yang dapat menggandakan nilai realisasi. Keempat, analisis lintas tahun yang menelusuri perpindahan cluster setiap OPD belum banyak dilakukan, sehingga dinamika kinerja tidak terlihat.

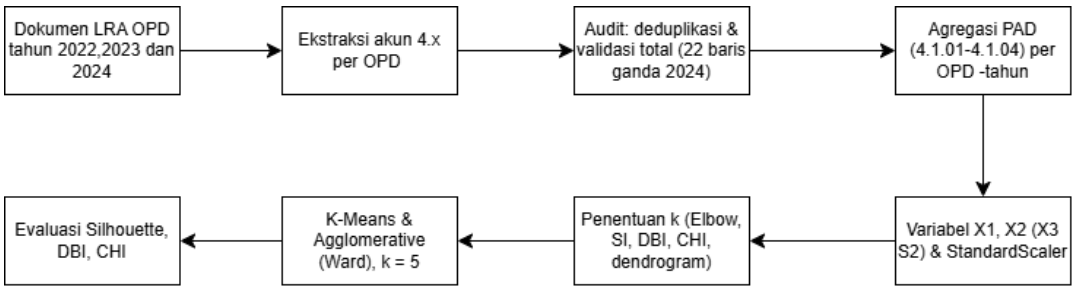
K-Means dipilih karena efisien dan banyak digunakan untuk data numerik berdimensi rendah [13], [14], sedangkan Agglomerative Clustering dipilih karena menghasilkan hierarki yang dapat divisualisasikan sebagai dendrogram dan tidak bergantung pada inisialisasi pusat cluster [15]. Kriteria Ward pada metode agglomeratif meminimalkan kenaikan jumlah kuadrat galat dalam cluster, sehingga fungsi objektifnya sebanding dengan K-Means dan perbandingan keduanya menjadi adil [16]. Evaluasi dilakukan dengan tiga indeks validitas internal yang saling melengkapi, yaitu Silhouette Score, Davies-Bouldin Index, dan Calinski-Harabasz Index.

Berdasarkan uraian tersebut, tujuan penelitian ini adalah (a) membangun dataset kinerja realisasi PAD OPD Kota Surabaya tahun 2022–2024 yang telah diaudit dari dokumen sumber; (b) membandingkan kinerja K-Means dan Agglomerative Clustering pada dataset dan skenario pengujian yang sama menggunakan Silhouette Score, Davies-Bouldin Index, dan Calinski-Harabasz Index; serta (c) menginterpretasikan karakteristik setiap cluster dan perpindahan cluster OPD selama 2022–2024. Hasil penelitian diharapkan memberi gambaran empiris tentang algoritma yang lebih sesuai untuk data kinerja pendapatan berukuran kecil dan menjadi masukan bagi pemerintah daerah dalam menyusun strategi optimalisasi PAD berbasis kelompok OPD.

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Penelitian dilaksanakan melalui delapan tahapan yang ditunjukkan pada Gambar 1, yaitu pengumpulan dokumen, ekstraksi akun pendapatan, audit data, agregasi PAD, pembentukan dan standardisasi variabel, penentuan jumlah cluster, penerapan kedua algoritma, serta evaluasi dan interpretasi. Seluruh tahapan diimplementasikan dalam tiga skrip Python yang dapat dijalankan ulang, sehingga setiap angka dalam artikel dapat ditelusuri ke halaman dan baris dokumen sumber.



Gambar 1. Diagram alur penelitian

2.2 Sumber Data dan Ekstraksi

Sumber data adalah tiga dokumen Pemerintah Kota Surabaya berjudul Rincian APBD menurut Urusan Pemerintahan Daerah, Organisasi, Program, Kegiatan, Sub Kegiatan, Kelompok, dan Jenis Pendapatan, Belanja, dan Pembiayaan untuk tahun anggaran 2022 (259 halaman), 2023 (269 halaman), dan 2024 (267 halaman) [4]. Teks diekstraksi menggunakan pdftotext dengan mode tata letak, kemudian baris akun pendapatan berpola 4.x.xx diurai dengan ekspresi reguler beserta kode organisasi, label unit pendapatan,

Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

nomor halaman, dan nomor baris. Ekstraksi menghasilkan 129 baris akun (34 baris pada 2022, 35 pada 2023, dan 60 pada 2024).

2.3 Audit dan Pra-pemrosesan Data

Audit mencakup konsistensi kode dan nama OPD, perubahan nomenklatur, data kosong, duplikasi, dan kemungkinan penghitungan ganda. Duplikasi didefinisikan sebagai baris dengan tahun, kode OPD, kode akun, dan nilai realisasi yang identik; dari setiap kelompok kembar dipertahankan satu baris yang memuat pagu anggaran terbesar. Hasil deduplikasi divalidasi terhadap baris JUMLAH PENDAPATAN DAERAH setiap OPD. Akun 4.2 (pendapatan transfer) dan 4.3 dikeluarkan karena bukan komponen PAD. PAD setiap OPD-tahun kemudian dihitung sebagai jumlah akun 4.1.01–4.1.04. Rekaman OPD-tahun dengan anggaran PAD nol dikeluarkan karena persentase realisasi tidak terdefinisi; penanganan ini dipilih daripada imputasi agar tidak menciptakan nilai yang tidak tercatat dalam dokumen.

Pengecualian anggaran nol diterapkan pada tingkat OPD-tahun, yaitu total anggaran PAD sama dengan nol. Baris akun dengan anggaran nol tetapi realisasi positif pada OPD yang total anggarannya positif tetap dipertahankan karena hanya tercatat satu kali dalam dokumen, misalnya retribusi RSUD Bhakti Dharma Husada 2024 sebesar Rp6,46 miliar. Baris tanpa pasangan ini, serta akun transfer antar daerah BPKAD yang hanya tercatat sekali, menjelaskan mengapa rasio total dokumen terhadap hasil pembersihan pada RSUD Bhakti Dharma Husada (1,96), Disbudporapar (1,98), dan BPKAD (1,75) tidak tepat 2,00.

