DATA ANALYTICS

Unit 3: Concept of stationarity, DF and ADF test and transformation

Jyothi R.

Department of Computer Science and Engineering

Concept of stationarity, DF and ADF test and transformation

- **Concept of stationarity, DF and ADF test and transformation of non stationary process to a stationary one.**

Contents:

1. **Dickey Fuller Test**
2. **Augmented Dickey−Fuller Test**
3. **Stationarity and differencing**
4. **Random Walk Model**
5. **Second-order differencing Seasonal differencing**

Dickey Fuller Test

- $\psi = 0$ is same as $\beta = 1$. So, the Dickey−Fuller test can be written in terms of ψ as

H_0: $\psi = 0$ (the time series is non-stationary)

H_A: $\psi < 0$ (the time series is stationary)

- The test statistic is given by as Equation-2
- DF Test Statistic = $\dfrac{\psi}{S_e(\psi)}$
- where S_e is the standard error. Note that DF test statistic is not *t*-statistic since the null hypothesis is on non-stationary process.
- Critical values are derived based on simulation

Augmented Dickey−Fuller Test

- Dickey−Fuller test is valid only when the residual ε_{t+1} follows a white noise.
- When ε_{t+1} is not white noise, the actual series may not be AR(1); it may have more significant lags.
- To address this issue, we augment *p*-lags of the dependent variable *Y*.
- The model in as Equation-2 can be written as in Equation-3

$$\Delta Y_t = \psi Y_t + \sum_{i=0}^{p} \alpha_i \Delta Y_{t-i} + \varepsilon_{t+1}$$

- The above equation can be now tested for non-stationarity.
- Again the null and alternative hypotheses are
- H_0: $\psi = 0$ (the time series is non-stationary)
- H_0: $\psi < 0$ (the time series is stationary)

Transforming Non-Stationary Process to Stationary Process Using Differencing

- The first step in ARIMA is to identify the order of differencing (*d*) required to convert a non-stationary process into a stationary process.
- Many time-series data will be non-stationary due to factors such as trend and seasonality.
- If the non-stationary behaviour is due to trend, then it can be converted into a stationary process by de-trending the data.
- De-trending is usually achieved by fitting a trend line and subtracting it from the time series; this is known as **trend stationarity**.
- When the reason is not due to trend stationarity, then differencing the original time series may be useful for converting the non-stationary process into a stationary process (called **difference stationarity**).

Transforming Non-Stationary Process to Stationary Process Using Differencing

- The first difference (d = 1) is the difference between consecutive values of the time series (Y_t and Y_{t-1}).
- That is, the first difference ΔY_t is given by

$$\Delta y_t = Y_t - Y_{t-1} \qquad \text{---------------Equation-4}$$

- The second difference (d = 2) is the difference of the first differences and is given by

$$\nabla^2 Y_t = \nabla(\nabla Y_t) = Y_t - 2Y_{t-1} + Y_{t-2} \quad \text{-------- } \textit{Equation -5}$$

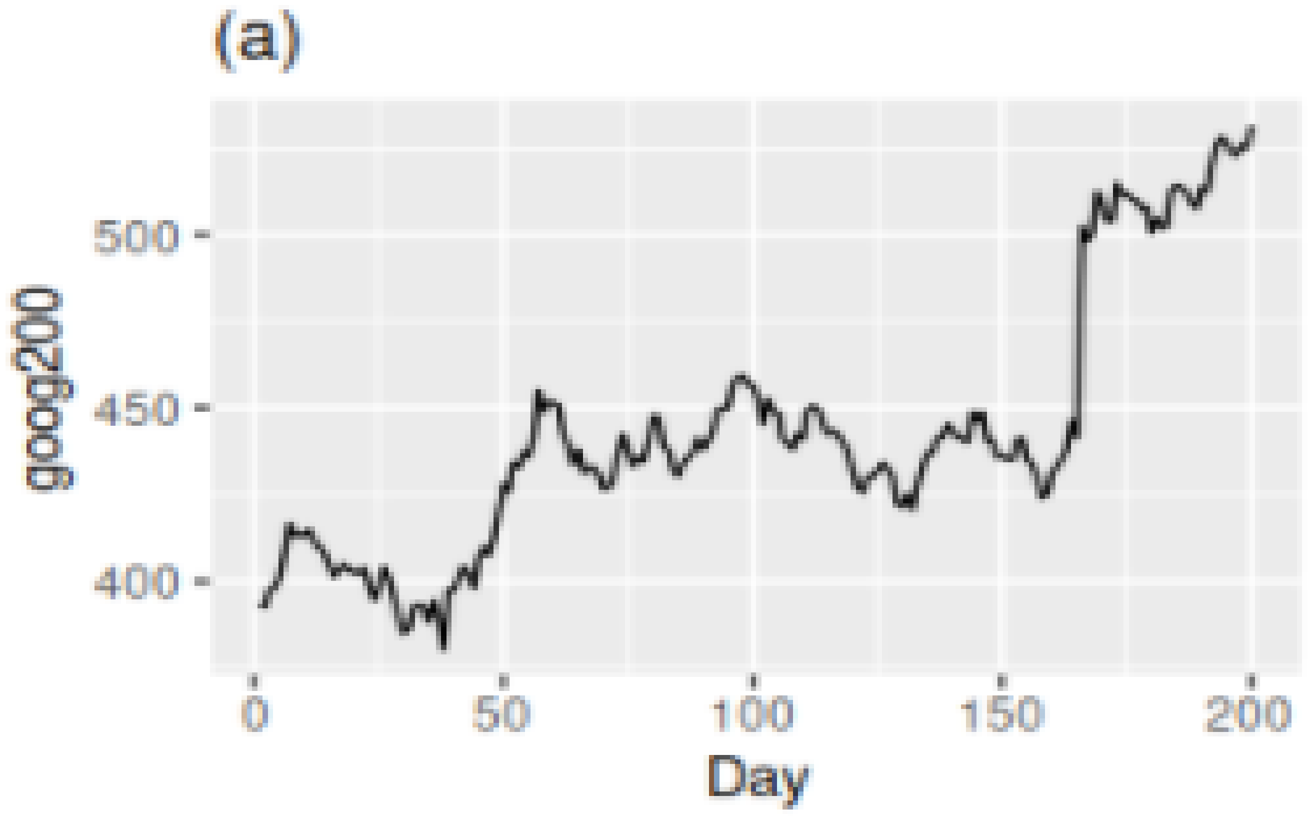
- In most cases, $d \leq 2$ will be sufficient to convert a non-stationary process to a stationary process.

Stationarity and differencing

- A stationary time series is one whose properties do not depend on the time at which the series is observed.
- Time series with trends, or with seasonality, are not stationary — the trend and seasonality will affect the value of the time series at different times.
- On the other hand, a white noise series is stationary — it does not matter when you observe it, it should look much the same at any point in time.
- Some cases can be confusing — a time series with cyclic behaviour but with no trend or seasonality is stationary.
- This is because the cycles are not of a fixed length, so before we observe the series we cannot be sure where the peaks and troughs of the cycles will be.
- In general, a stationary time series will have no predictable patterns in the long-term.

Which of these series are stationary?

- Time plots will show the series to be roughly horizontal although some cyclic behaviour is possible, with constant variance.
- Figure1. (a) Google stock price for 200 consecutive days;



Which of these series are stationary?

