Lecture 14

Time Series Analysis

DSA 8020 Statistical Methods II

April 12-16, 2021

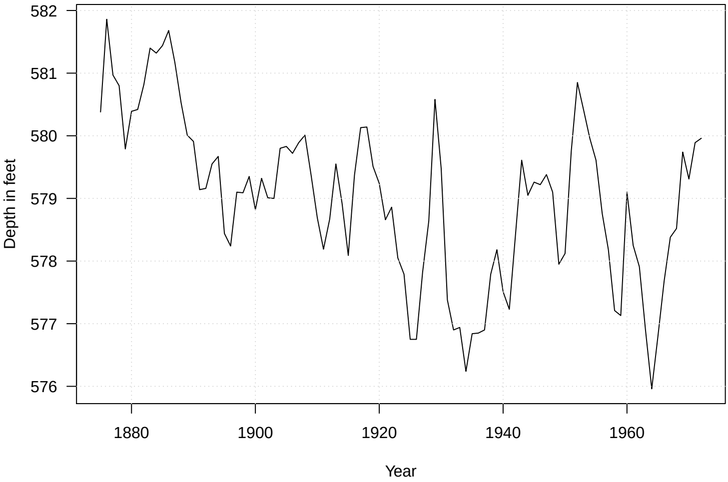
Whitney Huang

Clemson University

Agenda

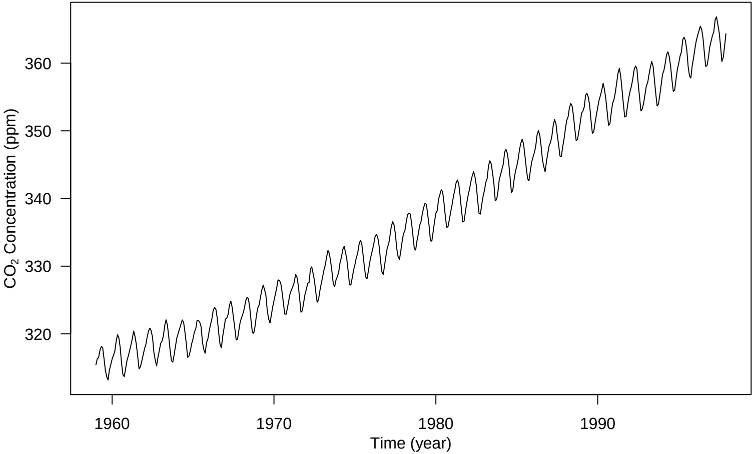
Level of Lake Huron 1875–1972

Annual measurements of the level of Lake Huron in feet. [Source: Brockwell & Davis, 1991]

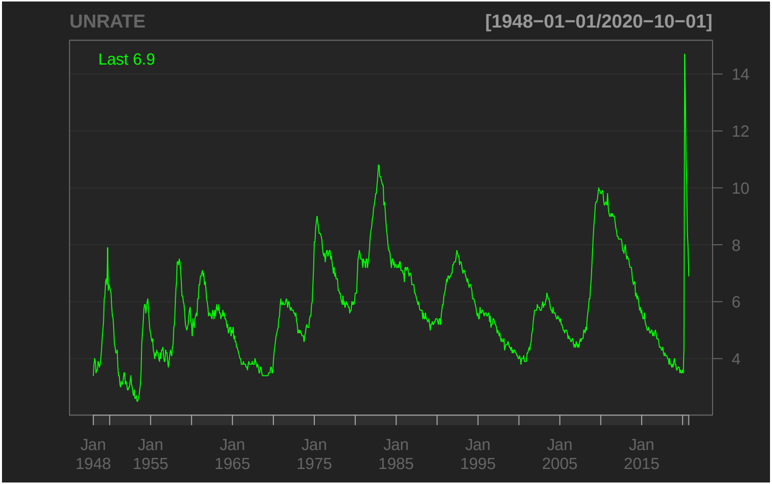


Mauna Loa Atmospheric CO_2 Concentration

Monthly atmospheric concentrations of CO_2 at the Mauna Loa Observatory [Source: Keeling & Whorf, Scripps Institution of Oceanography (SIO)]



US Unemployment Rate 1948 Jan. – 2020 Oct.