2.4 Variabel Penelitian

Tiga kandidat variabel dirumuskan sesuai persamaan (1)–(3), dengan R adalah realisasi PAD, A adalah anggaran PAD, i adalah OPD, dan t adalah tahun. Persentase realisasi (X1) mengukur capaian target, kontribusi (X2) mengukur peran relatif OPD terhadap total realisasi PAD seluruh OPD pada tahun yang sama, dan pertumbuhan (X3) mengukur perubahan realisasi terhadap tahun sebelumnya.

$$X_{1,i,t} = \frac{R_{i,t}}{A_{i,t}} \times 100\% \quad (1)$$

$$X_{2,i,t} = \frac{R_{i,t}}{\sum_{j=1}^{m} R_{j,t}} \times 100\% \quad (2)$$

$$X_{3,i,t} = \frac{R_{i,t} - R_{i,t-1}}{R_{i,t-1}} \times 100\% \quad (3)$$

Variabel X3 untuk tahun 2022 tidak dapat dihitung karena data realisasi 2021 tidak tersedia. Oleh karena itu, skenario utama (S1) memakai X1 dan X2 untuk seluruh 44 observasi 2022–2024, skenario S2 memakai X1, X2, dan X3 untuk 29 observasi 2023–2024, dan skenario S3 memakai X1 dan log10(X2) sebagai uji sensitivitas terhadap kemencengan X2. Nilai absolut realisasi tidak dimasukkan sebagai variabel terpisah karena bersifat redundan dengan X2 pada tahun yang sama. Seluruh variabel distandardisasi dengan StandardScaler menurut persamaan (4), dengan μ dan σ adalah rata-rata dan simpangan baku variabel.

$$z = \frac{x - \mu}{\sigma} \quad (4)$$

2.5 Algoritma Clustering

K-Means mempartisi n observasi ke dalam k cluster dengan meminimalkan jumlah kuadrat jarak Euclidean setiap observasi terhadap centroid clusternya, sebagaimana persamaan (5) [13], [14]. Implementasi menggunakan inisialisasi k-means++ dengan n_init = 50 dan random_state = 42 untuk menghindari solusi lokal. Agglomerative Clustering memulai setiap observasi sebagai cluster tunggal, lalu menggabungkan dua cluster secara berulang [15]. Kriteria Ward pada persamaan (6) menggabungkan pasangan cluster A dan B yang menghasilkan kenaikan galat terkecil, dengan n adalah ukuran cluster dan μ adalah centroid [16], [17]. Linkage complete, average, dan single juga diuji sebagai analisis sensitivitas.

$$J = \sum_{c=1}^{k} \sum_{x \in C_c} \|x - \mu_c\|^2 \quad (5)$$

$$\Delta(A, B) = \frac{n_A n_B}{n_A + n_B} \|\mu_A - \mu_B\|^2 \quad (6)$$

Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

2.6 Penentuan Jumlah Cluster dan Evaluasi

Nilai k diuji pada rentang 2–8 untuk kedua algoritma dengan data terstandardisasi yang sama. Penentuan k mempertimbangkan kurva Elbow inersia K-Means, ketiga indeks validitas, dendrogram Ward, dan keterinterpretasian cluster. Silhouette Score pada persamaan (7) membandingkan rata-rata jarak observasi ke anggota clusternya sendiri, a(i), dengan rata-rata jarak ke cluster terdekat lainnya, b(i); nilainya berkisar −1 sampai 1 dan semakin tinggi semakin baik [18], [19]. Davies-Bouldin Index pada persamaan (8) mengukur rasio sebaran dalam cluster, s, terhadap jarak antarcentroid, d, sehingga semakin rendah semakin baik [20]. Calinski-Harabasz Index pada persamaan (9) membandingkan dispersi antarcluster, B, dengan dispersi dalam cluster, W, sehingga semakin tinggi semakin baik [21].

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{7}$$

$$\mathrm{DBI} = \frac{1}{k}\sum_{i=1}^{k} \max_{j \neq i}\left(\frac{s_i + s_j}{d_{ij}}\right) \tag{8}$$

$$\mathrm{CHI} = \frac{\frac{\mathrm{tr}(B_k)}{k-1}}{\frac{\mathrm{tr}(W_k)}{n-k}} \tag{9}$$

Kesepakatan partisi kedua algoritma diukur dengan Adjusted Rand Index (ARI), dan stabilitas K-Means diuji pada 30 nilai random_state. Label C1 diberikan pada cluster dengan rata-rata X2 tertinggi, sedangkan C2–C5 diurutkan menurut rata-rata X1 menurun. Pengolahan menggunakan Python dengan pustaka pandas, scikit-learn, SciPy, dan Matplotlib [22]–[24].

3. HASIL DAN PEMBAHASAN

3.1 Hasil Audit Data

Pemeriksaan nomenklatur menunjukkan bahwa ke-60 organisasi memiliki kode dan nama yang identik pada dokumen 2022, 2023, dan 2024, sehingga tidak diperlukan pemetaan perubahan nomenklatur. Dari 60 organisasi tersebut, hanya 16 OPD yang memiliki akun PAD. Kecamatan, sekretariat, dan badan pendukung lainnya tidak mencatat pendapatan, sehingga tidak dimasukkan dalam analisis. Temuan audit terpenting adalah adanya penghitungan ganda pada dokumen 2024. Pada 13 OPD, akun pendapatan muncul dua kali: satu kali pada blok tanpa label unit dengan anggaran bertanda "-" dan satu kali pada blok berlabel unit pendapatan dengan pagu anggaran, dengan nilai realisasi identik. Sebanyak 22 baris teridentifikasi sebagai duplikat, terdiri atas akun PAD senilai Rp1.303,06 miliar dan akun transfer senilai Rp2.813,61 miliar.