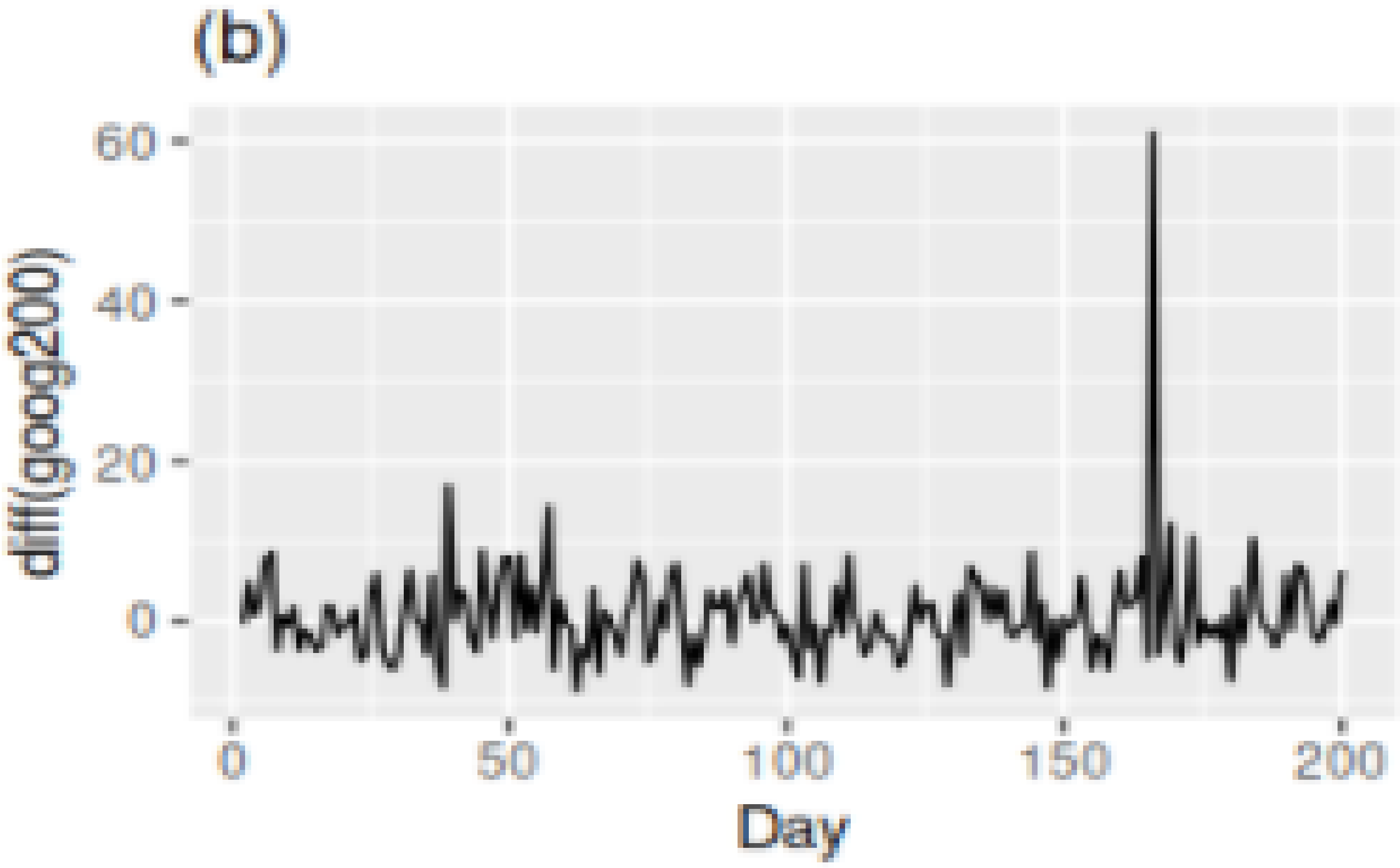


Figure 2: (b) Daily change in the Google stock price for 200 consecutive days;

Which of these series are stationary?

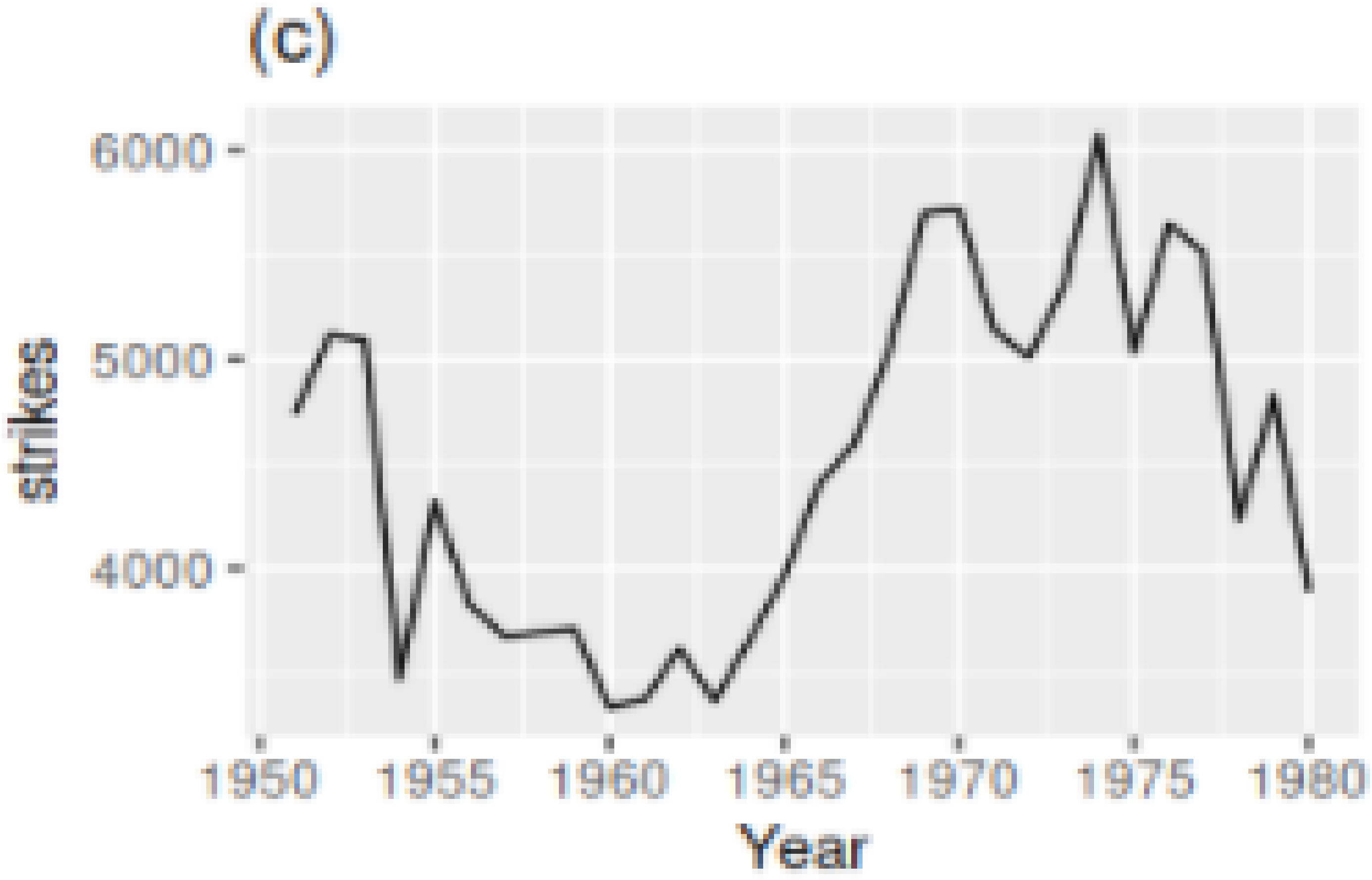


Figure 3 :(c) Annual number of strikes in the US;

Which of these series are stationary?

