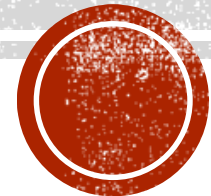


PERLOMBONGAN SIRI MASA

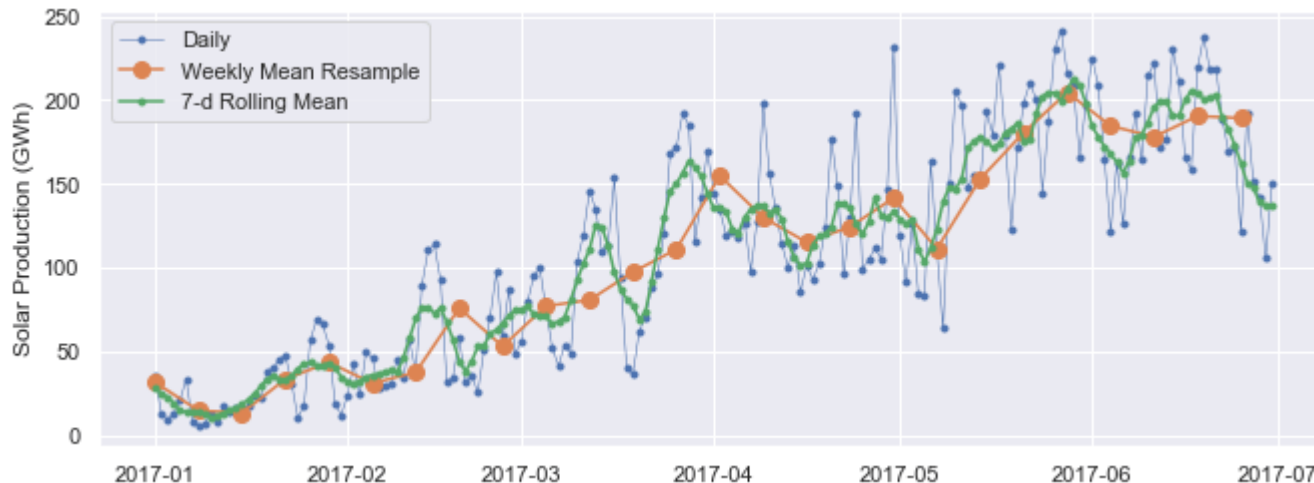
STQD6414 PERLOMBONGAN DATA



Prof. Madya Dr. Nurulkamal Masseran
Jabatan Sains Matematik
Universiti Kebangsaan Malaysia

Pengenalan:

- Data siri masa ialah himpunan cerapan yang diperolehi melalui pengukuran yang berulang dari semasa ke semasa.
- Data siri masa menjejaki maklumat suatu entiti dalam tempoh tertentu.
- Data siri masa adalah berbeza daripada data keratan rentas (*cross-sectional data*), yang mengumpul beberapa sampel berbeza pada masa yang sama.
- Secara umum, data siri masa berguna untuk menganalisis ciri dan tingkahlaku setiap jam, harian, mingguan atau tahunan yang berkaitan dengan entiti/peristiwa tertentu.



PENGENALAN:

- Contoh data siri masa:

- i) Rekod jualan runcit bulanan.
- ii) Pergerakan pasaran saham.
- iii) Peramalan cuaca.
- iv) Penunjuk ekonomi terhadap masa.
- v) Dan banyak lagi.

- Perlombongan siri masa meliputi empat topik utama (dalam domain masa):

i) Penguraian Siri Masa:

- Data siri masa boleh diuraikan kepada komponen trend, bermusiman, kitaran dan faktor rawak.



PENGENALAN:


ii) Peramalan Siri Masa:

- Membina model statistik dan kemudian menggunakannya untuk meramalkan nilai masa hadapan.
- Terdapat banyak model siri masa yang popular seperti model regresi siri masa, model ARIMA, model GARCH, model tak linear dan lain-lain.

iii) Pengkelompokan Siri Masa:

- Pengkelompokan siri masa ialah proses untuk membahagikan data siri masa berbilang kepada beberapa kelompok berdasarkan sifat kesamaan.
- Teknik umum pengkelompokan dibincangkan dalam kursus Pembelajaran Mesin.

iv) Pengelasan Siri Masa:

- Pengelasan siri masa bertujuan untuk membina model pengelasan berdasarkan data siri masa yang berlabel.
- Teknik umum klasifikasi data dibincangkan dalam Pembelajaran Mesin. 

MEMBINA OBJEK TARIKH DAN MASA:

- Pemasukan data siri masa dalam R biasanya dalam bentuk data harian, mingguan, bulanan, tahunan atau suku tahunan.
- R mengadaptasi format masa ISO 8601.
- Umumnya, fungsi `as.Date()` digunakan untuk membina objek tarikh dalam R.
- Aksara bagi tarikh perlu mematuhi format yang ditakrif menerusi symbol tertentu (dalam English), iaitu:
 - i) %Y: 4-digit year (2022)
 - ii) %y: 2-digit year (22)
 - iii) %m: 2-digit month (01)
 - iv) %d: 2-digit day of the month (12)
 - v) %A: weekday (Wednesday)
 - vi) %a: abbreviated weekday (Wed)
 - vii) %B: month (January)
 - viii) %b: abbreviated month (Jan)



KELAS DATA SIRI MASA DALAM R:

- Terdapat beberapa kelas untuk data siri masa dalam R. Antaranya:
 - i) Date.
 - ii) ts.
 - iii) timeSeries.
 - iv) Zoo.
 - v) Xts.
 - vi) POSIX.
 - vii) Dan banyak lagi.



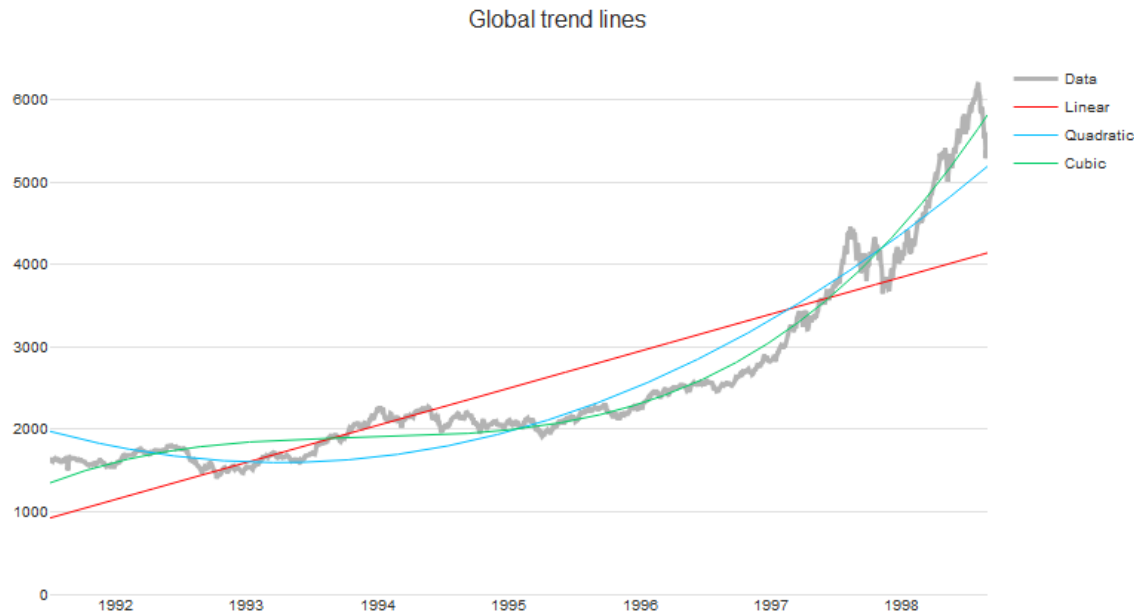
PENGURAIAAN SIRI MASA:

- Penguraian Siri Masa ialah kaedah untuk menguraikan data siri masa kepada beberapa komponen struktur iaitu:
 - i) Trend
 - ii) Kebermusiman (*Seasonality*)
 - iii) Kitaran (*Cyclical*)
- Dan juga komponen tak berstruktur:
 - i) Rawak.
- Penguraian siri masa memberikan maklumat tentang tingkah laku data siri masa bagi tujuan pemahaman yang lebih baik dalam menjalankan analisis dan ramalan siri masa.