Validasi memperlihatkan bahwa baris JUMLAH PENDAPATAN DAERAH pada dokumen 2024 justru sama dengan jumlah seluruh baris sebelum deduplikasi, sehingga total pada dokumen sumber ikut tergandakan. Untuk sepuluh OPD, rasio total dokumen terhadap total hasil pembersihan tepat 2,00, misalnya Dinas Kesehatan dengan total tertulis Rp355,66 miliar, padahal realisasi PAD setelah deduplikasi Rp177,83 miliar. Tanpa deduplikasi, realisasi PAD 2024 akan tercatat Rp7.408,14 miliar atau 21,3% lebih tinggi daripada nilai sebenarnya. Dokumen 2022 dan 2023 tidak memuat duplikasi; selisih anggaran hasil hitung terhadap total dokumen bernilai nol untuk seluruh OPD. Ringkasan audit disajikan pada Tabel 1.

Tabel 1. Ringkasan ekstraksi dan audit data

Uraian	2022	2023	2024
Jumlah halaman dokumen	259	269	267
Baris akun pendapatan (4.x.xx) terekstraksi	34	35	60
Baris duplikat (dihapus)	0	0	22
OPD dengan akun PAD	16	15	16
Rekaman dikeluarkan (anggaran PAD nol)	1	0	2
Observasi OPD-tahun dianalisis	15	15	14

Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

Anggaran PAD seluruh OPD (Rp miliar)	6.316,35	6.592,42	6.586,66
Realisasi PAD seluruh OPD (Rp miliar)	5.318,27	5.771,82	6.105,09
Rasio realisasi PAD tingkat kota (%)	84,20	87,55	92,69

Label unit pendapatan juga tidak selalu konsisten. Seluruh PAD Dinas Kesehatan dicatat di bawah label satu puskesmas yang berbeda setiap tahun (Puskesmas Bangkingan, Tembok Dukuh, dan Dr. Soetomo), sehingga agregasi dilakukan pada tingkat kode OPD, bukan label unit. Tiga rekaman dikeluarkan karena anggaran PAD bernilai nol, yaitu Dinas Perpustakaan dan Kearsipan 2022 (realisasi Rp5,32 juta), Dinas Koperasi Usaha Kecil dan Menengah dan Perdagangan 2024 (realisasi Rp12,58 juta), dan Dinas Perpustakaan dan Kearsipan 2024 (anggaran dan realisasi nol). Dengan demikian, dataset akhir berisi 44 observasi OPD-tahun dari 16 OPD, dengan 14 OPD teramati lengkap selama tiga tahun.

3.2 Karakteristik Variabel

Tabel 2 menunjukkan distribusi ketiga variabel. Rata-rata X1 meningkat dari 82,20% pada 2022 menjadi 100,51% pada 2023 dan 95,99% pada 2024. Sebaran X1 relatif simetris (kemencengan 0,31) dengan rentang 36,25–165,98%. Sebaliknya, X2 sangat menceng (kemencengan 3,49) karena Badan Pendapatan Daerah (Bapenda) yang mengelola pajak daerah menyumbang 78,55%, 79,24%, dan 78,55% realisasi PAD pada ketiga tahun, sementara median kontribusi OPD hanya sekitar 0,9–1,1%. Variabel X3 juga sangat menceng (kemencengan 4,98) akibat lonjakan realisasi Dinas Perindustrian dan Tenaga Kerja pada 2023 sebesar 1.271,39%, dari Rp0,49 miliar menjadi Rp6,72 miliar.

Tabel 2. Statistik deskriptif variabel per tahun: rata-rata / median (minimum s.d. maksimum)

Variabel	2022 (n = 15)	2023 (n = 15)	2024 (n = 14)
X1 (%)	82,20 / 80,39 (36,25 s.d. 128,89)	100,51 / 101,17 (48,73 s.d. 165,98)	95,99 / 96,18 (48,71 s.d. 145,97)
X2 (%)	6,67 / 0,87 (0,00 s.d. 78,55)	6,67 / 0,92 (0,01 s.d. 79,24)	7,14 / 1,07 (0,01 s.d. 78,55)
X3 (%)	–	105,35 / 10,35 (-43,97 s.d. 1.271,39)	24,19 / 7,05 (-27,74 s.d. 200,00)

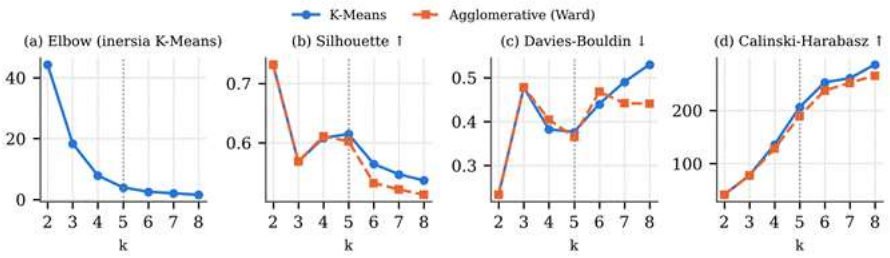
Catatan: X3 tahun 2022 tidak dapat dihitung karena tidak tersedia data realisasi 2021.

Korelasi Pearson antarvariabel sangat lemah (X1–X2 sebesar −0,052; X1–X3 sebesar −0,005; X2–X3 sebesar −0,086). Hal ini menunjukkan bahwa ketiga variabel membawa informasi yang berbeda dan tidak redundan. Namun, kemencengan ekstrem X2 berarti bahwa setelah standardisasi, keragaman X2 sebagian besar ditentukan oleh tiga observasi Bapenda. Implikasi ini perlu diperhatikan dalam interpretasi hasil.

