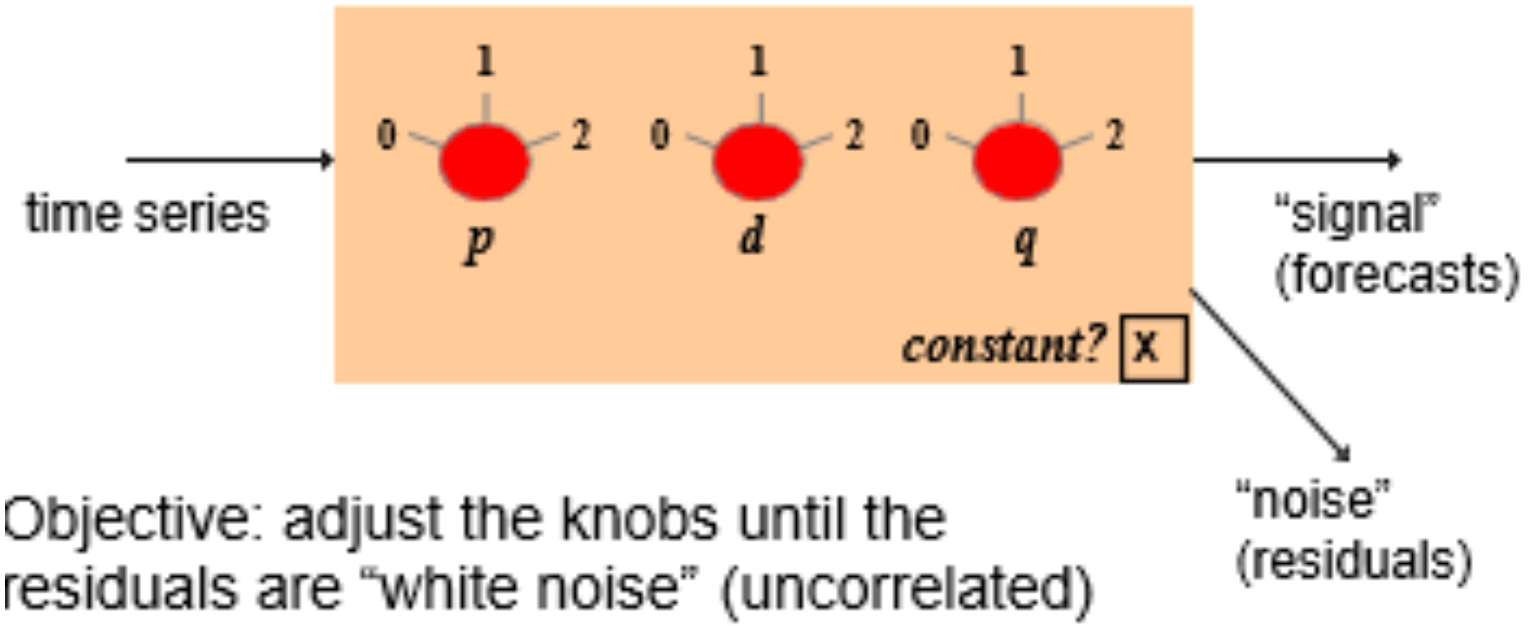
DATA ANALYTICS

Unit 3: SARIMA and ARIMAX

Jyothi R. Bharathi R

Department of Computer Science and Engineering

The ARIMA "filtering box" – Extensions(to build projects in Time Series)



Objective: adjust the knobs until the residuals are "white noise" (uncorrelated)

Seasonal ARIMA terminology

1. Autoregressive Integrated Moving Average, or ARIMA, is one of the most widely used forecasting methods for univariate time series data forecasting.

2. Although the method can handle data with a trend, it does not support time series with a seasonal component.

3. An extension to ARIMA that supports the direct modeling of the seasonal component of the series is called SARIMA.

4. Seasonal Autoregressive Integrated Moving Average, or SARIMA, method for time series forecasting with univariate data containing trends and seasonality.