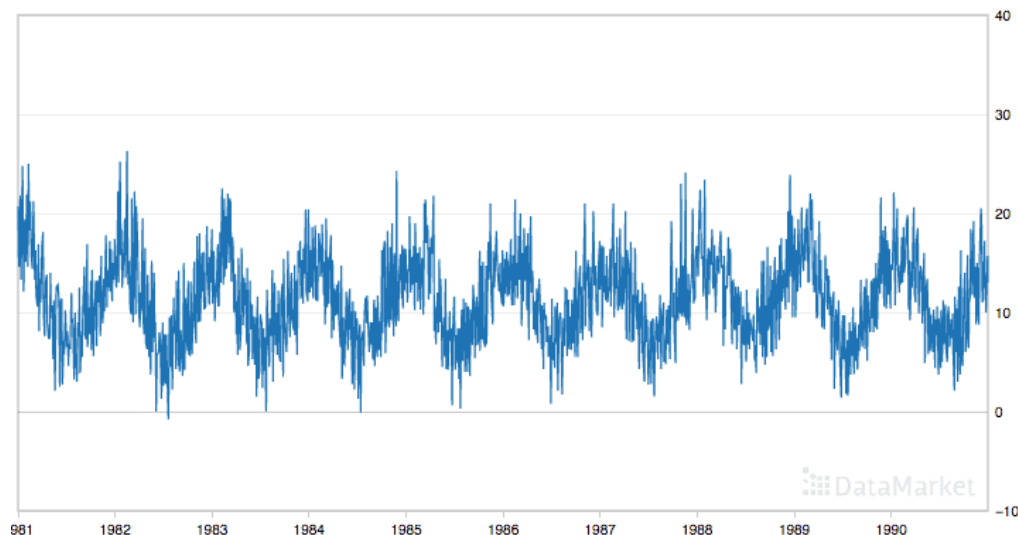
KOMPONEN TREND:

- Komponen trend merujuk kepada aliran meningkat atau menurun dalam data.
- Bergantung pada ciri data siri masa, trend boleh mempunyai bentuk pertumbuhan linear, polinomial atau eksponen.



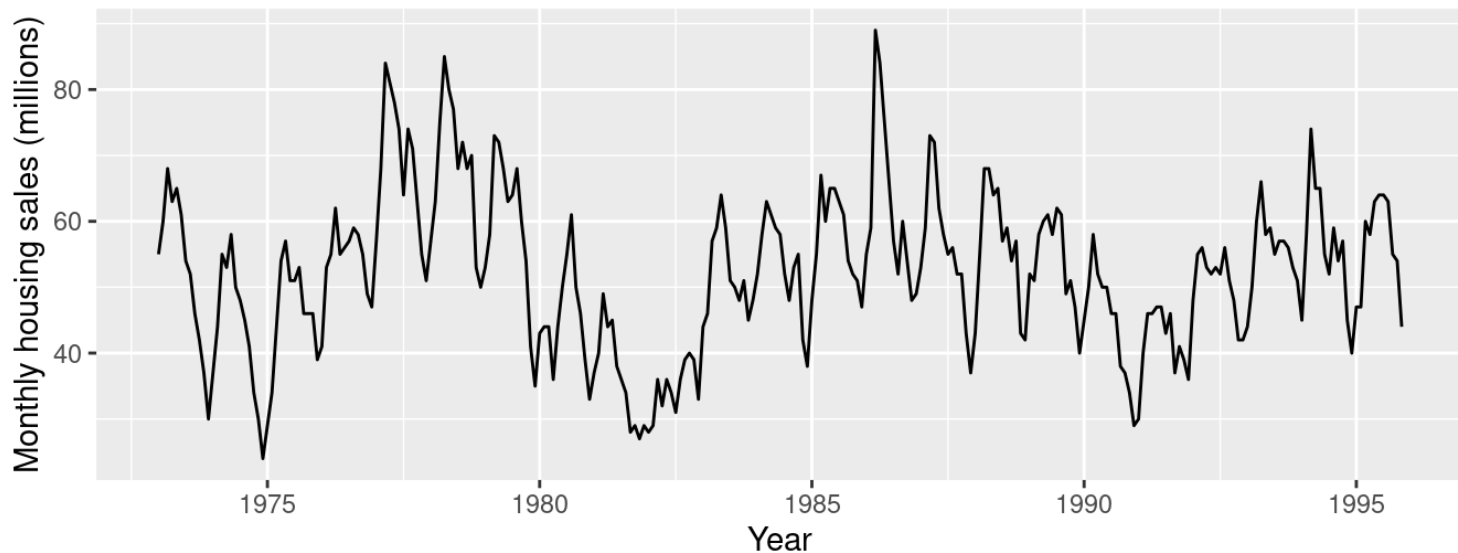
KOMPONEN BERMUSIM:

- Komponen bermusim ialah variasi bermusim yang berlaku secara berkala.
- Contoh:
- **Kebermusim per jam:** tingkahlaku data suhu cahaya matahari setiap jam sepanjang hari.
- **Kebermusiman mingguan:** tingkah laku kehadiran pelanggan di pasar raya dalam seminggu.
- **Kebermusiman bulanan:** tingkahlaku pembelian barangan keperluan setiap bulan.



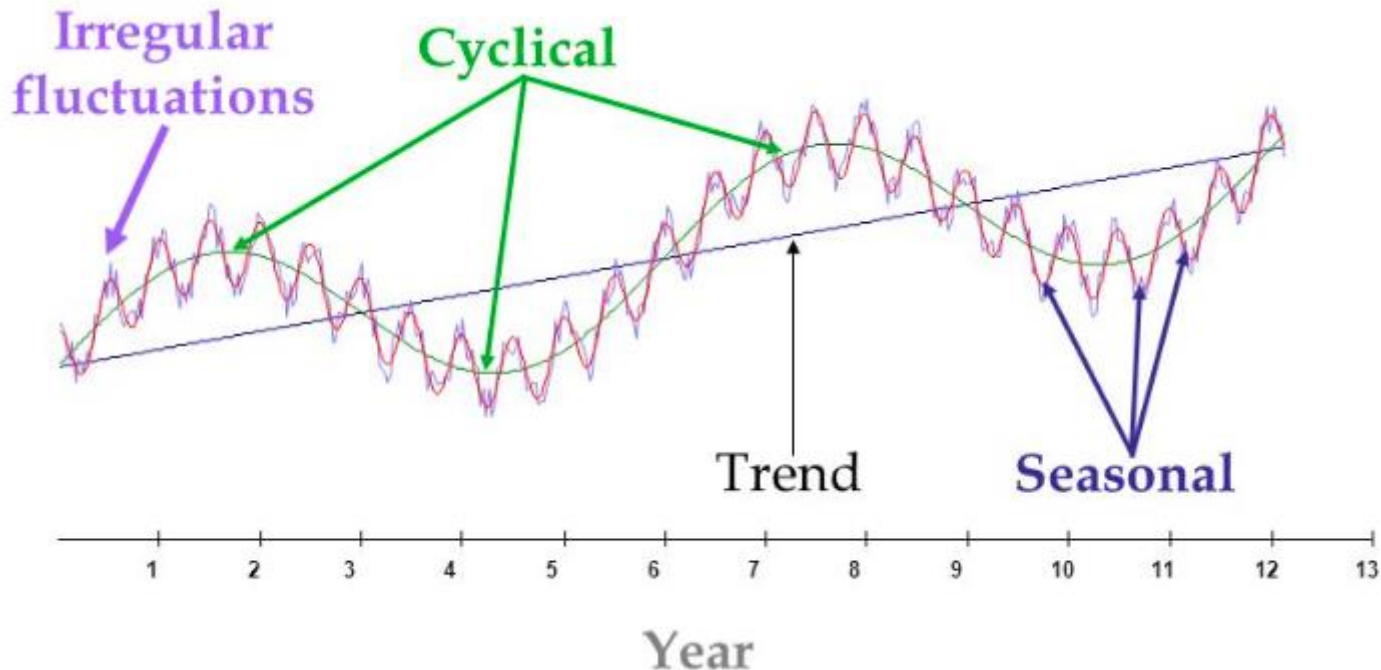
KOMPONEN KITARAN:

- Kitaran ialah jujukan peristiwa yang berulang dalam jangka masa yang panjang.
- Tidak seperti corak bermusim, kitaran tidak semestinya berlaku pada jarak masa yang sama, dan tempohnya boleh berubah dari kitaran ke kitaran.