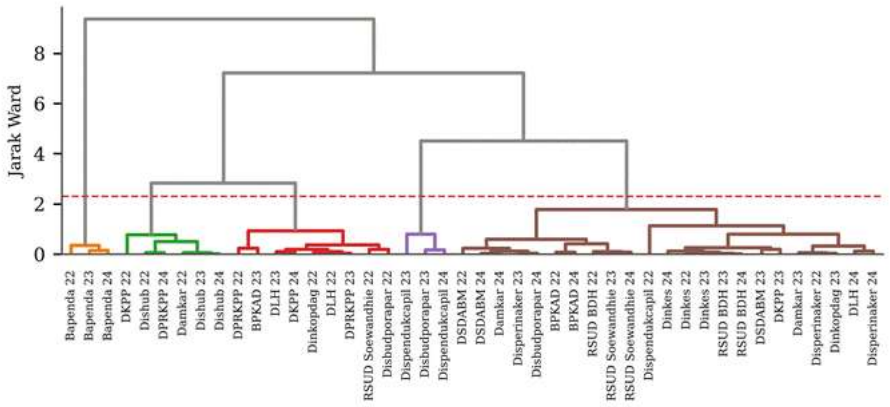
3.3 Penentuan Jumlah Cluster

Gambar 2 menyajikan hasil pengujian k = 2–8 pada skenario S1. Nilai k = 2 menghasilkan Silhouette tertinggi (0,732) dan DBI terendah (0,233) pada kedua algoritma. Namun, partisi tersebut hanya memisahkan tiga observasi Bapenda dari 41 observasi lainnya (ukuran 41–3), sehingga tidak membedakan kinerja OPD selain Bapenda. Pada k ≥ 3, K-Means mencapai Silhouette tertinggi pada k = 5 (0,615), sekaligus DBI terendah (0,376), dan Ward juga mencapai DBI terendahnya pada k = 5 (0,365). Kurva Elbow menunjukkan penurunan inersia yang tajam dari 44,31 (k = 2) menjadi 3,98 (k = 5), lalu melandai menjadi 2,57 pada k = 6. Nilai CHI meningkat monoton sehingga tidak menunjukkan titik optimum yang tegas. Ward memang memperoleh Silhouette sedikit lebih tinggi pada k = 4 (0,611) dibandingkan k = 5 (0,603), tetapi DBI dan CHI-nya lebih baik pada k = 5. Berdasarkan konvergensi indikator tersebut dan dendrogram pada Gambar 3, k = 5 ditetapkan sebagai jumlah cluster bersama untuk kedua algoritma.

Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)



Gambar 2. Kurva Elbow dan indeks validitas pada k = 2–8 (skenario S1)



Gambar 3. Dendrogram Agglomerative Ward; garis putus-putus menandai pemotongan pada k = 5

3.4 Perbandingan Kinerja K-Means dan Agglomerative Clustering

Tabel 3 merangkum hasil evaluasi kedua algoritma. Pada skenario utama S1 dengan k = 5, K-Means memperoleh Silhouette 0,6145 dan CHI 205,94, lebih tinggi daripada Agglomerative Ward (0,6025 dan 189,52). Sebaliknya, Ward memperoleh DBI yang lebih rendah (0,3650 berbanding 0,3763). Dengan demikian, K-Means unggul pada dua dari tiga indeks, tetapi selisihnya kecil: 0,012 pada Silhouette, 0,011 pada DBI, dan 16,43 pada CHI. Kedua partisi hampir identik, dengan ARI sebesar 0,939, dan hanya satu observasi yang ditempatkan berbeda, yaitu Dinas Kependudukan dan Pencatatan Sipil (Dispendukcapil) 2022 dengan X1 sebesar 128,89%. K-Means menempatkannya pada C2 bersama observasi berrealisasi sangat tinggi, sedangkan Ward menempatkannya pada C3. Pada partisi Ward, observasi ini memiliki Silhouette negatif. K-Means tidak menghasilkan Silhouette negatif.

Tabel 3. Perbandingan hasil evaluasi K-Means dan Agglomerative Clustering

Skenario	Algoritma	k	Silhouette ↑	DBI ↓	CHI ↑
S1 (X1,X2; 2022–2024; n=44)	K-Means	5	0,6145	0,3763	205,94
S1 (X1,X2; 2022–2024; n=44)	Agglomerative (Ward)	5	0,6025	0,3650	189,52
S2 (X1,X2,X3; 2023–2024; n=29)	K-Means	3	0,6916	0,1972	24,05
S2 (X1,X2,X3; 2023–2024; n=29)	Agglomerative (Ward)	3	0,6916	0,1972	24,05
S3 (X1, log10 X2; n=44) – sensitivitas	K-Means	5	0,5011	0,6302	51,89
S3 (X1, log10 X2; n=44) – sensitivitas	Agglomerative (Ward)	5	0,5034	0,6162	51,24

Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

Catatan: pada S1, ARI K-Means vs Ward = 0,939 (1 dari 44 observasi berbeda). Pada S2 kedua algoritma menghasilkan partisi identik.

