



Identifikasi Motif Time Series Data Mining dengan Euclid

Relita Buaton¹, Muhammad Zarlis², Herman Mawengkang², Syahril Effendi²

¹Graduate Program Of Computer Science,

²Department Of Computer Science

Faculty of Computer Science and Information Technology,

Universitas Sumatera Utara, Medan, Indonesia

Email : bbcbuaton@gmail.com

ABSTRAK

Perkembangan teknologi informasi yang didukung oleh teknologi media storage yang membawa perubahan besar terhadap ketersediaan gudang data. Ketersediaan gudang data sering diabaikan karena ketidakmampuan mengolah datanya sehingga tumpukan data sering dianggap sebagai sampah, yang seharusnya digunakan sebagai pendukung keputusan, salah satu bentuk gudang data yang sering dijumpai saat ini adalah time series. Oleh sebab itu perlu mengembangkan sebuah metode untuk meningkatkan penemuan motif dari time series data mining. Metode yang digunakan adalah dengan melakukan diskritisasi sehingga dihasilkan sub sequen time series, yang selanjutnya dilakukan clustering. Jarak yang digunakan adalah distance euclid. Hasil yang diperoleh adalah ditemukannya motif pada time series berdasarkan waktu yang dapat digunakan sebagai pendukung keputusan dan prediksi dimasa yang akan datang.

Kata kunci: motif time series data mining, distance euclid clustering

ABSTRACT

The development of information technology that is supported by storage media technology has brought about major changes to the availability of data warehouses. The availability of data warehouses is often overlooked because of the inability to process the data so that data stacks are often regarded as garbage, which should be used as decision support, one form of data warehouse that is often encountered today is time series. Therefore it is necessary to develop a method to improve the motive discovery of time series data mining. The method used is discretion so that it produces time series sub sequences, which are then clustered. The distance used is the euclid distance. The results obtained are the finding of motives in time series based on time that can be used as decision support and predictions in the future.

Keywords: motif time series data mining, distance euclid clustering

1. Pendahuluan

Perkembangan teknologi informasi dan didukung oleh perkembangan media penyimpanan (*storage*) membawa dampak perubahan besar terhadap penerapan teknologi informasi pada semua bidang. Hampir seluruh bidang saat ini menerapkan teknologi informasi sehingga menghasilkan tumpukan data yang sangat besar pada setiap instansi. Gudang data sering dianggap sebagai sampah karena ketidakmampuan menggali data/informasi yang terkandung didalamnya, pada hal gudang data tersebut dapat digali untuk menghasilkan informasi sebagai pendukung keputusan. Salah satu gudang data yang tersedia dan *trend* saat ini adalah data *time series*. Pada data *time series* perlu dicari pola, dimana pola tersebut dapat digunakan sebagai informasi untuk pendukung keputusan atau prediksi dimasa yang akan datang.

Menurut (Serr 2015), Deteksi pola yang sangat mirip dalam *time series* biasanya disebut dengan motif. Pendekatan terbaru untuk menemukan motif-motif serupa dengan

panjang berbeda telah diusulkan. Pada penelitian ini, ditunjukkan bahwa motif diperoleh berdasarkan kesamaan panjang variabel yang tidak dapat dibandingkan secara langsung, namun dapat dinormalisasi. Secara khusus, ditemukan bahwa ketidaksesuaian panjang motif ternormalisasi masih memiliki ketergantungan *intrinsic* pada panjang motif, dan perbedaan terendah sangat dipengaruhi oleh ketergantungan ini. Selain itu ditemukan bahwa dependensi tersebut umumnya tidak linier dan berubah dengan kumpulan data yang dianggap tidak memiliki ukuran yang sama. Berdasarkan temuan ini, diusulkan solusi untuk memberi peringkat pada motif-motif tersebut dan mengukur signifikansinya. Solusi ini bergantung pada model ruang *dissimilarity* yang ringkas namun akurat, menggunakan distribusi beta dengan tiga parameter yang bergantung pada panjang motif dengan cara non linear, diyakini bahwa ketidaksesuaian variabel-variabel dapat melampaui ruang *time series*, dan bahwa strategi pemodelan serupa seperti yang digunakan di sini dapat membantu dalam konteks yang lebih luas. Deteksi pola berulang atau sangat mirip dalam *time series*, telah terbukti sangat bermanfaat bagi para peneliti dan praktisi. Terdapat dua definisi motif pada *time series*, yakni :

