MS4S09

Data Mining and Statistical Modelling

Introduction to Time Series Analysis

What is a Time Series?

A *stochastic process* is a collection of random variables $\{X_t, t \in T\}$.

A *time series* is a stochastic process in which T is a set of time points, usually

$$T = \{0, \pm 1, \pm 2, \ldots\}, \{1, 2, 3, \ldots\}, [0, \infty), \text{or } (-\infty, \infty)$$

Note: The term “time series” is also used to refer to the realization of such a process (observed time series).

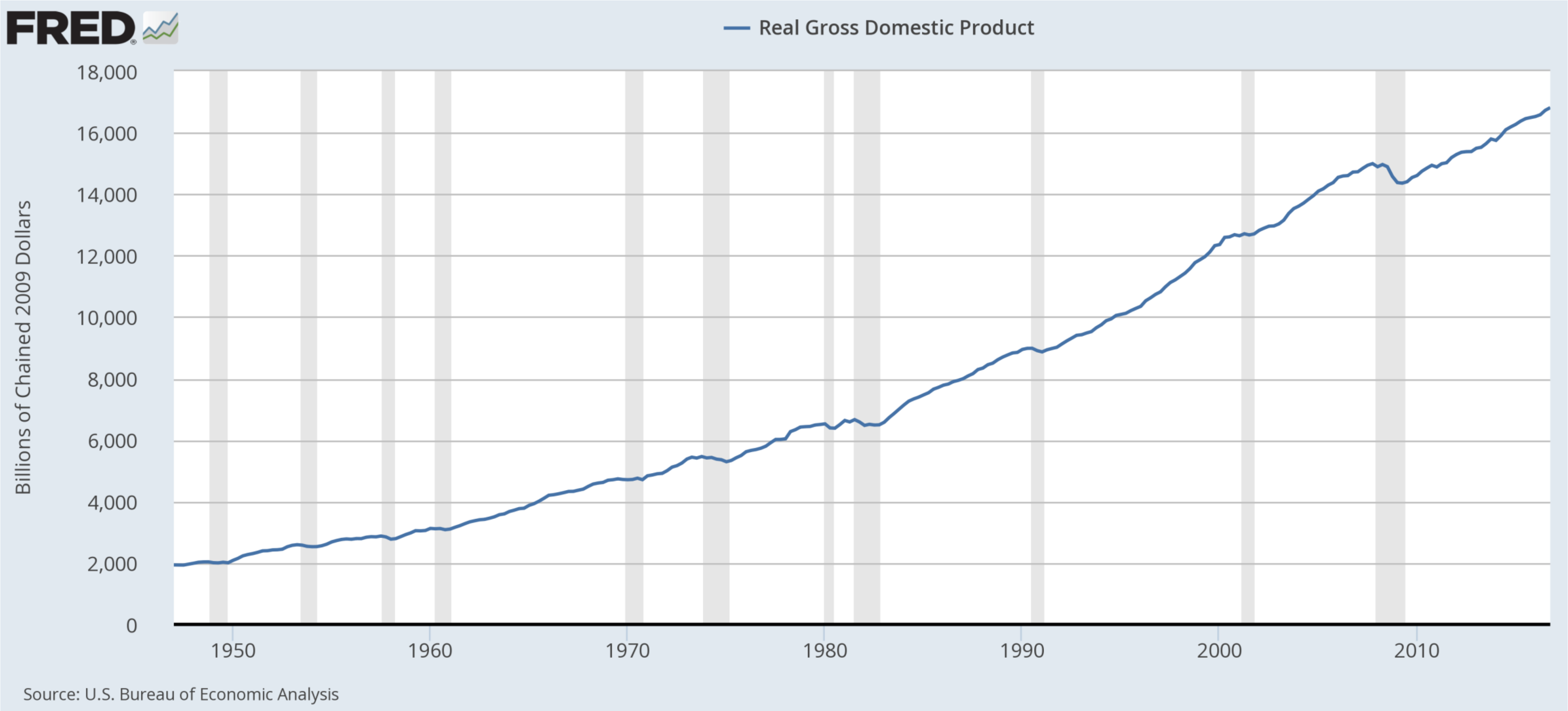
Example: Time Series

- Monthly sales of wine
- Monthly accidental deaths in Europe
- Daily Average Temperature in Cardiff
- Daily stock price of IBM stock
- UK monthly interest rates
- US Yearly GDP
- 1-minute intraday S&P500 return

