

Stat 443. Time Series and Forecasting.

Start with some data examples of time series for motivation.

Concepts at beginning of term:

- Time series data plots, trend and seasonality, autocorrelation.
- Operations of time series: deseasonalizing, smoothing, filtering, differencing.
- **Forecasting rules**: assessing and comparing different forecast rules.

For any data set, always start with exploratory data analysis to look for patterns and trends, etc.

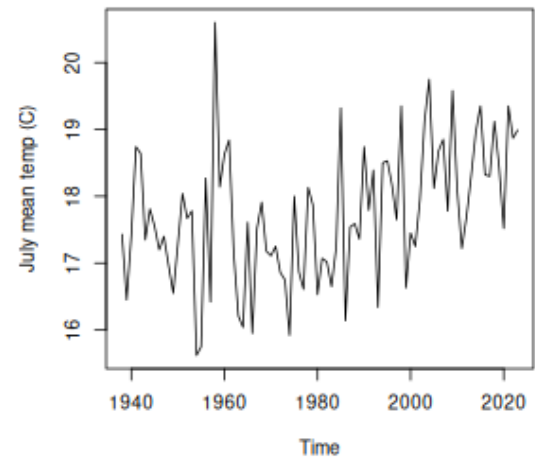
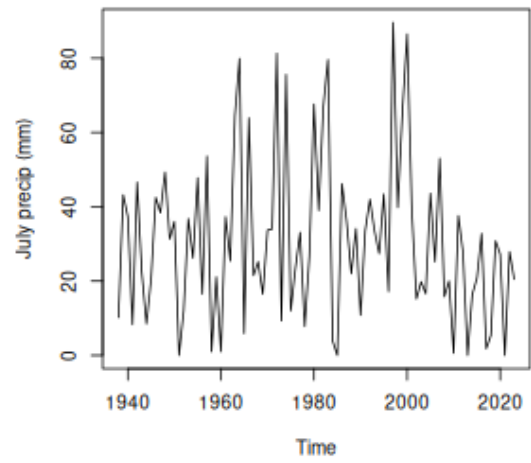
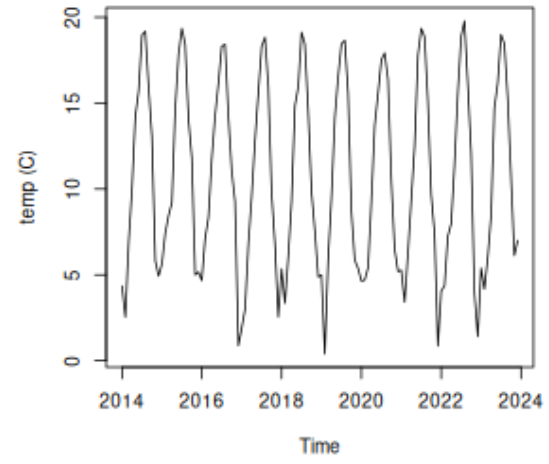
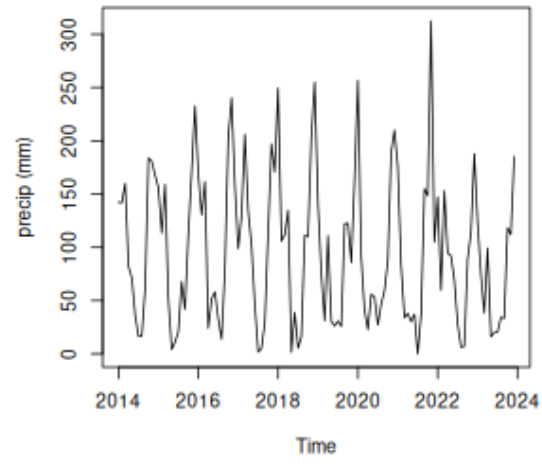
Example 1

Vancouver monthly mean temperature and total precipitation 1938 to 2023. Data obtained from <https://climate.weather.gc.ca> and wrangled to monthly variables.

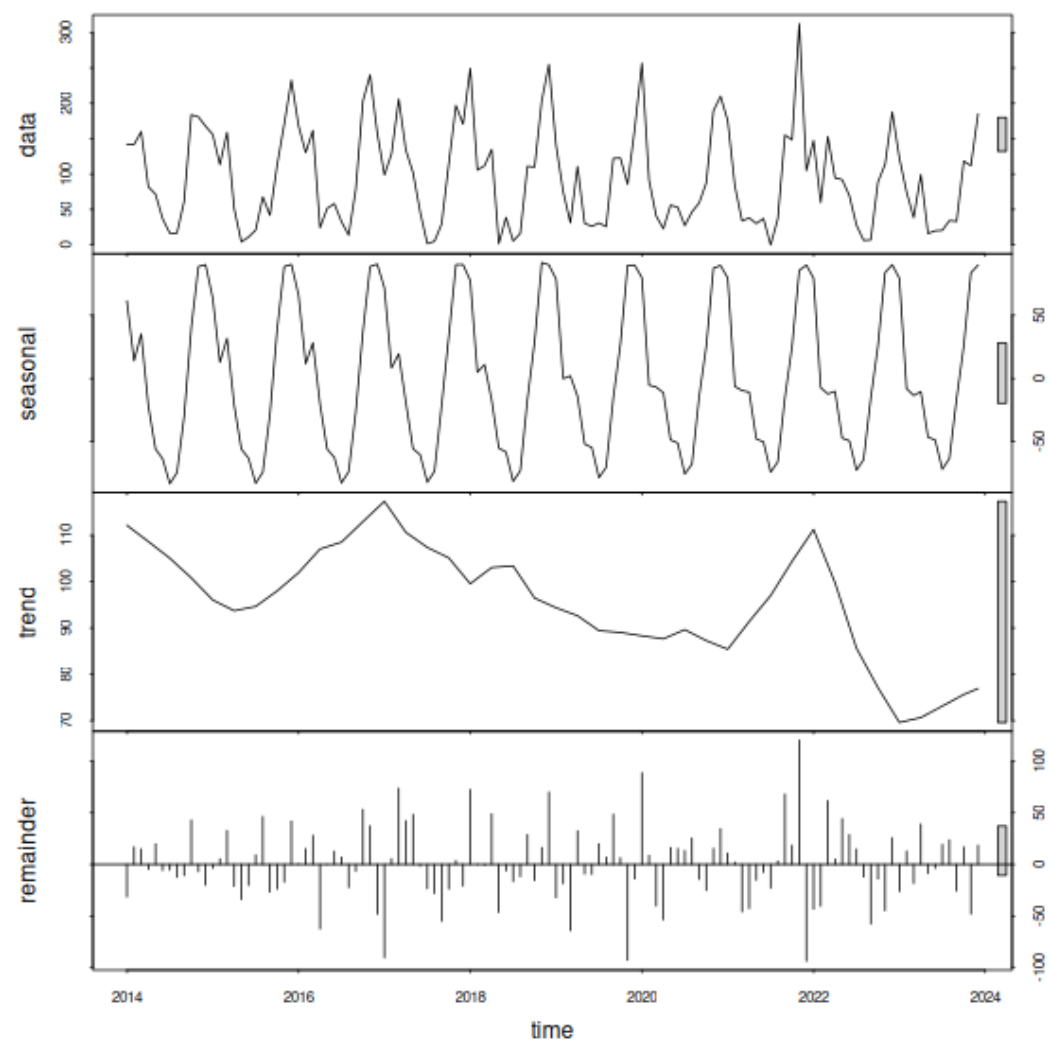
Wrangled data set at course web site: vanc-prec-temp.csv