1. Didasarkan pada gagasan frekuensi (Lin et al. 2002), pola menarik jika memiliki sejumlah pengulangan yang signifikan.
2. Didasarkan pada gagasan kesamaan (Mueen, Cash, and Westover 2002): sebuah pola menarik jika kejadiannya identik atau terlalu mirip terjadi secara acak.

Kedua definisi tersebut bersifat komplementer, karena pola yang sangat mirip tidak selalu diperlukan sering terjadi. Diluar definisi berbasis frekuensi, peringkat motif yang ditemukan dalam suatu *time series*, dianggap tidak penting. Adapun motif yang dianggap paling penting adalah motif dengan hitungan tertinggi pertama, motif dengan hitungan tertinggi kedua, motif dengan hitungan tertinggi ketiga dan seterusnya. Motif juga dapat dinilai dari signifikansi statistik dengan membandingkan jumlah yang diamati dan hitungan yang diharapkan di bawah *null* yang mencerminkan beberapa karakteristik *time series* (Castro and Azevedo 2008). Pada penelitian ini ditemukan kendala dalam penemuan motif dari suatu deret *time series*, yakni apabila pasangan motif dari suatu *time series* memiliki panjang yang berbeda karena tidak dapat secara langsung membandingkan kesamaan atau jarak. Solusi yang dapat dilakukan adalah :

- a. Menghitung ranking dari setiap panjang motif yang memungkinkan untuk menutupi motif yang tidak sesuai (Mueen and Chavoshi 2015)
- b. Menormalkan ukuran ketidaksamaan dengan panjang motif, atau menggunakan ukuran yang telah menggabungkan beberapa gagasan normalisasi. Misalnya dengan membagi jarak Euclidean dengan akar kuadrat dari panjangnya, atau mempertimbangkan ukuran korelasi Pearson. Dalam hal signifikansi motif, pendekatan berbasis kesamaan kurang berkembang dari pada yang berbasis frekuensi.
- c. Mengembangkan model komparasi ruang disimilaritas motif, menggunakan distribusi beta yang parameternya tidak linier bergantung pada panjang motif.

Kontribusi utama dari penelitian ini adalah untuk menunjukkan bahwa ketidaksamaan panjang motif *time series* tidak dapat dibandingkan secara langsung, dan dengan demikian tidak bisa diranking. Melalui kedua contoh motivasi dan analisis kuantitatif formal, maka dapat ditunjukkan bahwa ketidaksamaan panjang motif ternormalisasi memiliki ketergantungan non-linear pada panjang motif, bahwa ketergantungan ini berubah sesuai dengan data set dan ukuran ketidaksamaan, bahwa data set sangat mempengaruhi perbedaan terendah, yang justru menjadi fokus perhatian pada kesamaan berbasis penemuan motif. Kontribusi lain dari karya ini adalah model

ketidaksamaan ruang yang memungkinkan membandingkan motif dengan panjang yang berbeda dan menilai mereka signifikansi sehubungan dengan distribusi ketidaksamaan secara keseluruhan, mengusulkan prosedur agar sesuai dengan *three-parameter beta distribution* saat mengambil memperhitungkan kontinuitas lokal dan non-linearitas ketidaksamaan motif ruang.

