MATHunesa
Jurnal Ilmiah Matematika *Volume 11 No 03*
e-ISSN : 2716-506X | p-ISSN : 2301-9115 *Tahun 2023*

PENGELOMPOKAN BERDASARKAN GARIS KEMISKINAN PENDEKATAN *TIME SERIES BASED CLUSTERING* DI PROVINSI JAWA TIMUR

Rosalina Agista Riani
Program Studi Matematika, FMIPA, Universitas Negeri Surabaya
e-mail : rosalina.19001@mhs.unesa.ac.id

A'yunin Sofro
Program Studi Matematika, FMIPA, Universitas Negeri Surabaya
Penulis Korespondensi : ayuninsofro@unesa.ac.id

Abstrak

Provinsi Jawa Timur disebut sebagai provinsi terbesar di Pulau Jawa yang memiliki luas wilayah sebesar 48.037 km² dan banyaknya penduduk 41.416.407 jiwa. Banyaknya penduduk dapat menyebabkan masalah sosial seperti kemiskinan, salah satunya karena pembangunan sarana dan prasana tidak merata. Dalam tolak ukur kemiskinan terdapat faktor garis kemiskinan, yaitu pendapatan minimum yang harus dicapai seseorang untuk memperoleh standar hidup yang layak. Dari faktor tersebut dapat dilakukan pengelompokan untuk mengetahui Kabupaten/Kota manakah yang darurat akan faktor garis kemiskinan. Nantinya dapat menjadi informasi kepada masyarakat dan pemerintah terkait wilayah manakah yang perlu diperhatikan khusus terkait masalah kemiskinan. Sehingga perlu dilakukan penelitian untuk mengelompokan Kabupaten/Kota di Provinsi Jawa Timur terhadap garis kemiskinan dengan analisis *cluster*. Analisis *cluster* pada penelitian ini menggunakan metode ukuran jarak *Short Time Series* (STS) *distance, Autocorrelation Function* (ACF) *distance,* dan *Dynamic Time Warping* (DTW) *distance*. Untuk metode *clustering* yang digunakan yaitu metode hirarki *agglomerative* yang terdiri dari *single linkage, average linkage,* dan *complete linkage*. Hasil *cluster* yang terbentuk pada penelitian ini yaitu sebanyak 5 *cluster* dengan hasil *cluster* paling optimal yaitu metode ukuran jarak *Autocorrelation Function* (ACF) *distance* dengan metode *average linkage* yang memiliki nilai koefisien *silhouette* 0,8161.
Kata Kunci: Garis Kemiskinan, Analisis *Cluster,* Ukuran jarak

Abstract

East Java Province is the largest province on the island of Java with an area of 48,037 km2 and a population of 41,416,407 people. The large population can cause social problems such as poverty, one of which is because the development of facilities and infrastructure is not evenly distributed. In the poverty measurement, there is a poverty line factor, namely the minimum income that needs to be met by a person to obtain an adequate standard of living. From these factors, grouping can be done to find out which districts/cities have an emergency regarding the poverty line factor. Later it can provide information to the public and government regarding which areas need special attention related to the problem of poverty. So it is necessary to research to classify districts/cities in East Java province against the poverty line with cluster analysis. Cluster analysis in this study used the Dynamic Time Warping (DTW) distance, Short Time Series (STS) distance, and Autocorrelation Function (ACF) distance methods. For the clustering method, the agglomerative hierarchical method includes single linkage, average linkage, and complete linkage. The results of the clusters formed in this study were as many as 5 clusters with the most optimal cluster results, namely the Autocorrelation Function (ACF) distance measurement method with the average linkage method which has a silhouette coefficient value of 0.8161.
Keywords: *Rainfall, Cluster Analysis, Distance measures*

PENDAHULUAN

Provinsi Jawa Timur disebut sebagai provinsi terbesar di Pulau Jawa yang memiliki luas wilayah sebesar 48.037 km² dan banyaknya penduduk 41.416.407 jiwa berdasarkan data Publikasi Badan Pusat Statistik. Banyaknya penduduk dapat menyebabkan masalah sosial seperti kemiskinan. Hal tersebut dapat terjadi karena pembangunan sarana dan prasana tidak merata, penduduk terus bertambah sedangkan lapangan pekerjaan semakin sedikit.

Kemisikinan adalah kondisi dimana seseorang hidup dalam keadaan kurang dan dapat juga

diartikan sebagai kondisi seseorang tidak memiliki kemampuan, aset, dan akses kebutuhan mereka di waktu yang akan datang serta sangat rentan terhadap resiko dan tekanan seperti peningkatan harga kebutuhan yang mengalami peningkatan secara tiba-tiba (Pratama, 2014). Dalam tolak ukur kemiskinan terdapat faktor garis kemiskinan, yaitu pendapatan minimum yang harus dicapai seseorang untuk memperoleh standar hidup yang layak di suatu negara. Garis kemiskinan digunakan sebagai tolak ukur penduduk miskin dengan mempertimbangkan pembaruan sosio-ekonomi seperti program peningkatan kesejahteraan (Aprilia & Sembiring, 2021).

Analisis *cluster* adalah pengelompokkan suatu objek yang memiliki sifat sama tanpa menghilangkan struktur alami objek tersebut (Ayundari & Sutikno, 2019). Analisis *cluster* bertujuan untuk mengelompokkan data secara objektif ke dalam kelompok-kelompok yang homogen di mana kesamaan objek dalam kelompok diminimalkan dan ketidaksamaan antar kelompok dimaksimalkan (Liao, 2005). Analisis *cluster* semakin berkembang dalam penerapannya menggunakan data deret waktu (*time series*). Data deret waktu adalah data yang diperoleh dengan urutan pengamatan yang diambil secara berurutan dalam waktu yang memiliki struktur korelasi antara nilai data setiap deret waktu (Box et al., 2008).