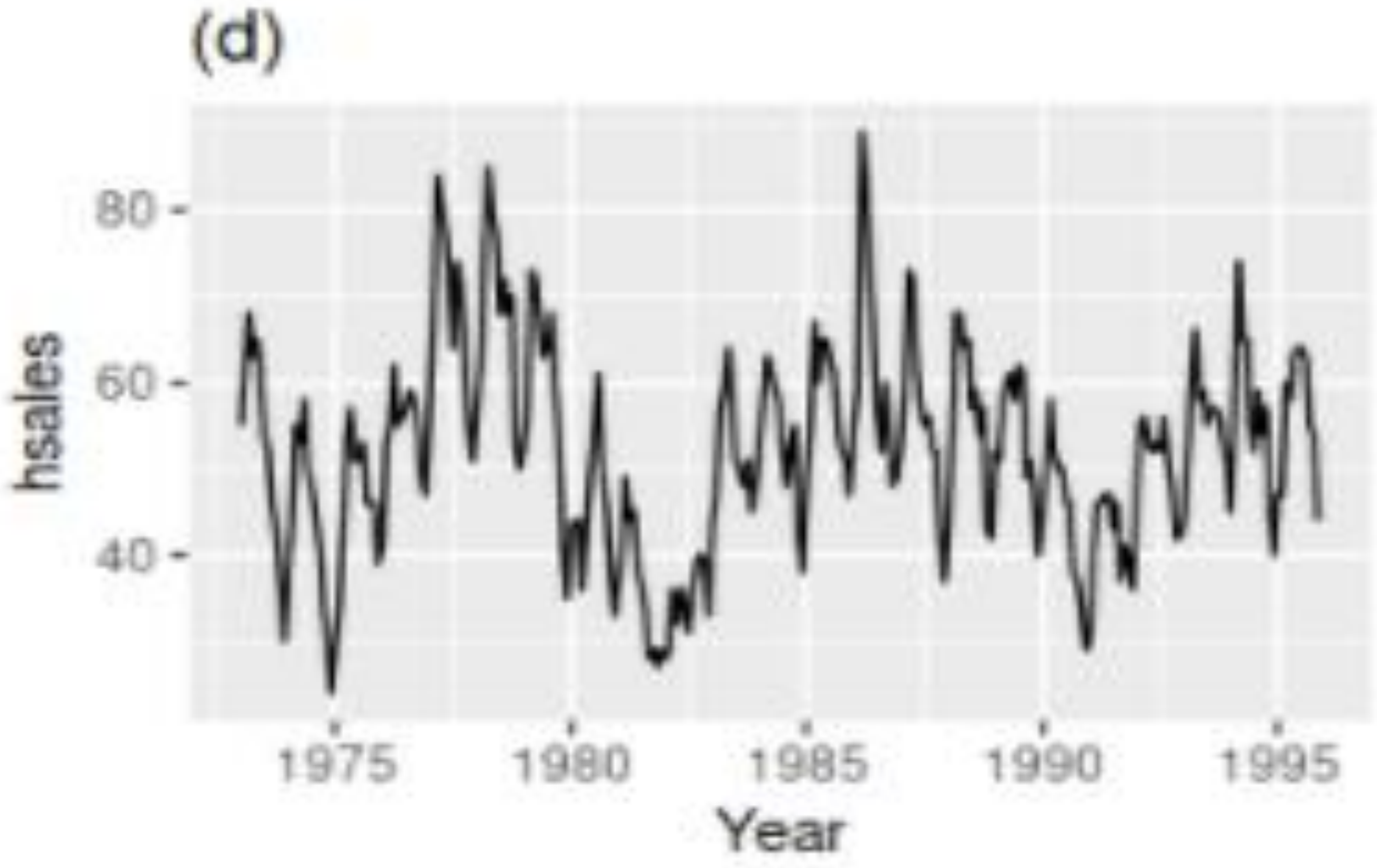


Figure 4: (d) Monthly sales of new one-family houses sold in the US;

Which of these series are stationary?

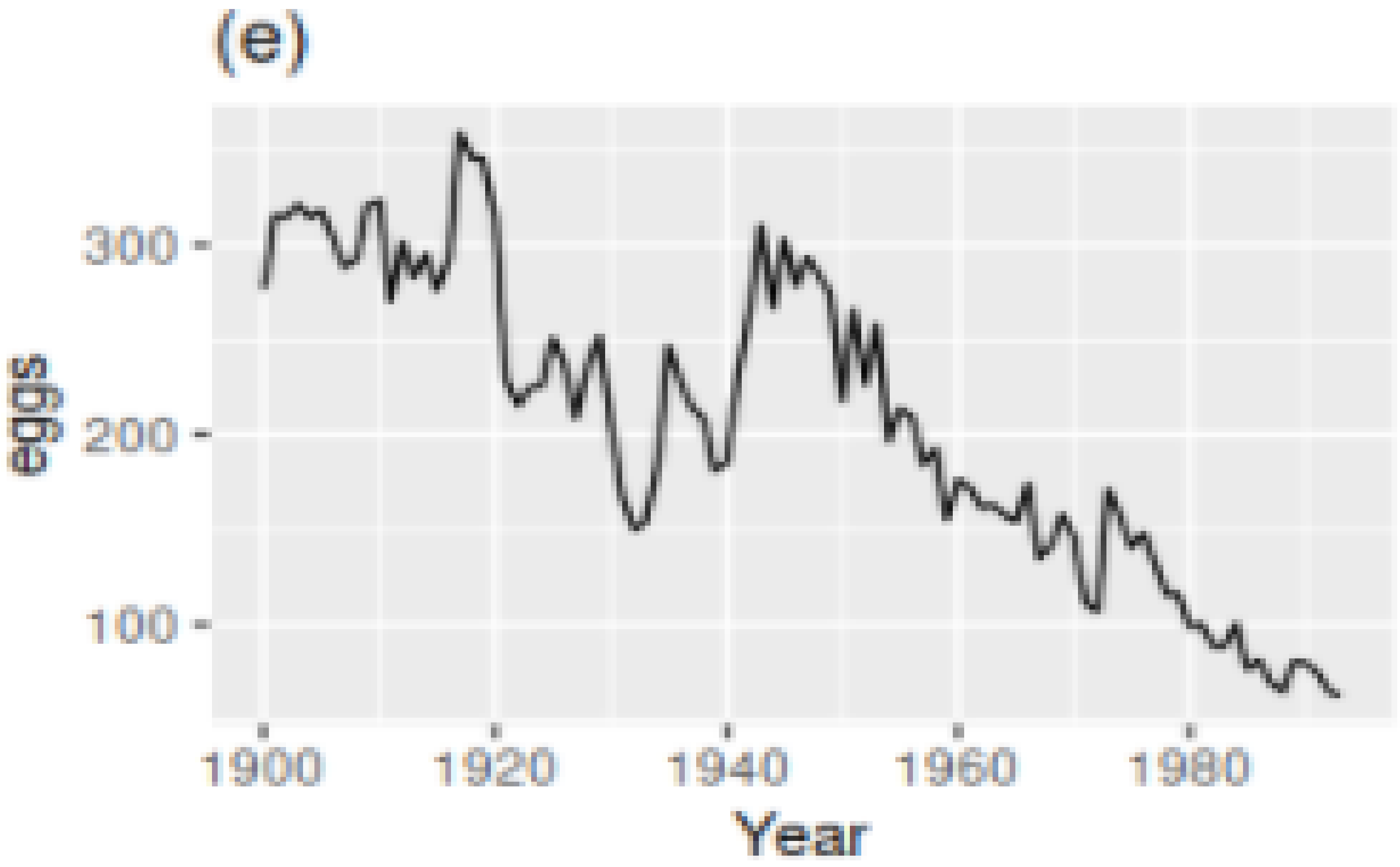


Figure 5: (e) Annual price of a dozen eggs in the US (constant dollars);

DATA ANALYTICS

Which of these series are stationary?