Seasonal ARIMA terminology

Why SARIMA?

Autoregressive Integrated Moving Average, or ARIMA, is a forecasting method for univariate time series data.

As its name suggests, it supports both an autoregressive and moving average elements. The integrated element refers to differencing allowing the method to support time series data with a trend.

A problem with ARIMA is that it does not support seasonal data. That is a time series with a repeating cycle.

ARIMA expects data that is either not seasonal or has the seasonal component removed, e.g. seasonally adjusted via methods such as seasonal differencing.

What is Seasonal ARIMA ?

Seasonal Autoregressive Integrated Moving Average, SARIMA or Seasonal ARIMA, is an extension of ARIMA that explicitly supports univariate time series data with a seasonal component.

It adds three new hyperparameters to specify the autoregression (AR), differencing (I) and moving average (MA) for the seasonal component of the series, as well as an additional parameter for the period of the seasonality.

Seasonal ARIMA terminology

- The seasonal part of an ARIMA model is summarized by three *additional* numbers:

  ***P* = # of *seasonal autoregressive* terms**

  D* = # of *seasonal differences

  ***Q* = # of *seasonal moving-average* terms**

- The complete model is called an "SARIMA(*p,d,q*)(*P,D,Q*)" model

How to Configure SARIMA ?

1. Configuring a SARIMA requires selecting hyperparameters for both the trend and seasonal elements of the series.
2. **Trend Elements**

   There are three trend elements that require configuration.
   They are the same as the ARIMA model; specifically:

   1. **p**: Trend autoregression order.
   2. **d**: Trend difference order.
   3. **q**: Trend moving average order.
3. **Seasonal Elements**

   There are four seasonal elements that are not part of ARIMA that must be configured; they are:

   1. **P**: Seasonal autoregressive order.
   2. **D**: Seasonal difference order.
   3. **Q**: Seasonal moving average order.
   4. **m**: The number of time steps for a single seasonal period.

SARIMA(p,d,q)(P,D,Q)