Analisis *cluster* deret waktu melakukan pengelompokkan objek berdasarkan pola deret waktunya. Namun pemilihan jarak dan metode *clustering* yang digunakan harus dapat memperhitungkan struktur data deret waktu yang memiliki sifat sangat dinamis (Liao, 2005). Menghitung jarak antara objek deret waktu adalah salah satu landasan dari algoritma *clustering* deret waktu (Sardá-Espinosa, 2019). Saat ini sudah banyak pengembangan terkait ukuran jarak seperti yang digunakan pada penelitian ini, yaitu *Short Time Series* (STS) *distance, Autocorrelation Function* (ACF) *distance,* dan *Dynamic Time Warping* (DTW) *distance.*

Analisis *cluster* terbagi atas dua metode, yaitu metode hirarki dan metode non-hirarki. Analisis *cluster* metode hirarki terbagi menjadi dua, yaitu metode penggabungan (*agglomerative method)* dan metode pemecahan (*divisive method).* Analisis *cluster* hirarki dengan *agglomerative method* terdiri dari *single linkage, complete linkage, average linkage,* dan metode *ward.* Dan untuk metode analisis *cluster* non-hirarki yaitu *k-means* (Johnson & Wichern, 2007).

Penelitian yang relevan dengan *time series based clustering* dilakukan oleh Pangestu dan Fitriani (2022) dengan judul "Perbandingan Perhitungan Jarak *Euclidean Distance, Manhattan Distance,* dan *Cosine Similarity* dalam Pengelompokkan Data Bibit Padi Menggunakan Algoritma *K-Means*". Penelitian yang relevan lainnya dilakukan oleh Artha dan Winarko (2016) dengan judul "Perbandingan *Eros, Euclidean Distance* dan *Dynamic Time Warping* dalam Klasifikasi Data *Multivariate Time Series* Menggunakan kNN". Pada penelitian sebelumnya hanya menggunakan salah satu metode hirarki dengan ukuran jarak yang sering digunakan. Untuk memperoleh hasil yang lebih baik, dapat dilakukan kombinasi analisis *cluster* metode hirarki dengan ukuran jarak terbarukan. Sehingga dapat diketahui metode hirarki dan ukuran jarak manakah yang paling baik dalam penerapan analisis *cluster* pada penelitian ini.

Penelitian ini menggunakan analisis *cluster* yang dilakukan untuk mengetahui pengelompokkan Kabupaten/Kota di provinsi Jawa Timur terhadap garis kemiskinan selama 2018 hingga 2022. Pengelompokkan Kabupaten/Kota di provinsi Jawa Timur perlu dilakukan sebagai bahan informasi Kabupaten/Kota manakah di Jawa Timur yang perlu dilakukan penanganan terlebih dahulu terkait kemiskinan. Data garis kemiskinan yang digunakan merupakan data *time series* sehingga dalam melakukan analisis *cluster* menggunakan ukuran jarak. Ukuran jarak yang digunakan yaitu, *Short Time Series* (STS) *distance, Autocorrelation Function* (ACF) *distance,* dan *Dynamic Time Warping* (DTW) *distance.* Untuk metode analisis *cluster* yang digunakan, yaitu metode hirarki *agglomerative* meliputi *single linkage, complete linkage,* dan *average linkage.*

Dari analisis *cluster* menggunakan tiga ukuran jarak, nantinya menggunakan metode *elbow* untuk menentukan jumlah *cluster,* kemudian dilanjutkan dengan penggunaan tiga metode hirarki *agglomerative.* Hasil dari masing-masing metode dan ukuran jarak tersebut nantinya di validasi dengan koefisien *silhouette* untuk menentukan hasil *cluster* paling optimal. Hasil paling optimal nantinya diketahui kelompok Kabupaten/Kota di Jawa timur berdasarkan tingkatan intensitas kemiskinan. Sehingga nantinya dapat dijadikan perkiraan

pemerintah Provinsi Jawa Timur untuk menanggulangi kemiskinan seperti bantuan sosial, diadakannya lapangan pekerjaan.

KAJIAN TEORI

Garis Kemiskinan

Garis kemiskinan adalah pendapatan minimum yang harus dicapai seseorang untuk memperoleh standar hidup yang layak di suatu negara. Garis kemiskinan digunakan sebagai tolak ukur penduduk miskin dengan mempertimbangkan pembaruan sosio-ekonomi seperti program peningkatan kesejahteraan (Aprilia & Sembiring, 2021). Garis kemiskinan adalah penjumlahan dari Garis Kemiskinan Makanan (GKM) dan Garis Kemiskinan Non Makanan (GKNM). Kategori penduduk miskin ialah penduduk yang memiliki rata-rata pengeluaran perkapita per bulan dibawah Garis Kemiskinan (Nugroho et al., 2020).

Analisis *cluster*

Menurut KBBI, *cluster* adalah beberapa benda atau hal yang berkelompok menjadi satu; gugus. Analisis *cluster* adalah pengelompokkan suatu objek yang memiliki sifat sama tanpa menghilangkan struktur alami objek tersebut (Ayundari & Sutikno, 2019) yang bertujuan untuk mengelompokkan data secara objektif ke dalam kelompok-kelompok yang homogen di mana kesamaan objek dalam kelompok diminimalkan dan ketidaksamaan antar kelompok dimaksimalkan (Liao, 2005).

Analisis *cluster* semakin berkembang dalam penerapannya menggunakan data deret waktu (*time series*). Data deret waktu adalah data yang diperoleh dengan urutan pengamatan yang diambil secara berurutan dalam waktu yang memiliki struktur korelasi antara nilai data setiap deret waktu (Box et al., 2008). Salah satu komponen kunci dalam analisis *cluster* deret waktu adalah pemilihan jarak dan metode *clustering* yang digunakan (Liao, 2005).

Dynamic Time Warping (DTW) *distance*

Dynamic Time Warping (DTW) *distance* adalah algoritma yang digunakan untuk menemukan jarak antara dua deret waktu yang memiliki panjang sama atau berbeda. *Dynamic Time Warping* (DTW) *distance* menggunakan teknik pemrograman dinamis untuk menemukan semua jalur yang memungkinkan dan menggunakan matriks jarak untuk memilih salah satu yang mengarah ke jarak minimum antara dua deret waktu, di mana setiap elemen dalam matriks adalah jarak kumulatif minimum dari tiga tetangga di sekitarnya. Misalkan terdapat dua deret waktu $X = x_1, x_2, \dots, x_i, \dots, x_n$ dengan panjang n dan $Y = y_1, y_2, \dots, y_j, \dots, y_m$ dengan panjang m. Pertama, membuat matriks d ukuran $n \times m$, di mana untuk setiap elemen (i, j) dalam matriks d merupakan perbedaan antara x_i dan y_j, yang dituliskan pada Persamaan (2.1) sebagai berikut:

$$d_{i,j} = |x_i - y_j| \qquad (2.1)$$

di mana $i = 1,2, \dots, n$ dan $j = 1,2, \dots, m$. Setelah mendapatkan jarak kumulatif berupa d_{ij}, kemudian tambahkan nilai minimum dari tiga elemen yang berdekatan dengan elemen (i, j) yaitu $\{d_{(i-1)(j-1)}, d_{(i-1)j}, d_{i(j-1)}\}$, di mana $0 < i \leq n$ dan $0 < j \leq m$ sehingga terbentuk matriks *E*. Dapat didefinisikan pada Persamaan (2.2) elemen (i, j) pada matriks *E* sebagai berikut:

$$E_{i,j} = d_{ij} + \min\{d_{(i-1)(j-1)}, d_{(i-1)j}, d_{i(j-1)}\} \qquad (2.2)$$