KOMPONEN RAWAK:

- Komponen rawak (*irregular*) ialah baki antara data siri masa dengan komponen berstruktur (trend, kebermusiman dan kitaran).
- Ia merujuk kepada variasi data yang tidak boleh dijelaskan oleh komponen berstruktur.
- Ianya juga menunjukkan corak atau peristiwa yang tidak sistematik dalam data.

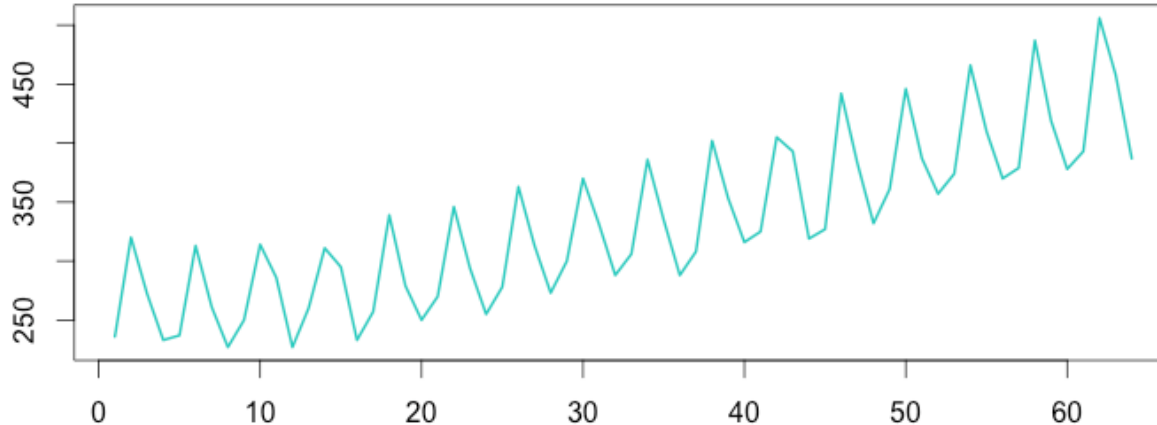


JENIS-JENIS PENGURAIAAN:

- Terdapat dua jenis penguraian siri masa, iaitu:

i) Penguraian Bertambah (*Additive Decomposition*):

- Struktur bertambah dalam siri masa wujud apabila terdapat pertumbuhan dalam trend, namun magnitud komponen bermusim secara umumnya hampir malar sepanjang masa.



- Struktur bertambah dalam siri masa boleh terangkan menerusi persamaan berikut:

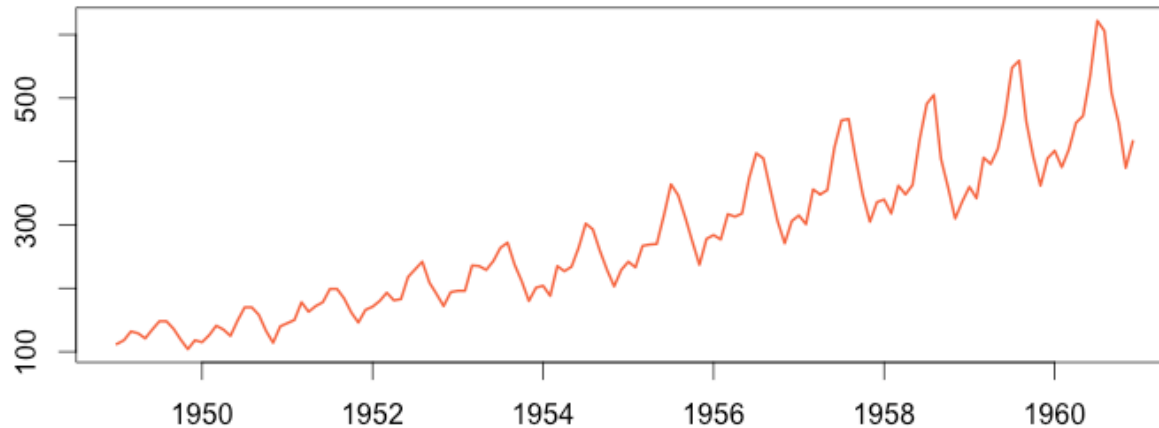
$$Y_t = T_t + S_t + C_t + I_t$$



JENIS-JENIS PENGURAIAAN:

ii) Penguraian Berganda (*Multiplicative Decomposition*):

- Struktur berganda dalam siri masa wujud apabila terdapat pertumbuhan dalam trend, namun magnitud komponen bermusim adalah tidak malar sepanjang masa.



- Struktur berganda dalam siri masa boleh diterangkan menerusi persamaan berikut:

$$Y_t = T_t \times S_t \times C_t \times I_t$$

- Struktur berganda dalam siri masa boleh dijemakan kepada struktur bertambah menerusi penjelmaan Box-Cox.

PERAMALAN SIRI MASA:

- Ramalan siri masa adalah bertujuan untuk meramal peristiwa masa depan berdasarkan data-data lepas.
- **Contoh:** meramal harga pembukaan stok saham berdasarkan prestasi saham masa lalu.
- Dua pendekatan utama dalam siri masa adalah pendekatan domain masa (*time domain*) dan pendekatann domain kekerapan (*frequency domain*) .
- Model asas untuk ramalan siri masa dalam pendekatan domain masa ialah model purata bergerak autoregressif (ARMA) dan model purata bergerak bersepadu autoregressif (ARIMA).



ASAS MODEL ARIMA:

- Model ARIMA terbentuk daripada gabungan model Autoregressif (AR) dan Purata Bergerak (MA)

(1) Model Autoregresif peringkat p , AR(p):

$$y_t = \delta + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t,$$

- y_t bergantung kepada p nilai-nilai cerapan yang lepas.

(2) Model Purata Bergerak peringkat q , MA(q):

$$y_t = \delta + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q},$$

- y_t bergantung kepada nilai-nilai q sebutan reja-reja yang lepas.



ASAS MODEL ARIMA:

(3) Gabungan Model AR dan MA p dan q menghasilkan model ARMA(p, q):

$$y_t = \delta + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} \\ + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q},$$

- y_t bergantung kepada p nilai-nilai lepas dari data dan q sebutan reja-reja lepas.
- Jika data tidak pegun, teknik pembezaan dilakukan terhadap data untuk menjadikan ianya pegun.
- Pembezaan peringkat i membentuk model ARIMA(p, i, q)



KEPEGUNAN SIRI MASA:

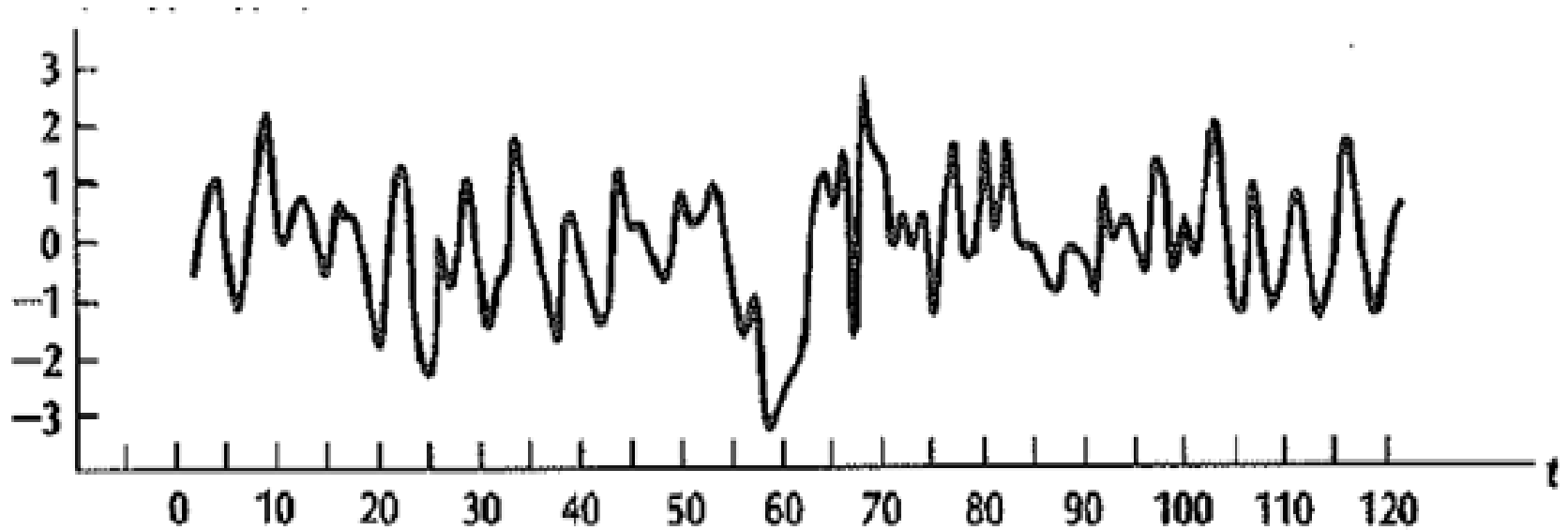
- Sifat “kepegunan” (*stationary*) ialah merupakan pra-syarat asas dalam aplikasi kebanyakan model statistik siri masa.
- Siri masa y_t dikatakan pegun jika ia memenuhi syarat berikut:

(1) $E(y_t) = u_y$ untuk semua t .