Lag-1 differencing
$y_t - y_{t-1}$

Used for removing trend
d= 0,1,2

Seasonal(lag-M differencing)
$y_t - y_{t-M}$

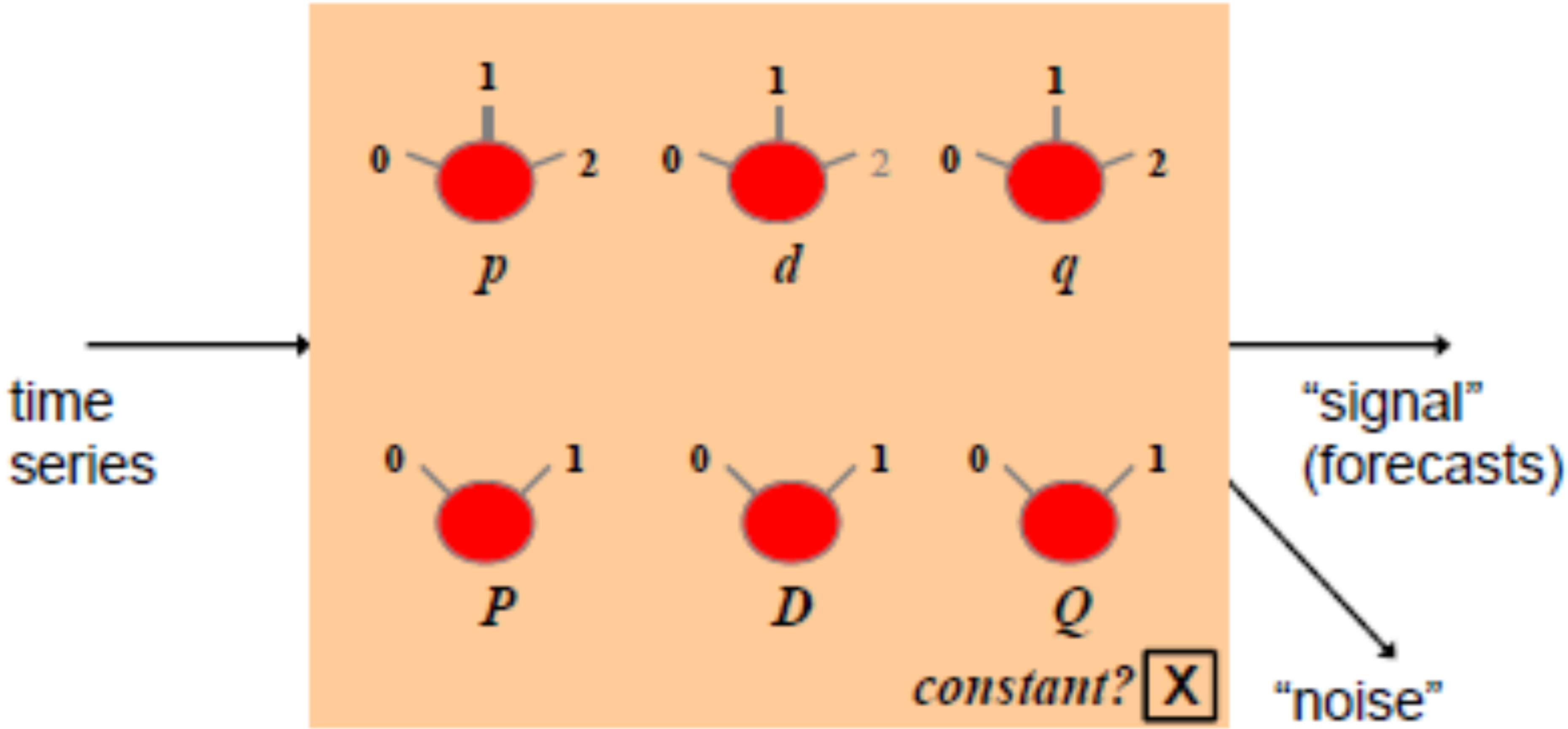
Used for removing seasonality
D= 0,1

Choosing parameters for SARIMA(p,d,q)(P,D,Q)

Visual Inspection

- seasonality and trend
- auto correlation chart
- partial auto correlation chart

The "filtering box" now has 6 knobs:



Note that *P*, *D*, and *Q* should never be larger than 1 !!

Seasonal differences

- How non-seasonal & seasonal differences are combined to stationarize the series:

If d=0, D=1: $y_t = Y_t - Y_{t\text{-}s}$

s is the seasonal period, e.g., s=12 for monthly data

If d=1, D=1: $y_t = (Y_t - Y_{t\text{-}1}) - (Y_{t\text{-}s} - Y_{t\text{-}s\text{-}1})$

$$= Y_t - Y_{t\text{-}1} - Y_{t\text{-}s} + Y_{t\text{-}s\text{-}1}$$

D should never be more than 1, and *d+D* should never be more than 2. Also, if *d+D* =2, the constant term should be suppressed.

SAR and SMA terms

- How SAR and SMA terms add coefficients to the model:

- Setting P =1 (i.e., SAR=1) adds a multiple of
  - y_{t-s} to the forecast for y_t

- Setting Q =1 (i.e., SMA=1) adds a multiple of
  - e_{t-s} to the forecast for y_t

  - Total number of SAR and SMA factors usually should not be more than 1 (i.e., either SAR=1 or SMA=1, not both)