Setelah mendapatkan matriks *E,* selanjutnya adalah menentukan jarak DTW antara dua deret waktu *X* dan *Y* dengan Persamaan (2.3) sebagai berikut (Niennattrakul & Ratanamahatana, 2007):

$$d_{DTW}(X, Y) = \min_{\forall w \in p} \left\{ \sum_{i,j=1}^{k} E_{i,j} \right\} \qquad (2.3)$$

di mana

p : kemungkinan dari sekumpulan semua *warping path*

$E_{i,j}$: elemen (i, j) pada matriks *E*

k : panjang dari *warping path*

Warping path adalah jalur yang terdiri dari jarak minimum dari elemen $E_{1,1}$ hingga $E_{n,m}$. Berikut adalah syarat yang harus dipenuhi dalam menentukan *warping path*, yaitu (Spiegel, 2015):

1. *Boundary conditions*
   $p_1 = (1,1)$ dan $p_k = (m, n)$.
   Kondisi dimana titik awal dan titik akhir dari *warping path* adalah dimulai dari elemen matriks (1,1) hingga elemen matriks terakhir (m, n). Hal ini bertujuan agar data

yang diolah seluruhnya mulai dari awal hingga akhir.

2. *Continuity conditions*
   $p_{k+1} - p_k \in \{(1,0),(0,1),(1,1)\}$
   Kondisi dimana penentuan *warping path* harus bertahap dengan indeks *i* dan *j* dengan selisih maksimal adalah 1 pada setiap langkahnya. Hal ini bertujuan agar data yang diolah tidak ada yang terlewat.
3. *Monotonicity conditions*
   $n_1 \leq n_2 \leq n_3 \leq \cdots n_k$ dan $m_1 \leq m_2 \leq m_3 \leq \cdots m_k$
   Kondisi dimana penentuan *warping path* berdasarkan urutan waktu agar tidak ada data yang diolah secara berulang.

Short Time Series* (STS) *distance

Short Time Series Distance (STS) *distance* diperkenalkan oleh Möller-Levet, Klawonn, Cho, dan Wolkenhauer (2003) yang digunakan untuk mengukur kesamaan data deret waktu *microarray* DNA. *Microarray* adalah pola yang diperoleh dengan menganalisis fungsi dan ekspresi beberapa gen secara bersamaan dalam satu percobaan. Moller memiliki tujuan untuk menentukan jarak yang mampu menangkap perbedaan dalam bentuk, ditentukan oleh perubahan ekspresif relatif dan informasi temporal yang sesuai. Misalkan terdapat dua data deret waktu $X = \{x_0, x_1, \dots, x_{N-1}\}$ dan $Y = \{y_0, y_1, \dots, y_{N-1}\}$, jarak STS didefinisikan pada Persamaan (2.4) sebagai berikut:

$$d_{STS}(X,Y) = \sqrt{\sum_{k=0}^{N-1}\left(\frac{y_{k+1}-y_k}{t_{k+1}-t_k} - \frac{x_{k+1}-x_k}{t_{k+1}-t_k}\right)} \quad (2.4)$$

di mana t_k ialah waktu dari setiap titik pada data X dan Y.

Autocorrelation Function* (ACF) *distance

Galeano dan Pella (2000) melakukan penelitian hubungan data deret waktu dengan menggunakan *Autocorrelation Function* (ACF) *distance.* Misalkan terdapat dua data deret waktu X dan Y. Di mana $\hat{\rho}_X = \left(\hat{\rho}_{1,X}, \hat{\rho}_{2,X} \dots, \hat{\rho}_{k,X}\right)^t$ dan $\hat{\rho}_Y = \left(\hat{\rho}_{1,Y}, \hat{\rho}_{2,Y}, \dots, \hat{\rho}_{k,Y}\right)^t$ adalah representasi vektor autokorelasi hasil estimasi dari lag 1 hingga lag k di mana $\hat{\rho}_{i,X} \approx 0$ dan $\hat{\rho}_{i,Y} \approx 0$ untuk $i > k$. *Autocorrelation Function* (ACF) adalah suatu fungsi yang digunakan untuk menjelaskan korelasi antara X_t dan $X_{(t+k)}$ dari suatu proses yang sama dan hanya dipisahkan oleh lag waktu ke-k. Misalkan terdapat data deret waktu $X = X_1, X_2, \dots, X_n$, maka ACF dituliskan pada Persamaan (2.5) berikut:

$$\rho_k = \frac{\sum_{t=1}^{n-k}(X_t - \bar{X})(X_{(t+k)} - \bar{X})}{\sum_{i=1}^{n}(X_t - \bar{X})^2} \quad (2.5)$$

Setelah diketahui vektor autokerelasi pada Persamaan (2.5), maka jarak antar dua deret waktu dapat dituliskan pada Persamaan (2.6) sebagai berikut:

$$d_{ACF}(X,Y) = \sqrt{\sum_{i=1}^{k}(\hat{\rho}_X - \hat{\rho}_Y)^2} \quad (2.6)$$

Di mana
$d_{ACF}(X,Y)$: jarak autokorelasi antara X dan Y
$\bar{X}$: rata-rata dari deret waktu

Metode Hirarki

Metode hirarki terbagi menjadi dua metode, yaitu metode *agglomerative* dan *divisive*. Pengelompokkan metode agglomerative dimulai dengan obyek-obyek yang terpisah. *Cluster* kemudian dibentuk dengan mengelompokkan obyek menjadi kelompok yang lebih besar. Proses ini berlanjut hingga semua obyek menjadi anggota *cluster*. Sedangkan pada metode *divisive* bekerja secara terbalik, pengelompokkan dimulai dengan mengelompokan semua obyek ke dalam satu grup. Kelompok – kelompok tersebut kemudian dipisahkan satu sama lain hingga masing-masing obyek berada dalam kelompoknya sendiri (Everitt et al., 2011).