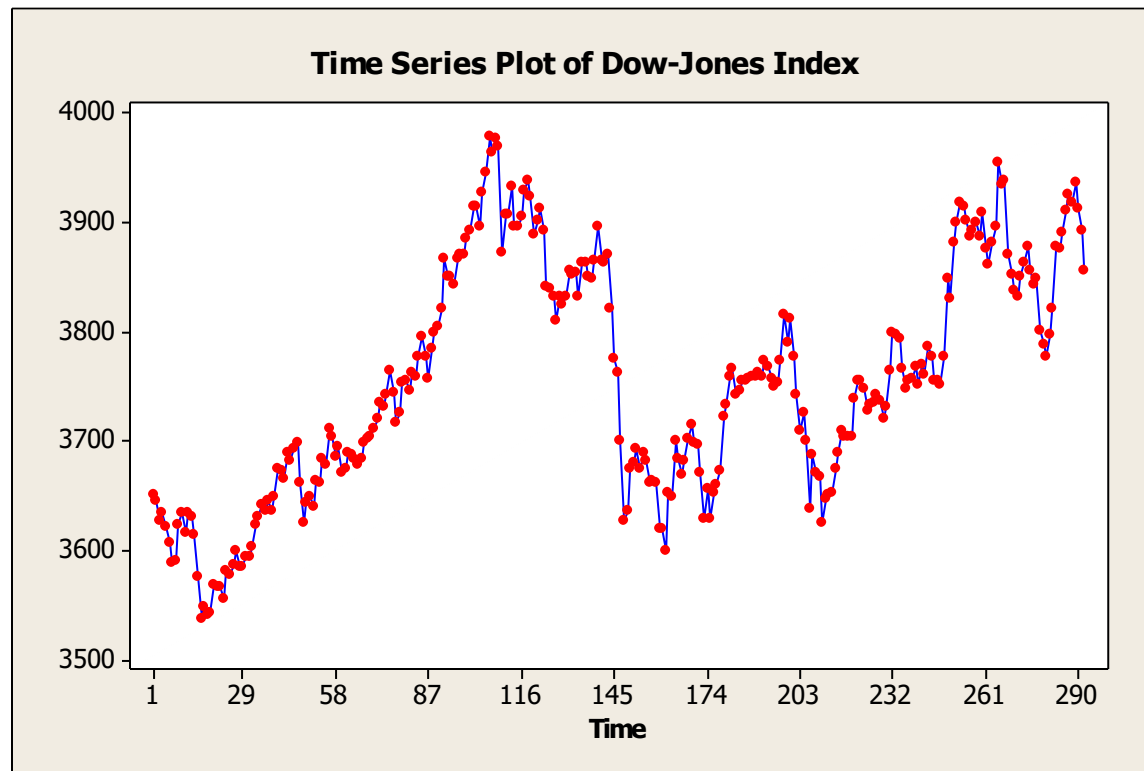
(2) $Var(y_t) = E[(y_t - u_y)^2] = \sigma_y^2$ untuk semua t .

(3) $Cov(y_t, y_{t-k}) = \gamma_k$ untuk semua t .

CONTOH: SIRI MASA PEGUN



CONTOH: SIRI MASA TAK PEGUN



PEMBEZAAN:

- Siri masa tak pegun boleh dijelmakan kepada siri masa pegun menerusi proses pembezaan (*differencing*).

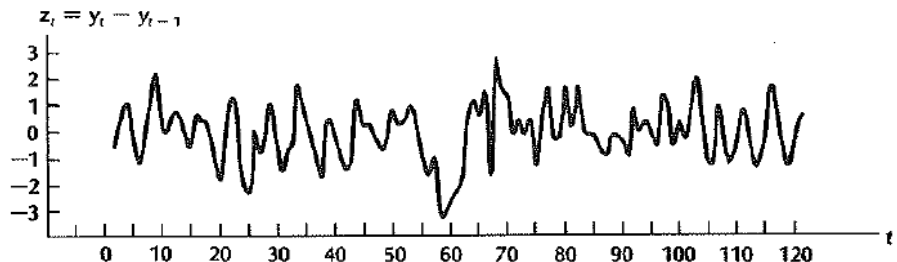
- Contoh:

y_t tak pegun , namun

$z_t = y_t - y_{t-1}$ adalah pegun



(a) Original values



(b) First differences

PEMBEZAAN:

- Proses pembezaan boleh dijalankan sehingga siri masa mencapai sifat pegun.

$$\Delta y_t = y_t - y_{t-1}$$

$$\Delta^2 y_t = \Delta(\Delta y_t) = \Delta(y_t - y_{t-1}) = y_t - 2y_{t-1} + y_{t-2}$$

- Bilangan pembezaan yang dijalankan untuk siri masa mencapai pegun dipanggil tertib integrasi, ditunjuk sebagai i .
- Secara umumnya, perbezaan tertib kedua adalah memadai.
- Pembezaan peringkat tinggi menjadikan model ARIMA lebih kompleks.

PENGECAMAN MODEL ARIMA:

- Setelah data pegun, kita boleh mengenal pasti model ARIMA yang sesuai menerusi pemeriksaan visual bagi fungsi autokorelasi (ACF) dan autokorelasi separa (PACF).
- Bagi siri masa y_1, y_2, \dots, y_n , autokorelasi bagi sampel pada lag- k ialah:

$$r_k = \frac{\sum_{t=1}^{n-k} (y_t - \bar{y})(y_{t+k} - \bar{y})}{\sum_{t=1}^n (y_t - \bar{y})^2}$$

- dengan $\bar{y} = \frac{\sum_{t=1}^n y_t}{n}$.

- Autokorelasi separa bagi sampel pada lag- k ialah:

$$r_{kk} = \begin{cases} r_1 & \text{if } k = 1, \\ \frac{r_k - \sum_{j=1}^{k-1} r_{k-1,j} \cdot r_{k-j}}{1 - \sum_{j=1}^{k-1} r_{k-1,j} \cdot r_k} & \text{if } k = 2, 3, \dots \end{cases}$$