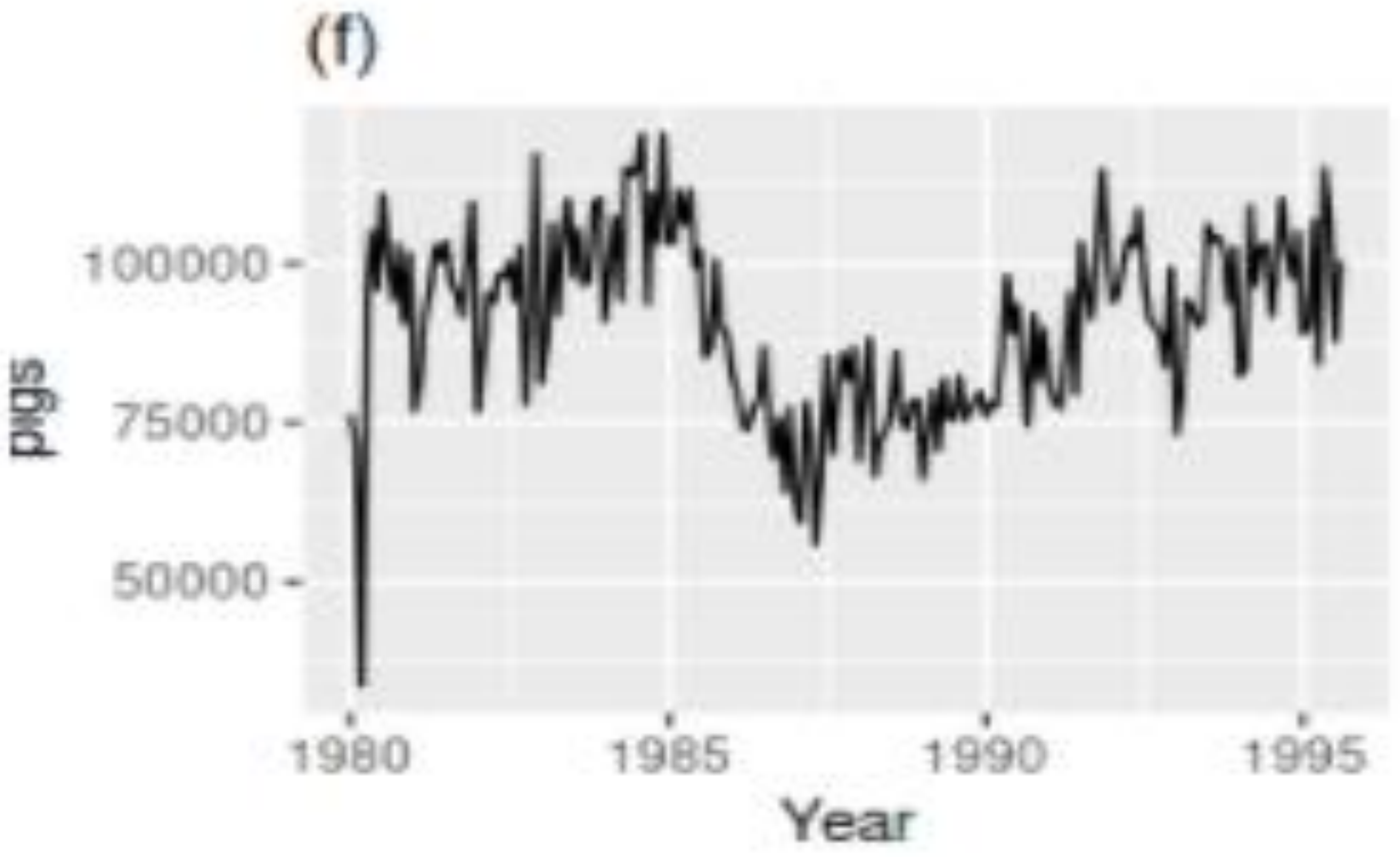


Figure 6: (f) Monthly total of pigs slaughtered in Victoria, Australia;

Which of these series are stationary?

