

MATH 8090: Introduction and Overview of the Course

Whitney Huang, Clemson University

8/18/2021

Contents

Time Series Data	1
Exploratory Time Series Analysis	2
Trend, Seasonality, and Noise	2
Trends	2
Seasonal components	2
Noises	3
Combining Trend, Seasonality, and Noise Together	3
Lake Huron Time Series Example	4
Time Series Models	8
Stationarity	8
Objectives of time series analysis	8
Modeling	8
Forecasting	9
Adjustment	9
Simulation	9
Control	10
References	10

Time Series Data

A time series is a set of observations $\{y_t : t \in T\}$ made sequentially in time t with the index set T . The observations can be taken at fixed (equidistant) time points (e.g., $T = \{0, 1, 2, \dots\}$) or at irregular times, or even over an entire time interval (e.g., $T = [0, T]$). Different types of time sampling require different approaches to the data analysis. In this course we will focus on some of the most commonly used methods for dealing with the first scenario, that is, *discrete-time* equal-spaced time series. A discrete-time time series might be intrinsically discrete (e.g., number of planes departing ATL every day) or might arise from a underlying *continuous-time* time series via

- sampling (e.g., instantaneous wind speed)
- aggregation (e.g., daily accumulated precipitation amount)

- extrema (e.g., daily maximum temperature)

It is worth pointing out that the methods for continuous-time time series (i.e., T is an entire interval) are useful in the setting where data were observed at irregular times.

Depending on the type of value that y_t can take, we have:

- real-valued: a value in \mathbb{R} or subset thereof (e.g., temperature)
- complex-valued (some applications in electrical engineering)
- non-negative integer (See Jia et al. (2021) for a recent development for modeling such kind of count time series data)
- categorical (e.g., outcome of scheduled basketball match: win, lose, cancellation)
- circular (e.g., wind direction. See Breckling (2012))

Exploratory Time Series Analysis

Suppose we have a time series data $\{y_t : t \in T\}$, we will start with a *time series plot* of y_t versus t . It is important to label the axes carefully (i.e., the time unit in x-axis, variable name and its unit in y-axis) and adjust the aspect ratio to get a nice looking time series plot. In this exploratory data analysis (EDA) stage, we look the following:

1. Are there abrupt changes? (e.g., shifts in mean and/or variance)
2. Are there *outliers*? (i.e., unusual values w.r.t. the rest of the data)
3. Is there a need to transform the data (e.g., should we take some transformation to stabilize the variance)?
4. Features of the time series: trend, seasonal components, and noise process