2. Metode Penemuan Motif

2.1. Distance Measures

Pengukuran jarak dalam clustering adalah mengukur kemiripan antar dua atau lebih objek, yang diperoleh dengan berbagai cara dan bentuk, diantaranya adalah:

1. *Euclidean Distance*, andaikan x_i dan y_j adalah sebagai vector P dimensi, maka perhitungan *distance Euclidean* dapat dihitung dengan menggunakan formula:

$$x = \sqrt{\sum_{k=1}^P (x_{ik} - y_{jk})^2} \quad (1)$$

2. *Dynamic Time Warping Distance (DTW)*

DTW algoritma digunakan sebagai perbandingan antara *discrete sequences* terhadap *continuous value sequence*. (Warren Liao 2005). Untuk pemodelan *time series* $Q = q_1, q_2 \dots q_i \dots q_n$ dan $R = r_1, r_2 \dots r_j \dots r_m$, DTW difungsikan untuk menghasilkan dan meminimalkan hasil atau perbedaan pada 2 time series tersebut. Sebuah matriks dengan ukuran $a \times b$, untuk elemen (i, j) yang meliputi distance $d(q_i, r_j)$ antara poin q_i dengan r_j . Dalam pengukuran ukuran matriks tersebut biasanya dengan menggunakan jarak distance Euclidean. Sebuah *warping path* $W = w_1, w_2 \dots w_k \dots w_m$, dimana $\max(a, b) \leq k \leq a + b - 1$ yaitu berbagai kumpulan elemen-elemen pada matriks yang memenuhi 3 kriteria yakni *boundary condition*, *continuity* dan *monotonicity*. Untuk menampilkan hasil jarak minimum yang terkandung dalam sebuah warping path dengan menggunakan formula berikut ini

$$d_{DTW} = \min \frac{\sum_{k=1}^k w_k}{K} \quad (2)$$

Untuk mengukur serta menentukan jarak akumulasi yakni dengan menggunakan formula

$$d_cum(i, j) = d(q_i, r_j) + \min\{d_cum(i-1, j-1), d_cum(i-1, j), d_cum(i, j-1)\} \quad (3)$$

Oates (Oates 1999) menerapkan rumus tersebut untuk menentukan *distinctive subsequence*, (Rakthanmanon et al. 2013) juga menerapkan pengukuran tersebut untuk menggali informasi yang terkandung dalam data time series yang signifikan

3. *Short Time Series Distances (STS)*

STS *distance* adalah hasil kuadrat jarak antara gradien terhadap dua atau lebih data *time series* (Möller-levet and Klawonn 2003), formula untuk menentukan jarak STS antar dua *time series* dituliskan sebagai berikut

$$d_{STS} = \sqrt{\sum_{k=1}^P \left(\frac{v_{j(k+1)} - v_{jk}}{t_{(k+1)} - t_k} - \frac{x_{i(k+1)} - x_{ik}}{t_{(k+1)} t_k} \right)^2} \quad (4)$$

t_k adalah time untuk data x_{ik} dan v_{jk}

4. *Minimum Description Length (MDL)*

Kriteria untuk pemilihan model tanpa mempertimbangkan kompleksitas batasan-batasan yang terdapat didalamnya dalam menghasilkan sampel data.

