PERAMALAN KELOMPOK DERET WAKTU MENGGUNAKAN JARAK DTW DAN K-MEANS *CLUSTERING* (Studi Kasus Data COVID-19 di 34 Provinsi Indonesia)

Ryan Adi Nugroho

Jurusan Statistika, FMIPA, Universitas Brawijaya
Email:ryanadinugroho90@gmail.com

Abstrak. Deret waktu adalah urutan titik data yang diukur pada titik waktu yang berurutan. Analisis deret waktu terdiri dari metode yang mencoba memahami deret waktu tersebut atau membuat prakiraan Metode pengelompokan atau clustering adalah teknik untuk mengelompokkan pengamatan serupa ke dalam sejumlah kelompok. Pengaplikasian metode cluster dengan data deret akan memberikan informasi yang lebih berguna daripada metode analisis deret waktu saja Penelitian ini, akan menggabungkan dua metode yaitu time series forecasting dan metode clustering dengan jarak DTW pada data kasus penambahan COVID-19 di 34 provinsi Indonesia. Data penambahan kasus dari 34 provinsi akan dikelompokkan berdasarkan karakteristik yang mirip menggunakan KMeans clustering dan jarak DTW. Deret pusat dari setiap kelompok akan menjadi perwakilan dari kelompok kemudian diramal menggunakan ARIMA sehingga tidak perlu meramal semua provinsi. Hasil penelitian menunjukkan ari pengelompokan dengan jarak DTW dan K-Means didapat 5 cluster dengan masing – masing anggota sebanyak 4, 8, 9, 6, 7 Provinsi dan cluster 1, cluster 2 dan cluster 4 akan mengalami penurunan jumlah kasus positif pada 30 hari kedepan dan untuk cluster 3 dan cluster 5 akan mengalami kenaikan jumlah kasus positif pada 30 hari kedepan.

Kata Kunci: Cluster, COVID-19, Dynamic Time Warping, Time Series.

1. PENDAHULUAN

Perkembangan teknologi semakin hari semakin pesat mengikuti perkembangan zaman. Adanya perkembangan teknologi juga diikuti oleh perkembangan informasi. Dalam ilmu statistika, dikenal adanya istilah Data Mining, yaitu kegiatan mengekstraksi atau menambang informasi dari data yang berukuran/berjumlah besar (Fadli, 2003). Salah satu metode dalam Data Mining yang cukup terkenal dan informatif dalam menggali sebuah informasi yaitu metode cluster. Metode pengelompokan atau clustering adalah teknik untuk mengelompokkan pengamatan serupa ke dalam sejumlah kelompok berdasarkan nilai-nilai yang diamati dari beberapa variabel untuk setiap individu (Sinharay, 2010).

Menurut Sinharay (2010) Deret waktu adalah urutan titik data yang diukur pada titik waktu yang berurutan. Analisis deret waktu terdiri dari metode yang mencoba memahami deret waktu tersebut atau membuat prakiraan. Untuk memahami karakteristik-karakteristik yang dimiliki oleh suatu data deret waktu, para peneliti telah mengadopsi metode-metode analisis data deret waktu (*time series analysis*) dengan tujuan menemukan suatu keteraturan atau pola yang dapat digunakan dalam peramalan kejadian mendatang.

Dynamic Time Warping adalah salah satu metode pengukuran jarak yang paling tepat digunakan untuk data deret atau sequence dibanding dengan metode pengukuran jarak yang lain seperti Euclidean atau Manhattan. Metode DTW sering digunakan dalam speech recognition untuk membandingkan dua kata yang sama dalam waktu yang berbeda.

Pada penelitian ini, akan menggabungkan dua metode yaitu time series forecasting dan metode clustering dengan jarak DTW. Penelitian ini akan mencoba untuk meramal deret waktu kasus penambahan yang kelompok provinsi di Indonesia. Data kasus penambahan COVID-19 provinsi akan dikelompokkan menggunakan K-Means dan jarak DTW, kemudian nilai tengah dari tiap – tiap kelompok akan diramalkan menggunakan metode *forecast time series* sebagai perwakilan dari setiap cluster atau perwakilan dari beberapa provinsi

2. METODOLOGI

2.1 Sumber Data

Data yang digunakan dalam penelitian ini adalah data sekunder yang didapatkan dari situs Kaggle tentang penyebaran kasus COVID-19 di seluruh dunia. Data yang digunakan adalah data tentang penambahan kasus positif di semua provinsi di Indonesia pada bulan Maret 2020 sampai Maret 2021.