Pada metode *agglomerative method* terdiri dari *single linkage, complete linkage,* dan *average linkage.* Metode *single linkage* merupakan metode yang menggunakan aturan jarak minimum antar *cluster*. Pada metode *single linkage* penentuan jarak antar *cluster* dilakukan dengan melihat jarak antar dua *cluster* dan memilih jarak paling dekat. Kemudian mencari jarak terdekat berikutnya, dan obyek berikutnya digabungkan ke dalam kelompok tersebut. Begitu seterusnya, hingga semua obyek digabungkan menjadi satu kelompok besar. Jika terdapat dua objek U dan V yang akan di kelompokkan, sehingga di peroleh *cluster* (UV) dengan jarak kedua objek tersebut dilambangan d_{UV}.

Untuk mencari jarak antara *cluster* (UV) dan *cluster* W atau *cluster* yang lain, maka persamaan yang digunakan untuk menentukan jarak keduanya dituliskan pada Persamaan (2.7) sebagai berikut:

$$d_{(UV)W} = \min(d_{UW}, d_{VW}) \qquad (2.7)$$

di mana nilai d_{UW}, dan d_{VW} adalah jarak minimum antara *cluster* U dan *cluster* W serta *cluster* V dan *cluster* W.

Metode *average linkage* adalah salah satu metode cluster hirarki yang pengelompokannya berdasarkan rata-rata antar objek. Metode ini dianggap lebih stabil dan tidak bias karena menggunakan rata-rata. Perhitungan jarak antar kelompok dituliskan pada Persamaan (2.8) sebagai berikut:

$$d_{(UV)W} = \frac{d_{(UW)} + d_{(VW)}}{n_{(UV)} n_W} \qquad (2.8)$$

dengan $n_{(UV)}$ yaitu banyak anggota *cluster* (UV) dan n_W yaitu banyak anggota *cluster* W.

Metode *complete linkage* merupakan metode dengan menggunakan aturan jarak maksimum antar *cluster*. Pengelompokkan dengan metode *complete linkage* dimulai dengan menentukan objek mana yang memiliki jarak terdekat, langkah selanjutnya adalah menggabungkan objek tersebut dengan melihat jarak yang jauh atau maksimal. Sehingga dapat dituliskan pada Persamaan (2.9) sebagai berikut:

$$d_{(UV)W} = \max(d_{UW}, d_{VW}) \qquad (2.9)$$

Dengan d_{UW} merupakan jarak terjauh dari cluster U dan W, sedangkan d_{VW} merupakan jarak terjauhdari cluster V dan W.

Metode *Elbow*

Metode *elbow* merupakan metode yang digunakan untuk menentukan berapa banyak *cluster* yang harus dipilih dengan melihat presentase hasil perbandingan antara jumlah cluster yang akan membentuk siku pada suatu titik (Bholowalia & Kumar, 2014). Metode *elbow* bekerja dengan cara memilih nilai *cluster* kemudian menambah nilai *cluster* tersebut untuk dijadikan model data dalam penentuan *cluster* terbaik dan presentasi perhitungan yang dihasilkan menjadi pembanding antara jumlah *cluster* yang ditambah. Hasil presentase yang berbeda dari setiap nilai *cluster* dapat ditunjukan dengan menggunakan grafik sebagai sumber informasinya. Jika nilai *cluster* pertama dengan nilai *cluster* kedua memberikan sudut dalam grafik atau nilainya mengalami penurunan paling besar maka nilai *cluster* tersebut yang terbaik (Putu et al., 2015).

Untuk mendapatkan perbandingan adalah dengan menghitung SSE (*Sum Square Error*) dari masing-masing nilai *cluster*. Semakin besar jumlah *cluster* K, maka nilai SSE akan semakin kecil. Algoritma mendapatkan nilai SSE seperti Persamaan (2.10) di bawah ini (Irwanto et al., 2012):

$$SSE = \sum_{K=1}^{K} \sum_{X_i} \|x_i - c_k\|^2 \qquad (2.10)$$

Di mana
K : jumlah *cluster* ke-c
x_i : jarak data objek ke-i
c_k : pusat *cluster* ke-i

Algoritma metode *elbow* dalam menentukan nilai K sebagai berikut (Putu et al., 2015):

1. Inisialisasi awal nilai K
2. Naikkan nilai K
3. Hitung hasil SSE dari setiap nilai K
4. Melihat hasil SSE dari nilai K yang turun secara drastis
5. Menetapkan nilai K yang berbentuk siku

Koefisien *Silhouette*

Koefisien *silhouette* merupakan salah satu ukuran ketepatan dalam pengelompokkan deret waktu dan juga biasanya digunakan untuk mengetahui kualitas pengelompokan. Langkah perhitungan koefisien *silhouette* di awali dari mencari a_i^j yaitu jarak rata-rata data ke-i dengan semua data di *cluster* yang sama. Kaufman & Rousseeuw (1991) menuliskan a_i^j dalam Persamaan (2.11) sebagai berikut:

$$a_i^j = \frac{1}{m_j - 1} \sum_{\substack{r=1 \\ r \neq 1}} d(x_i^j, x_r^j) \qquad (2.11)$$

di mana
j : *cluster*
i : index data $(i = 1,2, \dots, m_j)$

a_i^j : rata-rata jarak data ke-i dengan semua data di *cluster* yang sama
M_j : jumlah data dalam *cluster* ke-j
$d(x_i^j, x_r^j)$: jarak data ke-i dengan data ke-r dalam satu *cluster* j

Selanjutnya menghitung nilai b_i^j yaitu nilai minimum dari jarak rata-rata data ke-i dengan semua data di *cluster* berbeda. Maka, b_i^j dituliskan pada Persamaan (2.12) sebagai berikut:

$$b_i^j = \min_{\substack{n=1,\dots,k \\ n \neq j}} \left\{ \frac{1}{m_n} \sum_{\substack{r=1 \\ r \neq 1}}^{m_n} d(x_i^j, x_r^n) \right\} \quad (2.12)$$

di mana
j : *cluster*
i : index data ($i = 1,2, \dots, m_j$)
b_i^j : rata-rata jarak data ke-i dengan semua data di *cluster* yang berbeda
M_n : jumlah data dalam *cluster* ke-n
$d(x_i^j, x_r^j)$: jarak data ke-i dengan data ke-j dalam satu *cluster* n