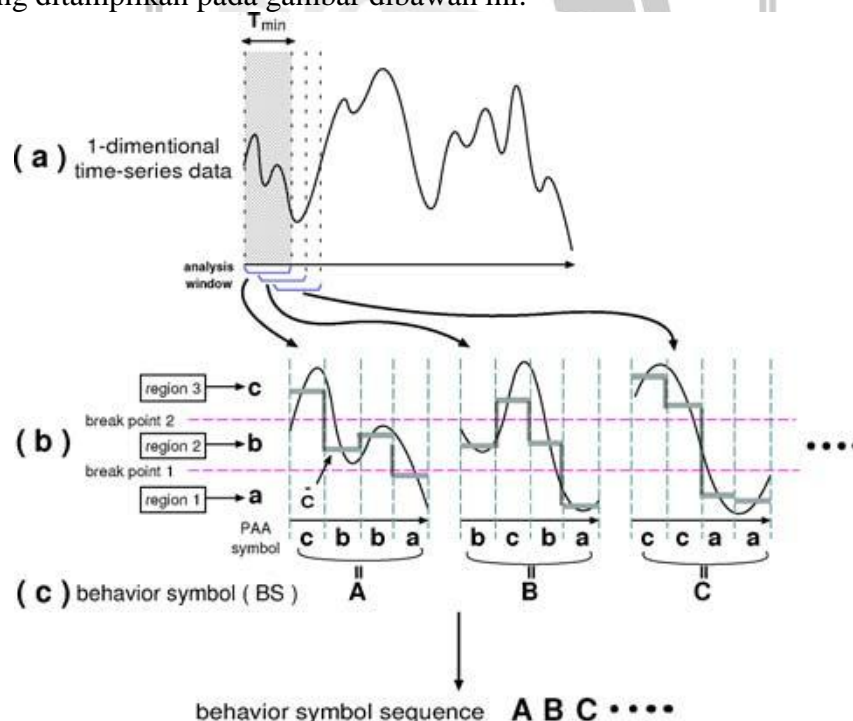
2.2. Motif Time Series

Metode untuk menemukan motif secara dinamis dalam mendeteksi motif pada data *time series* satu dimensi. Beberapa metode yang digunakan dalam menemukan motif adalah

1. *Minimum description length*(MDL) (Rissanen 1989)
2. *Akaike information criterion*(AIC) (Akaike 1969)
3. *Bayesian Information criterion*(BIC) (Schwarz 1981)

AIC mampu menemukan model terbaik dengan kemampuan hasil prediksi, dalam hal ini AIC menghasilkan pola yang sering tampil dan bukan berdasarkan hasil prediksi *time series*. BIC mampu menemukan model terbaik dengan menggunakan *teorema bayes* dan MDL mampu menghasilkan model terbaik dengan menerapkan dan meminimalkan panjang deskripsi kumpulan data secara keseluruhan. MDL adalah model terbaik untuk menghasilkan ringkasan atau rangkuman pada data *times series*. Tahapan yang digunakan adalah:

1. Transformasi *time series* menjadi beberapa urutan simbol
2. Melakukan ekstraksi pola dengan menerapkan prinsip MDL yang memiliki kelemahan bahwa pola yang mirip tidak tampil dalam *time series* dan terdapat beberapa *noise* didalamnya. Oleh sebab itu maka data *time series* diubah menjadi *sequence symbol* dengan menerapkan metode *reduce dimensionality* yang ditampilkan pada gambar dibawah ini.



Gambar 1. Hasil algoritma mengubah *time series* menjadi *sequence symbol* secara visual

- (a) Segmentasi grafik dengan menggeser analisa *window*
- (b) Mengubah hasil segmentasi grafik menjadi *symbol sequence*
- (c) Menghasilkan pola untuk setiap *symbol sequence*(Yoshiki Tanaka, Kazuhisa Iwamoto 2005)

2.3. Similaritas Deret Berkala

Hal penting dalam *time series* data mining adalah bagaimana menemukan motif dengan menentukan similarity yang terdapat dalam *time series data mining*. *Similarity* berkaitan erat dengan *clustering*. (Keogh and Kasetty, n.d.)(Fu 2011)(Serra and Arcos, n.d.)(Lauwers and Moor 2017). Himpunan *clustering* membutuhkan pengertian jarak untuk *time series* dengan panjang w . Ada beberapa hasil dan alternatif pilihan dalam menentukan ukuran jarak dalam penemuan motif. Alternatif yang paling sederhana yakni dengan menerapkan sistem sub sikuen dengan panjang w pada elemen R^w dan kemudian dengan menggunakan *distance Euclid* (yaitu, metric L_2). (Ding et al. 2008) telah menguji secara empiris dengan menerapkan *distance Euclid* sulit terkalahkan.