2.2. Metode Analisis

Tahap pertama adalah proses clustering dengan langkah – langkah sebagai berikut:

1. Penentuan jumlah *cluster* dengan metode *gap-statistic* dengan rumus
$$Gap(k) = E_n^*(\log(W_k)) - \log(W_k)$$
2. Inisialisasi deret pusat awal. Dilakukan dengan memilih secara acak deret waktu penambahan kasus COVID-19 per – provinsi

3. Menghitung jarak dari setiap data ke deret pusat yang telah ditetapkan sebelumnya. Penelitian ini menggunakan metode paling umum untuk menghitung jarak deret waktu yaitu dengan *Dynamic Time Warping* dengan rumus

$$D(i,j) = Dist(i,j) + \min [D(i+1,j), D(i,j+1), D(i+1,j+1)]$$

$$Dist(W) = \sum_{k=1}^{k=K} Dist(W_{ki}, W_{kj})$$

4. Mengelompokkan secara sementara data deret yang mempunyai jarak terdekat dengan deret pusat kedalam satu kelompok.
5. Perhitungan deret pusat baru dilakukan dengan mencari *mean* dari kelompok – kelompok yang terbentuk.
6. Mengulangi proses 3 sampai 6 sampai mendapatkan anggota cluster yang tetap.
7. Pengambilan perwakilan kelompok yang akan dilakukan *forecast time series*.

Perwakilan kelompok yang didapat akan diramal dengan model ARIMA dengan langkah – langkah sebagai berikut:

1. Melakukan pemeriksaan stasioneritas ragam menggunakan Uji Akar Unit *Dickey Fuller* dengan persamaan.

$$\Delta Y_t = \beta_1 + \beta_1 t + \delta Y_{t-1} + \sum_{t=1}^{m} \alpha_t \Delta Y_{t-1} + \varepsilon_t$$

2. Penetapan model sementara. Bisa dilakaukan menggunakan identifikasi orde model menggunakan plot ACF dan PACF.
3. pendugaan parameter menggunakan *Least Square Estimation* beserta uji signifikansi parameter.
4. Diagnostik sisaan model dengan uji *Ljung-Box* dengan uji statistik sebagai berikut.

$$Q = n(n+2) \sum_{i=1}^{K} (n-k)^{-1} \hat{\rho}_k^2$$

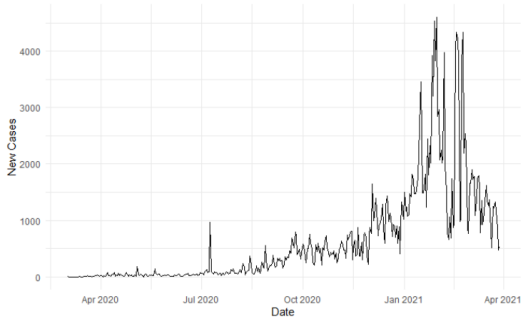
5. Pemilihan model terbaik dengan dasar nilai AIC terkecil dan *white noise* dengan rumus AIC sebagai berikut

$$AIC = -2\log(L) + 2k$$

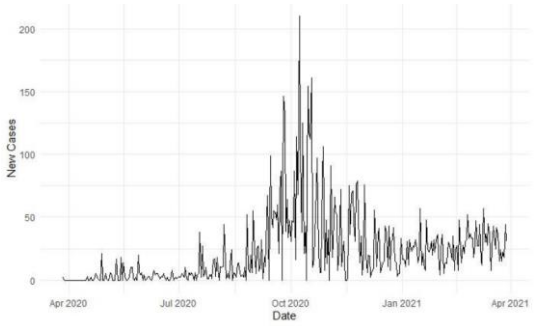
3. HASIL DAN PEMBAHASAN

3.1. Analisis Deskriptif

Secara umum, Indonesia memiliki 2 pola penyebaran kasus COVID-19 yaitu pola condong ke kanan atau tinggi di akhir dan tinggi di tengah seperti pada gambar.