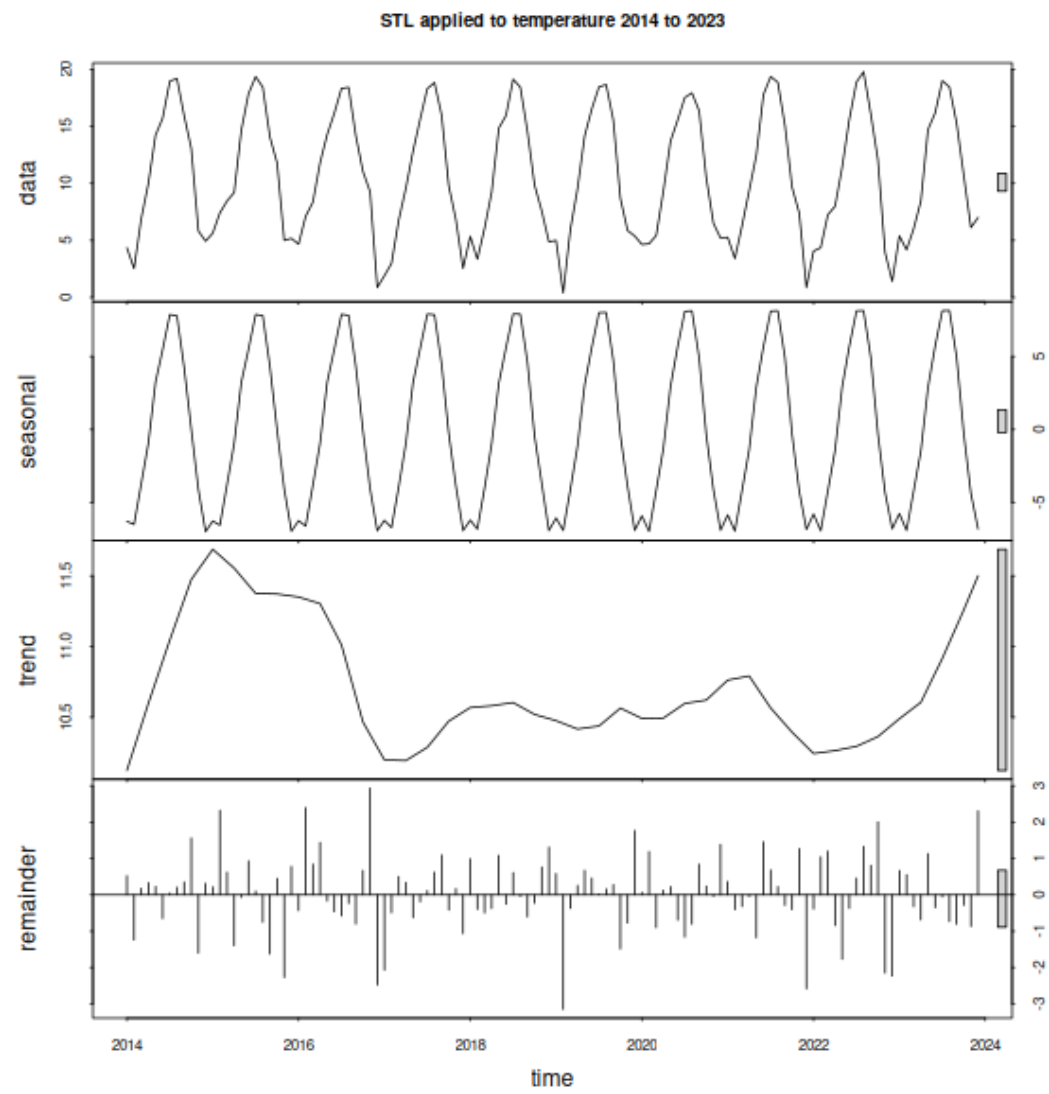
The next few slides have time series plots over a few years and plots for Julys since 1938. **What patterns do you notice?**

What might be forecasting rules for total monthly precipitation or average monthly temperature in future months?
What might be forecasting rules for future July monthly precipitation or July average monthly temperature?