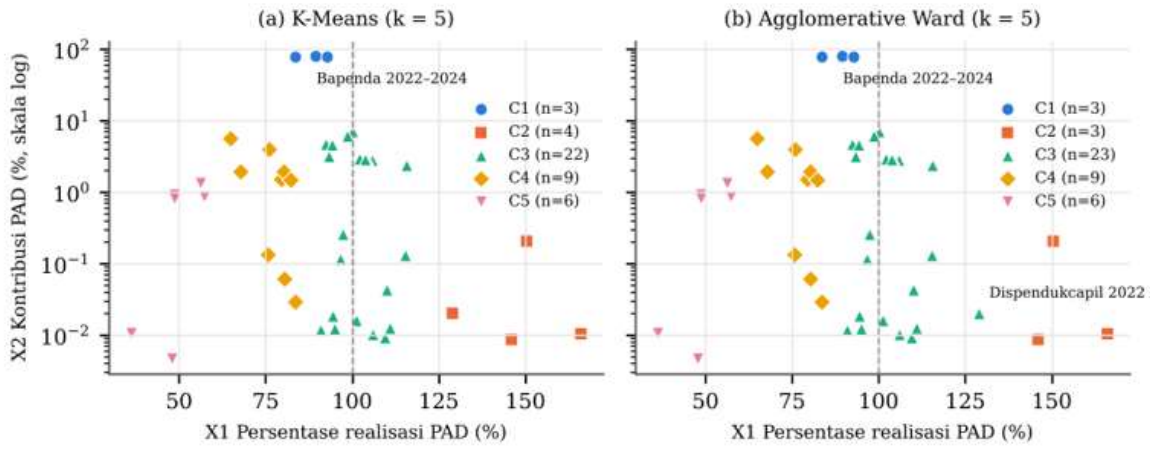
Hasil K-Means sangat stabil: partisi yang diperoleh dari 30 nilai random_state seluruhnya identik (ARI minimum 1,000). Uji sensitivitas linkage menunjukkan bahwa average linkage menghasilkan partisi yang sama persis dengan K-Means (ARI 1,000; Silhouette 0,6145). Complete linkage menghasilkan Silhouette 0,477 dan ARI 0,428, sedangkan single linkage menunjukkan efek *chaining* dengan cluster berukuran 37–3–2–1–1, Silhouette 0,375, dan CHI 26,01. Temuan ini sejalan dengan Abdalla yang menyatakan bahwa kinerja relatif kedua algoritma bergantung pada ukuran kemiripan dan kriteria penggabungan [10], serta dengan hasil Purnama dkk. yang menunjukkan selisih Silhouette antaralgoritma yang sangat kecil [11]. Karena Ward meminimalkan kriteria yang sama dengan K-Means secara bertahap tanpa realokasi [16], partisi keduanya cenderung berimpit. Perbedaan hanya muncul pada observasi yang berada di batas cluster.

Dari sisi praktis, kedua algoritma juga memiliki karakter yang berbeda. K-Means menghasilkan centroid yang langsung dapat dibaca sebagai profil cluster pada skala asli, misalnya centroid C5 pada X1 49,18% dan X2 0,66%. Model K-Means juga dapat dipakai untuk menempatkan data OPD tahun berikutnya ke cluster terdekat tanpa membangun ulang seluruh model. Agglomerative Ward tidak menyediakan mekanisme prediksi tersebut, tetapi dendrogramnya memberikan informasi tambahan tentang jarak penggabungan. Gambar 3 menunjukkan bahwa C1 (Bapenda) baru bergabung dengan cluster lain pada jarak penggabungan tertinggi (9,35), sementara C4 dan C5 bergabung pada jarak 2,82 serta C2 dan C3 pada jarak 4,49.

Pada skenario S2 (X1, X2, dan X3; 2023–2024), k optimum adalah 3 dengan Silhouette 0,692, DBI 0,197, dan CHI 24,05. Kedua algoritma menghasilkan partisi yang identik: Bapenda 2023–2024, Dinas Perindustrian dan Tenaga Kerja 2023, dan 26 observasi lainnya. Penambahan X3 lebih berfungsi sebagai pendeteksi pencilan pertumbuhan daripada pembeda kinerja, sehingga S2 tidak dipilih sebagai skenario utama. Pada skenario S3, transformasi log10(X2) membuat variasi kontribusi OPD kecil ikut berperan, tetapi pemisahan cluster melemah (Silhouette 0,501 untuk K-Means dan 0,503 untuk Ward). Pada S3, k = 5 dipertahankan agar sebanding dengan skenario utama. Secara internal, Silhouette dan DBI pada S3 sebenarnya terbaik pada k = 6 (Silhouette 0,542 untuk K-Means dan 0,547 untuk Ward). Pada skenario ini Ward unggul tipis pada Silhouette dan DBI, sedangkan K-Means unggul pada CHI. Hasil ini menunjukkan bahwa keunggulan algoritma bersifat kondisional terhadap representasi variabel.

3.5 Karakteristik Cluster

Gambar 4 memvisualisasikan hasil kedua algoritma pada ruang X1–X2, dan Tabel 4 menyajikan profil cluster K-Means. Centroid pada skala asli menunjukkan bahwa X1 menjadi pembeda utama antara C2–C5, sedangkan X2 memisahkan C1.



Gambar 4. Visualisasi hasil clustering pada ruang X1–X2: (a) K-Means, (b) Agglomerative Ward

Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

Tabel 4. Profil cluster K-Means pada skenario S1 (k = 5)

Cluster	Label	n	X1 rata-rata (min–maks) %	X2 rata-rata %	Porsi realisasi %	n 22/23/24	Anggota (OPD tahun)
C1	Penghimpun dominan	3	88,60 (83,57–92,77)	78,78	78,78	1/1/1	Bapenda 22–23–24
C2	Jauh melampaui target	4	147,82 (128,89–165,98)	0,06	0,08	1/2/1	Dispendukcapil 22–23–24; Disbudporapar 23
C3	Memenuhi target	22	101,91 (90,91–115,63)	1,95	14,42	5/8/9	Dinkes 22–23–24; RSUD BDH 22–23–24; DSDABM 22–23–24; BPKAD 22–24; Disperinaker 22–23–24; RSUD Soewandhie 23–24; Damkar 23–24; Dinkopdag 23; DKPP 23; DLH 24; Disbudporapar 24
C4	Di bawah target	9	76,71 (64,82–83,64)	1,84	5,38	5/3/1	RSUD Soewandhie 22; DPRKPP 22–23; DLH 22–23; Dinkopdag 22; Disbudporapar 22; BPKAD 23; DKPP 24
C5	Realisasi rendah	6	49,18 (36,25–57,33)	0,66	1,34	3/1/2	Damkar 22; Dishub 22–23–24; DKPP 22; DPRKPP 24

Singkatan: Dinkes = Dinas Kesehatan; RSUD BDH = RSUD Bhakti Dharma Husada; DSDABM = Dinas Sumber Daya Air dan Bina Marga; Damkar = Dinas Pemadam Kebakaran dan Penyelamatan; Disperinaker = Dinas Perindustrian dan Tenaga Kerja; Dinkopdag = Dinas Koperasi UKM dan Perdagangan; singkatan lain dijelaskan pada teks. Partisi Ward identik, kecuali Dispendukcapil 2022 yang masuk C3.