Gambar 1. Pola penyebaran pertama



Gambar 2. Pola penyebaran kedua

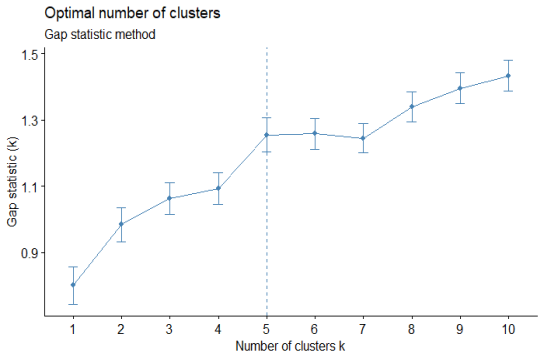
Penurunan kasus positif pada bulan – bulan tertentu mengindikasikan bahwa semakin hari pemerintah semakin bisa mengendalikan penyebaran COVID-19 pada provinsi di Indonesia, terbukti dengan suksesnya PPKM mikro (Pemberlakuan Pembatasan Kegiatan Masyarakat).

Pulau Jawa merupakan daerah dengan penyebaran tertinggi dari pulau lainnya dengan DKI Jakarta menjadi provinsi dengan penyebaran tertinggi dari daerah lainnya diikuti dengan Jawa Barat, Jawa Tengah dan Jawa Timur. Pulau Jawa mempunyai tingkat penyebaran yang paling tinggi karena menjadi pusat perindustrian dan perdagangan di Indonesia.

Untuk provinsi dengan tingkat penyebaran (new case) terendah adalah Provinsi Maluku Utara diikuti dengan Gorontalo, Sulawesi Barat dan Bengkulu. Provinsi ini mempunyai tingkat penyebaran yang rendah karena kegiatan masyarakat dan kedatangan mancanegara tidak begitu tinggi.

3.2. Analisis *Cluster*

Dari penentuan jumlah *cluster* menggunakan *gap-statistic* diperoleh jumlah *cluster* optimum sebanyak 5 dengan gambar sebagai berikut



Gambar 3. Jumlah *cluster* optimal

Dari K = 5, maka terbentuk anggota *cluster* sebagai berikut.

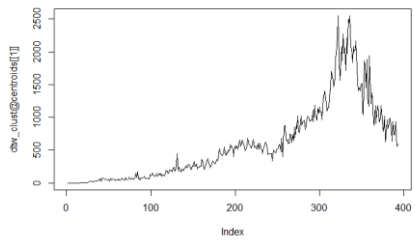
Tabel 1. Anggota *cluster*

Cluster	Anggota (Provinsi)
1	DKI Jakarta, Jawa Barat, Jawa Tengah, Jawa Timur
2	Riau, Banten, Kalimantan Timur, DI Yogyakarta, Sulawesi Selatan, Sumatera Barat, Bali, Kalimantan Selatan
3	Sumatera Utara, Papua, Kalimantan Tengah, Lampung, Sulawesi Tengah, Kalimantan Utara, Kepulauan Bangka Belitung, Sulawesi Utara, Nusa Tenggara Timur
4	Maluku, Maluku Utara, Sulawesi Barat, Kalimantan Barat, Nusa Tenggara Barat, Gorontalo
5	Sulawesi Tenggara, Jambi, Aceh, Papua Barat, Sumatera Selatan, Kepulauan Riau, Bengkulu

Dengan tingkat pemaparan yang paling tinggi adalah *cluster* 1 diikuti oleh *cluster* 2, *cluster* 3, *cluster* 5 dan tingkat pemaparan paling rendah dialami oleh *cluster* 4. Setelah *cluster* terbentuk maka Langkah selanjutnya adalah mengambil *means* (rata – rata) dari setiap *cluster* yang akan menjadi perwakilan dari anggota setiap *cluster*.