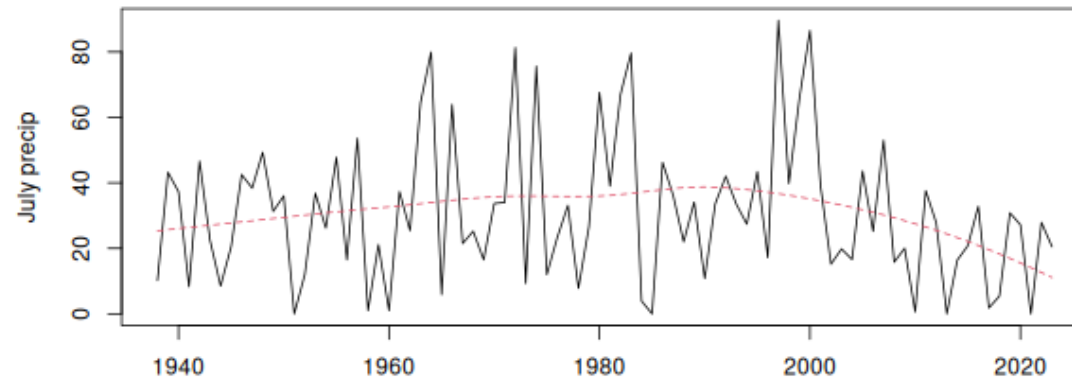


STL applied to precipitation 2014 to 2023

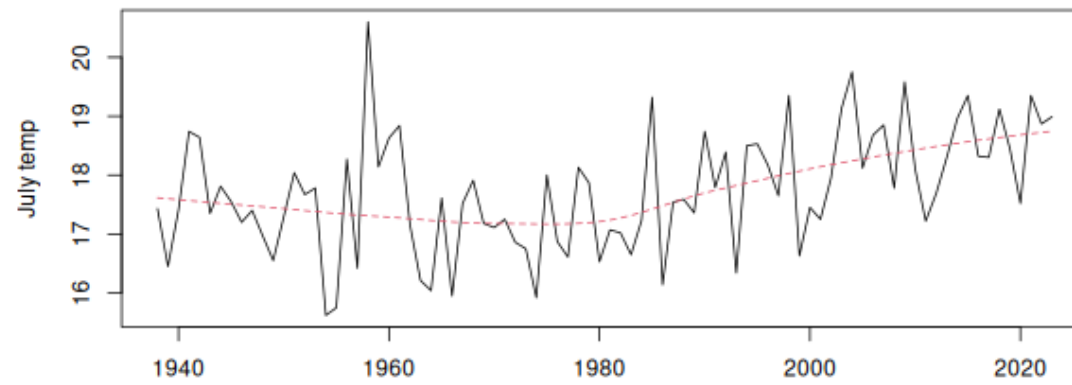




loess plot



loess plot



Time series data $y_1, \dots, y_t, \dots, y_T$; for example, monthly average temperature for month $1, \dots, t, \dots$

Time series plot (t, y_t) over t ; join consecutive points with line segments.

Autocorrelation function (acf) plot: $(k, \hat{\rho}_k)$, $k = 0, 1, 2, \dots$, where $\hat{\rho}_k$ is sample serial correlation of lag k ; it measures *strength of dependence of values that are k time units apart*.

Vancouver monthly average temperature or total precipitation: strong positive serial correlation for lags 1 and 12, strong negative serial correlation for lags 5, 6, 7. Interpret.

Decomposition: $y_t = s_t + a_t + r_t$ for seasonal, trend and irregular components. s_t relevant for periodicity in series, a_t relevant for increasing or non-monotonic trend, r_t for the remainder which has no seasonal or curvilinear pattern.

Note: H&A uses $y_t = S_t + T_t + R_t$, but lectures will reserve upper case for random variables. Also T is used for length of series.

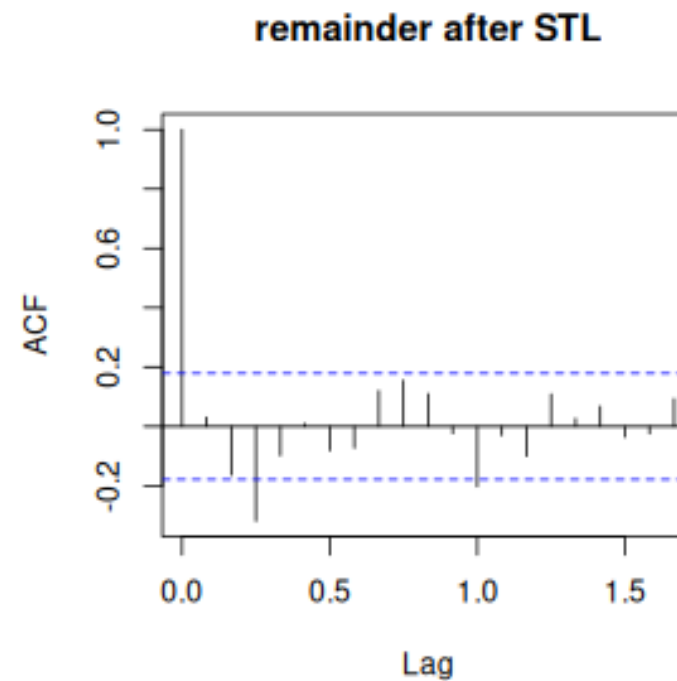
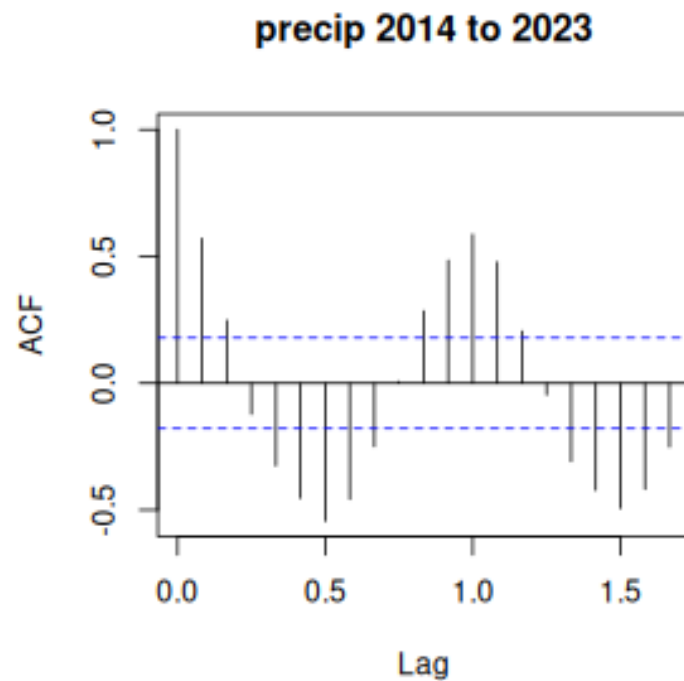
Operations in time series

- For monthly time series data, the seasonal term could be $s_t = \text{average over the common month } s_t = (\cdots + y_{t-12} + y_t + y_{t+12} + \cdots)/[T/12]$.
- For a smoothed trend, a moving average filter is $\tilde{y}_t = (y_{t-2} + y_{t-1} + y_t + y_{t+1} + y_{t+2})/5$ as an example.
- For time series that have periods of ups and downs, differencing such as $y'_t = y_t - y_{t-1}$ can be useful.