Dalam hal ini, model yang digunakan adalah, untuk $\hat{x} = (x_1, \dots, x_w)$ dan $\hat{y} = (y_1, \dots, y_w)$ didefinisikan

$$d(\hat{x}, \hat{y}) = (\sum_i (x_i - y_i)^2)^{1/2} \quad (5)$$

Sebagai hasil *metric* untuk mengclusterkan *metric* lainnya yang mencakup *metric* umum L_p dan dimodelkan dengan

$$L_p(\hat{x}, \hat{y}) = (\sum_i (x_i - y_i)^2)^{1/p} \quad (6)$$

3. Hasil dan Pembahasan

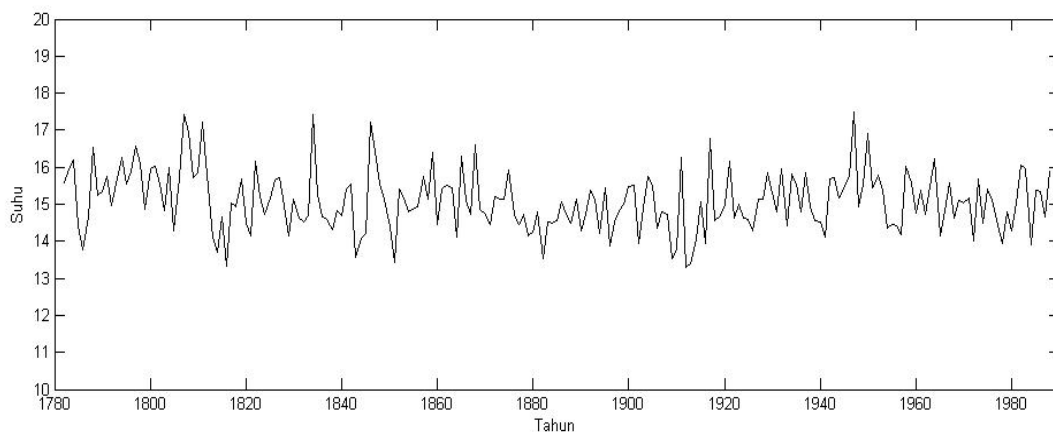
Data yang digunakan sebagai pengujian yakni data suhu rata-rata hasil pengukuran temperatur suhu

Tabel 1. Data *Time Series* Temperatur

Tahun	Suhu (°C)	Tahun	Suhu (°C)	Tahun	Suhu (°C)	Tahun	Suhu (°C)
1782	15,58	1885	14,57	1833	14,68	1936	14,8
1783	15,96	1886	15,06	1834	17,41	1937	15,84
1784	16,18	1887	14,72	1835	15,29	1938	14,93
1785	14,43	1888	14,5	1836	14,66	1939	14,54
1786	13,78	1889	15,13	1837	14,59	1940	14,51
1787	14,66	1890	14,3	1838	14,32	1941	14,1
1788	16,53	1891	14,72	1839	14,82	1942	15,68
1789	15,25	1892	15,37	1840	14,68	1943	15,72
1790	15,35	1893	15,13	1841	15,41	1944	15,18
1791	15,75	1894	14,23	1842	15,56	1945	15,45
1792	14,96	1895	15,46	1843	13,58	1946	15,76
1793	15,72	1896	13,89	1844	14,07	1947	17,49
1794	16,27	1897	14,52	1845	14,21	1948	14,94
1795	15,56	1898	14,83	1846	17,23	1949	15,54
1796	15,9	1899	15,01	1847	16,44	1950	16,9
1797	16,58	1900	15,47	1848	15,55	1951	15,44
1798	16,09	1901	15,52	1849	15,12	1952	15,79
1799	14,88	1902	13,93	1850	14,37	1953	15,38
1800	15,97	1903	14,91	1851	13,44	1954	14,36
1801	16,02	1904	15,76	1852	15,42	1955	14,46
1802	15,55	1905	15,52	1853	15,16	1956	14,43