Trend, Seasonality, and Noise

Trends

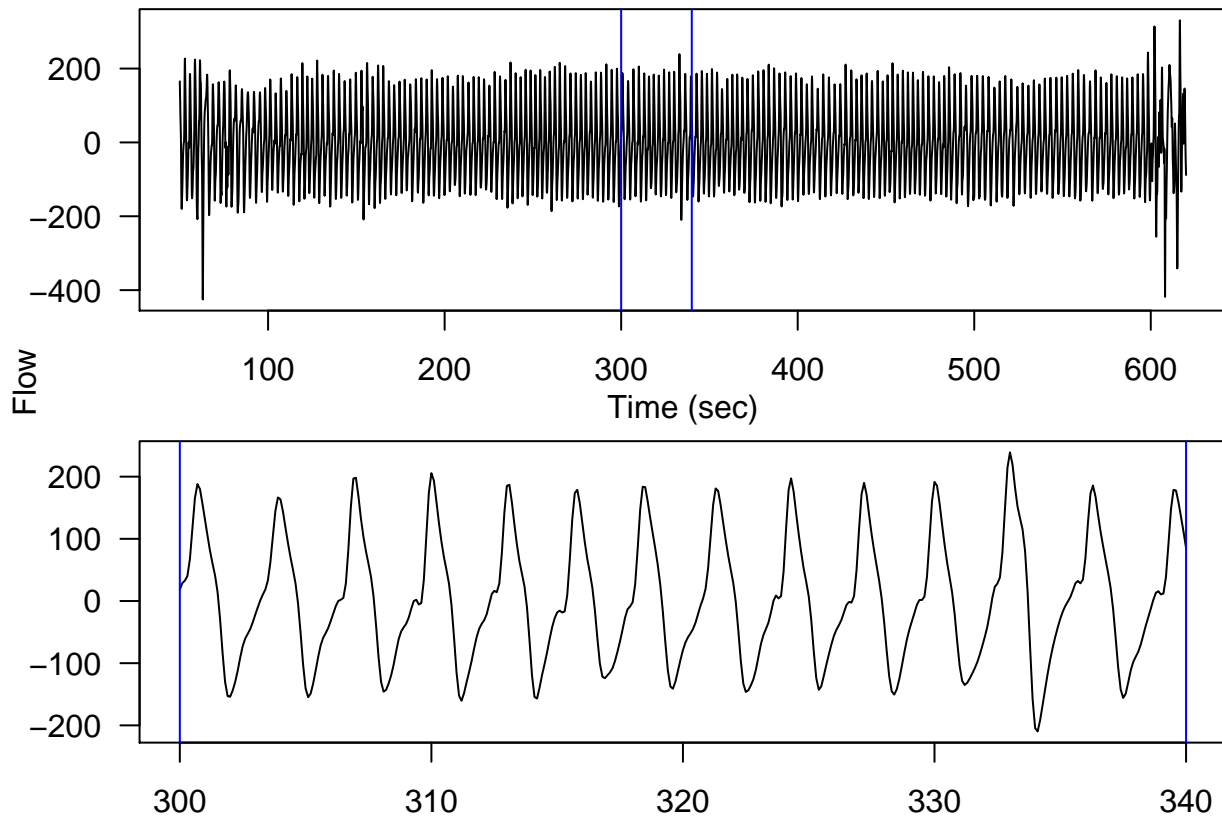
One can think of trend, μ_t , as continuous changes, usually in the mean, over longer time scales: *the essential idea of trend is that it shall be smooth*. Usually the form of the trend is unknown and needs to be specified and then estimated for the data. When the trend is estimated and then removed, we obtain a [detrended](#) series

Seasonal components

A seasonal component, s_t , constantly repeats itself in time, (i.e., $s_t = s_{t+kd}$ for all t and k). Sometimes the seasonality has a very clear structure form (e.g., a sin or cosine wave) while other times the structure is more complicated (see the following sleep airflow data example used in Huang et al. (2020)).

```
load("Flow_sub1.RData")
par(mfrow = c(2, 1), mar = c(2.6, 3.6, 0.8, 0.6))
id <- 500:6200
plot(id / 10, flow[id], type = "l", las = 1, xlab = "", ylab = "")
mtext("Time (sec)", 1, line = 2)
mtext("Flow", 2, line = 2.5, at = -650)
abline(v = c(3000, 3400) / 10, col = "blue")
```

```
id2 <- 3000:3400
plot(id2 / 10, flow[id2], type = "l", las = 1, xlab = "", ylab = "")
abline(v = c(3000, 3400) / 10, col = "blue")
```



Noises

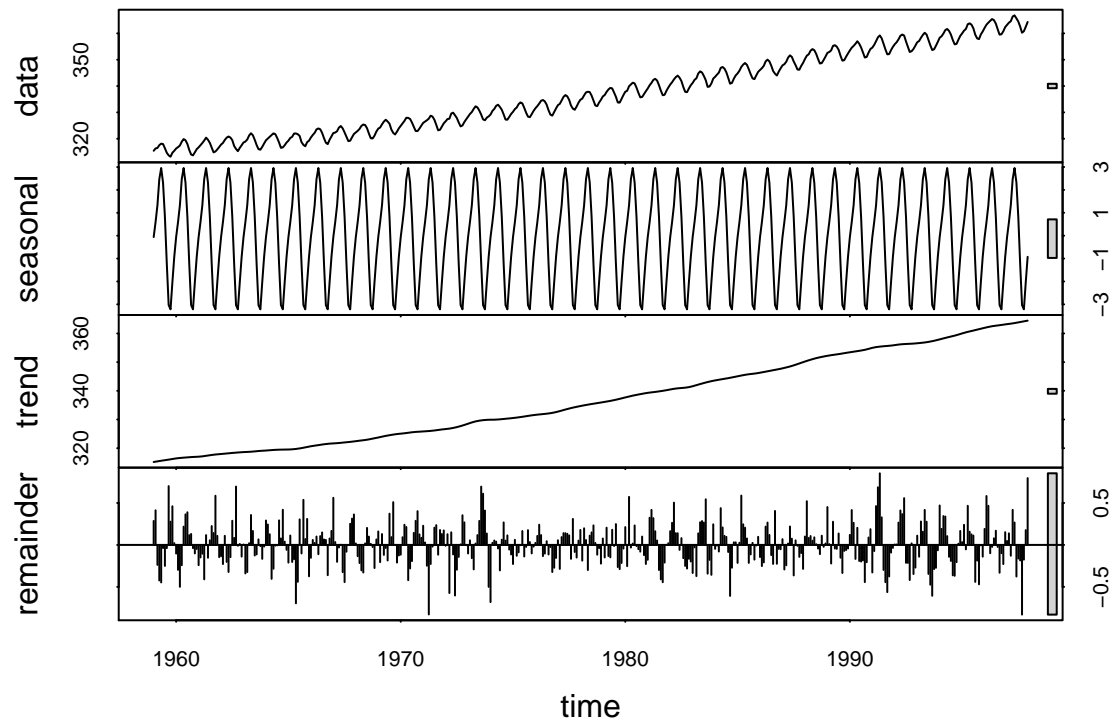
The noise process, η_t , is the component that is neither trend nor seasonality. We will focus on finding plausible (typically stationary) statistical models for this process.