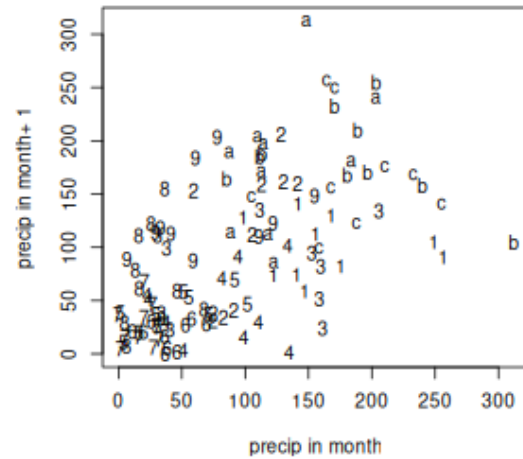
Can apply autocorrelation function (acf) to $y_t - s_t$, r_t , $y_t - \tilde{y}_t$, y'_t etc.

Examples forthcoming.

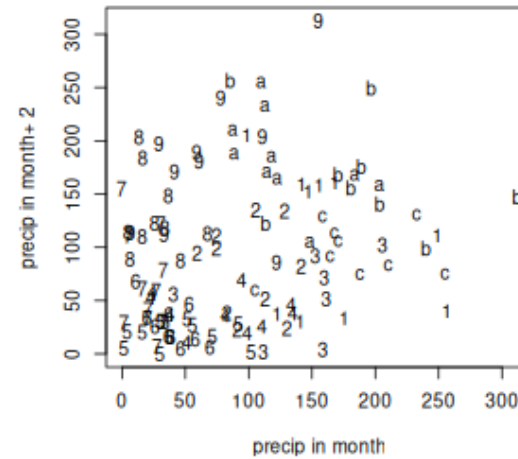
Plot of autocorrelation function (acf) for Vancouver monthly total precipitation: **interpret, compare with next slide.** Mathematical definition is given later in this set of slides.



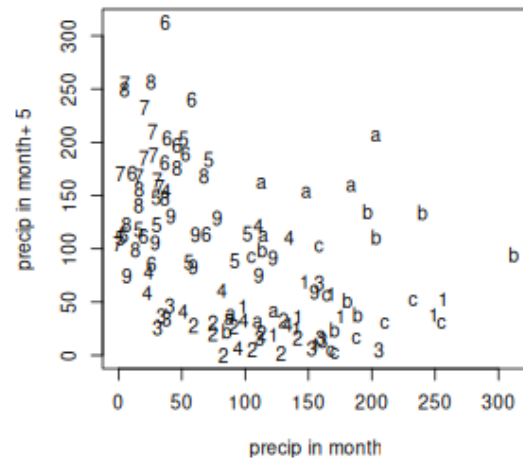
Monthly precip, lag= 1



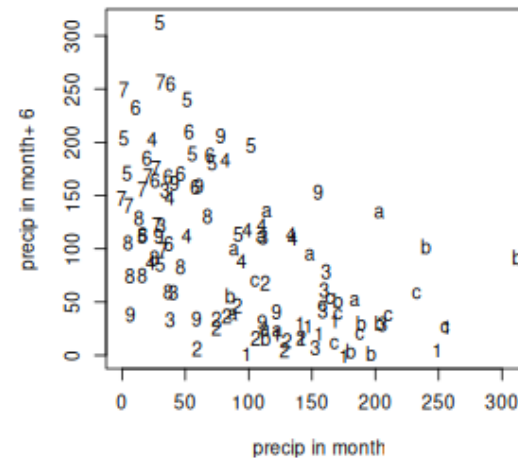
Monthly precip, lag= 2



Monthly precip, lag= 5



Monthly precip, lag= 6



Example 2. Canadian CPI growth rates:
source <https://fred.stlouisfed.org/categories/32268>

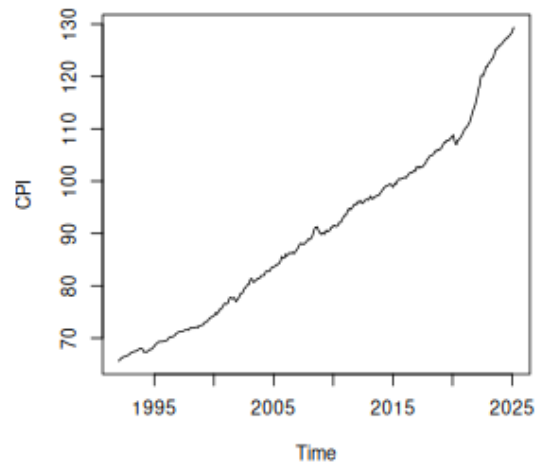
Data set at course web site: CANCPALTT01IXOBSAM.csv (CPI in 2015 dollars) seasonally adjusted.

Some macroeconomic variables, such as $z = \text{CPI} = \text{Consumer Price Index}$, tend to increase over time. Growth rate in CPI is a proxy for inflation.

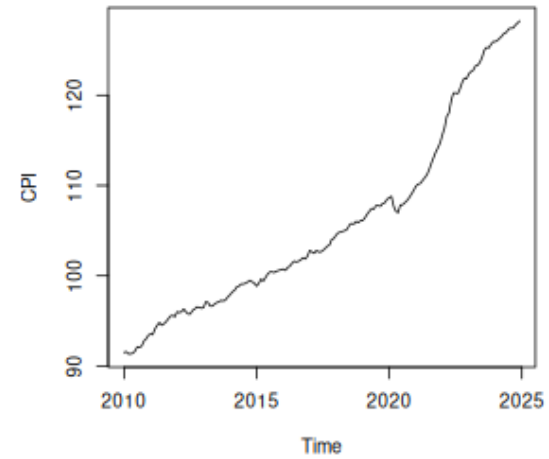
A common transformation for macroeconomic variables is $y_t = \log(z_t) - \log(z_{t-1})$. This transformed variable is sometimes called a growth rate (why this name?, what is an alternative y_t that could be called growth rate?). It can be positive or negative depending on an increase or decrease in consecutive time units; and it removes the trend.