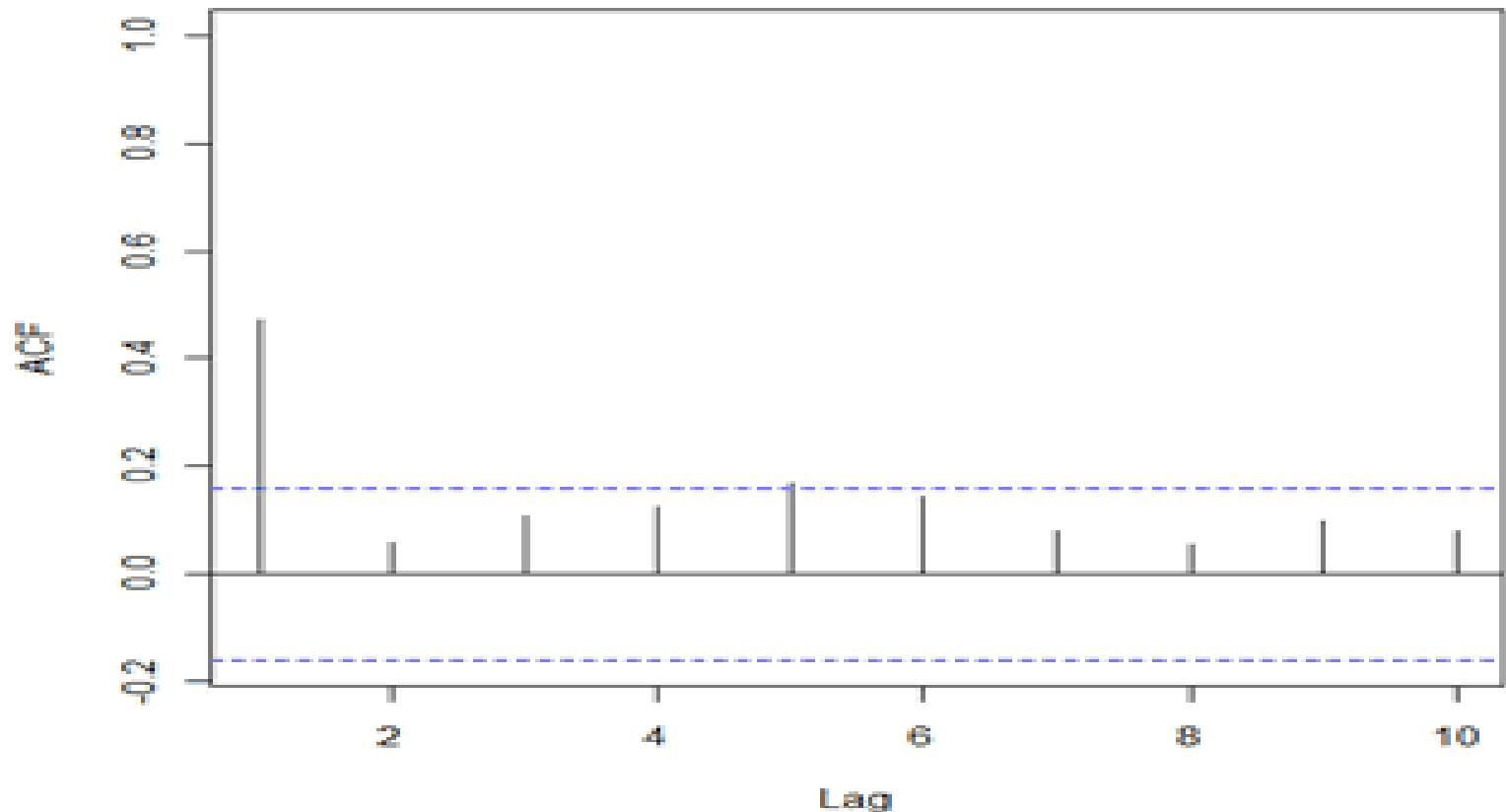
- dengan $r_{kj} = r_{k-1,j} - r_{kk} r_{k-1,k-j}$ untuk $j = 1, 2, \dots, k-1$.

TINGKAH LAKU ACF DAN PACF:

Model	AC	PAC
Autoregressive of order p $y_t = \delta + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_p y_{t-p} + \varepsilon_t$	Dies down	Cuts off after lag p
Moving Average of order q $y_t = \delta + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \dots - \theta_q \varepsilon_{t-q}$	Cuts off after lag q	Dies down
Mixed Autoregressive-Moving Average of order (p,q) $y_t = \delta + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \dots - \theta_q \varepsilon_{t-q}$	Dies down	Dies down

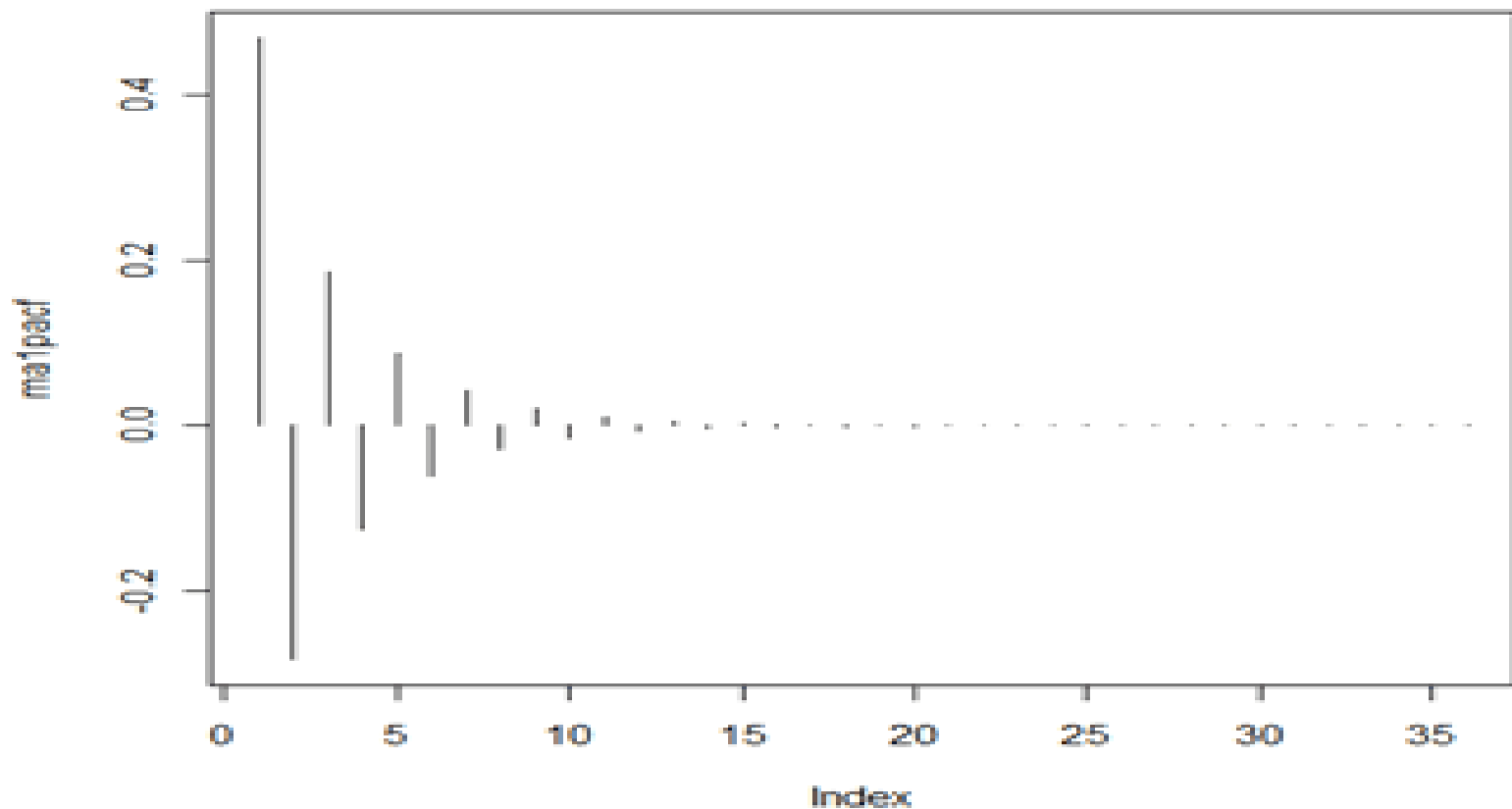
CONTOH: MODEL PURATA BERGERAK TERTIB-1, MA(1)