Combining Trend, Seasonality, and Noise Together

There are two common approaches to combine these components:

1. The *additive model*, $y_t = \mu_t + s_t + \eta_t, t = 1, \dots, T$.

```
# Seasonal and Trend decomposition using Loess (STL)
par(mar = c(4, 3.6, 0.8, 0.6))
stl <- stl(co2, s.window = "periodic")
plot(stl, las = 1)
```



2. The *multiplicative model*, $y_t = \mu_t s_t \eta_t, t = 1, \dots, T$.

Note that if all the variables are *positive* then we obtain the additive model by taking logarithms:

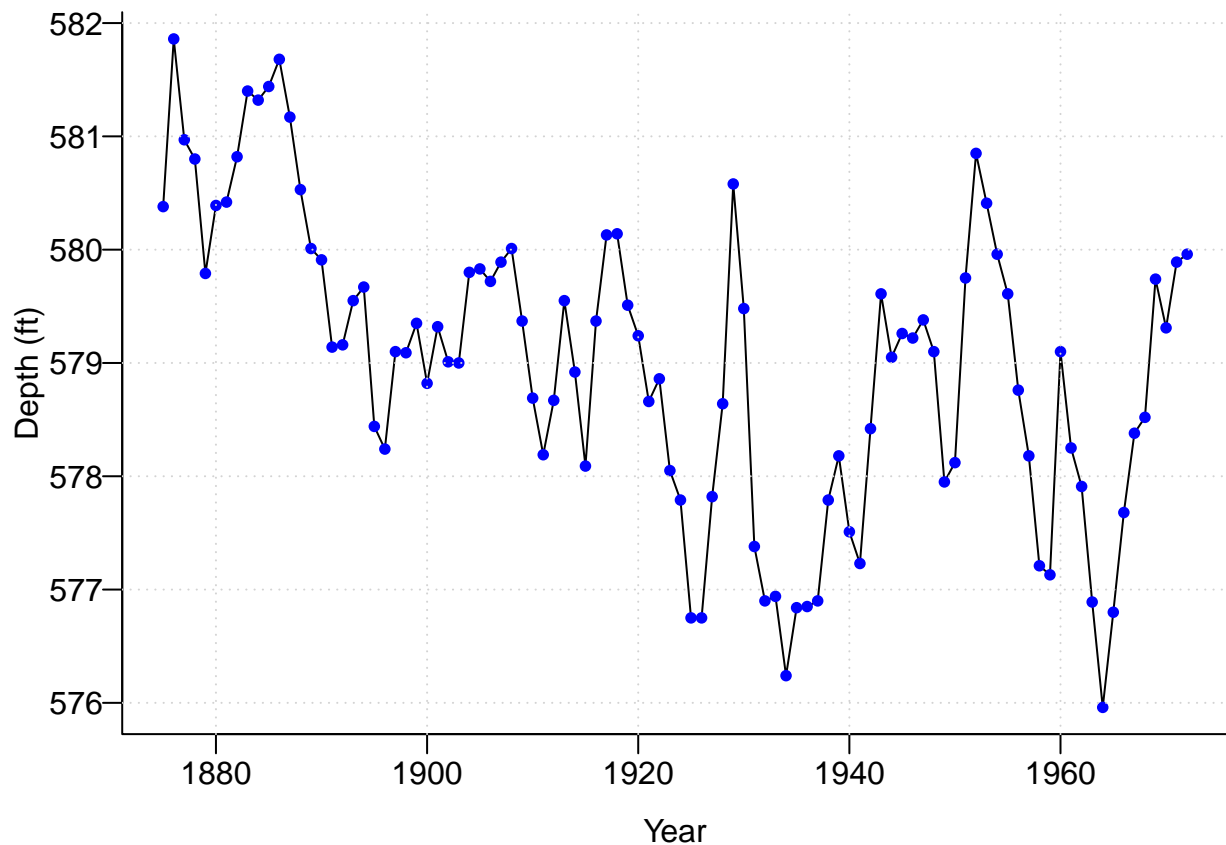
$$\log(y_t) = \log(\mu_t) + \log(s_t) + \log(\eta_t), t = 1, \dots, T.$$