What patterns do you notice in the plots in the next slide? Forecasting rules?

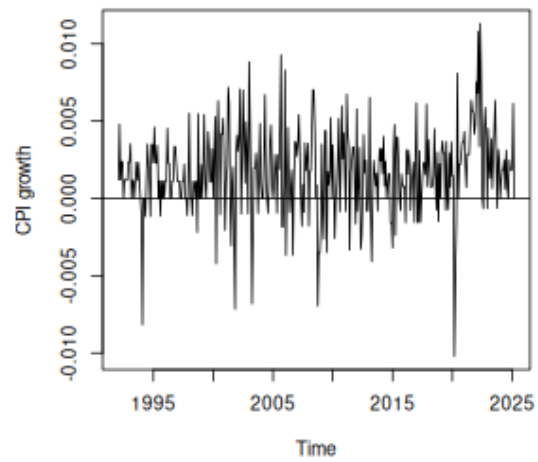
Canada CPI, 2015=100, Seasonally Adj



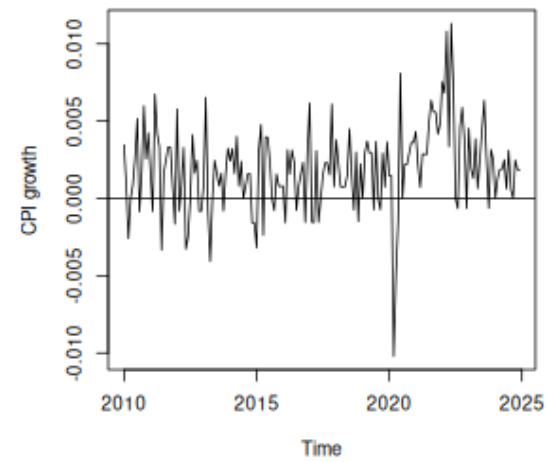
Canada CPI, 2010 to 2024



Canada CPI growth, 1992 to 2024



Canada CPI growth, 2010 to 2024

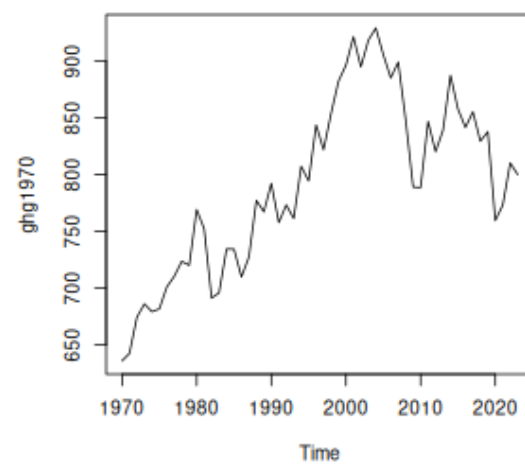
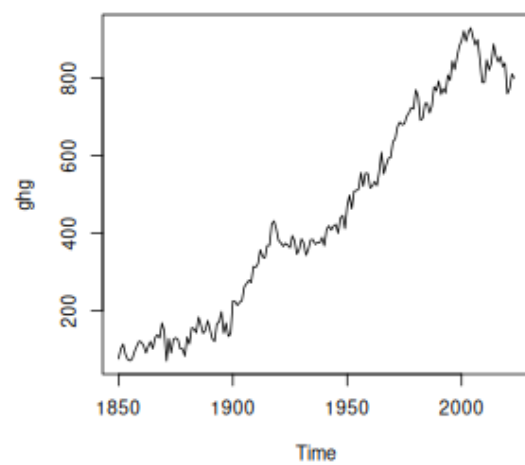


Example 3. Greenhouse gas emissions.

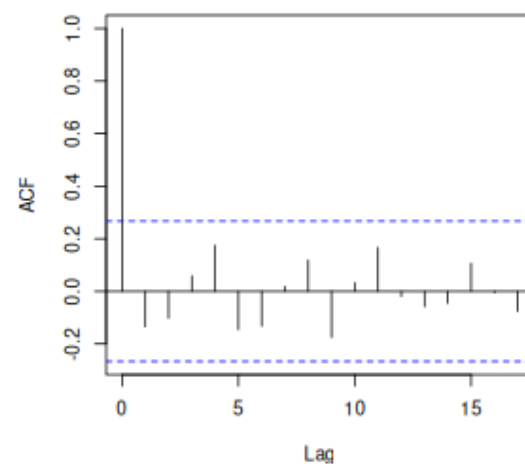
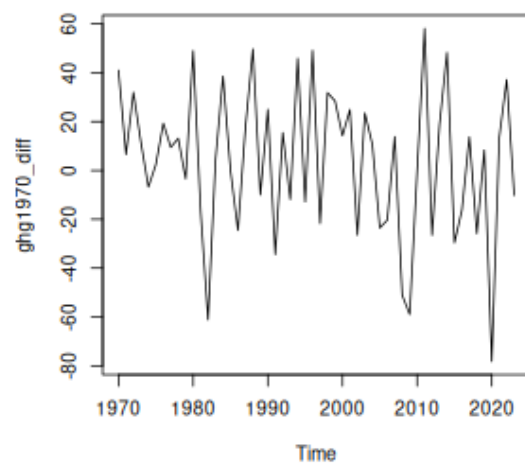
Source: <https://ourworldindata.org/co2-dataset-source>

Wrangled data set at course web site: GHGtotal_canada.csv

What patterns do you notice in the plots in the next slide? Forecasting rules?