Cluster C1, penghimpun dominan, hanya berisi Bapenda pada ketiga tahun. Cluster ini menyumbang 78,78% realisasi PAD seluruh observasi dengan rata-rata X1 88,60% dan rentang 83,57–92,77%. Meskipun belum mencapai target, capaian Bapenda meningkat setiap tahun. Karena porsinya yang dominan, kesenjangan target PAD kota sangat dipengaruhi oleh kinerja pajak daerah. Cluster C2, realisasi jauh melampaui target, berisi empat observasi dengan rata-rata X1 147,82% tetapi kontribusi sangat kecil (rata-rata 0,06%). Anggotanya adalah Dispendukcapil 2022–2024 dan Dinas Kebudayaan, Kepemudaan dan Olah Raga serta Pariwisata (Disbudporapar) 2023. Capaian yang konsisten di atas 128% menunjukkan bahwa target pada OPD ini berpotensi ditetapkan terlalu rendah dibandingkan realisasinya.

Cluster C3, memenuhi target, merupakan cluster terbesar dengan 22 observasi, rata-rata X1 101,91% (90,91–115,63%), dan kontribusi 14,42% dari realisasi PAD seluruh observasi. Cluster ini memuat OPD layanan dengan penerimaan besar dan stabil, seperti Dinas Kesehatan, RSUD Bhakti Dharma Husada, dan RSUD dr. Mohamad Soewandhie, serta Badan Pengelolaan Keuangan dan Aset Daerah (BPKAD) pada 2022 dan 2024. Cluster C4, di bawah target, berisi sembilan observasi dengan rata-rata X1 76,71% (64,82–83,64%). Anggotanya antara lain Dinas Perumahan Rakyat dan Kawasan Permukiman serta Pertanahan (DPRKPP) 2022–2023, Dinas Lingkungan Hidup (DLH) 2022–2023, dan BPKAD 2023. Cluster C5, realisasi rendah, berisi enam observasi dengan rata-rata X1 49,18% (36,25–57,33%). Anggotanya adalah Dinas Perhubungan pada ketiga tahun, Dinas Pemadam Kebakaran dan Penyelamatan 2022, Dinas Ketahanan Pangan dan Pertanian (DKPP) 2022, dan DPRKPP 2024. Kualitas internal setiap cluster K-

Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

Means cukup baik, dengan rata-rata Silhouette per cluster 0,940 (C1), 0,525 (C2), 0,579 (C3), 0,594 (C4), dan 0,673 (C5).

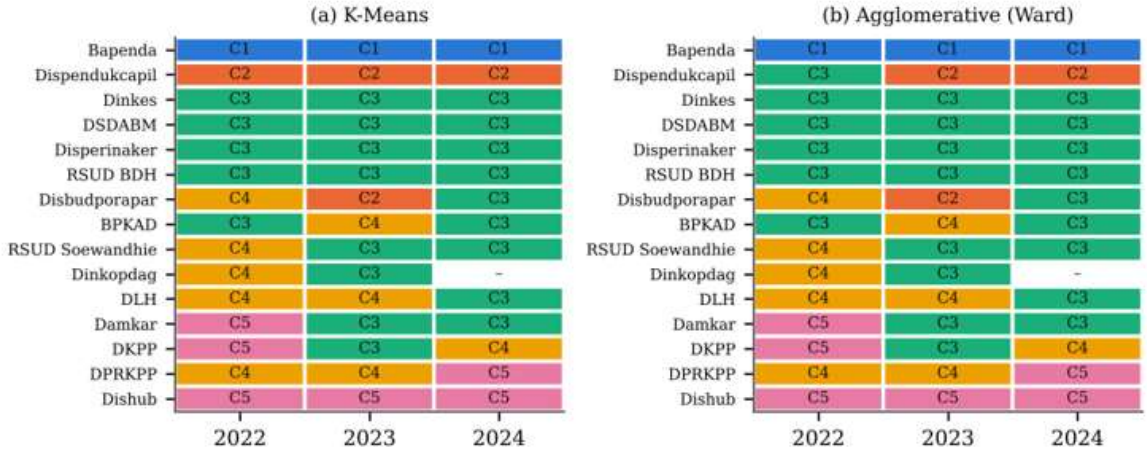
Interpretasi cluster perlu dibaca bersama skala pengukuran kedua variabel. Setelah standardisasi, tiga observasi Bapenda memiliki skor-z X2 sekitar 3,7, sedangkan 41 observasi lain berada pada rentang skor-z yang sempit, yaitu −0,35 sampai 0,01. Akibatnya, setelah Bapenda terpisah sebagai C1, pembagian C2 sampai C5 hampir sepenuhnya ditentukan oleh X1. Hal ini terlihat dari rentang X1 antarcluster yang tidak saling tumpang tindih: C5 berada di bawah 58%, C4 pada 64–84%, C3 pada 90–116%, dan C2 di atas 128%. Sebaliknya, rentang X2 pada C3 dan C4 relatif serupa, yaitu 0,01–7,06% dan 0,03–5,62%. Temuan ini sejalan dengan pola pada studi APBD provinsi, yaitu objek dengan nilai ekstrem cenderung membentuk cluster tunggal ketika data tidak ditransformasi [8]. Oleh sebab itu, skenario S3 dengan transformasi logaritmik dilaporkan sebagai pembanding agar pengguna hasil dapat menilai konsekuensi pilihan representasi variabel.

3.6 Perpindahan Cluster OPD 2022–2024

Distribusi anggota K-Means per tahun menunjukkan pergeseran ke arah capaian yang lebih baik. Proporsi observasi pada C3 meningkat dari 33,3% (5 dari 15) pada 2022 menjadi 53,3% (8 dari 15) pada 2023 dan 64,3% (9 dari 14) pada 2024, sedangkan proporsi C4 menurun dari 33,3% menjadi 20,0% dan 7,1%. Pola ini sejalan dengan kenaikan rasio realisasi PAD tingkat kota dari 84,20% menjadi 92,69%, tetapi kenaikan rasio kota terutama digerakkan Bapenda (sekitar 78,5% realisasi), sedangkan rata-rata X1 tak berbobot justru turun pada 2024. Gambar 5 menyajikan lintasan cluster setiap OPD.