Keep in mind that we need to transform back so that we can interpret on the *original measurement scale*.

Lake Huron Time Series Example

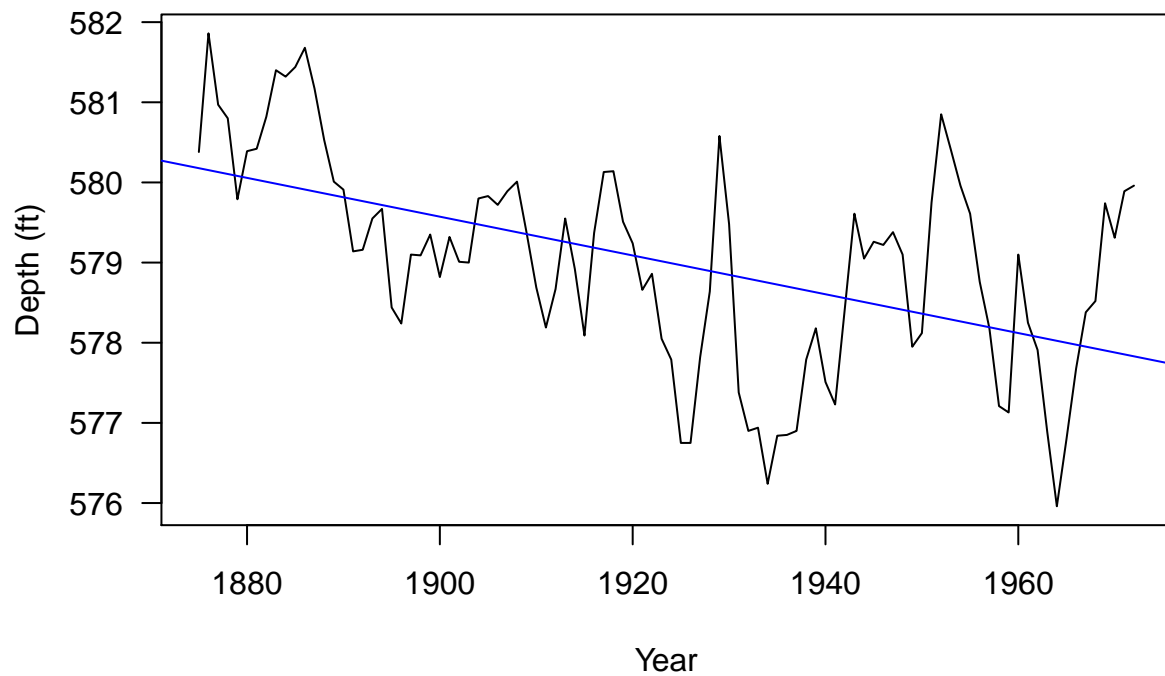
This example is taken from the book by Brockwell et al. (2002). The data consist of annual measurement of the depth, in feet (ft), of Lake Huron from 1875-1972.

```
par(mar = c(3.2, 3.2, 0.5, 0.5), mgp = c(2, 0.5, 0), bty = "L")
data(LakeHuron)
plot(LakeHuron, ylab = "Depth (ft)", xlab = "Year", las = 1)
points(LakeHuron, cex = 0.8, col = "blue", pch = 16)
grid()
```