Series ghg1970_diff

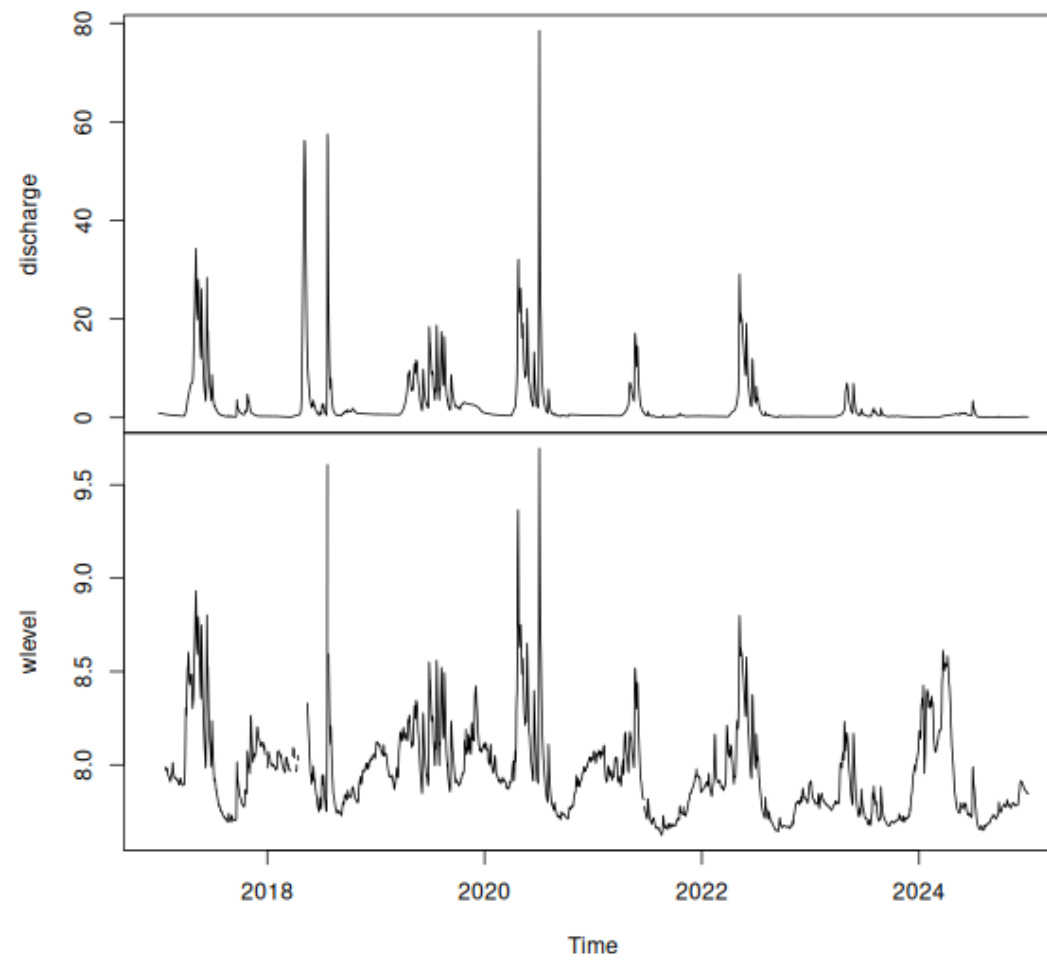


Example 4. Dawson Creek river : daily discharge (m^3/s) and water level (m),
Source: <https://wateroffice.ec.gc.ca> , search “Kiskatinaw” (name of river in Dawson Creek, northern BC, with droughts in recent years).

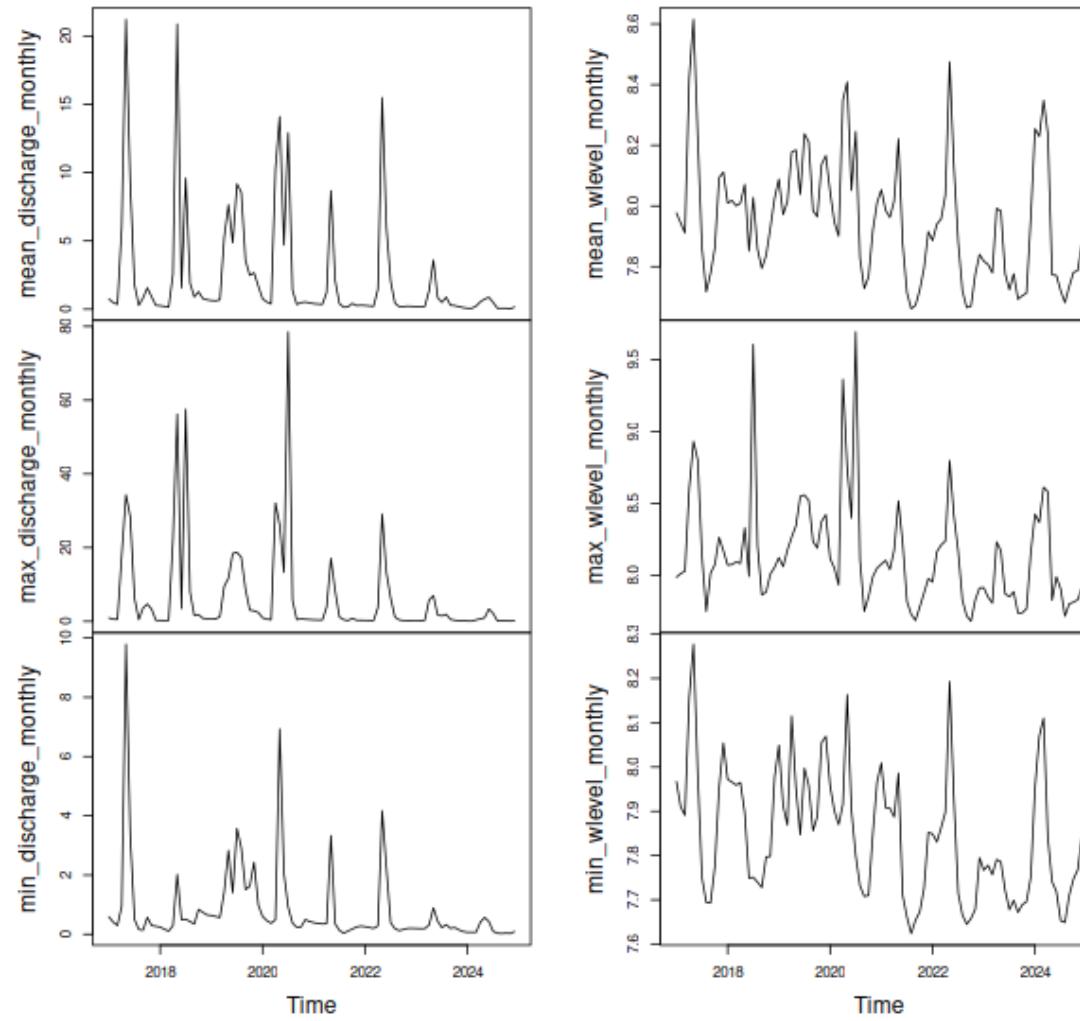
Wrangled data set at course web site: dawson_creek.csv for daily 2017-01-01 to 2024-12-31