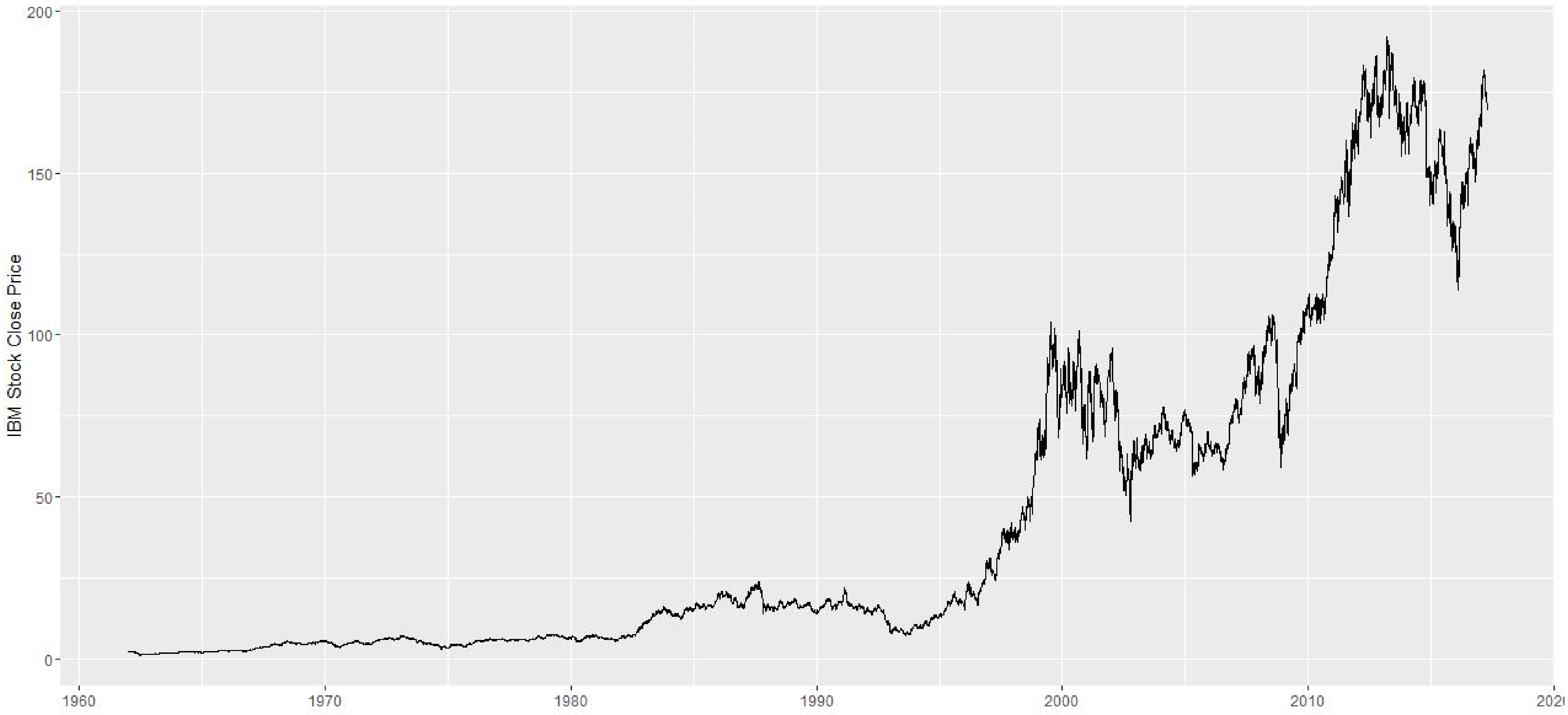
Time Series: Characteristics

- **Trend**: long-term increase or decrease in the data over time
- **Seasonality**: influenced by seasonal factors (e.g. quarter of the year, month, or day of the week)
- **Periodicity**: exact repetition in regular pattern (seasonal series often called periodic, although they do not exactly repeat themselves)
- **Cyclical trend**: data exhibit rises and falls that are not of a fixed period
- **Heteroscedasticity**: varying variance with time
- **Dependence**: positive (successive observations are similar) or negative (successive observations are dissimilar)