ACF for simulated sample data

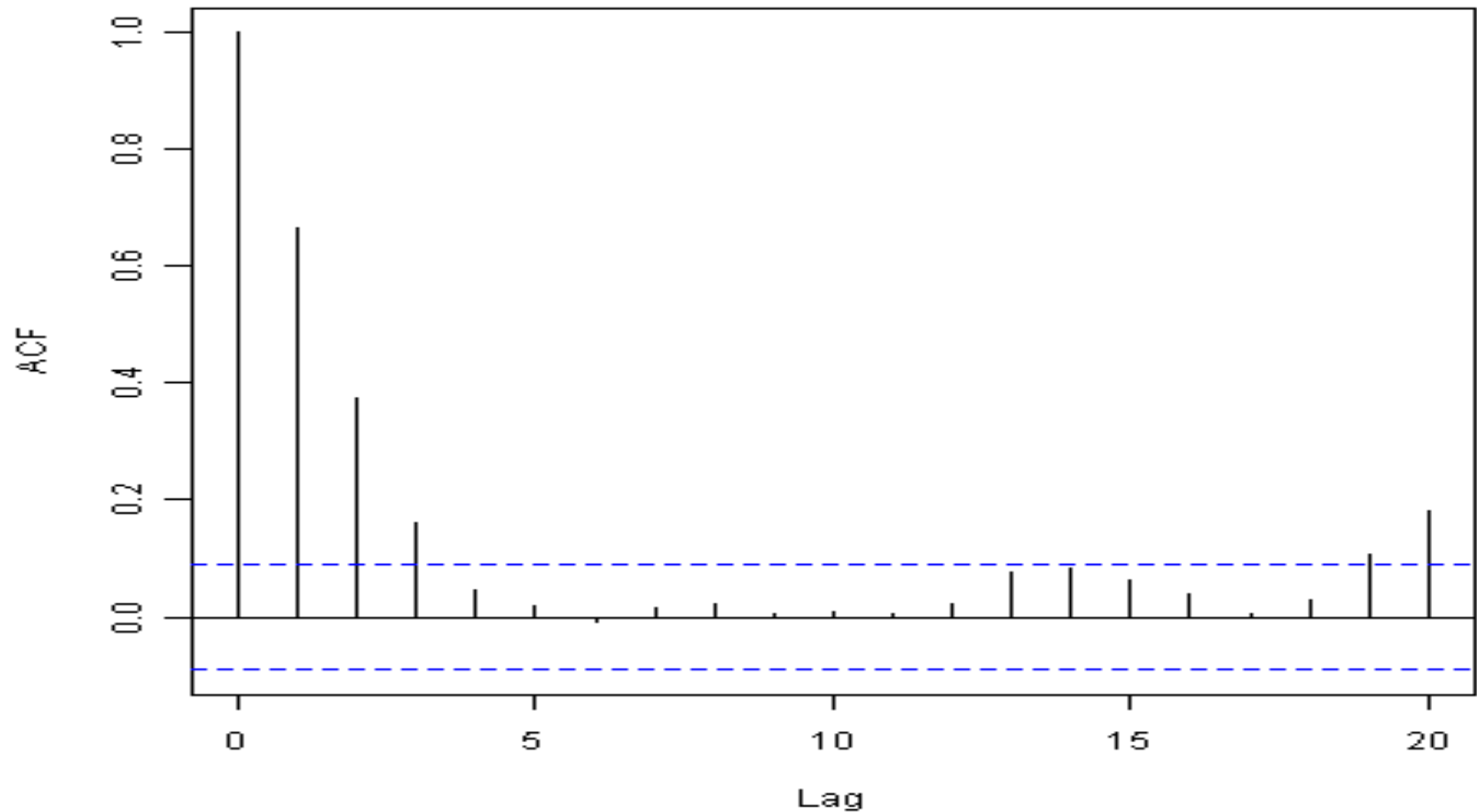


CONTOH: MODEL PURATA BERGERAK TERTIB-1, MA(1)

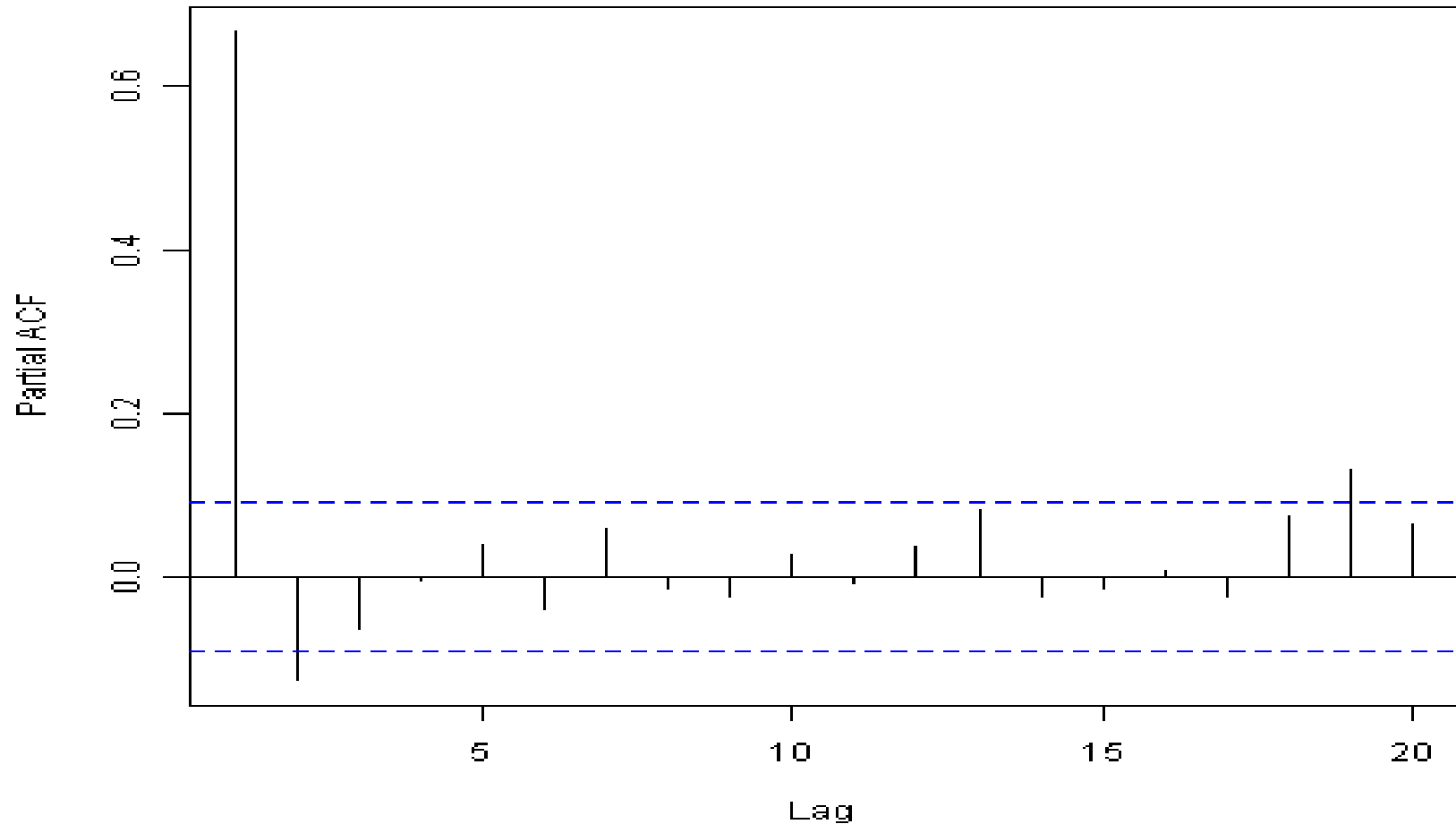
Theoretical PACF of MA(1) with $\theta = 0.7$



CONTOH: MODEL AUTOREGRESIF TERTIB-2, AR(2)

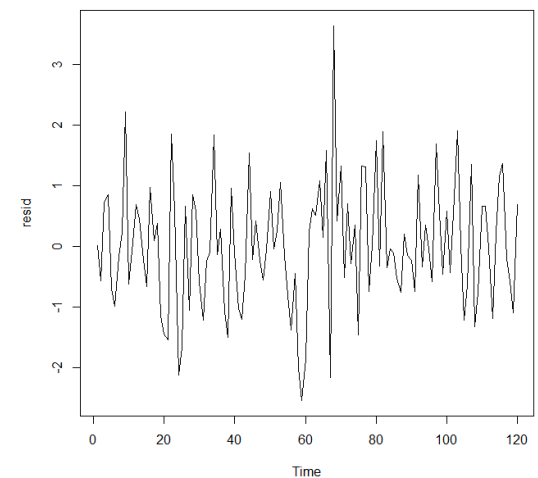
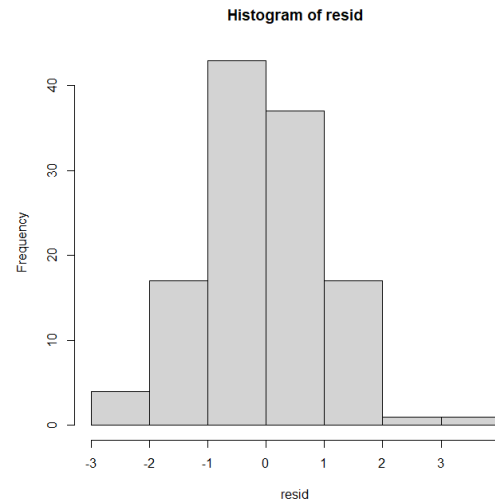
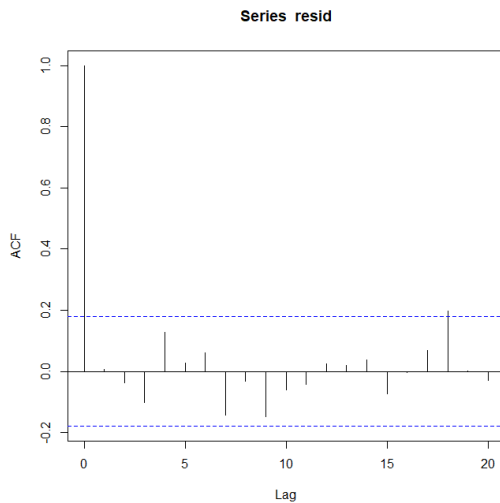