Any trends? News reports indicate a summer drought in recent years.

dawson



DawsonCreek, KiskatinawR

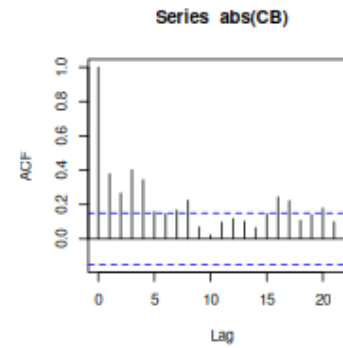
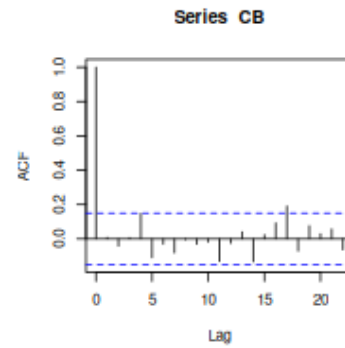
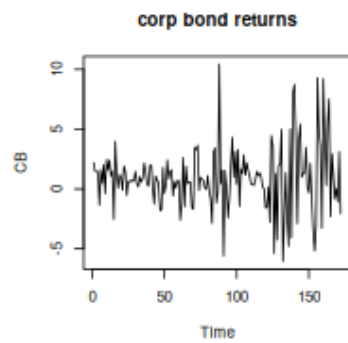
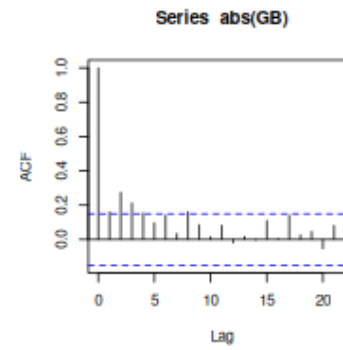
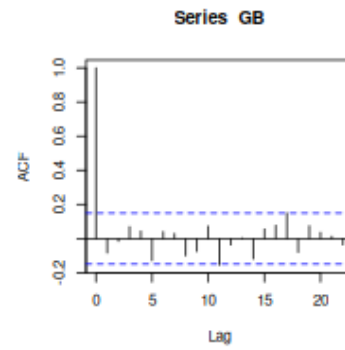
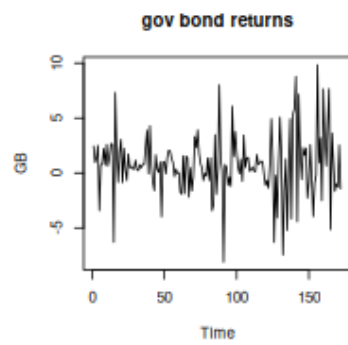
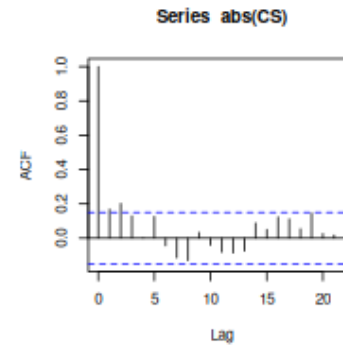
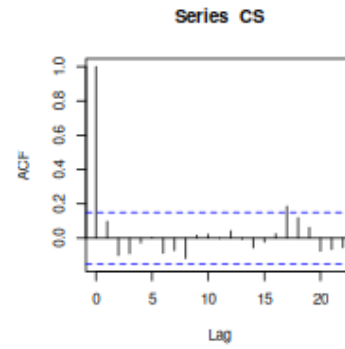
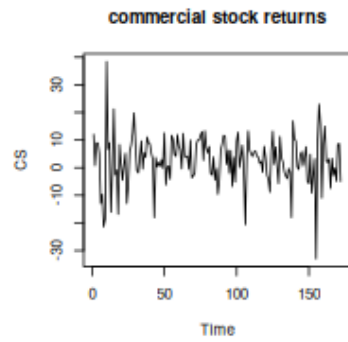


Example 5. Monthly returns on commercial stocks, government bonds, corporate bonds. A positive return in a month implies a “profit”; a negative return indicates a “loss” (if you have the investment).

Data set at course web site: CSGBCBRL36to78.csv,
Source is Grauer and Hakansson (1982). Finance Analysts Journal.

What patterns do you notice? Forecasting rules?

Interpretations of the various acf plots will come later in the term.



Notation: reference book by Hyndman and Athanasopoulos does not have good notation to distinguish random variables in stochastic models and realized values in data sets.

See documentation [sat443-notation-bigpicture.pdf](#) at course web site for details.

Stat 443: t for time index, T for length of time series. Upper case roman letters for random variables, lower case roman letters for realized data values, Greek letters for parameters (exception is ϵ for innovation random variable), overhead hat or caret for estimated quantity.

Observed data $y_1, \dots, y_t, \dots, y_T$ of a variable y at time points $1, \dots, T$. The time unit is usually one of day, week, month, quarter, year. Sometimes there are vectors of exogenous variables $\mathbf{x}_1, \dots, \mathbf{x}_T$. For example, y =monthly sales at a clothing store, \mathbf{x} =(average temperature, unemployment rate, amount of advertising, ...)

(Sample) Autocorrelation function (acf) $\hat{\rho}_k = \hat{\rho}(k) : k = 1, 2, \dots$: this has the serial correlations of lag 1, lag 2, lag 3, ..., lag k , etc.

Realized time series y_1, \dots, y_T .