3.3. Peramalan *Time Series* *Cluster* 1

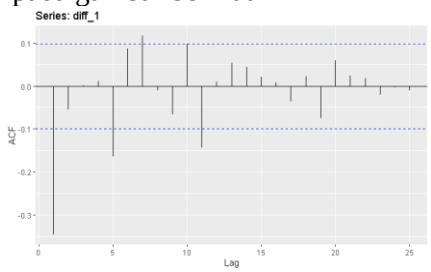
Nilai tengah dari *cluster* 1 yang terbentuk adalah sebagai berikut.



Gambar 4. Plot *means cluster* 1

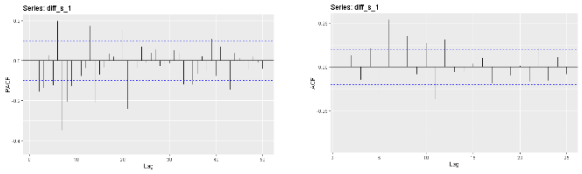
Dari pengecekan stasioneritas terhadap ragam, didapat nilai $\lambda = 0.273$, setelah menerapkan transformasi Box-Cox maka didapat data yang sudah stasioner terhadap ragam.

Dari pengecekan stasioneritas terhadap rata – rata didapat plot ACF yang sudah stasioner seperti pada gambar berikut



Gambar 5. Plot ACF *cluster* 1

Penentuan model tentatif menggunakan plot ACF dan PACF. Dari plot ACF dan PACF didapat model seasonal 7 degan plot sebagai berikut.



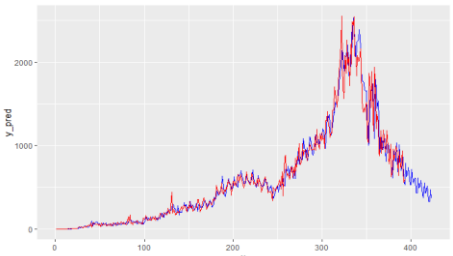
Gambar 6. Plot ACF dan PACF *cluster* 1

Maka didapat model tentative tertinggi yaitu (3, 1, 1)(3, 1, 1)7. Dari model tentatif selanjutnya pendugaan parameter dan signifikansi parameter. Didapat model dengan parameter signifikan sebagai berikut.

Tabel 2. Model dengan parameter signifikan

Model dengan Parameter Signifikan			
(0,1,1)(0,1,1)7	(1,1,0)(1,1,0)7	(2,1,0)(1,1,0)7	(3,1,1)(1,1,0)7
(0,1,1)(1,1,0)7	(1,1,0)(1,1,1)7	(2,1,0)(2,1,0)7	(3,1,1)(2,1,0)7
(0,1,1)(2,1,0)7	(1,1,0)(2,1,0)7	(2,1,0)(3,1,0)7	
(0,1,1)(3,1,0)7	(1,1,0)(3,1,0)7	(2,1,1)(1,1,0)7	
(1,1,0)(0,1,1)7	(2,1,0)(0,1,1)7	(3,1,0)(2,1,0)7	

Dari model diatas didapat model yang memenuhi asumsi *white-noise* dan memiliki nilai AIC terkecil adalah model (0,1,1)(3,1,0)7. Selanjutnya peramalan menghasilkan plot seperti berikut.

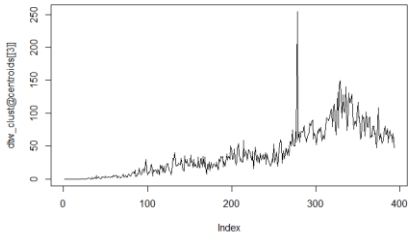


Gambar 7. Hasil prediksi *cluster* 1

Dari hasil peramalan menunjukkan bahwa nilai tengah (*means*) *cluster* 1 akan mengalami penurunan dengan pola mingguan yang berarti semua anggota *cluster* 1 akan mengalami penurunan jumlah kasus

***Cluster* 2**

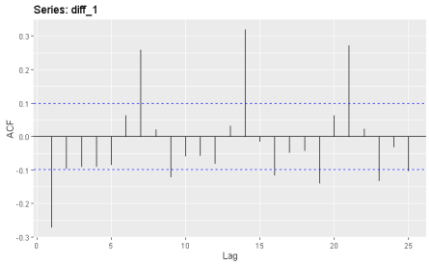
Nilai tengah dari *cluster* 2 yang terbentuk adalah sebagai berikut.