Model-fitting steps

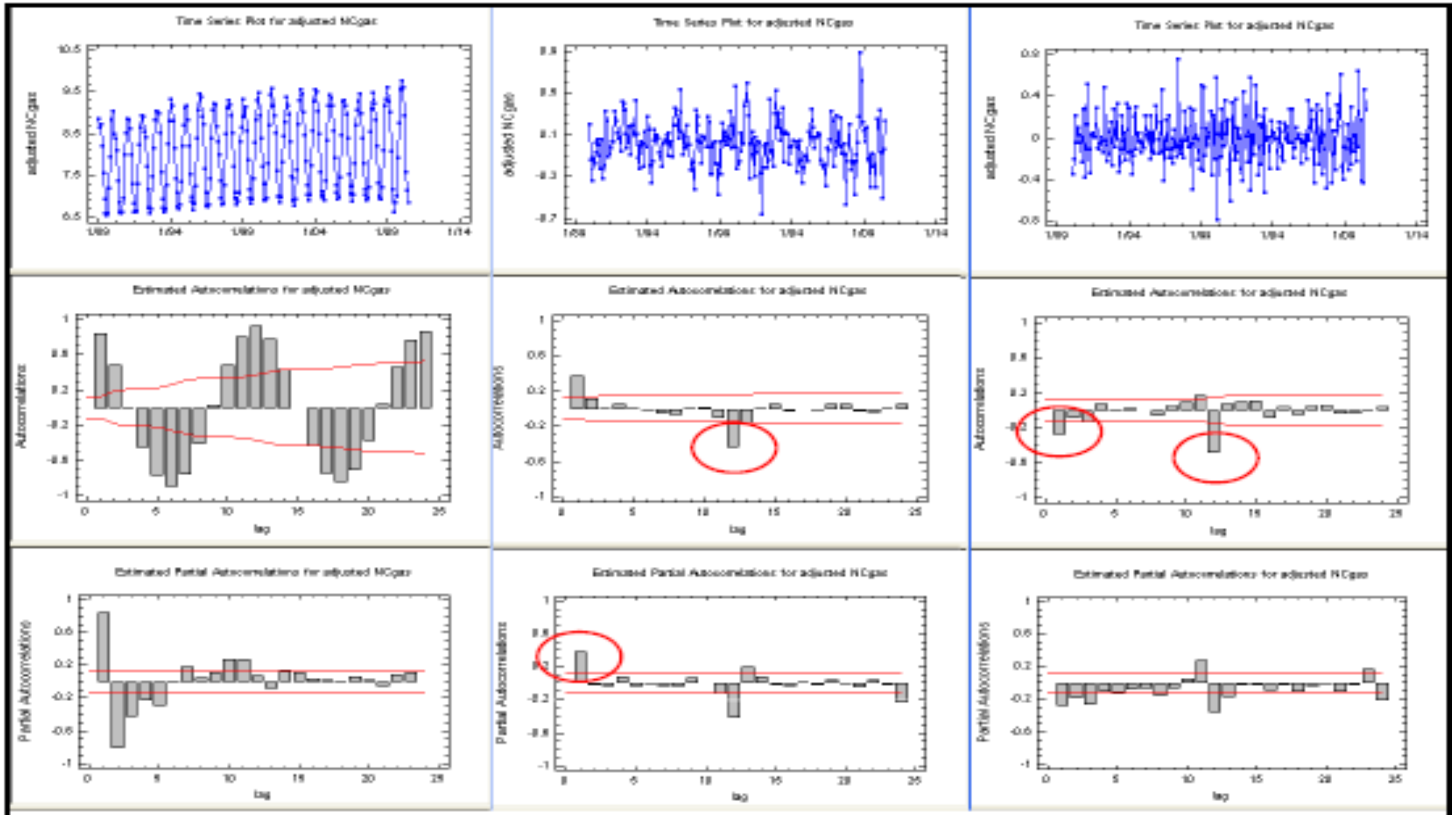
- Start by trying various combinations of one seasonal difference and/or one non-seasonal difference to stationarize the series and remove gross features of seasonal pattern.

- If the seasonal pattern is strong and stable, you MUST use a seasonal difference (otherwise it will "die out" in long-term forecasts)

Model-fitting steps, continued

- After differencing, inspect the ACF and PACF at multiples of the seasonal period (s):
  - Positive spikes in ACF at lag s, 2s, 3s…, single positive spike in PACF at lag
  - s =>SAR=1
  - Negative spike in ACF at lag s, negative spikes in PACF at lags
  - s, 2s, 3s,… => SMA=1
  - SMA=1 often works well in conjunction with a seasonal difference.
  - Same principles as for non-seasonal models, except focused on what happens at multiples of lag s in ACF and PACF.