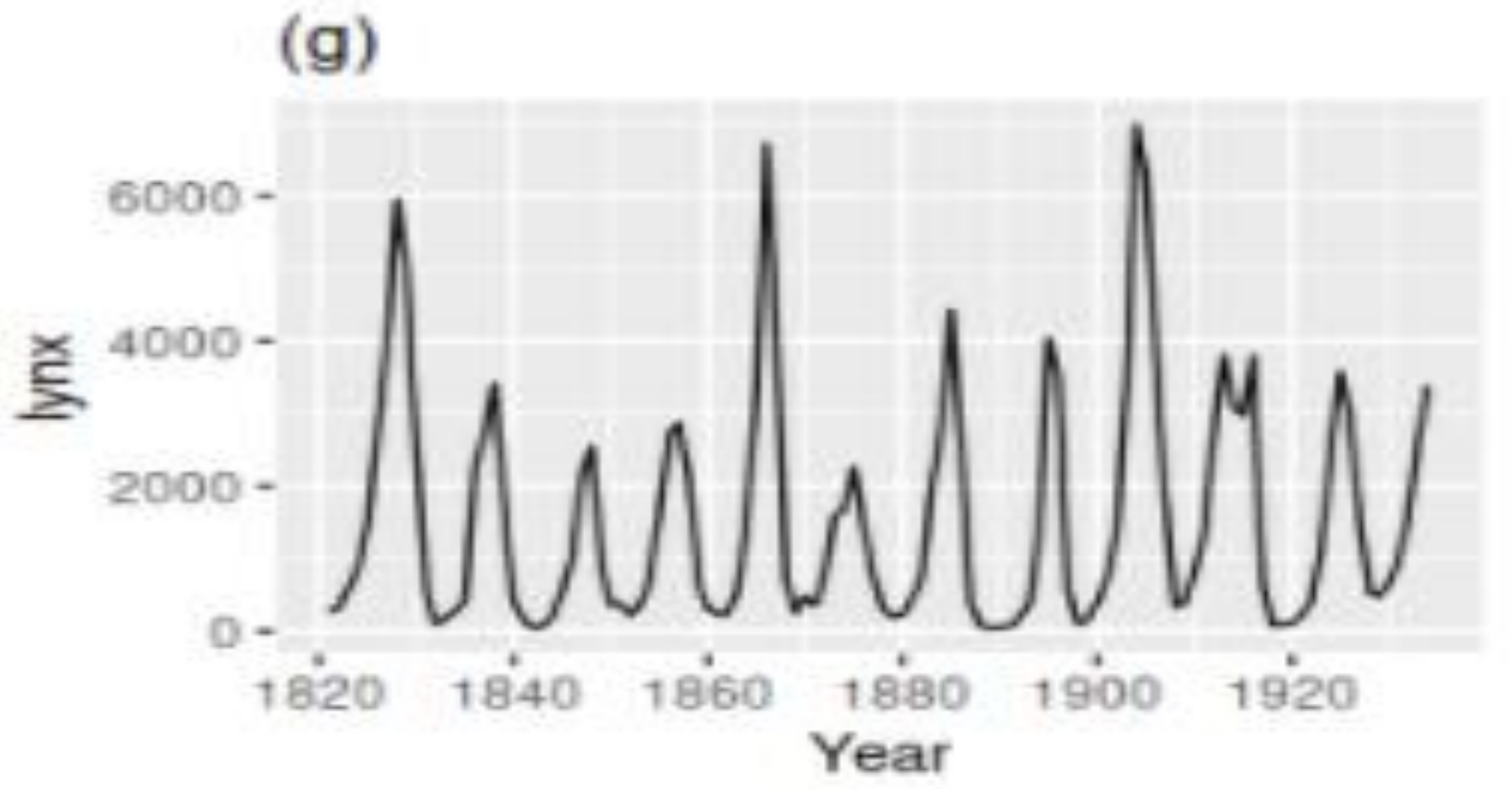


Figure 7: (g) Annual total of lynx trapped in the McKenzie River district of north-west Canada;

Which of these series are stationary?

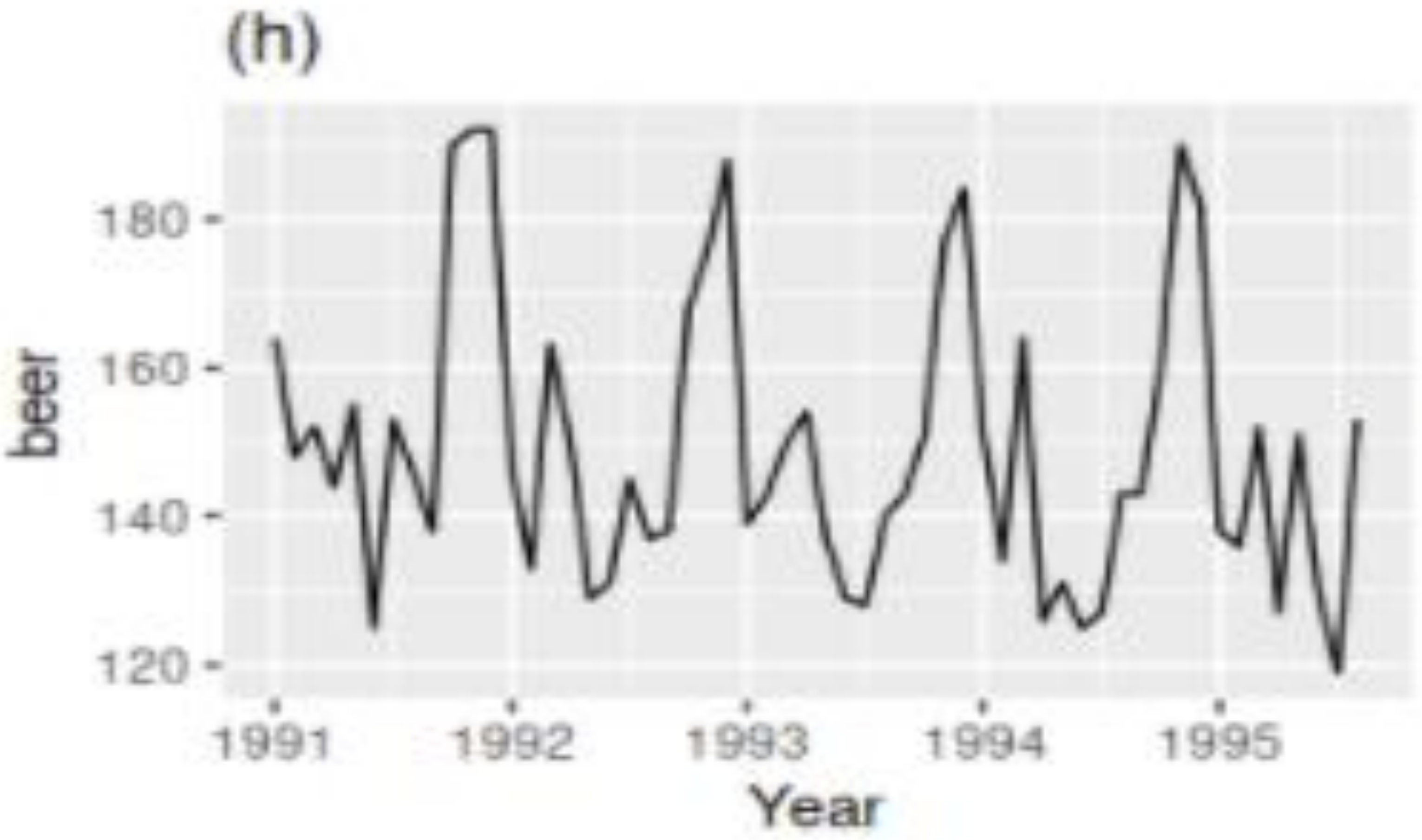


Figure 8: (h) Monthly Australian beer production;

Which of these series are stationary?