Dari 14 OPD yang teramati lengkap, tujuh OPD (50%) berpindah cluster minimal satu kali menurut K-Means, dengan 10 perpindahan dari 28 pasangan tahun berurutan (35,7%). Menurut Ward, delapan OPD berpindah dengan 11 perpindahan (39,3%). OPD yang stabil adalah Bapenda (C1), Dispendukcapil (C2), Dinas Kesehatan, RSUD Bhakti Dharma Husada, Dinas Sumber Daya Air dan Bina Marga, Dinas Perindustrian dan Tenaga Kerja (C3), serta Dinas Perhubungan (C5). Dinas Perhubungan memerlukan perhatian khusus karena berada di cluster realisasi rendah selama tiga tahun berturut-turut dengan X1 sebesar 57,33%, 48,73%, dan 48,71%.

Perpindahan ke arah perbaikan terlihat pada Dinas Pemadam Kebakaran dan Penyelamatan (C5 ke C3 dan bertahan), RSUD dr. Mohamad Soewandhie (C4 ke C3), DLH (C4 ke C3 pada 2024), dan Dinas Koperasi Usaha Kecil dan Menengah dan Perdagangan (C4 ke C3 pada 2023). Sebaliknya, DPRKPP memburuk dari C4 ke C5 pada 2024 dengan X1 turun dari 80,30% menjadi 56,17%. BPKAD (C3–C4–C3) dan DKPP (C5–C3–C4) berfluktuasi, sedangkan Disbudporapar bergerak dari C4 ke C2 lalu ke C3. Matriks transisi dihitung dari 29 pasangan tahun berurutan seluruh OPD, termasuk Dinas Koperasi UKM dan Perdagangan 2022–2023 (C4 ke C3), sehingga terdapat 11 perpindahan (37,9%) menurut K-Means dan 12 (41,4%) menurut Ward. Matriks transisi K-Means memperlihatkan bahwa dari 12 pasangan yang berawal di C3, sebanyak 10 tetap di C3. Dari delapan pasangan yang berawal di C4, empat naik ke C3, dua tetap, satu ke C2, dan satu turun ke C5. Dari empat pasangan yang berawal di C5, dua naik ke C3.



Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

Gambar 5. Lintasan cluster OPD 2022–2024: (a) K-Means, (b) Agglomerative Ward; tanda "–" berarti rekaman dikeluarkan

Secara praktis, hasil ini memberikan beberapa implikasi. OPD di C5 yang persisten, terutama Dinas Perhubungan, serta OPD yang menurun seperti DPRKPP memerlukan evaluasi penetapan target dan efektivitas pemungutan retribusi. OPD di C2 memerlukan kalibrasi target agar lebih realistis. Bapenda di C1 tetap menjadi penentu utama karena setiap kenaikan satu poin persentase capaian pajak daerah berdampak jauh lebih besar terhadap PAD kota dibandingkan perbaikan di OPD lain. Penelitian ini memiliki beberapa keterbatasan. Pertama, jumlah observasi kecil (44 OPD-tahun). Kedua, X2 sangat menceng sehingga keragamannya didominasi Bapenda. Ketiga, X3 tidak tersedia untuk 2022. Keempat, akurasi data bergantung pada ekstraksi teks PDF, meskipun telah divalidasi terhadap total setiap OPD. Penyebab lonjakan atau penurunan realisasi, seperti kemungkinan pengalihan objek retribusi antar-OPD, tidak dapat dipastikan dari dokumen yang tersedia dan perlu dikonfirmasi dengan data administratif lainnya.

4. KESIMPULAN

Penelitian ini berhasil membangun dataset kinerja realisasi PAD OPD Kota Surabaya tahun 2022–2024 yang telah diaudit dari tiga dokumen rincian realisasi APBD. Audit menemukan 22 baris pendapatan ganda pada dokumen 2024 yang menyebabkan total pendapatan pada dokumen sumber tergandakan dan, jika tidak dikoreksi, akan menaikkan realisasi PAD 2024 sebesar 21,3%. Setelah pembersihan, 44 observasi OPD-tahun dari 16 OPD dianalisis menggunakan variabel persentase realisasi dan kontribusi PAD yang distandardisasi. Pada dataset dan skenario yang sama dengan k = 5, K-Means memperoleh Silhouette 0,615, Davies-Bouldin Index 0,376, dan Calinski-Harabasz Index 205,94, sedangkan Agglomerative Ward memperoleh 0,603, 0,365, dan 189,52. K-Means unggul pada dua dari tiga indeks, stabil terhadap inisialisasi, dan tidak menghasilkan observasi dengan Silhouette negatif, sehingga lebih direkomendasikan untuk data ini. Namun, selisih kinerja kedua algoritma kecil dan kesepakatan partisinya sangat tinggi (ARI 0,939). Kelima cluster yang terbentuk, yaitu penghimpun dominan, realisasi jauh melampaui target, memenuhi target, di bawah target, dan realisasi rendah, memberikan pemetaan kinerja yang jelas. Analisis lintas tahun menunjukkan bahwa separuh OPD yang lengkap tiga tahun berpindah cluster, dengan kecenderungan bergeser ke cluster memenuhi target, sementara Dinas Perhubungan konsisten berada pada cluster realisasi rendah. Penelitian selanjutnya disarankan memperluas periode data agar variabel pertumbuhan dapat digunakan pada seluruh tahun, merinci analisis hingga jenis retribusi atau objek pendapatan, serta menguji algoritma yang lebih tahan terhadap pencilan, seperti K-Medoids atau DBSCAN.