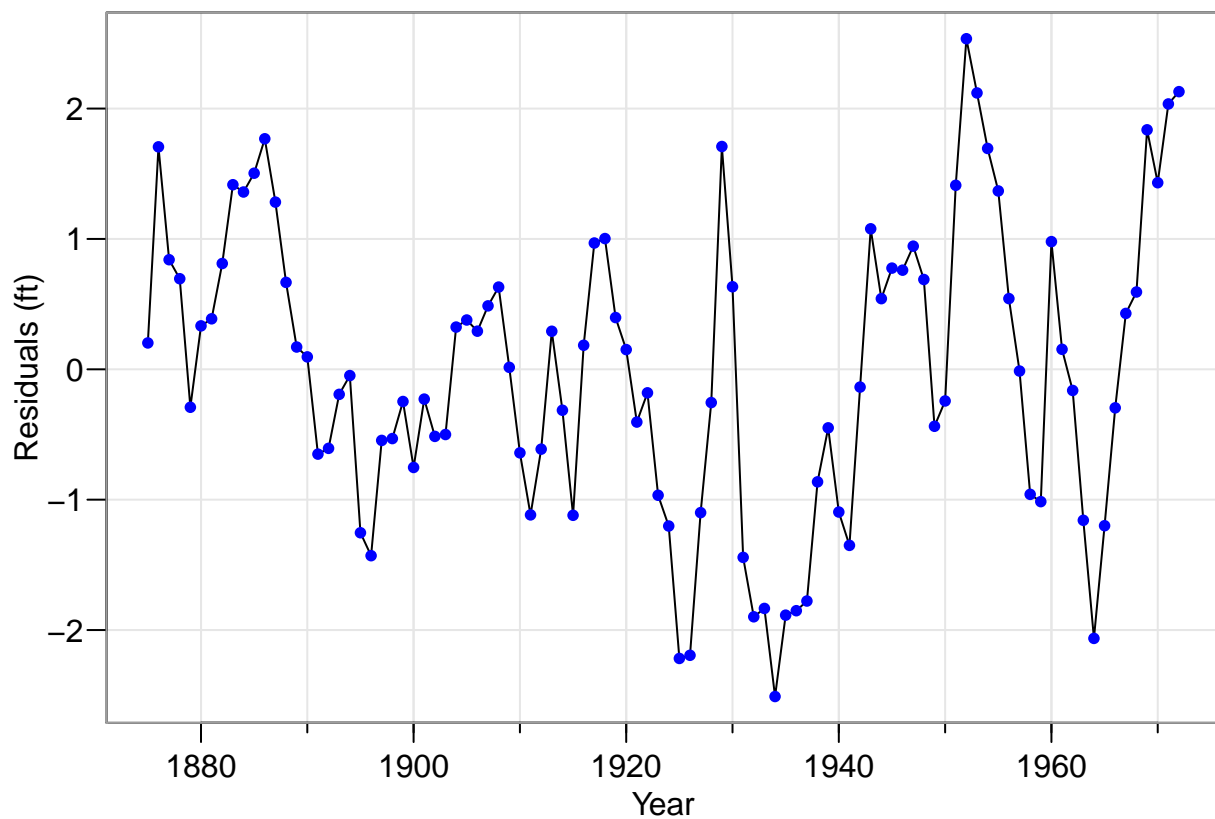


This is an example of *discrete-time* equal-spaced *real-valued* time series. Some key features of this time series are 1) decreasing trend; 2) some *random* fluctuations around the decreasing trend. Let's conduct a simple analysis to this data to get a better idea about the nature of these random *noises*. Specifically, we are going to assume a linear trend for this time series so we can remove the trend by performing a *simple linear regression*.

```
library(astsa)
yr <- time(LakeHuron)
lm <- lm(LakeHuron ~ yr)
plot(LakeHuron, ylab = "Depth (ft)", xlab = "Year", las = 1)
abline(lm, col = "blue")
```



```
lm$residuals <- ts(lm$residuals, start = 1875, end = 1972)
tsplot(lm$residuals, ylab = "Residuals (ft)", xlab = "Year", las = 1)
points(lm$residuals, cex = 0.8, col = "blue", pch = 16)
```