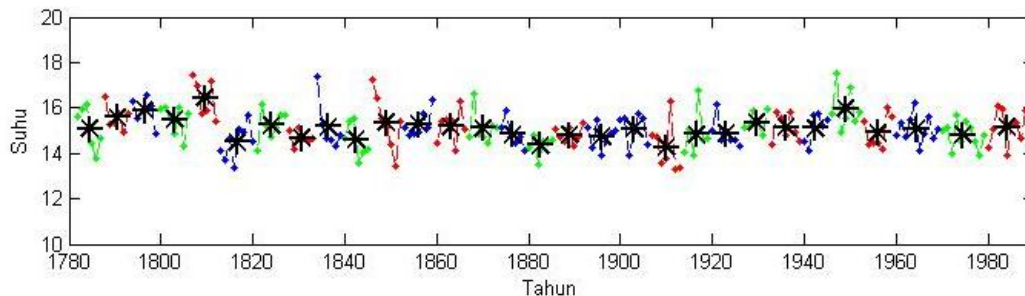
1803	14,84	1906	14,36	1854	14,81	1957	14,18
1804	15,99	1907	14,8	1855	14,88	1958	16,01
1805	14,29	1908	14,74	1856	14,93	1959	15,58
1806	15,72	1909	13,54	1857	15,74	1960	14,77
1807	17,43	1910	13,76	1858	15,14	1961	15,36
1808	16,94	1911	16,27	1859	16,39	1962	14,73
1809	15,73	1912	13,3	1860	14,46	1963	15,47
1810	15,87	1913	13,4	1861	15,4	1964	16,24
1811	17,21	1914	14,08	1862	15,5	1965	14,13
1812	15,43	1915	15,06	1863	15,45	1966	14,84
1813	14,1	1916	13,93	1864	14,1	1967	15,58
1814	13,72	1917	16,78	1865	16,3	1968	14,63
1815	14,65	1918	14,56	1866	15,09	1969	15,1
1816	13,34	1919	14,65	1867	14,72	1970	15,02
1817	15,04	1920	15	1868	16,62	1971	15,16
1818	14,93	1921	16,16	1869	14,87	1972	14
1819	15,68	1922	14,62	1870	14,72	1973	15,68
1820	14,5	1923	14,99	1871	14,44	1974	14,5
1821	14,13	1924	14,62	1872	15,22	1975	15,4
1822	16,17	1925	14,59	1873	15,14	1976	15,12
1823	15,22	1926	14,3	1874	15,14	1977	14,54
1824	14,71	1927	15,15	1875	15,91	1978	13,94
1825	15,16	1928	15,13	1876	14,71	1979	14,78
1826	15,65	1929	15,87	1877	14,46	1980	14,28
1827	15,71	1930	15,31	1878	14,71	1981	15,04
1828	14,98	1931	14,78	1879	14,13	1982	16,06
1829	14,15	1932	15,94	1880	14,24	1983	15,94
1830	15,12	1933	14,42	1881	14,78	1984	13,91
1831	14,64	1934	15,83	1882	13,53	1985	15,36
1832	14,51	1935	15,55	1883	14,53	1986	15,34
1884	14,48	1987	14,67	1988	15,88		

Data pada tabel 1 selanjutnya diplot menjadi grafik time series yang ditunjukkan pada gambar 2



Gambar 2. Grafik time series temperatur suhu rata-rata

Gambar 2. menunjukkan bahwa grafik plot data time series suhu rata-rata terhadap waktu(dalam tahun), dapat dilihat bahwa secara grafik terjadi perubahan bentuk sesuai dengan perubahan waktu setiap tahunnya. Selanjutnya dilakukan diskritisasi menjadi beberapa segmen,dalam percobaan ini segmentasi dilakukan untuk 10 window dan menjadi pusat analisis *cluster* dengan pengukuran jarak euclid