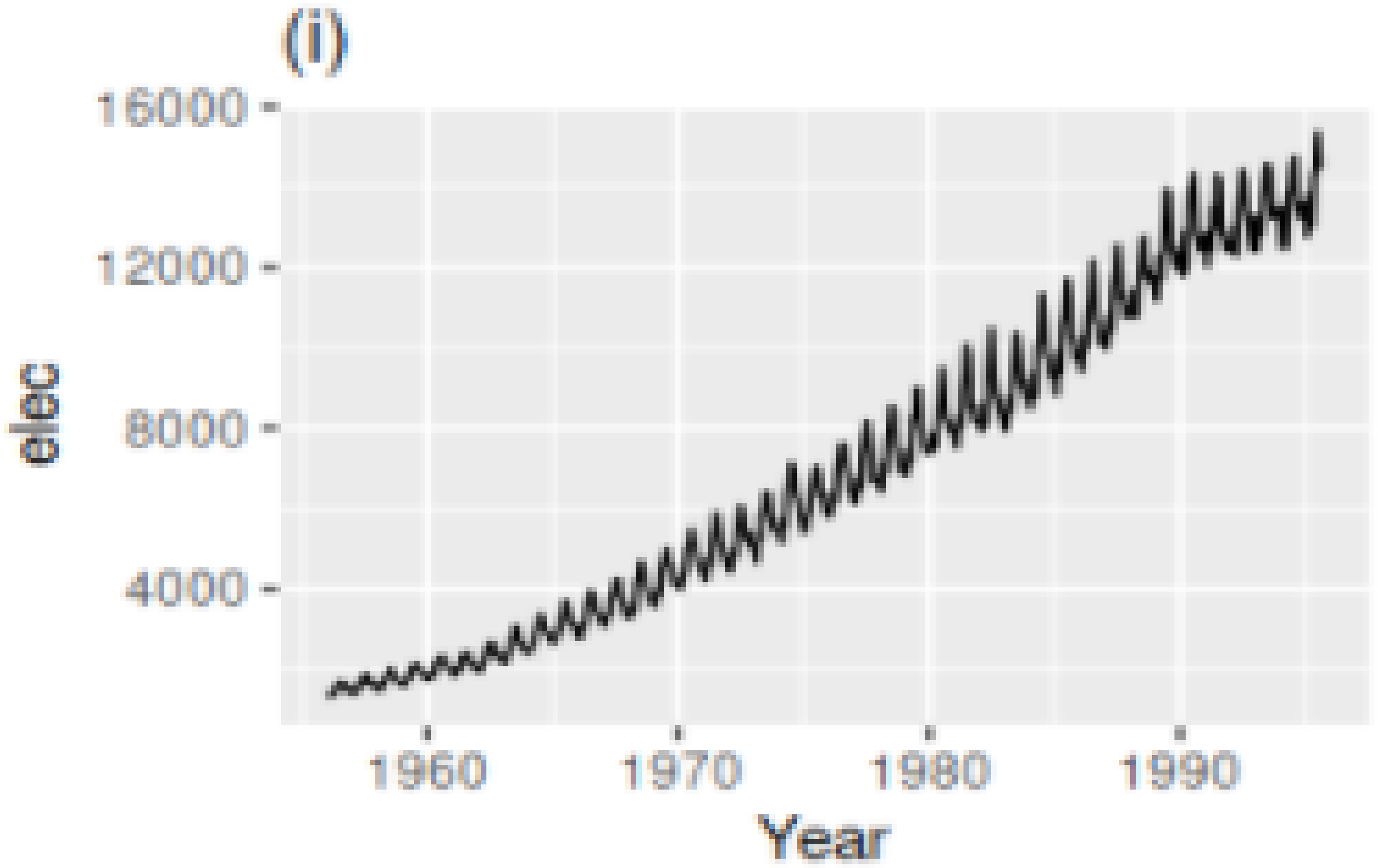


Figure 9:(i) Monthly Australian electricity production.

Which of these series are stationary?

- Consider the nine series plotted in Figure 1 to 9:
- Which of these do you think are stationary?
- Obvious seasonality rules out series (d), (h) and (i).
- Trends and changing levels rules out series (a), (c), (e), (f) and (i).
- Increasing variance also rules out (i).
- That leaves only (b) and (g) as stationary series.
- At first glance, the strong cycles in series (g) might appear to make it non-stationary.
- But these cycles are aperiodic — they are caused when the lynx population becomes too large for the available feed, so that they stop breeding and the population falls to low numbers, then the regeneration of their food sources allows the population to grow again, and so on.
- In the long-term, the timing of these cycles is not predictable.
- Hence the series is stationary.

Differencing

- In Figure 1 to 9: The Google stock price was non-stationary in panel (a),
- But the daily changes were stationary in panel (b). This shows one way to make a non-stationary time series stationary — compute the differences between consecutive observations. This is known as **differencing**.
- Transformations such as logarithms can help to stabilise the variance of a time series.
- Differencing can help stabilise the mean of a time series by removing changes in the level of a time series, and therefore eliminating or reducing trend and seasonality.
- By looking at the time plot of the data, the ACF plot is also useful for identifying non-stationary time series.
- For a stationary time series, the ACF will drop to zero relatively quickly, while the ACF of non-stationary data decreases slowly.
- Also, for non-stationary data, the value of r_1 is often large and positive.

Differencing

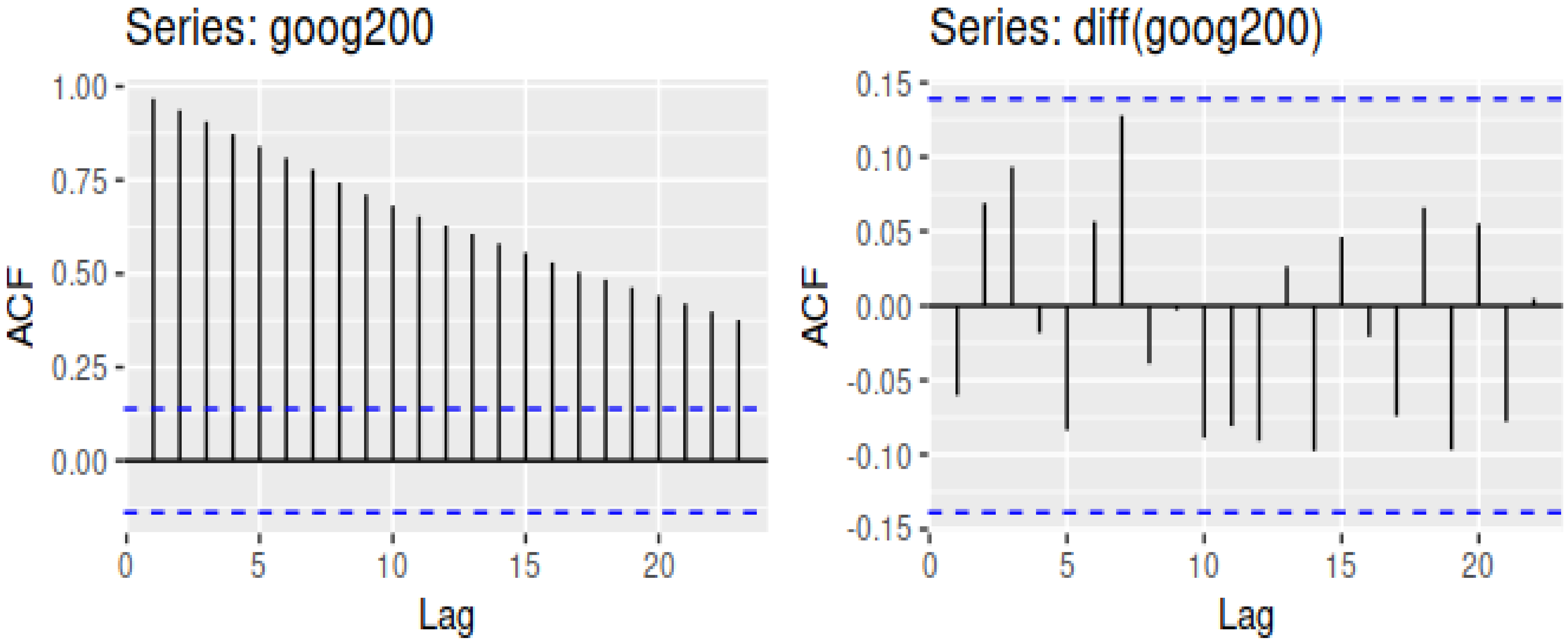


Figure 10:The ACF of the Google stock price (left) and of the daily changes in Google stock price (right).

Differencing

Figure 8.2:The ACF of the Google stock price (left) and of the daily changes in Google stock price (right).

- The ACF of the differenced Google stock price looks just like that of a white noise series.
- There are no autocorrelations lying outside the 95% limits, and
- The Ljung -Box $Q*Q*$ statistic has a *p*-value of 0.355 (for $h=10h=10$).
- This suggests that the *daily change* in the Google stock price is essentially a random amount which is uncorrelated with that of previous days.

Random Walk Model

- The differenced series is the *change* between consecutive observations in the original series, and can be written as
- $y'_t = y_t - y_{t-1}$.
- $yt' = y_t - y_{t-1}$.
- The differenced series will have only T−1T−1 values, since it is not possible to calculate a difference y'_1 for the first observation.
- When the differenced series is white noise, the model for the original series can be written as $y_t - y_{t-1} = \varepsilon_t$, $y_t - y_{t-1} = \varepsilon_t$, where ε_t denotes white noise.
- Rearranging this leads to the "random walk" model $y_t = y_{t-1} + \varepsilon_t$. $y_t = y_{t-1} + \varepsilon_t$.

Random Walk Model

- Random walk models are widely used for non-stationary data, particularly financial and economic data.

Random walks typically have:

- long periods of apparent trends up or down
- sudden and unpredictable changes in direction.