Gambar 8. Plot *means cluster* 2

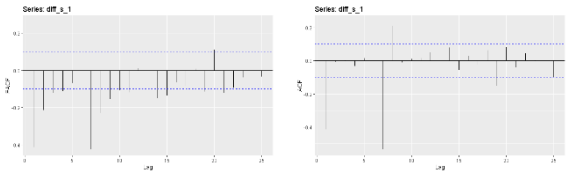
Dari pengecekan stasioneritas terhadap ragam, didapat nilai $\lambda = 0.318$, setelah menerapkan transformasi Box-Cox maka didapat data yang sudah stasioner terhadap ragam.

Dari pengecekan stasioneritas terhadap rata – rata didapat plot ACF yang sudah stasioner seperti pada gambar berikut



Gambar 9. Plot ACF *cluster* 2

Penentuan model tentatif menggunakan plot ACF dan PACF. Dari plot ACF dan PACF didapat model seasonal 7 degan plot sebagai berikut.



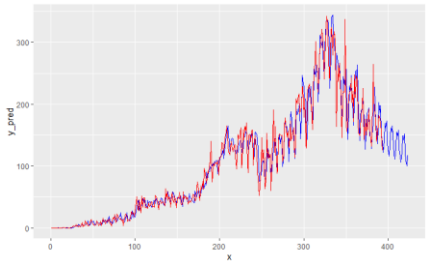
Gambar 10. Plot ACF dan PACF *cluster* 2

Maka didapat model tentative tertinggi yaitu (4, 1, 1)(3, 1, 1)7. Dari model tentatif selanjutnya pendugaan parameter dan signifikansi parameter. Didapat model dengan parameter signifikan sebagai berikut

Tabel 3. Model dengan parameter signifikan

Model dengan Parameter Signifikan			
(0,1,1)(0,1,1)7	(1,1,0)(2,1,0)7	(2,1,0)(3,1,0)7	(4,1,0)(0,1,1)7
(0,1,1)(1,1,0)7	(1,1,0)(3,1,0)7	(2,1,1)(1,1,0)7	(4,1,0)(1,1,0)7
(0,1,1)(2,1,0)7	(1,1,1)(0,1,1)7	(3,1,0)(0,1,1)7	(4,1,0)(2,1,0)7
(0,1,1)(3,1,0)7	(2,1,0)(0,1,1)7	(3,1,0)(1,1,0)7	(4,1,0)(3,1,0)7
(1,1,0)(0,1,1)7	(2,1,0)(1,1,0)7	(3,1,0)(2,1,0)7	
(1,1,0)(1,1,0)7	(2,1,0)(2,1,0)7	(3,1,0)(3,1,0)7	

Dari model diatas didapat model yang memenuhi asumsi *white-noise* dan memiliki nilai AIC terkecil adalah model (0,1,1)(0,1,0)7. Selanjutnya peramalan menghasilkan plot seperti berikut.

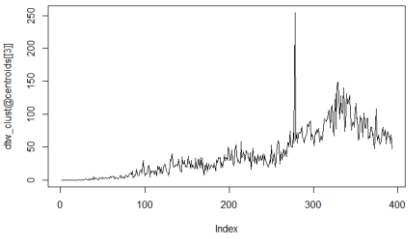


Gambar 11. Hasil prediksi *cluster* 2

Dari hasil peramalan menunjukkan bahwa nilai tengah (*means*) *cluster* 2 akan mengalami penurunan dengan pola mingguan yang berarti semua anggota *cluster* 2 akan mengalami penurunan jumlah kasus.