Time Series Data & Models

- A time series is a set of observations made sequentially in time
- Time series analysis is the area of statistics which deals with the analysis of dependency between different observations in time series data
- A time series model is a probabilistic model that describes ways that the series data $\{y_t\}$ could have been generated
- More specifically, a time series model is usually a probability model for $\{Y_t : t \in T\}$, a collection of random variables indexed in time

Some Objectives of Time Series Analysis

- Find a statistical model that adequately explains the dependence observed in a time series
- To conduct statistical inferences, e.g., Is there evidence of a decreasing trend in the Lake Huron depths?
- To forecast future values of the time series based on those we have already observed

Features of Times Series

- Trends
  - One can think of trend, μ_t as continuous changes, usually in the mean, over longer time scales
  - Usually the form of the trend is unknown and needs to be estimated. When the trend is removed, we obtain a detrended series
- Seasonal or periodic components
  - A seasonal component s_t constantly repeats itself in time, i.e., $s_t = s_{t+kd}$
  - We need to estimate the form and/or the period d of the seasonal component to deseasonalize the series
- The "noise" process
  - The noise process, η_t, is the component that is neither trend nor seasonality
  - We will focus on finding plausible (typically stationary) statistical models for this process

Combining Trend, Seasonality, and Noise Together

There are two commonly used approaches

- Additive model:

$$y_t = \mu_t + s_t + \eta_t$$

- Multiplicative model:

$$y_t = \mu_t s_t \eta_t$$

If all $\{y_t\}$ are positive then we obtain the additive model by taking logarithms:

$$\log y_t = \log \mu_t + \log s_t + \log \eta_t$$

Mean and Autocovariance

- The mean function of $\{Y_t\}$ is

$$\mu_t = \mathrm{E}[Y_t], \quad t \in T$$

- The autocovariance function of $\{Y_t\}$ is

$$\gamma(t, t') = \mathrm{Cov}(Y_t, Y_{t'}) = \mathrm{E}[(Y_t - \mu_t)(Y_{t'} - \mu_{t'})], \quad t, t' \in T$$

When $t = t'$ we obtain $\gamma(t, t') = \mathrm{Cov}(Y_t, Y_t) = \mathrm{Var}(Y_t) = \sigma_t^2$, the variance function of Y_t

Autocorrelation Function

The autocorrelation function (ACF) of $\{Y_t\}$ is

$$\rho(t,t') = \mathrm{Corr}(Y_t, Y_{t'}) = \frac{\gamma(t,t')}{\sqrt{\gamma(t,t)\gamma(t',t')}}$$

It measures the strength of linear association between Y_t and $Y_{t'}$

Properties:

1. $-1 \le \rho(t,t') \le 1, \quad t,t' \in T$
2. $\rho(t,t') = \rho(t',t), \quad \forall t,t' \in T; \rho(t,t) = 1, \quad \forall t \in T$
3. $\rho(t,t')$ is a non-negative definite function

Stationary Processes

We will still try to keep our models for $\{\eta_t\}$ as simple as possible by assuming stationarity, meaning that some characteristic of $\{\eta_t\}$ does not depend on the time points, only on the "time lag" between time points:

- $\mathrm{E}[\eta_t] = 0, \quad \forall t \in T$
- $\mathrm{Cov}(\eta_t, \eta_{t'}) = \gamma(t' - t) = \mathrm{Cov}(\eta_{t+s}, \eta_{t'+s})$

$\Rightarrow$ autocorrelation function (ACF):

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)}$$

Autoregressive Moving Average (ARMA) Models

Let $\{Z_t\}$ be independent and identical random variables that follow $\mathrm{N}(0, \sigma^2)$

- Moving Average Processes (MA(q)):
  $\eta_t = Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2} \cdots + \theta_q Z_{t-q}$

Autoregressive Moving Average (ARMA) Models

Let $\{Z_t\}$ be independent and identical random variables that follow $\mathrm{N}(0, \sigma^2)$

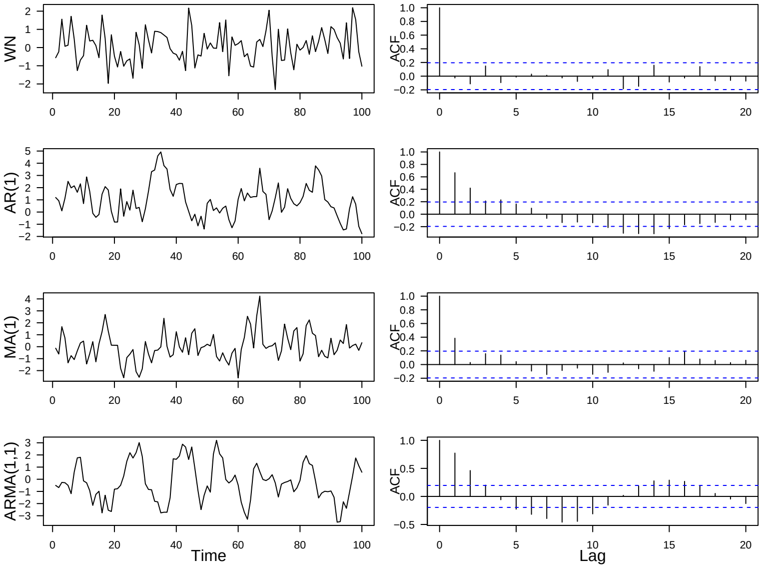
- Moving Average Processes (MA(q)):
  $\eta_t = Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2} \cdots + \theta_q Z_{t-q}$
- Autoregressive Processes (AR(p)):
  $\eta_t = \phi_1 \eta_{t-1} + \phi_2 \eta_{t-2} + \cdots + \phi_p \eta_{t-p} + Z_t$