Setelah a_i^j dan b_i^j diketahui, langkah selanjutnya menghitung SI_i^j yang dituliskan pada Persamaan (2.13) sebagai berikut:

$$SI_i^j = \frac{b_i^j - a_i^j}{\max\{a_i^j, b_i^j\}} \quad (2.13)$$

di mana
SI_i^j : *Silhouette Index* data ke-i dalam satu *cluster*
a_i^j : rata-rata jarak data ke-i dengan semua data di *cluster* yang sama
b_i^j : rata-rata jarak data ke-i dengan semua data di *cluster* yang berbeda

Nilai koefisien *silhouette* dari setiap objek dalam suatu *cluster* merupakan suatu ukuran yang menunjukkan seberapa dekat kemiripan data dikelompokkan di dalam satu *cluster* tersebut. Nilai SI_i^j berada pada rentang :

$$-1 < SI_i^j \leq 1$$

Nilai SI_i^j mendekati -1 menunjukkan bahwa jarak antar objek dalama_i^j jauh lebih besar dibandingkan b_i^j , sehingga dikatakan bahwa terjadi salah pengelompokkan atau keragu-raguan dalam pengelompokkan yang dilakukan. Sedangkan jika nilai SI_i^j mendekati 1 menunjukkan bahwa jarak antar objek dalam a_i^j jauh lebih kecil dibandingkan b_i^j, sehingga dikatakan pengelompokkan dilakukan dengan baik.

Kemudian menggunakan Persamaan (2.14) untuk perhitungan mendapatkan SI_j sebagai berikut:

$$SI_j = \frac{1}{m_j} \sum_{i=1}^{m_j} SI_i^j \quad (2.14)$$

di mana
SI_j : rata-rata *Silhouette Index cluster j*
SI_i^j : *Silhouette Index* data ke-i dalam satu *cluster*
M_j : jumlah data dalam *cluster* ke-j
i : index data ($i = 1,2, \dots, m_j$)

Untuk rumus perhitungan mendapatkan nilai SI global dituliskan pada Persamaan (2.15) berikut:

$$SI = \frac{1}{k} \sum_{j=1}^{k} SI_j \quad (2.15)$$

di mana
SI : rata-rata *Silhouette Index* dari dataset
SI_j : rata-rata *Silhouette Index cluster j*
k : jumlah *cluster*

Langkah yang terakhir yaitu menentukan koefisien *silhouette* (SC) yang diperoleh dengan mencari nilai maksimal dari *Silhouette Index Global* dari jumlah *cluster* 2 sampai jumlah *cluster* ke-$q-1$. SC dituliskan pada Persamaan (2.16) sebagai berikut:

$$SC = {}^{max}_{k}\, SI(k) \quad (2.16)$$

di mana
SC : Koefisien *Silhouette*
SI : *Silhouette Index Global*
k : *cluster* ke-k ($k = 2,3, \dots, q-1$) dengan q adalah jumlah *cluster*

Kriteria koefisien *silhouette* yang ditetapkan oleh Kaufman & Rousseeuw (1991) disajikan pada tabel 1 sebagai berikut :

Tabel 1. Kriteria Koefisien *Silhouette*

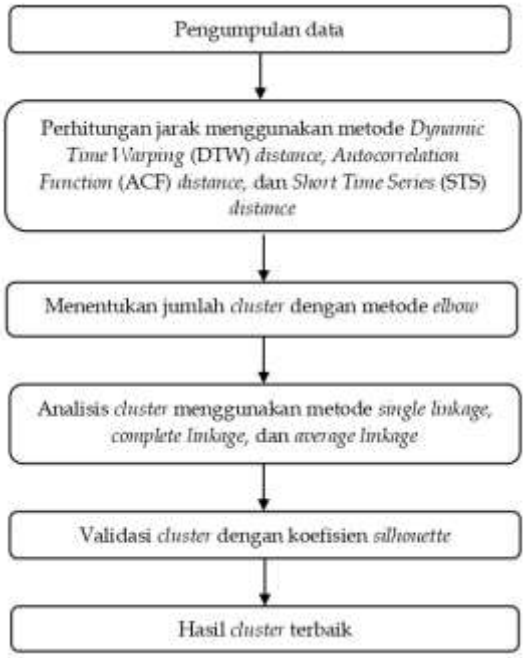
Nilai Koefisien *Silhouette*	**Kriteria *Cluster***
0,71 – 1,00	*Strong*
0,51 – 0,70	*Good*
0,26 – 0,50	*Weak*
0,00 – 0,25	*Bad*

METODE

Data yang digunakan dalam penelitian ini ialah data sekunder garis kemiskinan Kabupaten/Kota di Provinsi Jawa Timur tahun 2018 hingga 2022 yang berasal dari publikasi Badan Pusat Statistik (BPS) Jawa Timur pada website (https://jatim.bps.go.id/).

Diagram alir penelitian

Berikut adalah diagram alir berisi tahapan yang akan dilakukan pada penelitian ini:



Gambar 1. Diagram Alir Penelitian

Berdasarkan gambar 1, tahapan awal pada penelitian ini yaitu pengumpulan data garis kemiskinan 38 Kabupaten/Kota di Provinsi Jawa Timur selama 5 tahun mulai dari 2018 hingga 2022. Tahapan selanjutnya yaitu melalukan perhitungan jarak dengan menggunakan tiga metode ukuran jarak, yaitu *Dynamic Time Warping* (DTW) *distance, Autocorrelation Function* (ACF) *distance,* dan *Short Time Series* (STS) *distance.* Kemudian menentukan jumlah *cluster* terbaik untuk masing-masing metode analisis *cluster* menggunakan metode *elbow.* Langkah selanjutnya melakukan analisis *cluster* menggunakan metode hirarki *agglomerative* yang terdiri dari metode *single linkage, average linkage,* dan *complete linkage.* Setelah memperoleh hasil analisis *cluster,* dilakukanlah validasi *cluster* untuk mengetahui metode ukuran jarak dan metode analisis *cluster* terbaik dengan hasil *cluster* paling optimal.