Model-fitting steps, continued



Original series: nonstationary   Seas. diff: need AR(1) & SMA(1)   Both diff: need MA(1) & SMA(1)

A common seasonal ARIMA model

- Often you find that the "correct" order of differencing is d=1 and D=1.
- With one difference of each type, the autocorr. often negative at both lag 1 and lag s.
- This suggests an SARIMA(0,1,1)(0,1,1) model, a common seasonal ARIMA model.
- Similar to Winters' model in estimating time-varying trend and time-varying seasonal pattern

Bottom-line suggestion

- When fitting a time series with a strong seasonal pattern, you generally should try

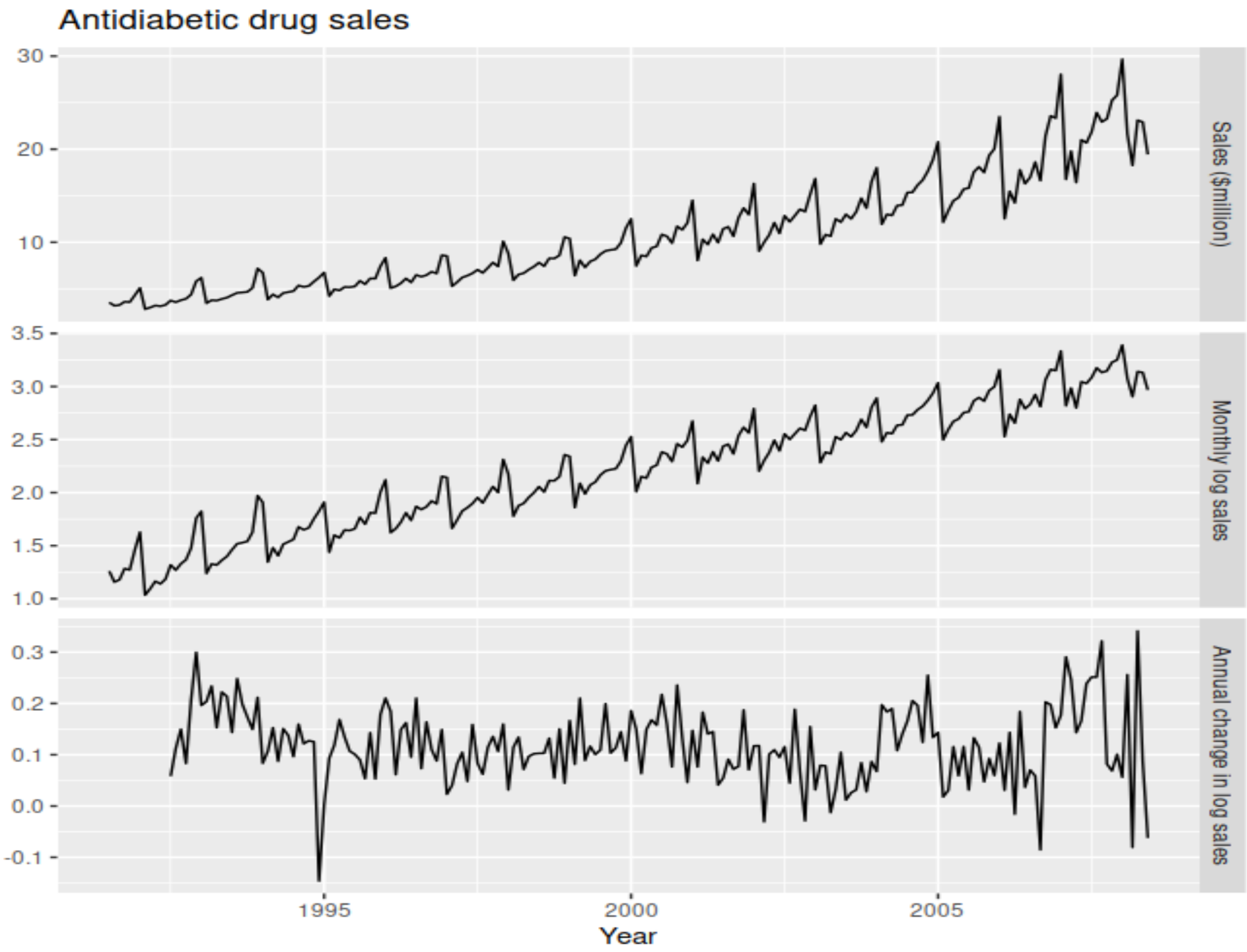
  ARIMA(0,1,q)(0,1,1) model (q=1 or 2)