***Cluster* 3**

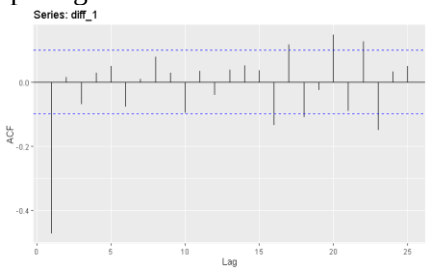
Nilai tengah dari *cluster* 3 yang terbentuk adalah sebagai berikut.



Gambar 12. Plot *means cluster* 3

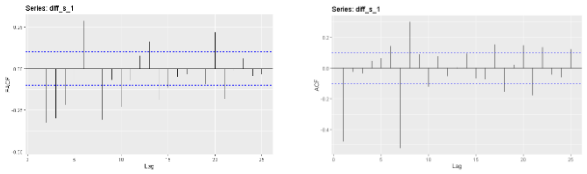
Dari pengecekan stasioneritas terhadap ragam, didapat nilai $\lambda = 0.207$, setelah menerapkan transformasi Box-Cox maka didapat data yang sudah stasioner terhadap ragam.

Dari pengecekan stasioneritas terhadap rata – rata didapat plot ACF yang sudah stasioner seperti pada gambar berikut



Gambar 13. Plot ACF *cluster* 3

Penentuan model tentatif menggunakan plot ACF dan PACF. Dari plot ACF dan PACF didapat model seasonal 7 degan plot sebagai berikut.



Gambar 14. Plot ACF dan PACF *cluster* 3

Maka didapat model tentative tertinggi yaitu (4, 1, 1)(3, 1, 1)7. Dari model tentatif selanjutnya pendugaan parameter dan signifikansi parameter. Didapat model dengan parameter signifikan sebagai berikut

Tabel 4. Model dengan parameter signifikan

Model dengan Parameter Signifikan			
(0,1,1)(0,1,1)7	(1,1,0)(1,1,0)7	(2,1,0)(2,1,0)7	(3,1,0)(3,1,0)7
(0,1,1)(1,1,0)7	(1,1,0)(2,1,0)7	(2,1,0)(3,1,0)7	(4,1,0)(0,1,1)7
(0,1,1)(2,1,0)7	(1,1,0)(3,1,0)7	(3,1,0)(0,1,1)7	(4,1,0)(1,1,0)7
(0,1,1)(3,1,0)7	(2,1,0)(0,1,1)7	(3,1,0)(1,1,0)7	(4,1,0)(2,1,0)7
(1,1,0)(0,1,1)7	(2,1,0)(1,1,0)7	(3,1,0)(2,1,0)7	(4,1,0)(3,1,0)7

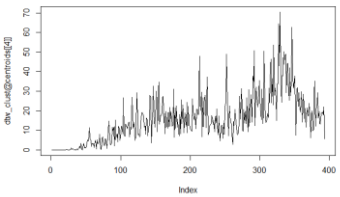
Dari model diatas tidak memenuhi asumsi *white-noise* dikarenakan terdapat pencilan, sehingga model dengan AIC terkecil dipilih, yaitu model (0,1,1)(0,1,0)7. Selanjutnya peramalan menghasilkan plot seperti berikut.

Gambar 15. Hasil prediksi *cluster* 3

Dari hasil peramalan menunjukkan bahwa nilai tengah (*means*) *cluster* 3 akan mengalami kenaikan dengan pola mingguan yang berarti semua anggota *cluster* 3 akan mengalami kenaikan jumlah kasus.

***Cluster* 4**

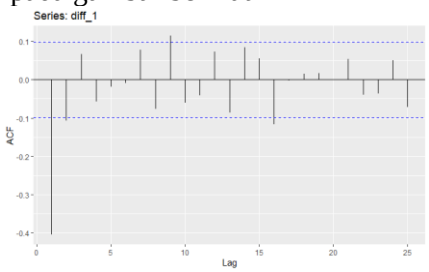
Nilai tengah dari *cluster* 4 yang terbentuk adalah sebagai berikut.