Autoregressive Moving Average (ARMA) Models

Let $\{Z_t\}$ be independent and identical random variables that follow $\mathrm{N}(0, \sigma^2)$

- Moving Average Processes (MA(q)):
  $\eta_t = Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2} \cdots + \theta_q Z_{t-q}$
- Autoregressive Processes (AR(p)):
  $\eta_t = \phi_1 \eta_{t-1} + \phi_2 \eta_{t-2} + \cdots + \phi_p \eta_{t-p} + Z_t$
- Autoregressive Moving Average Processes ARMA(p,q):
  $\eta_t =$
  $\phi_1 \eta_{t-1} + \phi_2 \eta_{t-2} + \cdots + \phi_p \eta_{t-p} + Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2} + \cdots + \theta_q Z_{t-q}$

Autocorrelation Plot

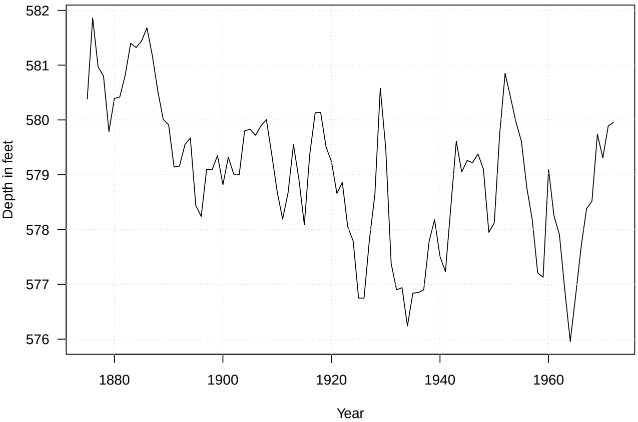


Lake Huron Case Study

Source: https://www.worldatlas.com/articles/what-states-border-lake-huron.html

- Detrending
- Model fitting and selection
- Forecasting

Annual Measurements of the Level of Lake Huron

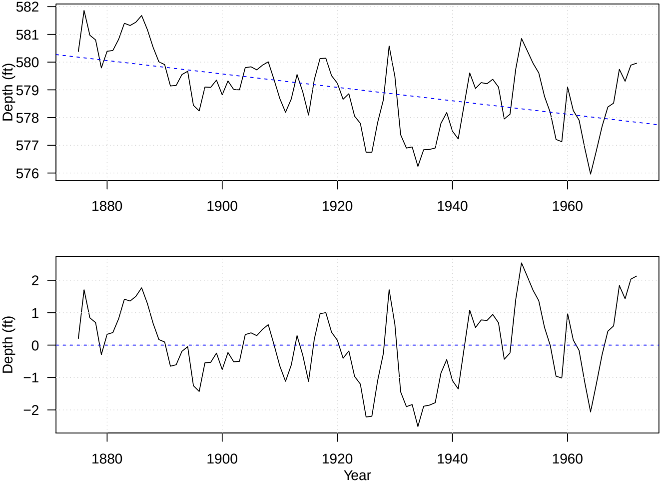


There seems to be a decreasing trend ⇒ need to estimate the trend to get the detrended series