HASIL DAN PEMBAHASAN

Pada bagian ini dijelaskan mengenai analisis pengelompokan 38 Kabupaten/Kota di Provinsi Jawa Timur terhadap garis kemiskikan dari Januari 2018 hingga Desember 2022.

Perhitungan jarak

Setelah melakukan pengumpulan data, dilakukanlah perhitungan dengan metode ukuran jarak. Setiap metode ukuran jarak mengukur masing-masing jarak antar dua deret waktu, yang artinya setiap satu Kabupaten/Kota akan dihitung jaraknya dengan satu Kabupaten/Kota lainnya dari Kabupaten/Kota ke-1 hingga Kabupaten/Kota ke-38, sehingga nantinya akan terbentuk matriks 38x38 untuk setiap metode ukuran jarak.

Berikut disajikan tabel perhitungan jarak dari masing-masing metode ukuran jarak.

Tabel 2. Hasil Perhitungan Jarak *Dynamic Time Warping* (DTW) *Distance*

	1	2	3	…	38
1	0				
2	3,58415	0			
3	5,37512	1,60741	0		
…	…	…	…	…	…
38	25,87122	22,09323	20,39875	…	0

Tabel 3. Hasil Perhitungan Jarak *Autocorrelation Function* (ACF) *Distance*

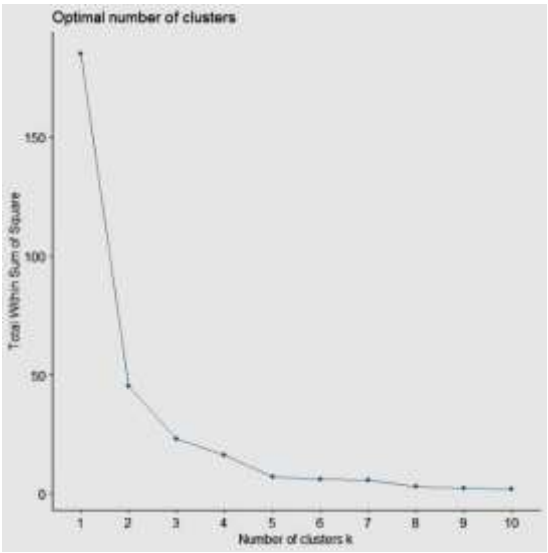
	1	2	3	…	38
1	0				
2	0,16030	0			
3	0,22667	0,24708	0		
…	…	…	…	…	…
38	0,37671	0,23760	0,37968	…	0

Tabel 4. Hasil Perhitungan Jarak *Short Time Series* (STS) *Distance*

	1	2	3	…	38
1	0				
2	0,02920	0			
3	0,08888	0,07041	0		
…	…	…	…	…	…
38	0,11829	0,12166	0,18127	…	0

Penentuan Jumlah *Cluster*

Setelah melakukan perhitungan jarak, yang dilakukan adalah melakukan menentukan jumlah *cluster* dengan metode *elbow*. Hasil penentuan jumlah *cluster* terbaik disajikan pada gambar 2 berikut:



Gambar 2. Hasil Metode *Elbow*

Hasil dari metode *elbow* menunjukan bahwa pengelompokan terbaik pada penelitian ini dikelompokan menjadi 5 *cluster* yang menggunakan metode hirarki *single linkage, average linkage,* dan *complete linkage* dengan masing-masing menggunakan metode ukuran jarak *Dynamic Time Warping* (DTW) *distance, Autocorrelation Function* (ACF) *distance,* dan *Short Time Series* (STS) *distance.*

Hasil Analisis *Cluster*

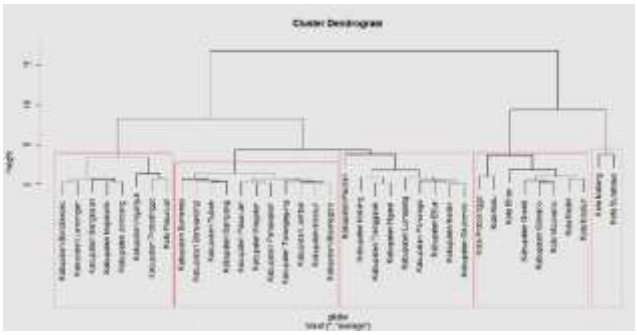
Setelah melakukan pengukuran jarak dan mengetahui banyaknya *cluster* yang terbentuk, langkah selanjutnya melakukan analisis *cluster* menggunakan tiga metode ukuran jarak dan tiga metode analisis *cluster* hirarki *agglomerative*, yaitu *single linkage, average linkage*, dan *complete linkage*. Kemudian masing-masing hasil tersebut dicari hasil *cluster* paling optimal dengan membandingkan nilai koefisien *silhouette.*

Perbandingan nilai koefisien *silhouette* jarak *Dynamic Time Warping* (DTW) *distance* dengan masing-masing metode analisis *cluster* hirarki *agglomerative* disajikan dalam Tabel 5 berikut:

Tabel 5. Nilai Koefisien *Silhouette* Analisis *Cluster* dengan Jarak *Dynamic Time Warping* (DTW) *Distance*

Metode Analisis *Cluster*	**Nilai Koefisien *Silhouette***
Single Linkage	0,7635
Average Linkage	0,8145
Complete Linkage	0,7926

Pada Tabel 5 menunjukan bahwa hasil analisis *cluster* paling optimal pada jarak *Dynamic Time Warping* (DTW) *distance* dengan metode *average linkage,* karena memiliki nilai koefisien *silhouette* terbaik yaitu 0,8145 yang masuk kedalam kategori *strong*. Dendogram hasil analisis *cluster* paling optimal pada jarak *Dynamic Time Warping* (DTW) *distance* dengan metode *average linkage* disajikan pada Gambar 3 berikut:

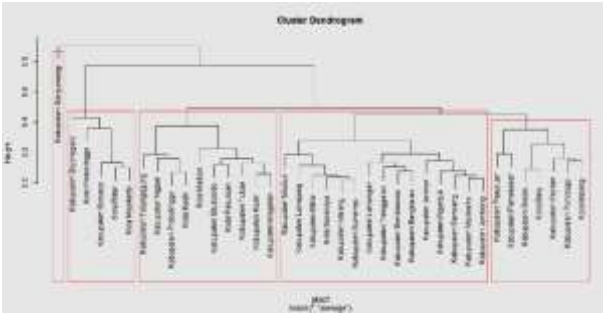


Gambar 3. Dendogram Hasil Analisis *Cluster* Paling Optimal pada Jarak *Dynamic Time Warping* (DTW) *Distance* dengan Metode *Average Linkage*