Gambar 16. Plot *means cluster* 4

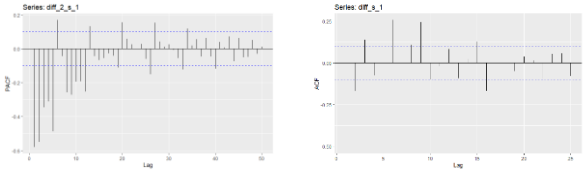
Dari pengecekan stasioneritas terhadap ragam, didapat nilai $\lambda = 0.217$, setelah menerapkan transformasi Box-Cox maka didapat data yang sudah stasioner terhadap ragam.

Dari pengecekan stasioneritas terhadap rata – rata didapat plot ACF yang sudah stasioner seperti pada gambar berikut



Gambar 16. Plot ACF *cluster* 4

Penentuan model tentatif menggunakan plot ACF dan PACF. Dari plot ACF dan PACF didapat model seasonal 7 degan plot sebagai berikut.



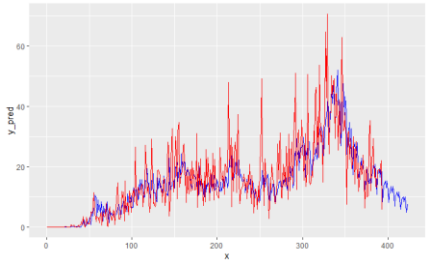
Gambar 17. Plot ACF dan PACF *cluster* 4

Maka didapat model tentative tertinggi yaitu (6, 2, 3)(0, 1, 1)7. Dari model tentatif selanjutnya pendugaan parameter dan signifikansi parameter. Didapat model dengan parameter signifikan sebagai berikut

Tabel 5. Model dengan parameter signifikan

Model dengan Parameter Signifikan		
(0,2,1)(0,1,1)7	(2,2,0)(0,1,1)7	(4,2,0)(0,1,1)7
(0,2,2)(0,1,1)7	(2,2,1)(0,1,1)7	(4,2,1)(0,1,1)7
(0,2,3)(0,1,1)7	(2,2,2)(0,1,1)7	(5,2,0)(0,1,1)7
(1,2,0)(0,1,1)7	(2,2,3)(0,1,1)7	(5,2,1)(0,1,1)7
(1,2,1)(0,1,1)7	(3,2,0)(0,1,1)7	(6,2,0)(0,1,1)7
(1,2,2)(0,1,1)7	(3,2,1)(0,1,1)7	(6,2,1)(0,1,1)7

Dari model diatas didapat model yang memenuhi asumsi *white-noise* dan memiliki nilai AIC terkecil adalah model (0,1,1)(0,1,0)7. Selanjutnya peramalan menghasilkan plot seperti berikut.



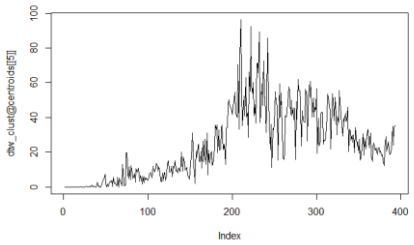
Gambar 18. Hasil prediksi *cluster* 4

Dari hasil peramalan menunjukkan bahwa nilai

tengah (*means*) *cluster* 4 akan mengalami penurunan dengan pola mingguan yang berarti semua anggota *cluster* 4 akan mengalami penurunan jumlah kasus

***Cluster* 5**

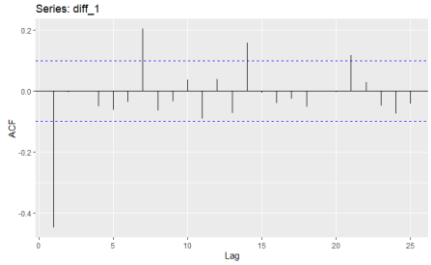
Nilai tengah dari *cluster* 5 yang terbentuk adalah sebagai berikut.