Plots of the Trend and Residuals

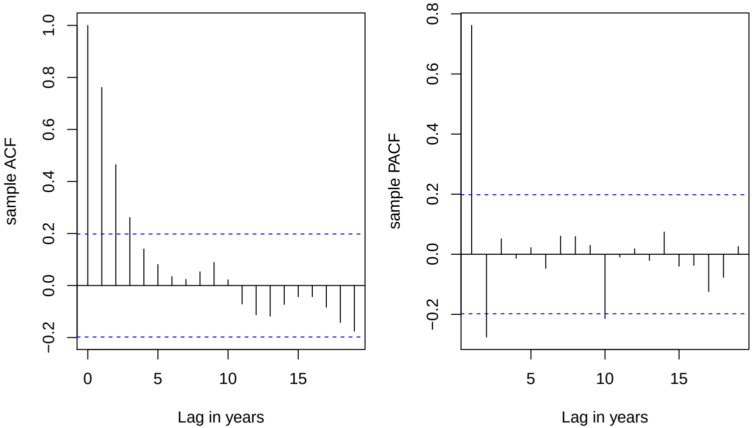
$$y_t = \overbrace{\mu_t}^{\text{trend}} + \underbrace{\eta_t}_{\text{residual}},$$

where we **assume** $\mu_t = \alpha + \beta t$, i.e., a linear trend in time

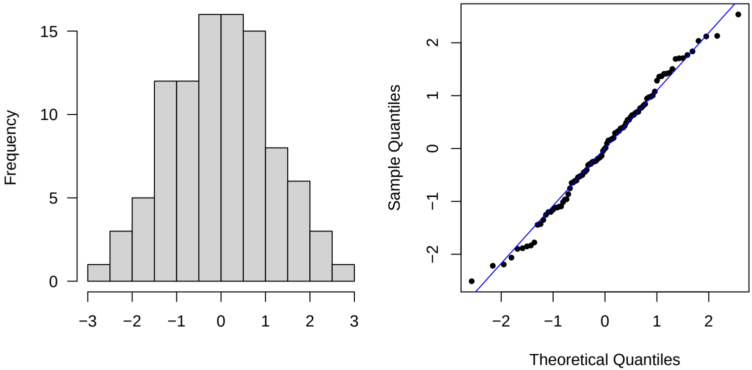


ACF and PACF Plots

- Tapering pattern in ACF ⇒ need to include AR terms
- Significant PACF values at the first 2 lags ⇒ a AR(2) may be appropriate



Assessing Normality Assumption for η_t



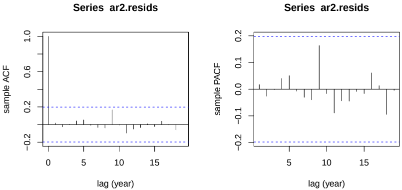
Fitting AR(2)

```
> (ar2.model <- arima(deTrend, order = c(2, 0, 0)))

Call:
arima(x = deTrend, order = c(2, 0, 0))

Coefficients:
         ar1      ar2  intercept
      1.0047  -0.2919     0.0196
s.e.  0.0977   0.1004     0.2351

sigma^2 estimated as 0.4571:  log likelihood = -101.25,  aic = 210.5
```




```
> Box.test(ar2.resids, type = "Ljung-Box")

        Box-Ljung test

data:  ar2.resids
X-squared = 0.029966, df = 1, p-value = 0.8626
```

Model Selection

We can conduct model selection by using, for example, AIC

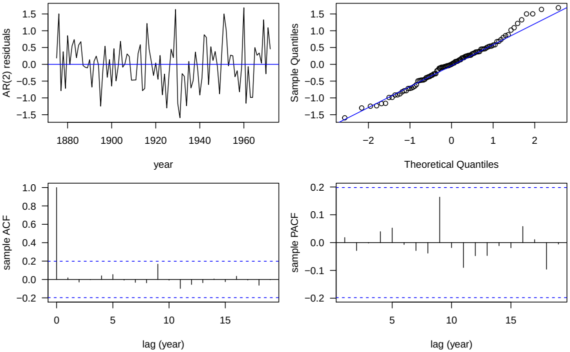
```
> ar1.model <- arima(deTrend, order = c(1, 0, 0))
> ar2.model <- arima(deTrend, order = c(2, 0, 0))
> arma21.model <- arima(deTrend, order = c(2, 0, 1))
> AIC(ar1.model); AIC(ar2.model); AIC(arma21.model)
[1] 216.5835
[1] 210.5032
[1] 212.1784
```

Fitting AR(2) + a Linear Trend

```
> library(forecast)
> (fit <- Arima(LakeHuron, order = c(2, 0, 0), include.drift = T))
Series: LakeHuron
ARIMA(2,0,0) with drift

Coefficients:
         ar1      ar2  intercept    drift
      1.0048  -0.2913   580.0915  -0.0216
s.e.  0.0976   0.1004     0.4636   0.0081

sigma^2 estimated as 0.476:  log likelihood=-101.2
AIC=212.4   AICc=213.05   BIC=225.32
```



10-Year-Ahead Forecasts