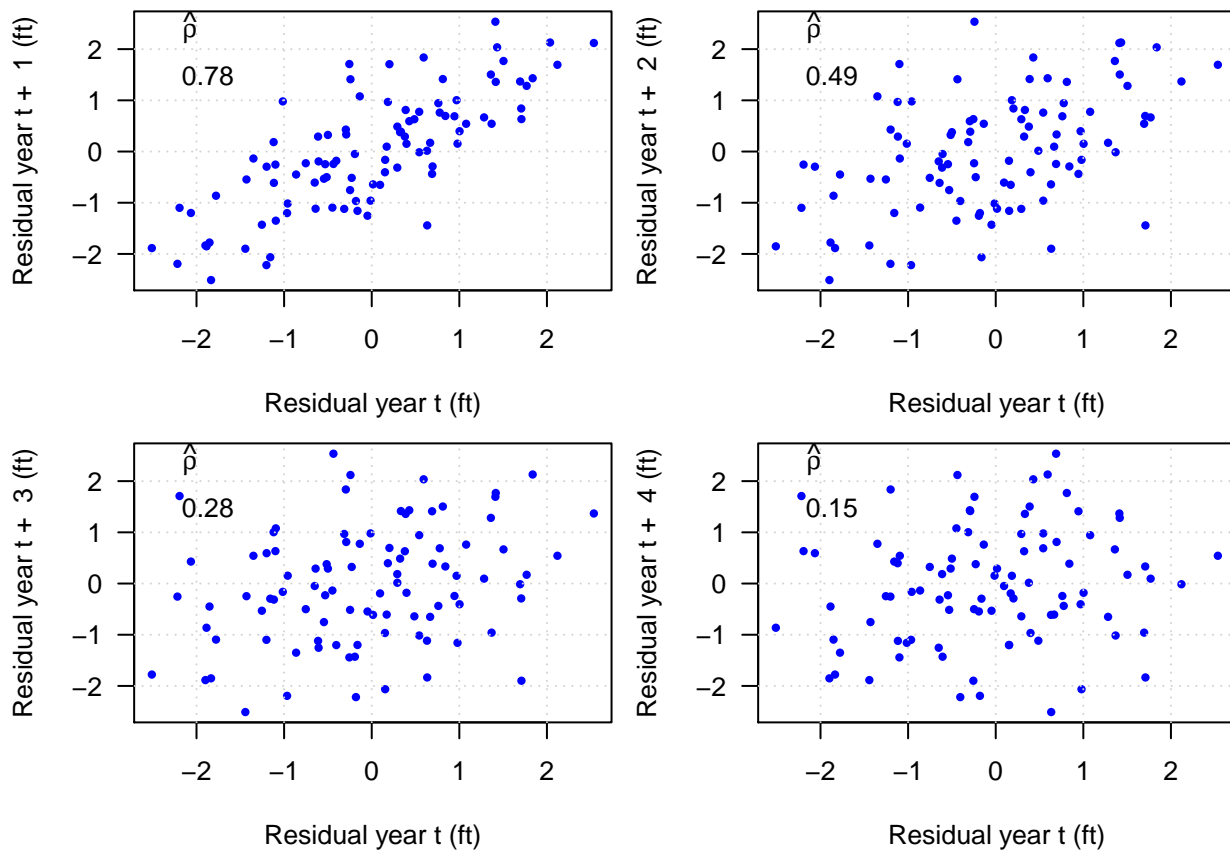


These residual values exhibit some temporal dependence structure, that is, the nearby (in time) values tend to be more alike than those far part values. To see this, let's make a few time lag plots.

```

n <- length(LakeHuron)
h <- 1:4
par(mfrow = c(2, 2), mar = c(4, 4, 0.8, 0.6))
for (i in h){
  plot(lm$residuals[-(n - i + 1):n], lm$residuals[-(1:i)], pch = 16, col = "blue", cex = 0.7,
       las = 1, xlab = "Residual year t (ft)",
       ylab = paste("Residual year t + ", h[i], "(ft)"))
  legend("topleft", legend = round(cor(lm$residuals[-(n - i + 1):n], lm$residuals[-(1:i)]), 2),
        title = expression(hat(rho)), bty = "n")
  grid()
}