Gambar 19. Plot *means cluster* 5

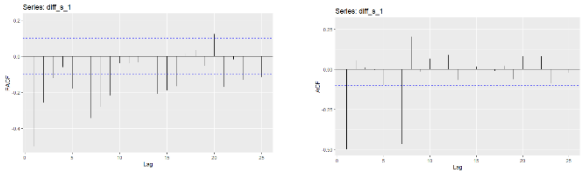
Dari pengecekan stasioneritas terhadap ragam, didapat nilai $\lambda = 0.285$, setelah menerapkan transformasi Box-Cox maka didapat data yang sudah stasioner terhadap ragam.

Dari pengecekan stasioneritas terhadap rata – rata didapat plot ACF yang sudah stasioner seperti pada gambar berikut



Gambar 20. Plot ACF *cluster* 5

Penentuan model tentatif menggunakan plot ACF dan PACF. Dari plot ACF dan PACF didapat model seasonal 7 degan plot sebagai berikut.



Gambar 21. Plot ACF dan PACF *cluster* 5

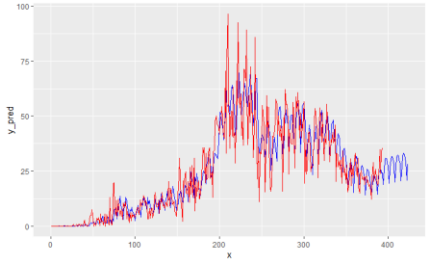
Maka didapat model tentative tertinggi yaitu (3, 1, 1)(3, 1, 1)7. Dari model tentatif selanjutnya pendugaan parameter dan signifikansi parameter. Didapat model dengan parameter signifikan sebagai berikut

Tabel 6. Model dengan parameter signifikan

Model dengan Parameter Signifikan			
(0,1,1)(0,1,1)7	(1,1,0)(1,1,0)7	(2,1,0)(2,1,0)7	(3,1,0)(3,1,0)7
(0,1,1)(1,1,0)7	(1,1,0)(2,1,0)7	(2,1,0)(3,1,0)7	(3,1,1)(2,1,0)7
(0,1,1)(2,1,0)7	(1,1,0)(3,1,0)7	(3,1,0)(0,1,1)7	(2,1,1)(2,1,0)7

Model dengan Parameter Signifikan			
(0,1,1)(3,1,0)7	(2,1,0)(0,1,1)7	(3,1,0)(1,1,0)7	
(1,1,0)(0,1,1)7	(2,1,0)(1,1,0)7	(3,1,0)(2,1,0)7	
(1,1,0)(1,1,1)7	(1,1,1)(1,1,0)7	(2,1,1)(1,1,0)7	

Dari model diatas didapat model yang memenuhi asumsi *white-noise* dan memiliki nilai AIC terkecil adalah model (0,1,1)(0,1,0)7. Selanjutnya peramalan menghasilkan plot seperti berikut.



Gambar 22. Hasil prediksi *cluster* 5

Dari hasil peramalan menunjukkan bahwa nilai tengah (*means*) *cluster* 5 akan mengalami kenaikan dengan pola mingguan yang berarti semua anggota *cluster* 5 akan mengalami kenaikan jumlah kasus

4. KESIMPULAN

Berdasarkan hasil analisis dapat disimpulkan bahwa:

1. Dari pengelompokan dengan jarak DTW dan K-Means didapat 5 *cluster* dengan masing – masing anggota sebanyak 4, 8, 9, 6, 7 anggota
2. Dari hasi peramalan titik tengah masing – masing *cluster* didapatkan bahwa *cluster* 1, *cluster* 2 dan *cluster* 4 akan mengalami penurunan jumlah kasus positif pada 30 hari kedepan dan untuk *cluster* 3 dan *cluster* 5 akan mengalami kenaikan jumlah kasus positif pada 30 hari kedepan.

DAFTAR PUSTAKA

Fadli, Ari. (2003). "Konsep Data Minning." *Konsep Data Mining*: 1-9.

Sinharay, S. (2010) An Overview of Statistics in Education. In: Peterson, P., et al., Eds., International Encyclopedia of Education, 3rd Edition, Elsevier Ltd., Amsterdam, 1-11. https://doi.org/10.1016/B978-0-08-044894-7.01719-X