Tabel 6. Nilai Koefisien *Silhouette* Analisis *Cluster* dengan Jarak *Autocorrelation Function* (ACF) *Distance*

Metode Analisis *Cluster*	**Nilai Koefisien *Silhouette***
Single Linkage	0,7641
Average Linkage	0,8161
Complete Linkage	0,7941

Pada Tabel 6 menunjukan bahwa hasil analisis *cluster* paling optimal pada jarak *Autocorrelation Function* (ACF) *distance* dengan metode *average linkage,* karena memiliki nilai koefisien *silhouette* terbaik yaitu 0,8161 yang masuk kedalam kategori *strong*. Dendogram hasil analisis *cluster* paling optimal pada jarak *Autocorrelation Function* (ACF) *distance* dengan metode *average linkage* disajikan pada Gambar 4 berikut:

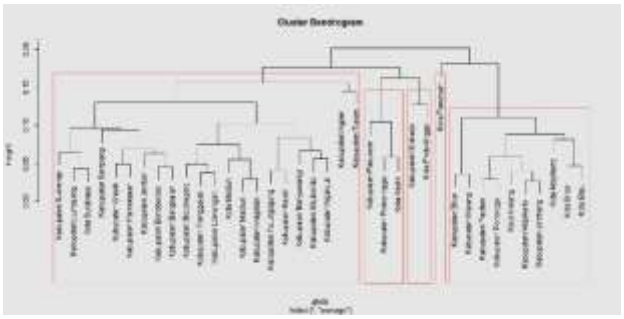


Gambar 4. Dendogram Hasil Analisis *Cluster* Paling Optimal pada Jarak *Autocorrelation Function* (ACF) *Distance* dengan Metode *Average Linkage*

Tabel 7. Nilai Koefisien *Silhouette* Analisis *Cluster* dengan Jarak *Short Time Series* (STS) *Distance*

Metode Analisis *Cluster*	**Nilai Koefisien *Silhouette***
Single Linkage	0,4892
Average Linkage	0,7087
Complete Linkage	0,6131

Pada Tabel 7 menunjukan bahwa hasil analisis *cluster* paling optimal pada jarak *Short Time Series* (STS) *distance* dengan metode *average linkage,* karena memiliki nilai koefisien *silhouette* terbaik yaitu 0,7087 yang masuk kedalam kategori *good.* Dendogram hasil analisis *cluster* paling optimal pada jarak *Short Time Series* (STS) *distance* dengan metode *average linkage* disajikan pada Gambar 5 berikut:



Gambar 5. Dendogram Hasil Analisis *Cluster* Paling Optimal pada Jarak *Short Time Series* (STS) *Distance* dengan Metode *Average Linkage*

Validasi *Cluster* dengan Koefisien *Silhouette*

Setelah mengetahui hasil *cluster* paling optimal dari masing-masing metode ukuran jarak, selanjutnya melakukan validasi hasil *cluster* paling optimal dari keseluruhan metode ukuran jarak dan metode analisis *cluster* menggunakan koefisien *silhouette*. Perbandingan koefisien *silhouette* disajikan pada Tabel 8 berikut:

Tabel 8. Perbandingan Nilai Koefisien *Silhouette*

Metode	**Nilai Koefisien *Silhouette***
jarak *Dynamic Time Warping* (DTW) *distance* dengan metode *average linkage*	0,8145
jarak *Autocorrelation Function* (ACF) *distance* dengan metode *average linkage*	0,8161
jarak Short Time Series *(STS)* distance *dengan metode* average linkage	0,7087

Dari hasil perbandingan nilai koefisien *silhouette* pada Tabel diketahui bahwa hasil *cluster* paling optimal ialah metode ukuran jarak *Autocorrelation Function* (ACF) *distance* dengan metode *average linkage* dengan nilai koefisien *silhouette* 0,8161 yang mana masuk kedalam kategori *strong*. Hasil *cluster* metode ukuran jarak *Autocorrelation Function* (ACF) *distance* dengan metode *average linkage* disajikan pada Tabel 9 berikut:

Tabel 9. Hasil Analisis *Cluster* Paling Optimal

Cluster	**Kabupaten/Kota**
1	Kab. Pacitan, Kab. Ponorogo, Kab. Pasuruan, Kab. Gresik, Kab. Pamekasan, Kota Malang, Kota Batu
2	Kab. Trenggalek, Kab. Blitar, Kab. Malang, Kab. Lumajang, Kab. Jember, Kab. Bondowoso, Kab. Mojokerto, Kab. Jombang, Kab. Nganjuk, Kab. Madiun, Kab. Lamongan, Kab. Bangkalan, Kab. Sampang, Kab. Sumenep, Kota Surabaya
3	Kab. Tulungagung, Kab. Kediri, Kab. Situbondo, Kab. Probolinggo, Kab. Magetan, Kab. Ngawi, Kab. Tuban, Kota Kediri, Kota Pasuruan, Kota Madiun
4	Kab. Banyuwangi
5	Kab. Sidoarjo, Kab. Bojonegoro, Kota Blitar, Kota Probolinggo, Kota Mojokerto

Dari hasil *cluster* yang disajikan pada Tabel pengelompokan terbentuk menjadi 5 *cluster.* Berdasarkan data garis kemiskinan di Provinsi Jawa Timur pada tahun 2018 hingga 2022, hasil

pengelompokan menjadi 5 *cluster* dapat dikategorikan untuk *cluster* 1 adalah Kabupaten/Kota yang memiliki kondisi sangat darurat terhadap kemiskinan, *clsuter* 2 adalah Kabupaten/Kota yang memiliki kondisi darurat, *clsuter* 3 adalah Kabupaten/Kota yang memiliki kondisi cukup darurat, *clsuter* 4 adalah Kabupaten/Kota yang memiliki kondisi cukup aman, dan *clsuter* 5 adalah Kabupaten/Kota yang memiliki kondisi aman.