CONTOH: MODEL AUTOREGRESIF TERTIB-2, AR(2)



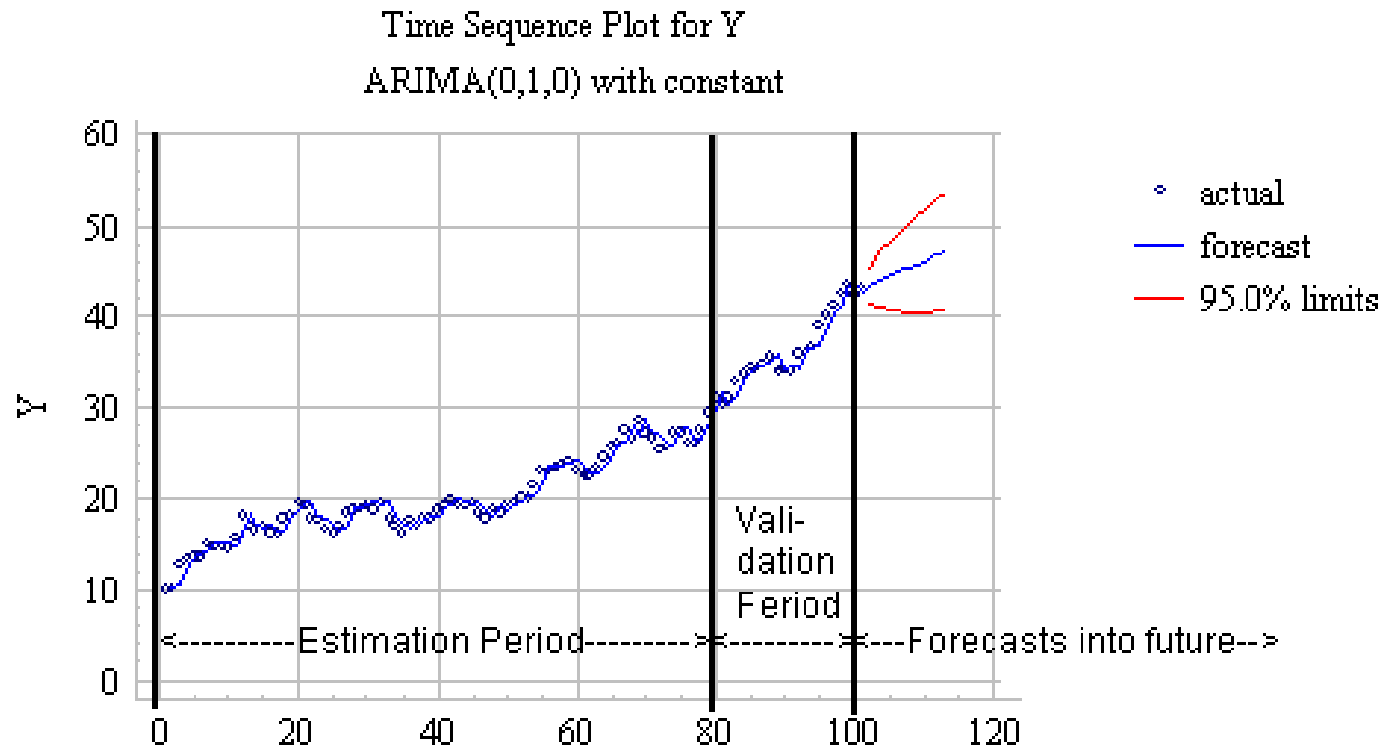
DIAGNOSTIK MODEL ARIMA:

- Model ARIMA yang telah disuaikan terhadap data perlu disemak andaian-andaian matematik yang perlu dipenuhi oleh model ARIMA.
- Ini dibuat menerusi analisis reja (*residual analysis*).
- Analisis Reja:
 - Reja adalah tak berkorelasi.
 - Reja tertabur secara normal.
 - Varians bagi reja adalah malar terhadap masa.



PERAMALAN SIRI MASA:

- Model ARIMA yang dikenal pasti sesuai untuk menerangkan data siri masa boleh digunakan untuk peramalan data masa hadapan.
- Namun, peramalan hanya sesuai untuk jangka pendek.
- Peramalan jangka panjang memberikan ketidakpastian yang sangat tinggi.



RINGKASAN PEMODELAN SIRI MASA MENERUSI MODEL ARIMA:

- 1) Plotkan Siri Masa dan lihat sama ada data pegun atau tidak.
- 2) Tentukan model ARIMA berdasarkan plot ACF dan PACF.
- 3) Suaikan model ARIMA terhadap data.
- 4) Jalankan analisis reja untuk pengesahan model.
- 5) Gunakan model ARIMA tersuai untuk mendapatkan nilai peramalan.
- 6) Dapatkan selang-keyakinan peramalan.

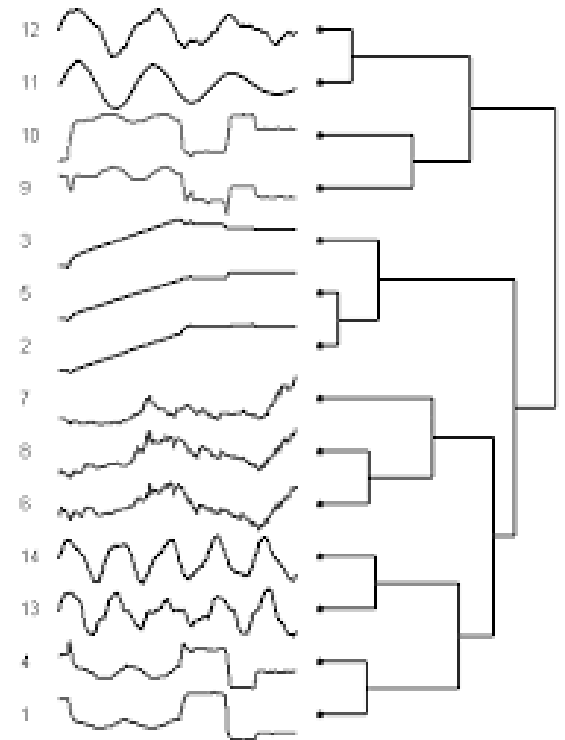


PENGKELOMPOKAN SIRI MASA:

- Pengkelompokan siri masa ialah pemetakan data siri masa berganda kepada kelompok-kelompok tertentu berdasarkan sifat kesamaan atau jarak.
- Siri masa dalam kelompok yang sama akan mempunyai ciri kesamaan yang tinggi.
- Siri masa dalam kelompok yang berbeza akan mempunyai ciri kesamaan yang rendah.