REFERENCES

[1] Pemerintah Republik Indonesia, *Undang-Undang Republik Indonesia Nomor 1 Tahun 2022 tentang Hubungan Keuangan antara Pemerintah Pusat dan Pemerintahan Daerah*. Jakarta, Indonesia, 2022. [Online]. Available: https://peraturan.bpk.go.id/Details/195696/uu-no-1-tahun-2022

[2] Kementerian Dalam Negeri Republik Indonesia, *Peraturan Menteri Dalam Negeri Nomor 77 Tahun 2020 tentang Pedoman Teknis Pengelolaan Keuangan Daerah*. Jakarta, Indonesia, 2020. [Online]. Available: https://peraturan.bpk.go.id/details/162792/permendagri-no-77-tahun-2020

[3] S. P. Faradilla and N. Hanifa, "Analisis Rasio Efektivitas PAD, Efisiensi PAD dan Kemandirian Kota Surabaya," *INDEPENDENT: Journal of Economics*, vol. 4, no. 3, pp. 98–104, 2024.

[4] Pemerintah Kota Surabaya, "Rincian APBD Menurut Urusan Pemerintahan Daerah, Organisasi, Program, Kegiatan, Sub Kegiatan, Kelompok, dan Jenis Pendapatan, Belanja, dan Pembiayaan, Tahun Anggaran 2022, 2023, dan 2024 (Realisasi)," Pemerintah Kota Surabaya, Surabaya, Indonesia, 2023–2025.

[5] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Waltham, MA, USA: Morgan Kaufmann, 2012.

Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

[6] A. E. Ezugwu *et al.*, "A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects," *Engineering Applications of Artificial Intelligence*, vol. 110, p. 104743, 2022, doi: 10.1016/j.engappai.2022.104743.

[7] F. A. I. S. Aji, S. Achmadi, and F. Ariwibisono, "Penerapan Metode Clustering pada Analisis Realisasi Pendapatan Asli Daerah dengan Algoritma K-Means," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 5, no. 2, pp. 443–451, 2021, doi: 10.36040/jati.v5i2.3741.

[8] T. Rahmadianissa, "Clustering Menggunakan Algoritma K-Means dan DBSCAN pada Anggaran Pendapatan dan Belanja Daerah di Indonesia," Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Riau, Pekanbaru, Indonesia, 2022. [Online]. Available: https://repository.unri.ac.id/handle/123456789/10905

[9] R. Wulaningrum, V. E. Satya, M. Kadafi, and A. Azizah, "Clustering of Provincial Government in Indonesia Based on Fixed Asset Ownership and Economic Growth," in *Proc. Int. Conf. Applied Science and Technology on Social Science 2021 (iCAST-SS 2021)*, Atlantis Press, 2022, pp. 471–476, doi: 10.2991/assehr.k.220301.077.

[10] H. I. Abdalla, "A Brief Comparison of K-means and Agglomerative Hierarchical Clustering Algorithms on Small Datasets," in *Proc. 2021 Int. Conf. Wireless Communications, Networking and Applications*, Lecture Notes in Electrical Engineering. Singapore: Springer, 2022, pp. 623–632, doi: 10.1007/978-981-19-2456-9_64.

[11] D. G. Purnama, Safrizal, and C. B. Santoso, "A Comparison of K-Means, Hierarchical Clustering, and K-Medoids for Market Segmentation Based on Silhouette Score Evaluation," *JSAI (Journal Scientific and Applied Informatics)*, vol. 9, no. 2, pp. 294–305, 2026, doi: 10.36085/jsai.v9i2.10849.

[12] M. Sholeh and K. Aeni, "Perbandingan Evaluasi Metode Davies Bouldin, Elbow dan Silhouette pada Model Clustering dengan Menggunakan Algoritma K-Means," *STRING (Satuan Tulisan Riset dan Inovasi Teknologi)*, vol. 8, no. 1, pp. 56–65, 2023, doi: 10.30998/string.v8i1.16388.

[13] A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming, "K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data," *Information Sciences*, vol. 622, pp. 178–210, 2023, doi: 10.1016/j.ins.2022.11.139.

[14] K. P. Sinaga and M.-S. Yang, "Unsupervised K-Means Clustering Algorithm," *IEEE Access*, vol. 8, pp. 80716–80727, 2020, doi: 10.1109/ACCESS.2020.2988796.

[15] X. Ran, Y. Xi, Y. Lu, X. Wang, and Z. Lu, "Comprehensive survey on hierarchical clustering algorithms and the recent developments," *Artificial Intelligence Review*, vol. 56, no. 8, pp. 8219–8264, 2023, doi: 10.1007/s10462-022-10366-3.

[16] F. Murtagh and P. Legendre, "Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion?," *Journal of Classification*, vol. 31, no. 3, pp. 274–295, 2014, doi: 10.1007/s00357-014-9161-z.

[17] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*. Hoboken, NJ, USA: John Wiley & Sons, 1990, doi: 10.1002/9780470316801.

[18] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987, doi: 10.1016/0377-0427(87)90125-7.

[19] M. Shutaywi and N. N. Kachouie, "Silhouette Analysis for Performance Evaluation in Machine Learning with Applications to Clustering," *Entropy*, vol. 23, no. 6, p. 759, 2021, doi: 10.3390/e23060759.

[20] D. L. Davies and D. W. Bouldin, "A Cluster Separation Measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, no. 2, pp. 224–227, 1979, doi: 10.1109/TPAMI.1979.4766909.

[21] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics*, vol. 3, no. 1, pp. 1–27, 1974, doi: 10.1080/03610927408827101.

Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

[22] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[23] scikit-learn developers, "Clustering," *scikit-learn User Guide*. [Online]. Available: https://scikit-learn.org/stable/modules/clustering.html

[24] W. McKinney, *Python for Data Analysis: Data Wrangling with pandas, NumPy, and Jupyter*, 3rd ed. Sebastopol, CA, USA: O'Reilly Media, 2022.

Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)