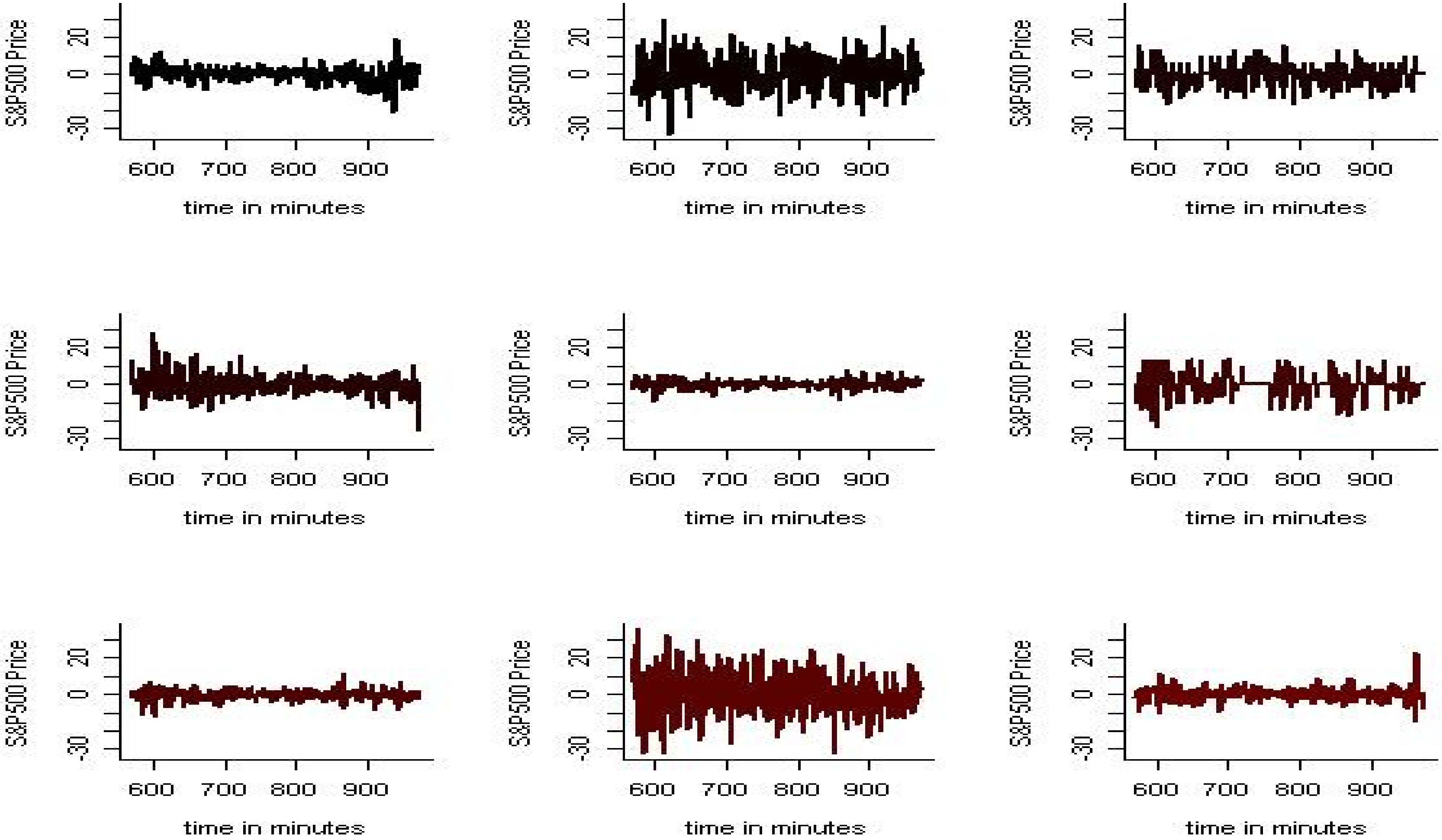
Example: GDP


FRED
Real Gross Domestic Product
Billions of Chained 2009 Dollars
18,000
16,000
14,000
12,000
10,000
8,000
6,000
4,000
2,000
0
1950
1960
1970
1980
1990
2000
2010
Source: U.S. Bureau of Economic Analysis

Example: Daily IBM Stock Price



Example: S&P500 Intraday



Is Time Series Analysis Necessary?