Gambar 3. Hasil *cluster* dengan jarak euclid

Tabel 2. Motif yang ditemukan pada setiap segmentasi

Segmen	Tahun	Pusat <i>cluster</i>	Pola Yang Ditemukan		
			Perubahan suhu	Pola perubahan Suhu	Waktu (Tahun)
I	1780-1800	1784-15,1 °C	Suhu cenderung 15,1 °C sampai dengan 15,9 °C	Start	
		1790-15,6 °C		Meningkat	6
		1796-15,9 °C		Meningkat	6
II	1801-1820	1803-15,5 °C	Suhu cenderung 14,5 °C sampai dengan 16,4 °C	Menurun	7
		1809-16,4 °C		Meningkat	6
		1816-14,5 °C		Menurun	7
III	1821-1840	1824-15,3 °C	Suhu cenderung 14,7 °C sampai dengan 15,3 °C	Meningkat	8
		1830-14,7 °C		Menurun	6
		1836-15,2 °C		Meningkat	6
IV	1841-1860	1842-14,6 °C	Suhu cenderung 14,6 °C sampai dengan 15,4 °C	Menurun	6
		1849-15,4 °C		Meningkat	7
		1856-15,3 °C		Menurun	7
V	1861-1880	1863-15,2 °C	Suhu cenderung 14,8 °C sampai dengan 15,2 °C	Menurun	7
		1870-15,1 °C		Menurun	7
		1876-14,8 °C		Menurun	6
VI	1881-1900	1882-14,4 °C	Suhu cenderung 14,4 °C sampai dengan 14,8 °C	Menurun	6
		1889-14,8 °C		Meningkat	7
		1896-14,7 °C		Menurun	7

Segmen	Tahun	Pusat cluster	Pola Yang Ditemukan		
			Perubahan suhu	Pola perubahan Suhu	Waktu (Tahun)
VII	1901-1920	1903-15,1 °C	Suhu cenderung 14,3 °C sampai dengan 15,1 °C	Meningkat	7
		1910-14,3 °C		Menurun	7
		1916-14,8 °C		Meningkat	6
VIII	1921-1940	1923-14,9 °C	Suhu cenderung 14,9 °C sampai dengan 15,4 °C	Meningkat	7
		1929-15,4 °C		Menurun	6
		1936-15,1 °C		Menurun	7
IX	1941-1960	1942-15,1 °C	Suhu cenderung 115,1 °C sampai dengan 16,0 °C	Normal	6
		1949-16,0 °C		Meningkat	7
		1956-15,1 °C		Menurun	7
X	1961-1980	1964-15,1 °C	Suhu cenderung 14,8 °C sampai dengan 15,2 °C	Normal	8
		1974-14,8 °C		Menurun	10
		1984-15,2 °C		Meningkat	10

Tabel 2 menyajikan informasi tentang motif yang dihasilkan berdasarkan pengukuran jarak euclid pada gambar 3. Jumlah baris pada kolom segmen sesuai dengan jumlah hasil segmentasi dan kolom tahun berdasarkan interval interval window yang telah ditetapkan. Masing-masing window memiliki 3 interval tahun yang dihasilkan dari perhitungan jarak similarity dan setiap interval tahun memiliki data peningkatan atau penurunan suhu rata-rata, data ini diperoleh dari titik-titik similarity pada gambar 3. Perubahan motif terjadi seiring dengan perubahan waktu dalam satuan tahun yang ditunjukkan pada kolom waktu. Terjadi peningkatan dan penurunan suhu rata-rata setiap window, informasi ini termasuk hal yang menarik untuk dijadikan pengetahuan dalam pengambilan keputusan atau prediksi dimasa yang akan datang.