  ARIMA(p,0,0)(0,1,1)+c model (p=1, 2 or 3)

  … in addition to other models (e.g., Random Walk, Single Exponential Smoothing, etc., with seasonal adjustment; or Winters Method)

- If there is a significant trend and/or the seasonal pattern is multiplicative, you should also try a natural log transformation.

Seasonal differencing



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

The logarithms stabilise the variance, while the seasonal differences remove the seasonality and trend.

The Figureshows 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.

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

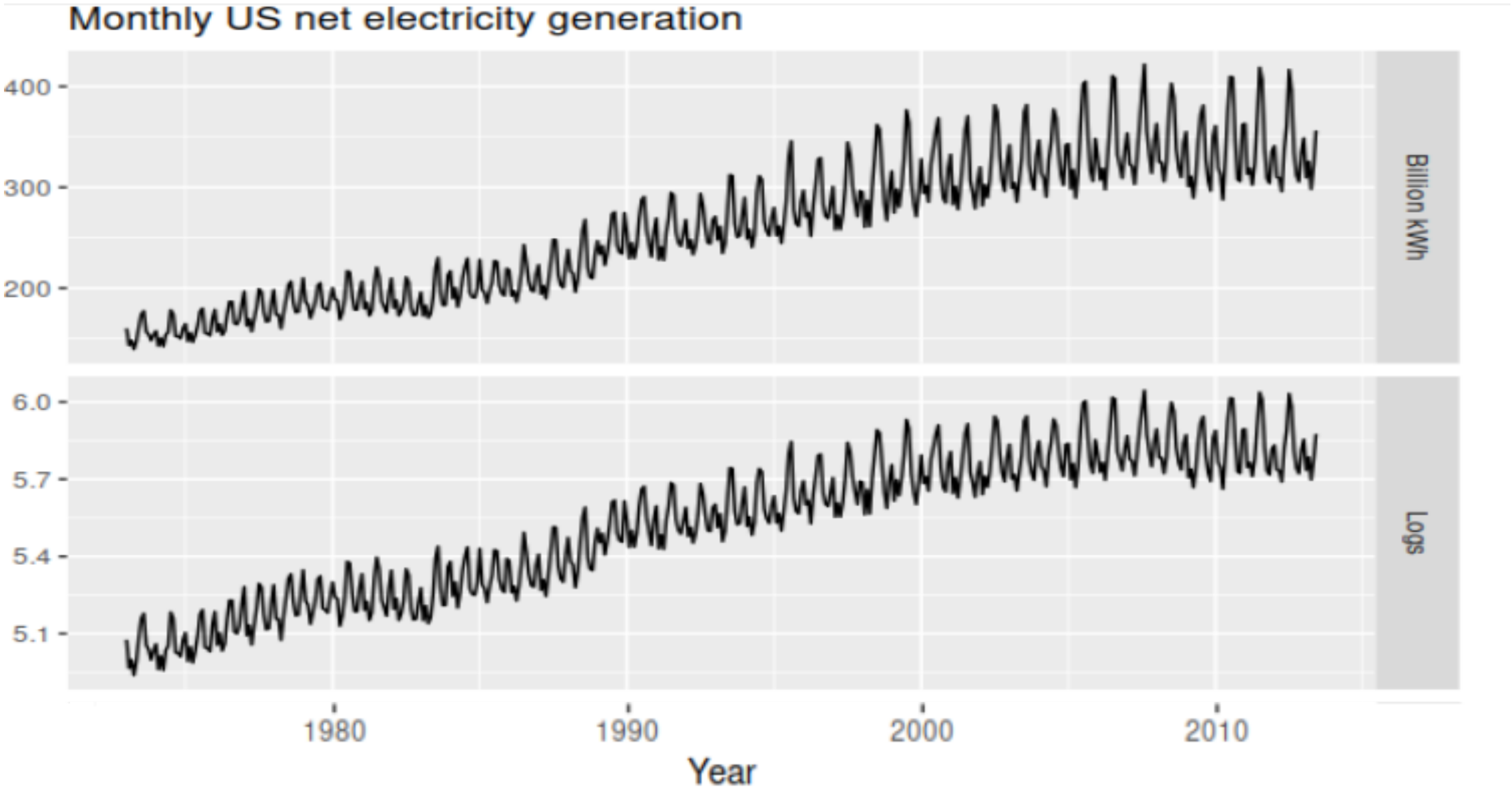
Seasonal differencing

- Sometimes it is necessary to take both a seasonal difference and a first difference to obtain stationary data, as is shown in Figure (Next slides).

- 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).