- **Time Series ⇒ Dependence**
- Data redundancy: number of degrees of freedom is smaller than T (T is the number of observations)
- Data sampling: $Y_t, t = 1, .., T$ concentrated about a small part of the probability space
- **Ignoring dependence leads to**
- Inefficient estimates of regression parameters
- Poor predictions
- Standard errors unrealistically small (too narrow CI ⇒ improper inferences)

Time Series: Objectives

Description

Plot the data and obtain simple descriptive measures of the main properties of the series.

Explanation

Find a model to describe the time dependence in data.

Forecasting

Given a finite sample from the series (observations), forecast the next value or the next several values.

Control/Tuning

After forecasting, adjust various control/tune parameters.

Time Series Analysis: Approaches

Time domain approach

Assume that correlation between adjacent points in time can be explained through dependence of the current value on past values.

Frequency domain approach

Characteristics of interest relate to periodic (systematic) sinusoidal variations in the data, often caused by biological, physical, or environmental phenomena.

Decomposition:
Trend Estimation

Time Series: Basics

Data: Y_t, where t indexes time, e.g. minute, hour, day, month

Model: Y_t = m_t+s_t+X_t

- m_t is a trend component;
- s_t is a seasonality component with known periodicity d ($s_t = s_{t+d}$) such that $\sum_{j=1}^{d} s_j = 0$
- X_t is a stationary component, i.e. its probability distribution does not change when shifted in time

Approach: m_t and s_t are first estimated and subtracted from Y_t to have left the stationary process X_t to be model using time series modeling approaches.

Time Series: Trend Estimation

Elimination of Trend (no Seasonality)

1. Estimate trend and remove it, or
2. Difference the data to remove the trend directly.

Estimation Methods

1. Moving Average
2. Parametric Regression (Linear, Quadratic, etc.)
3. Non-Parametric Regression

Trend: Moving Average

Estimate the trend for t with a width of the moving window *d*:

If the width is $d = 2q$, use

$$\hat{m}_t = \frac{1}{d}\left[\frac{x_{t-q}}{2} + x_{t-q+1} + x_{t-q+2} + \ldots + x_{t+q-1} + \frac{x_{t+q}}{2}\right].$$

If the width is $d = 2q + 1$, use

$$\hat{m}_t = \frac{1}{d}\sum_{j=-q}^{q} x_{t+j}$$

The width selection reflects the bias-variance trade-off:

- If width large, then the trend is smooth (i.e. low variability)
- If width small, then the trend is not smooth (i.e. low bias)

Trend: Parametric Regression

- Estimate the trend m_t assuming a polynomial in t:
$$m_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \cdots + \beta_p t^p$$
- Commonly use small order polynomial (p=1 or 2)
- Estimation approach: Fit a linear regression model where the predicting variables are $(t, t^2, \ldots, t^p)$
- Which terms to keep? Use model selection to select among the predicting variables. Cautious! Strong correlation among the predicting variables.

Trend: Non- Parametric Regression

Estimate the trend m_t with t in $\{t_1, t_2, \ldots, t_n\}$:

1. **Kernel Regression**

   $m_t = m(t) = \sum_{i=1}^{n} l_i(t) X_{t_i}$ where $l_i(t)$ a weight function depending on a kernel function.

2. **Local Polynomial Regression**

   - An extension of the kernel regression and the polynomial regression: fit a local polynomial within a width of a data point

3. **Other Approaches**

   - Splines regression
   - Wavelets
   - Orthogonal basis function decomposition

Example

Data Example:
Temperature in Atlanta, Georgia

Data: Average monthly temperature records starting in 1879 until 2016.

- Available from the iWeatherNet.com
- The Weather Bureau (now the National Weather Service) began keeping weather records for Atlanta for 138 years since October 1, 1878.
- Provided in Fahrenheit degrees

Question: Do we find an increasing trend in temperature in Atlanta?