Berdasarkan Tabel, hasil *cluster* 1 terdiri dari Kab. Pacitan, Kab. Ponorogo, Kab. Pasuruan, Kab. Gresik, Kab. Pamekasan, Kota Malang, Kota Batu. *Cluster* 2 yaitu Kab. Trenggalek, Kab. Blitar, Kab. Malang, Kab. Lumajang, Kab. Jember, Kab. Bondowoso, Kab. Mojokerto, Kab. Jombang, Kab. Nganjuk, Kab. Madiun, Kab. Lamongan, Kab. Bangkalan, Kab. Sampang, Kab. Sumenep, Kota Surabaya. Untuk *Cluster* 3 Kab. Tulungagung, Kab. Kediri, Kab. Situbondo, Kab. Probolinggo, Kab. Magetan, Kab. Ngawi, Kab. Tuban, Kota Kediri, Kota Pasuruan, Kota Madiun. *Cluster* 4 hanya diisi oleh Kab. Banyuwangi dan pada *Cluster* 5 yaitu Kab. Sidoarjo, Kab. Bojonegoro, Kota Blitar, Kota Probolinggo, Kota Mojokerto.

Hasil dari penelitian ini diharapkan dapat menjadi informasi bagi masyarakat dan pemerintah terkait wilayah manakah yang perlu diperhatikan khusus terkait masalah kemiskinan.

PENUTUP

SIMPULAN

Berdasarkan hasil analisis *cluster* menggunakan metode ukuran jarak *Dynamic Time Warping* (DTW) *distance, Autocorrelation Function* (ACF) *distance,* dan *Short Time Series* (STS) *distance,* dengan metode analisis *cluster* yaitu metode *single linkage, average linkage,* dan *complete linkage*, diperoleh hasil *cluster* paling optimal menggunakan metode ukuran jarak *Autocorrelation Function* (ACF) *distance* dengan metode *average linkage* dengan nilai koefisien *silhouette* 0,8161 yang mana masuk kedalam kategori *strong.*

SARAN

Untuk penelitian selanjutnya terkait analisis *cluster* dapat dilakukan pengembangan terkait metode ukuran jarak yang digunakan, seperti metode *Global Alignment Kernel,* metode *Frechet,* dan lain-lain. Untuk metode *clustering* dapat menggunakan metode non hirarki, sehingga dapat mengetahui hasil metode ukuran jarak dan metode *clustering* terbaik.

DAFTAR PUSTAKA

Aprilia, K., & Sembiring, F. (2021). Analisis Garis Kemiskinan Makanan Menggunakan Metode Algoritma K-Means Clustering. *Informasi Dan Manajemen Informatika*, *2*(4), 1–10.

Artha, K. S., & Winarko, E. (2016). Perbandingan Eros, Euclidean Distance dan Dynamic Time Warping dalam Klasifikasi Data Multivariate Time Series Menggunakan kNN. *Prosiding Seminar Nasional Pendidikan Teknik Informatika (SENAPATI 2016)*, *Senapati*, 223–228.

Ayundari, I., & Sutikno. (2019). Penentuan Zona Musim di Mojokerto Menurut Karakteristik Curah Hujan Dengan Metode Time Series Based Clustering. *INFERENSI*, *2*(2), 63–70.

Bholowalia, P., & Kumar, A. (2014). EBK-Means: A Clustering Technique based on Elbow Method and K-Means in WSN. *International Journal of Computer Applications*, *105*(9), 975–8887.

Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (2008). Time Series Analysis : Forecasting and Control Fourth Edition. In *Handbook Of Medical Statistics*. New Jersey : John Wiley and Sons Inc. https://doi.org/10.1142/9789813148963_0009

Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster Analysis: Fifth Edition*. John Wiley and Son, Ltd.

Galeano, P., & Pella, D. (2000). Multivariate Analysis in Vector Time Series. *Resenhas, the Journal of the Institute of Mathematics and Statistics of the University of Sao Paolo*, *4*, 383–403.

Irwanto, Purwananto, Y., & Soelaiman, R. (2012). Optimasi Kinerja Algoritma Klasterisasi K-Means. *Jurnal Teknik ITS*, *1*(1), 197–202.

Johnson, R. A., & Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis*. Pearson Education, Inc.

Kaufman, L., & Rousseeuw, P. J. (1991). Finding Groups in Data: An Introduction to Cluster Analysis. In *Biometrics* (Vol. 47, Issue 2). https://doi.org/10.2307/2532178

Liao, T. W. (2005). Clustering of time series data—a survey. *Pattern Recognition*, *38*(11), 1857–1874.

Möller-Levet, C. S., Klawonn, F., Cho, K. H., & Wolkenhauer, O. (2003). Fuzzy clustering of short time-series and unevenly distributed sampling points. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in*

Bioinformatics), *2810*(0), 330–340. https://doi.org/10.1007/978-3-540-45231-7_31

Mori, U., Mendiburu, A., & Lozano, J. A. (2016). *Distance Measures for Time Series in R : The TSdist Package*.

Niennattrakul, V., & Ratanamahatana, C. A. (2007). On clustering multimedia time series data using k-means and dynamic time warping. *Proceedings - 2007 International Conference on Multimedia and Ubiquitous Engineering, MUE 2007*, 733–738. https://doi.org/10.1109/MUE.2007.165

Nugroho, D., Asmanto, P., & Adji, A. (2020). Leading Indicators Kemiskinan Di Indonesia: Penerapan pada Outlook Jangka Pendek. In *The Nasional Team For The Acceleration Of Poverty Reduction (TNP2K)* (Vol. 92, Issue 11).

Pangestu, M. S., & Fitriani, M. A. (2022). Perbandingan Perhitungan Jarak Euclidean Distance, Manhattan Distance, dan Cosine Similarity dalam Pengelompokan Data Bibit Padi Menggunakan Algoritma K-Means. *Sainteks*, *19*(2), 141. https://doi.org/10.30595/sainteks.v19i2.14495

Pratama, Y. C. (2014). Analisis Faktor-Faktor Yang Mempengaruhi Kemiskinan di Indonesia. *Jurnal Bisnis Dan Manajemen*, *4*(2), 1–8. https://doi.org/10.36917/japabis.v1i2.9

Putu, N., Merliana, E., & Santoso, A. J. (2015). *Analisa Penentuan Jumlah Cluster Terbaik pada Metode K-Means*. 978–979.

Sardá-Espinosa, A. (2019). Time-series clustering in R Using the dtwclust package. *R Journal*, *11*(1), 1–45. https://doi.org/10.32614/rj-2019-023

Spiegel, S. (2015). *Time Series Distance Measures: Segmentation, Classification, and Clustering of Temporal Data*. 211.