Lag 1 serial correlation is roughly the sample correlation of the $T - 1$ pairs $(y_1, y_2), (y_2, y_3), \dots, (y_{T-1}, y_T)$.

Lag 2 serial correlation is roughly the sample correlation of the $T - 2$ pairs $(y_1, y_3), (y_2, y_4), \dots, (y_{T-2}, y_T)$.

Lag 3 serial correlation is roughly the sample correlation of the $T - 3$ pairs $(y_1, y_4), (y_2, y_5), \dots, (y_{T-3}, y_T)$.

Let $\bar{y} = T^{-1} \sum_{i=1}^T y_t$ be the sample mean, $s_y^2 = (T-1)^{-1} \sum_{i=1}^T (y_t - \bar{y})^2$ be the sample variance. Actual definition in textbooks and R: for $k = 1, 2, \dots$,

$$\hat{\rho}_k = \frac{(T-1)^{-1} \sum_{t=1}^{T-k} (y_t - \bar{y})(y_{t+k} - \bar{y})}{s_y^2}$$

The numerator of $\hat{\rho}_k$ is the autocovariance of lag k .

Why called autocorrelation, autocovariance, serial?

Why is autocorrelation more informative than autocovariance?

Why roughly as an adverb in previous slide?

Forecasting :

$\hat{y}_{t+1|t}$ is the forecast of Y_{t+1} after observing y_1, \dots, y_t (1-step forecast).

$\hat{y}_{t+h|t}$ is the forecast of Y_{t+h} after observing y_1, \dots, y_t (h -step forecast).

What are some reasonable forecast rules for the time series data examples presented in previous slides?

Suppose we have two rules $\hat{y}_{t+h|t}^{(1)}$ and $\hat{y}_{t+h|t}^{(2)}$.

How do we make a comparison to decide which rule is better?

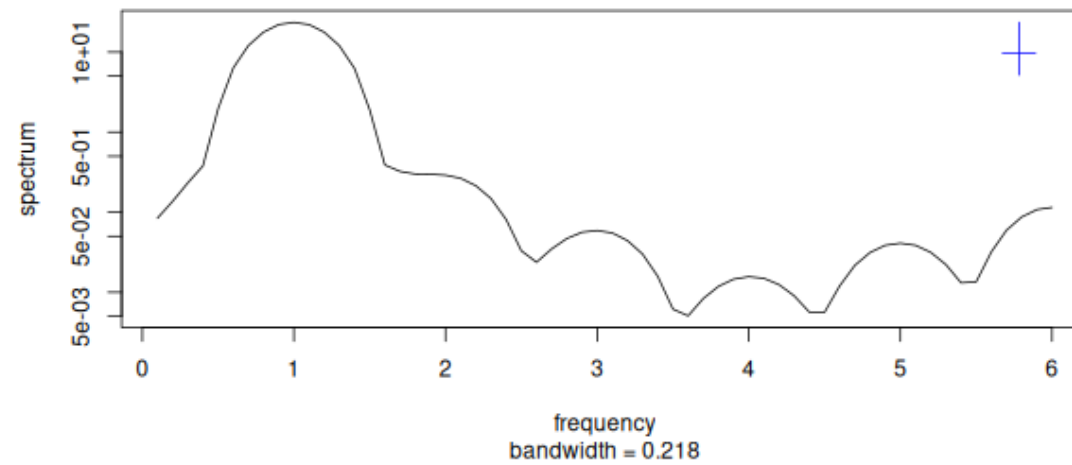
Sources of data for weather, climate change, macro-economics, finance are given in this document to provide you with possible ideas and resources for the [term project](#).

Suggestion: Choose topic for project based on your other interests. Preferably choose a topic that is scientifically meaningful.

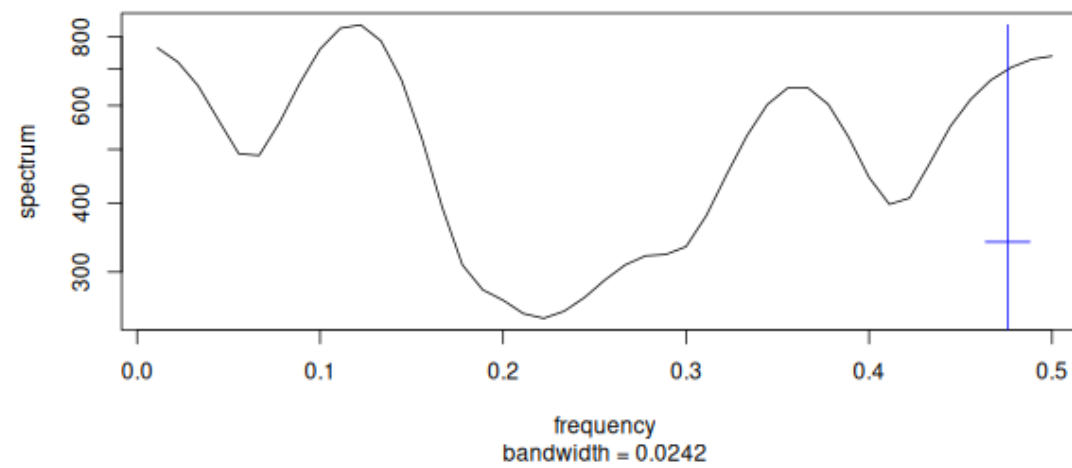
More plots that will be discussed at a later date.

Smoothed periodogram of spectral density to assess possible periodic behaviour.

temperature_seasonal_STL



July precipitation



Statistical software : R and SAS

Previously, the instructor used SAS (cloud version for University use) as well as R. This term: only R.

But neither software has built-in functions for comparing out-of-sample 1-step forecasting rules. Software in R/SAS have 1-step to 10-step forecasts at the end of the input time series data.

Different organizations (e.g., Statistics Canada) move towards open source software such as R and python for statistical analyses.

Self-study: For R, you can look into the following R functions:

- `ts` to create time series object
- `window` for subset of a time series
- `acf` for plot of autocorrelation function and values of serial correlations
- `diff` for differencing a time series
- `stl` for seasonal-trend-loess smoothing
- `loess` for trend-loess smoothing without seasonality
- `filter` for moving average (smoothing) filter (no example in the plots in this document)

See a code file at course web site.