```



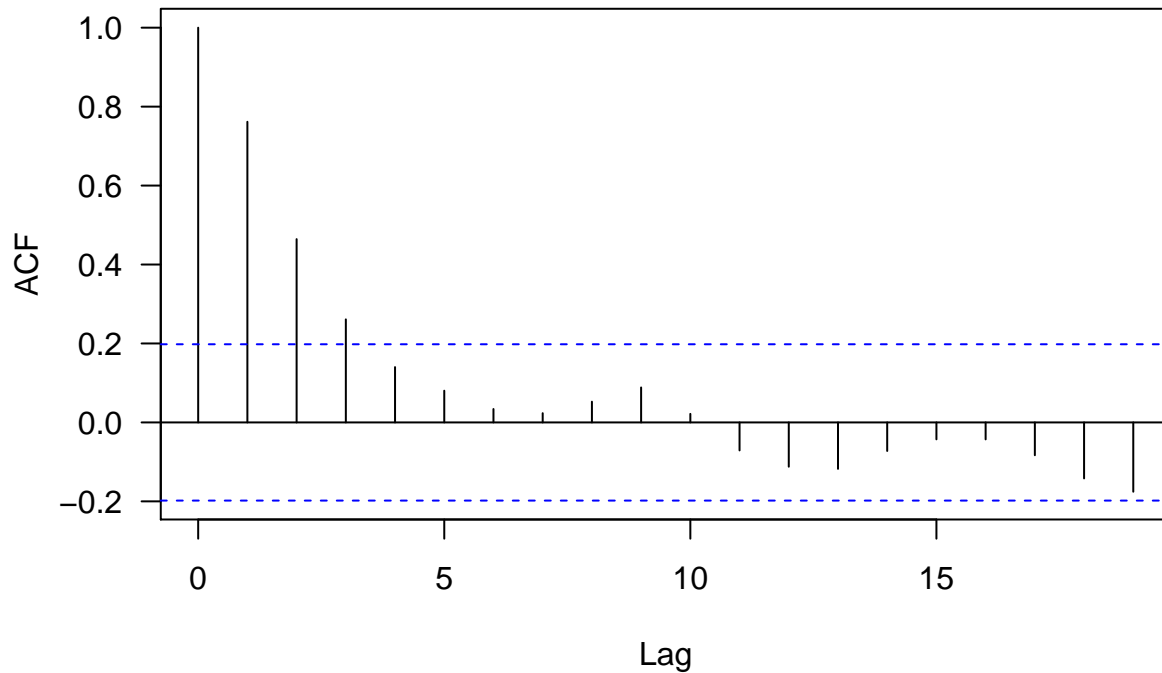
Later in this course we will learn how to use the autocovariance plot to explore such temporal dependence structure.

```

acf(lm$residuals, las = 1)

```

Series lm\$residuals



[Time series analysis](#) is the area of statistics which (largely) deals with the analysis of *dependency* between observations (i.e., $\{\eta_t\}$ in time series data).

Time Series Models

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](#).

Stationarity

Here we have a very challenging task as we need to *estimate* the characteristics of a T -dimensional random vector with only a single realization (i.e., one realized value at each time point). Therefore, we will restrict our model class by assuming **stationarity**. Stationarity means that some characteristic of the distribution of a time series does not depend on t , only on the *distance* between time points. While most time series are not stationary, one either remove or model the non-stationary parts (e.g., de-trend or de-seasonalization) so that we are only left with a stationary component (e.g., $\{\eta_t\}$). We typically further assume that the process is [second order stationary](#), that is, the mean function (as a function of t) is a constant and the covariance function is a function of the *distance* between time points only.