- The forecasts from a random walk model are equal to the last observation, as future movements are unpredictable, and are equally likely to be up or down.
- Thus, the random walk model underpins naïve forecasts,

Random Walk Model

- A closely related model allows the differences to have a non-zero mean.

Then

- $y_t - y_{t-1} = c + \varepsilon_t$ or $y_t = c + y_{t-1} + \varepsilon_t$.
- $y_t - y_{t-1} = c + \varepsilon_t$ or $y_t = c + y_{t-1} + \varepsilon_t$.
- The value of cc is the average of the changes between consecutive observations.
- If cc is positive, then the average change is an increase in the value of y_t.
- Thus, y_t will tend to drift upwards.
- However, if c is negative, y_t will tend to drift downwards.
- This is the model behind the drift method

Second-order differencing

- Occasionally the differenced data will not appear to be stationary and it may be necessary to difference the data a second time to obtain a stationary series:
- \[\begin
- {align*}
- y''_{t}
- &= y'_{t} - y'_{t - 1}
- \\ &= (y_t - y_{t-1}) - (y_{t-1}-y_{t-2})
- \\ &= y_t - 2y_{t-1} +y_{t-2}.
- \end{align*}\]
-
- In this case, \(y_t''\) will have \(T-2\) values.
- Then, we would model the “change in the changes” of the original data.
- In practice, it is almost never necessary to go beyond second-order differences.
- .

Seasonal differencing

- A seasonal difference is the difference between an observation and the previous observation from the same season.
- So $$y'_t = y_t - y_{t-m},$$ where $m=$ the number of seasons.
- These are also called “lag-m differences”, as we subtract the observation after a lag of m periods.
- If seasonally differenced data appear to be white noise, then an appropriate model for the original data is $$y_t = y_{t-m}+\varepsilon_t.$$ Forecasts from this model are equal to the last observation from the relevant season.
- That is, this model gives seasonal naïve forecasts.

Seasonal differencing

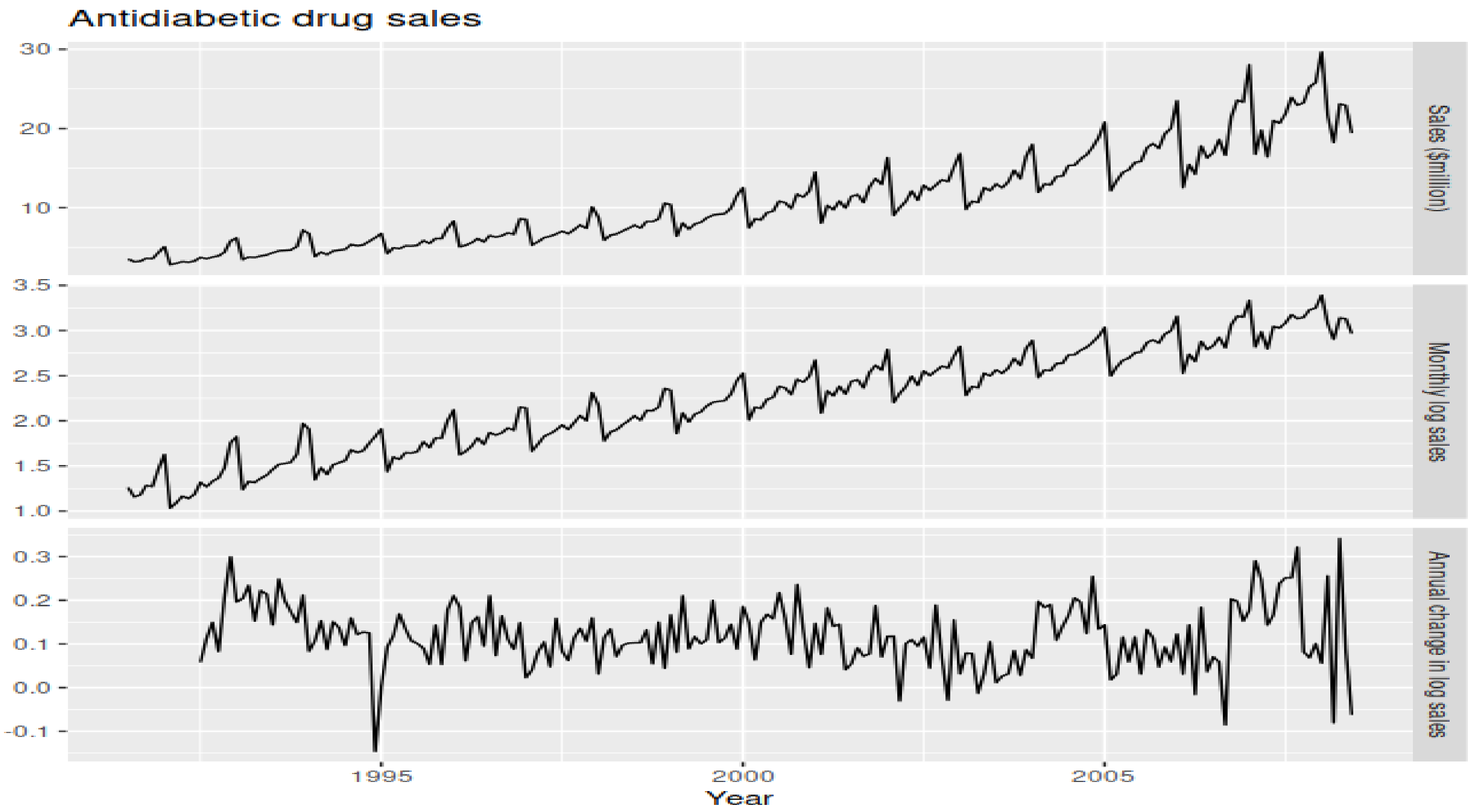


Figure 11: Logs and seasonal differences of the A10 (antidiabetic) sales data.