4. Kesimpulan

Berdasarkan analisa pengetahuan dari setiap window maka dapat ditentukan pola atau motif yang berpotensi menarik dengan melakukan pruning , motif yang cenderung memiliki redundansi, berdasarkan analisa pengetahuan dari setiap window maka dapat ditentukan motif adalah suhu meningkat dan akan menurun dalam 6 sampai dengan 7 tahun sebesar 14,5 sampai dengan 15,5°C. Hal ini membuktikan bahwa dengan menggunakan jarak euclid maka mampu menemukan motif dari data *time series* data mining.

5. Daftar Pustaka

Akaike, Hirotugu. 1969. "Fitting Autoregressive Models For Prediction." *Annals of the Institute of Statistical Mathematics* 21:243–247.

- Castro, Nuno, and Paulo J Azevedo. 2008. "Time Series Motifs Statistical Significance."
- Ding, Hui, Goce Trajcevski, Peter Scheuermann, Xiaoyue Wang, and Eamonn Keogh. 2008. "Querying and Mining of Time Series Data : Experimental Comparison of Representations and Distance Measures."
- Fu, Tak-chung. 2011. "Engineering Applications of Artificial Intelligence A Review on Time Series Data Mining." *Engineering Applications of Artificial Intelligence* 24 (1). Elsevier:164–81. <https://doi.org/10.1016/j.engappai.2010.09.007>.
- Keogh, Eamonn, and Shruti Kasetty. n.d. "On the Need for Time Series Data Mining Benchmarks : A Survey and Empirical Demonstration."
- Lauwers, Oliver, and Bart De Moor. 2017. "A Time Series Distance Measure for Efficient Clustering of Input / Output Signals by Their Underlying Dynamics" 1 (2):286–91.
- Lin, Jessica, Eamonn Keogh, Stefano Lonardi, and Pranav Patel. 2002. "Finding Motifs in Time Series." *Proc. of the 2nd Workshop on Temporal Data Mining*, 53–68. <https://doi.org/10.1.1.19.6629>.
- Möller-levet, Carla Sofia, and Frank Klawonn. 2003. "Fuzzy Clustering of Short Time-Series and Unevenly Distributed Sampling Points" 2810 (June 2014). <https://doi.org/10.1007/978-3-540-45231-7>.
- Mueen, Abdullah, Sydney Cash, and Brandon Westover. 2002. "Exact Discovery of Time Series Motifs."
- Mueen, Abdullah, and Nikan Chavoshi. 2015. "Enumeration of Time Series Motifs of All Lengths." *Knowledge and Information Systems*. Springer London, 105–32. <https://doi.org/10.1007/s10115-014-0793-4>.
- Oates, Tim. 1999. "Identifying Distinctive Subsequences in Multivariate Time Series by Clustering." *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '99*, 322–26. <https://doi.org/10.1145/312129.312268>.
- Rakthanmanon, Thanawin, Bilson Campana, Abdullah Mueen, Gustavo Batista, Brandon Westover, Qiang Zhu, Jesin Zakaria, and Eamonn Keogh. 2013. "Addressing Big Data Time Series: Mining Trillions of Time Series Subsequences Under Dynamic Time Warping." *Transactions on Knowledge Discovery from Data (TKDD)* 7 (3):3047–51. <https://doi.org/10.1145/2500489>.
- Rissanen, Jorma. 1989. "Stochastic Complexity in Statistical Inquiry." *World Scientific Series in Computer Science*.
- Serr, Joan. 2015. "Ranking and Significance of Variable-Length Similarity-Based Time Series Motifs."
- Serra, Joan, and Josep Ll Arcos. n.d. "An Empirical Evaluation of Similarity Measures for Time Series Classification."
- Warren Liao, T. 2005. "Clustering of Time Series Data - A Survey." *Pattern Recognition* 38 (11):1857–74. <https://doi.org/10.1016/j.patcog.2005.01.025>.
- Yoshiki Tanaka, Kazuhisa Iwamoto, Kuniaki Uehara. 2005. "Discovery of Time-Series Motif from Multi-Dimensional Data Based on MDL Principle." *Springer Science + Business Media, Inc. Manufactured in The Netherlands*, no. 2000:269–300.