Seasonal differencing

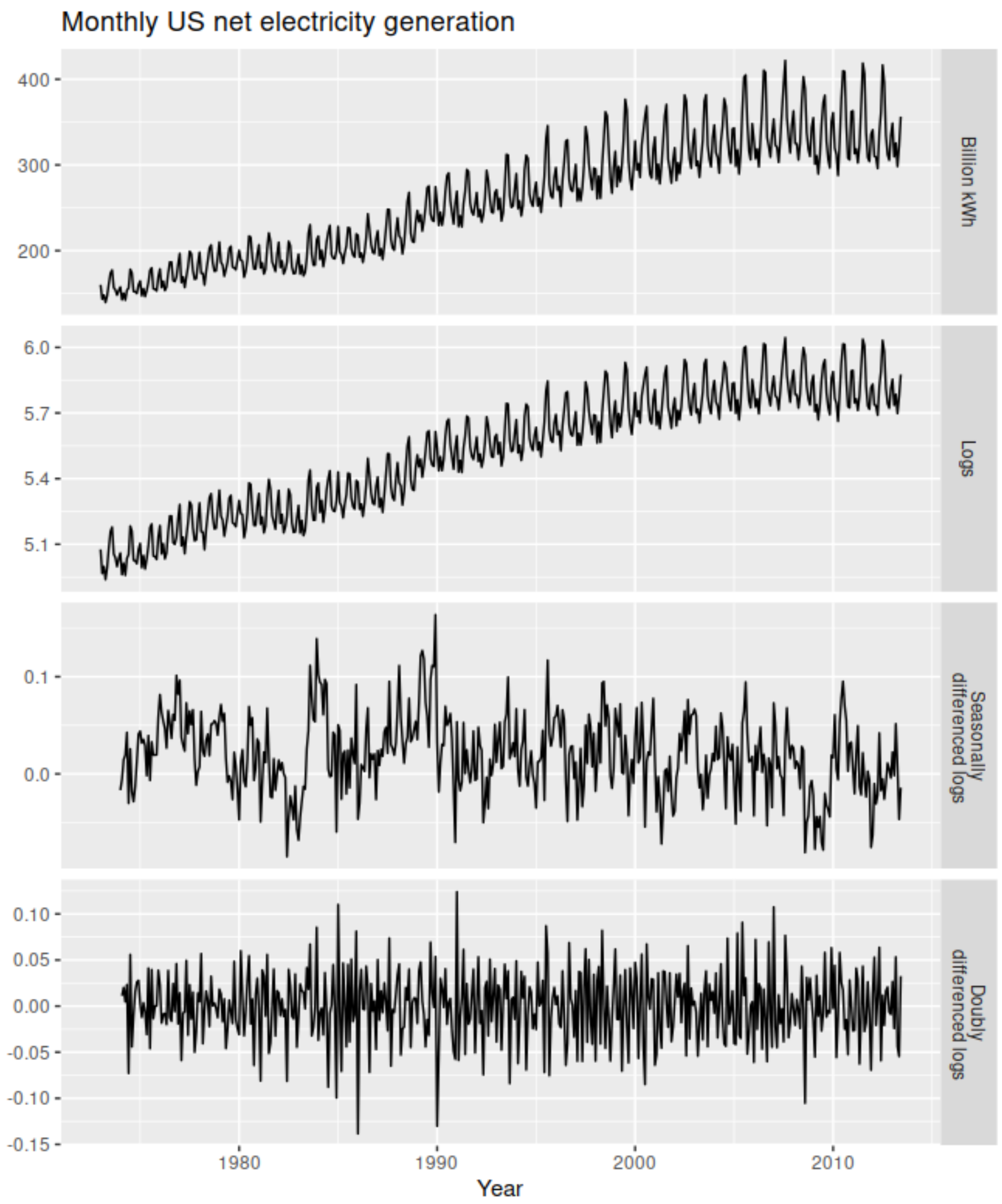


https://otexts.com/fpp2/stationarity.html

Figure 12: Monthly Net Electricity Generation in the US

Seasonal differencing

Top panel: US net electricity generation (billion kWh). Other panels show the same data after transforming and differencing.



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

Identifying Seasonal model

Take-aways

- Seasonal ARIMA models (especially the (0,1,*q*)x(0,1,1) and
- (*p*,0,0)x(0,1,1)+c models) compare favorably with other seasonal models and often yield better short-term forecasts.

- Advantages: solid underlying theory, stable estimation of time-varying trends and seasonal patterns, relatively few parameters.

- Drawbacks: no explicit seasonal indices, hard to interpret coefficients or explain “how the model works”, danger of overfitting or mis-identification if not used with care.

The ‘X’ Factor (ARX, ARIMAX, etc.)

- ‘X’ = exogenous variables or explanatory variables
- ARIMAX stands for *autoregressive integrated moving average with exogenous variables.

Important considerations:

- What other factors influence the forecast?

- How do we process this additional data to make it amenable for inclusion in our model?

The ‘X’ Factor (ARX, ARIMAX, etc.)

- An Autoregressive Integrated Moving Average with Explanatory Variable (ARIMAX) model can be viewed as a multiple regression model with one or more autoregressive (AR) terms and/or one or more moving average (MA) terms.

- This method is suitable for forecasting when data is stationary/non stationary, and multivariate with any type of data pattern, i.e., level/trend /seasonality/cyclicity.

- ARIMA is suitable for datasets that are univariate

- ARIMAX is suitable for analysis where there are additional explanatory variables (multivariate) in categorical and/or numeric format.

ARIMAX- Example

- To understand ARIMAX Forecasting, let's look at an example of annual GDP values in India. This dataset is suitable for the ARIMAX algorithm because there is more than one variable affecting the GDP – in other words, the dataset is multivariate.

- A company wants to forecast its product line growth for the new couple of years, based on data from the past thirty (30) years. The predictor variables for this use case would be yearly consumer inflation rate, yearly GDP data and yearly population growth rate.

- Predicting Dengue Spread Using Seasonal ARIMAX Model and Meteorological Data
  https://towardsdatascience.com/predicting-dengue-spread-using-seasonal-arimax-model-on-meteorology-data-3f35979ec5d

Image Courtesy

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

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

https://www.elegantjbi.com/blog/what-is-arimax-forecasting-and-how-is-it-used-for-enterprise-analysis.htm

THANK YOU

Jyothi R

Assistant Professor, Department of Computer Science

jyothir@pes.edu