Objectives of time series analysis

Modeling

Find a statistical model that adequately explains the observed time series data. For example, identify a model which can account for the fact that the depths of Lake Huron are correlated with different years and

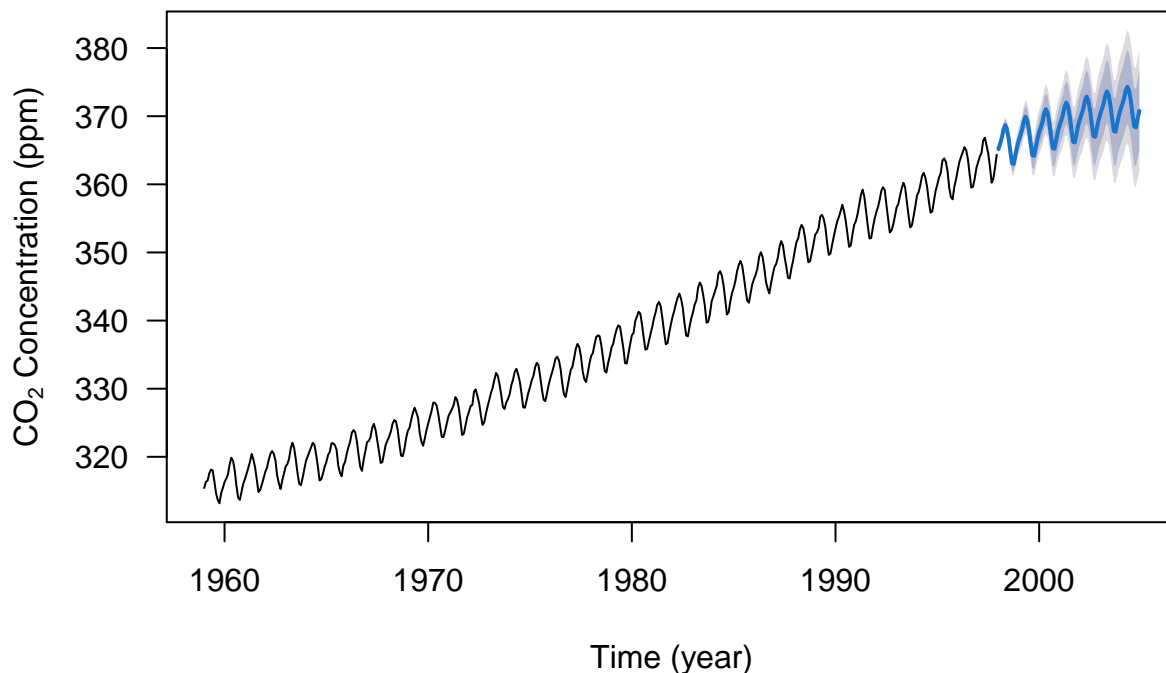
with a decreasing long-term trend. The fitted model can be used for further *statistical inference*, for instant, to answer the question like *Is there evidence of decreasing trend in the Lake Huron depths?*

Forecasting

This is perhaps the most common objective. We observe a time series of given length and wish to *predict* or *forecast* future values of the time series based on those we have already observed. A statistical model allows us to make a probabilistic forecast (i.e., provides both point and interval estimates)

```
library(forecast)
TBATSfit <- tbats(co2, use.box.cox = F, use.trend = F, use.damped.trend = F,
                 seasonal.periods = 12)
plot(forecast(TBATSfit, 84), las = 1, xlab = "Time (year)",
     ylab = expression(paste("CO"[2], " Concentration (ppm)")))
```

Forecasts from TBATS(1, {3,1}, −, {<12,5>})



Adjustment

An example would be seasonal adjustment, where the seasonal component is estimated and then removed in order to better understand the underlying trend.

Simulation

We can use a time series model (which adequately describes a physical process) as a surrogate to simulate repeatedly in order to approximate how the physical process behaves.

Control

We can adjust various input (control) parameters so that the time series fits closer to a given standard (many examples from statistical quality control).

References

- Breckling, Jens. 2012. *The Analysis of Directional Time Series: Applications to Wind Speed and Direction*. Vol. 61. Springer Science & Business Media.
- Brockwell, Peter J, Peter J Brockwell, Richard A Davis, and Richard A Davis. 2002. *Introduction to Time Series and Forecasting*. Springer.
- Huang, Whitney K, Yu-Min Chung, Yu-Bo Wang, Jeff E Mandel, and Hau-Tieng Wu. 2020. “Airflow Recovery from Thoracic and Abdominal Movements Using Synchrosqueezing Transform and Locally Stationary Gaussian Process Regression.” *arXiv Preprint arXiv:2008.04473*.
- Jia, Yisu, Stefanos Kechagias, James Livsey, Robert Lund, and Vlasdas Pipiras. 2021. “Latent Gaussian Count Time Series.” *Journal of the American Statistical Association*, no. just-accepted: 1–28.