- Antara ciri kelompok bagi siri masa ialah:

- (i) Normal
- (ii) Bermusim
- (iii) Kitaran
- (iv) Trend meningkat
- (v) Trend menurun
- (vi) Anjakan ke atas (*Upward shift*)
- (vii) Anjakan ke bawah (*Downward shift*)

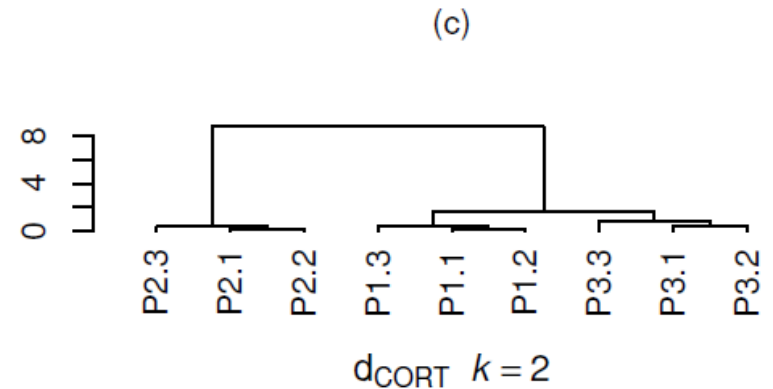
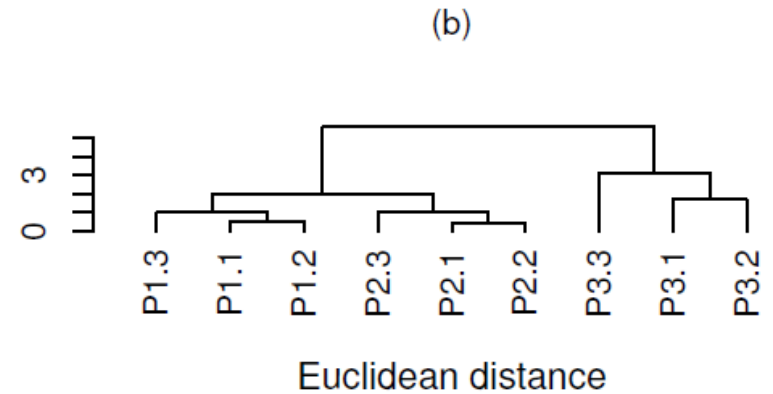
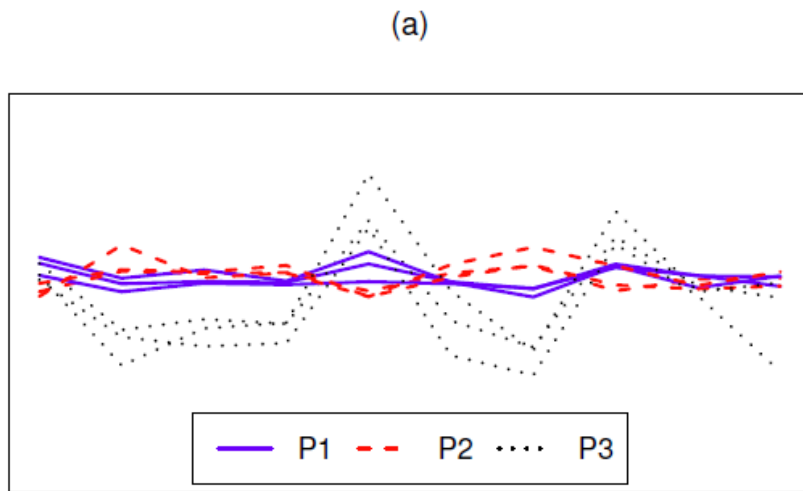


PENGUKURAN PERBEZAAN:

- Terdapat empat pendekatan utama untuk mengukur jarak dalam pengelompokan siri masa, iaitu:
 - i) Pendekatan tanpa-model.
 - ii) Pendekatan berasaskan model.
 - iii) Pendekatan berasaskan Kekompleksan.
 - iv) Pendekatan berasaskan Peramalan.



CONTOH PENGKELOMPOKAN SIRI MASA:

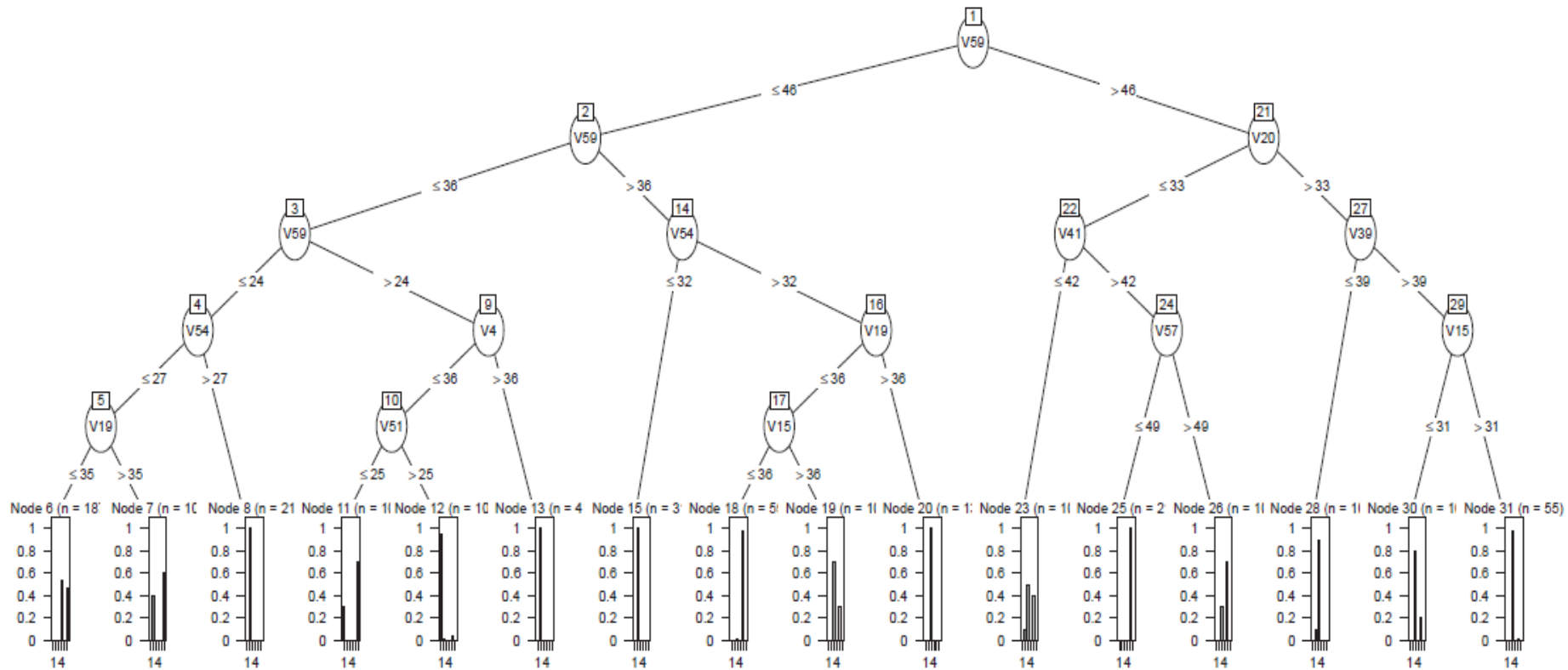


KLASIFIKASI SIRI MASA:

- Klasifikasi siri masa bertujuan untuk membina model klasifikasi berdasarkan siri masa berlabel.
- Seterusnya model tersebut akan digunakan untuk meramalkan label siri masa tanpa label.
- Ciri-ciri baru yang diekstrak dari siri masa boleh membantu meningkatkan prestasi model klasifikasi.
- Antara teknik yang digunakan untuk pengekstrakan ciri dalam Siri Masa ialah:
 - i) Penguraian Nilai Singular (SVD)
 - ii) Penjelmaan Diskret Fourier (DFT)
 - iii) Penjelmaan Diskret wavelet (DWT)



CONTOH KLASIFIKASI SIRI MASA:



RUJUKAN:

- Aggarwal, C.C. (2015). *Data Mining: The Textbook*. New York: Springer.
- Bowerman, B.L., O'Connel, R.T., Koehler, A.B. (2005). *Forecasting, time series, and regression: an applied approach*. 4th edition. Belmont: thompson Learning.
- Chatfield, C., Xing, H. (2019). *The Analysis Of Time Series: An Introduction with R*. Taylor and Francis.
- Maharaj, E.N., D'Urso, P., Caiado, J. (2019). *Time Series Clustering and Classification*. Chapman and Hall
- Montero, P., Vilar, J.A. (2014). TSclust: An R Package for Time Series Clustering. *Journal of Statistical Software* 62 (1): 1-43.
- Sardá-Espinosa, A. (2019). Time-Series Clustering in R Using the dtwclust Package. *The R Journal* (11/01): 1-22.
- Shumway, R., Stoffer, D. (2019). *Time Series: A Data Analysis Approach Using R*. CRC Press
- Woodward, W.A., Gray, H.L., Elliott, A.C. (2021). *Applied Time Series Analysis with R*. 2nd edition. CRC Press.



TOPIK SETERUSNYA:

Perlombongan Data Jujukan