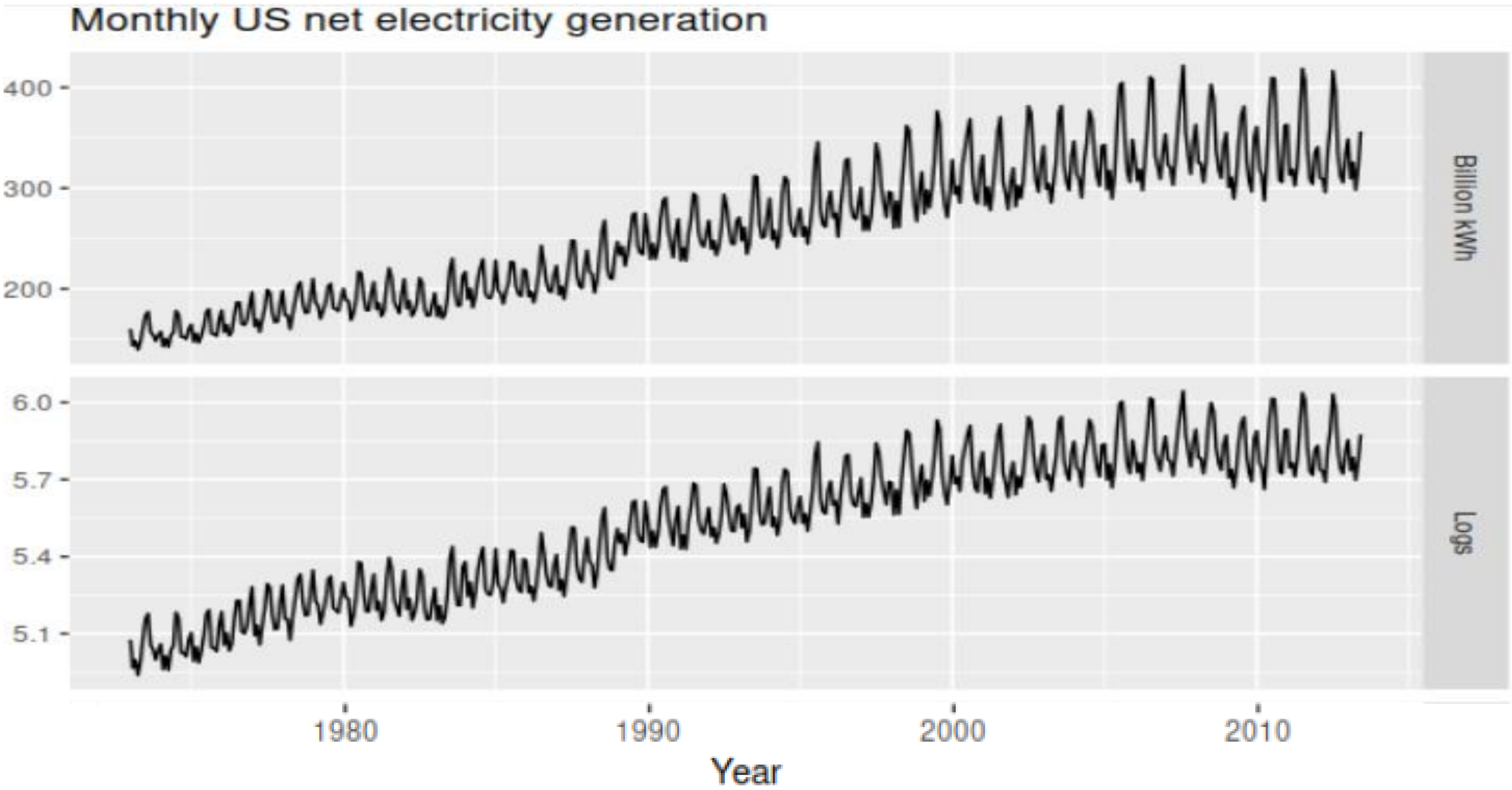
Seasonal differencing

- The bottom panel in Figure11, shows the seasonal differences of the logarithm of the monthly scripts for A10 (antidiabetic) drugs sold in Australia.
- The transformation and differencing have made the series look relatively stationary.
- The logarithms stabilise the variance, while the seasonal differences remove the seasonality and trend.
- To distinguish seasonal differences from ordinary differences, we refer to ordinary differences as “first differences”, meaning differences at lag 1.

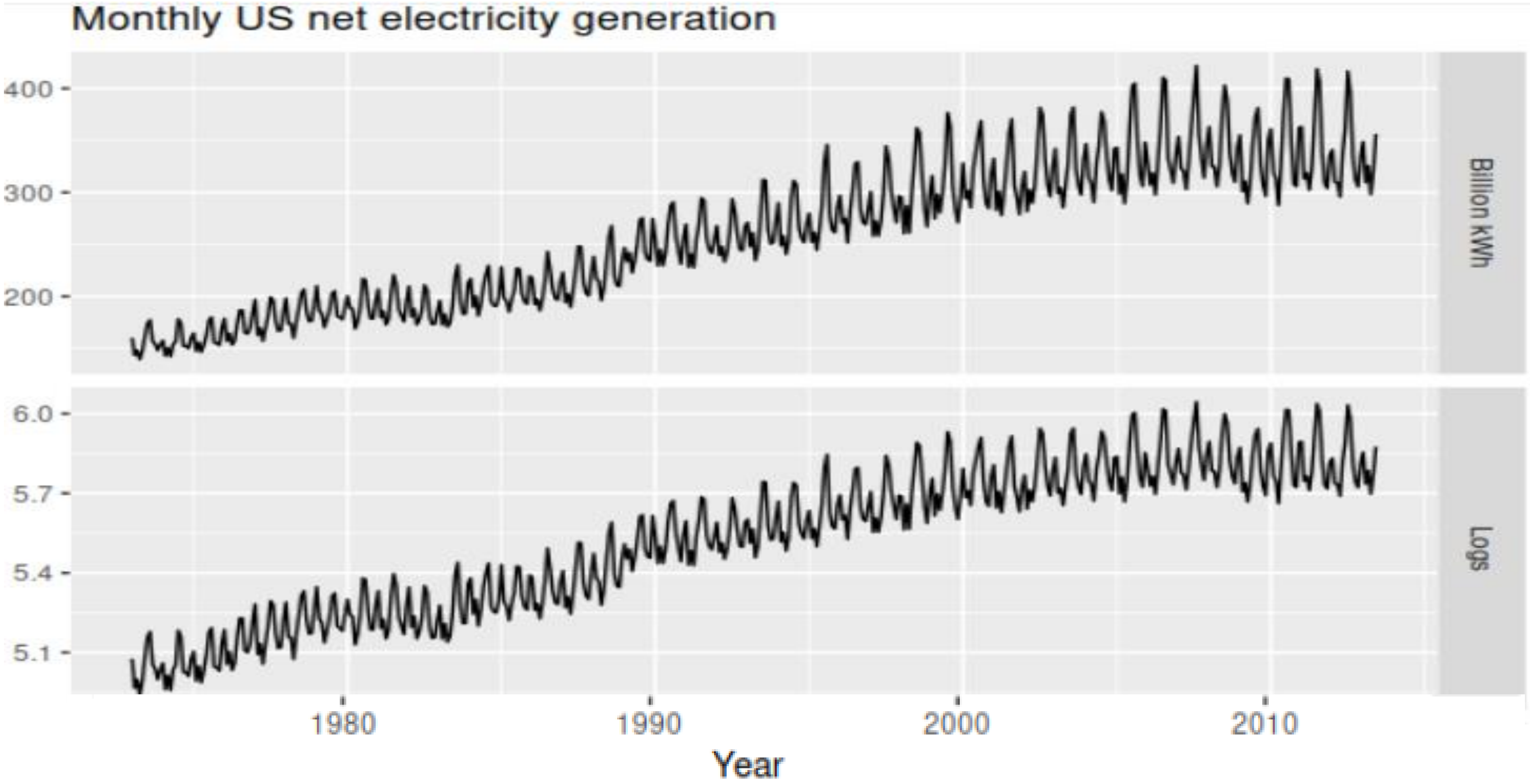
Seasonal differencing

- Sometimes it is necessary to take both a seasonal difference and a first difference to obtain stationary data, as is shown in Figure 12.
- Here, the data are first transformed using logarithms (second panel),
- then seasonal differences are calculated (third panel).
- The data still seem somewhat non-stationary, and so a further lot of first differences are computed (bottom panel).
- Figure 12: Top panel: US net electricity generation (billion kWh).
- Other panels show the same data after transforming and differencing.

Seasonal differencing



Seasonal differencing



References

Text Book:

“Business Analytics, The Science of Data-Driven Making”, U. Dinesh Kumar, Wiley 2017

Chapter-13 Concept of stationarity, DF and ADF test and transforming non stationary process to a stationary one 13.14.1-13.14.3 in text

Image Courtesy

https://www.abs.gov.au/websitedbs/D3310114.nsf/home/Time+Series+Analysis:+The+Basics

https://otexts.com/fpp2/stationarity.html

THANK YOU

Jyothi R
Assistant Professor, Department of Computer Science
jyothir@pes.edu