Forecasting & Predictive Analytics

Guillaume Chevillon
chevillon@essec.edu

ESSEC Business School

October-December 2017
2nd set of slides
Time Series Decomposition

Questions

- Dynamic models
  - How can we observe the *trend and the cycle* in a variable?
  - What is the purpose of time series analysis?
  - How can we forecast?

And always:
What caveats are there to the standard answers?

DETERMINISTIC MODELS FOR TIME SERIES DECOMPOSITION AND FORECASTING

Overview

What is a Time Series?

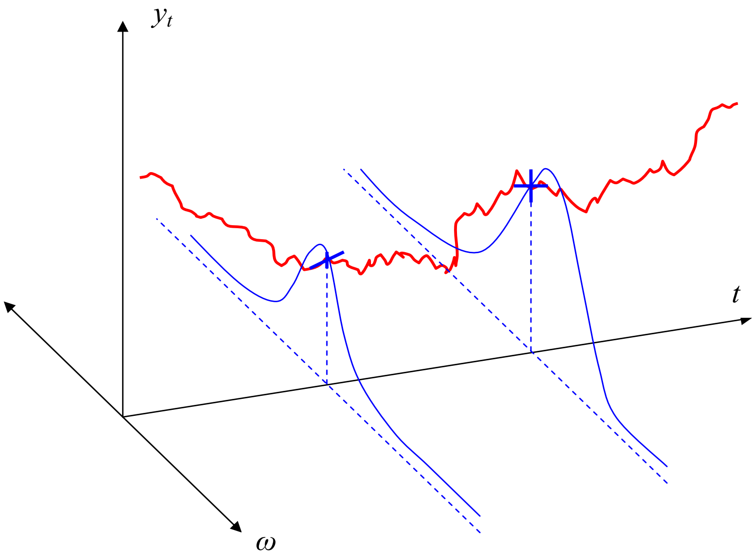
Time Series Analysis studies the *dynamics* of a variable. The latter is essential for three reasons:

1. Advances in econometrics have shown that it is only possible to relate variables that exhibit similar properties, in particular as concerns stability (or instability);
2. Mathematical properties of models that allow for estimation of the link between variables depends on their dynamics;
3. Causality or forecasting can only be understood if some pre-determination (i.e. it happens before) is allowed.

Remark

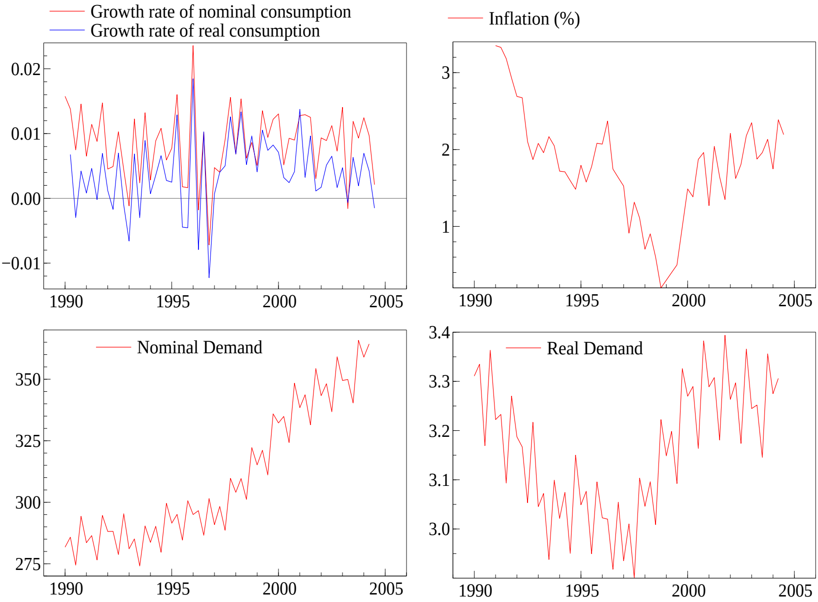
A Time Series is defined as a sequence of observations, or as a sequence of random variables . In this course, we will come to see time series as one or the other but it is important always to remember that they can be both.

A stochastic process is an ordered sequence of random variables $\{y_t(\omega), \omega \in \Omega, t \in \mathbb{T}\}$, such that for all $t \in \mathbb{T}$, y_t is a random variable on Ω and that for all $\omega \in \Omega$, $y_t(\omega)$ is a realization of the stochastic process on the indexation set $\mathbb{T}$. It is possible to study a time series according to its trajectory, as a function of t (i.e. for a given event ω, we focus on a *realization* of the stochastic process), or to its distribution at a given time t


y_t
t
ω

A time Series is hence **any succession of observations that correspond to the same variable**: it can be e.g. from

- macroeconomics (a country's GDP, inflation, exports...),
- microeconomics (a firm's sales, total number of employees, a person's income, her/his number of children...),
- financial (S&P500, DJIA, CAC40, the price of a put or call option, a stock return),
- meteorological (rainfall, number of sunny days per year),
- political (voters' turnout, number of votes received by a candidate...),
- demographic (average height in the population, age...).
- electromagnetic (transmission of sound signals, speech modeling, image analysis and transmission)
- geophysical (seismic data, oil extraction...),



French quarterly data for: (a) Nominal and real growth in consumption; (b) Inflation ; (c) Nominal demand; (d) Real demand.

What are the aims of this analysis?

- **Extract information:** dynamic models allow to analyze evolving tendencies of the data. It is for instance possible to decompose the series for GDP into a long run trend and the position of the business cycle. It is also possible to estimate unobservable components such as *volatility*.
- **Adjust for seasonal variations:** this leads to seasonally adjusted data.
- **Forecast:** this is the key reason why focus on time series data. By observing historical chronicles, we get information about some regularities that can be extrapolated into the future. It is even possible to obtain forecasts that are "robust" against some unexpected events.

How does this work in practice?

- **Aim:** observe the effects of the passage of time (trends, seasonality) to use (forecast, extract information) or correct (seasonal adjustments) some dynamic aspects.
- **Approach:** It is in practice impossible to know the distribution of a time series $\{y_t\}_{t\geq 0}$, we therefore focus on the modeling on the variable y_t **as a function of its past** through its **conditional distribution** (a priori constant through time): for a given $(y_{t-1}, y_{t-2}, ..., y_0)$, what can be said about y_t? We study the density of the variable:

$$y_t | y_{t-1}, y_{t-2}, ..., y_0.$$

conditional on the history of the process: $(y_{t-1}, y_{t-2}, ..., y_0)$.
Example: $\mathsf{E}\left[y_t | y_{t-1}\right]$? in $y_t = y_{t-1} + \varepsilon_t$

Result

The conditional approach gives an **Error-Prevision Decomposition**, according to which

$$y_t = \mathsf{E}\left[y_t | Y_{t-1}\right] + \epsilon_t,$$

where

$$\begin{cases} (i) & \widehat{y}_t = \mathsf{E}\left[y_t | Y_{t-1}\right] \text{ is the component} \\ & \text{of } y_t \text{ that can be predicted} \\ & \text{when the history of the process,} \\ & Y_{t-1} = \{y_{t-1}, y_{t-2}, y_{t-3}, ..., y_0\}\text{ , is known ; and} \\ (ii) & \epsilon_t \text{ represents informations} \\ & \text{that are not modeled:} \\ & \text{the model error.} \end{cases}$$

Some usual processes

- A **White Noise Process** is a process whose distribution is identical at all times, whose expectation is zero and that is dynamically uncorrelated:

$$u_t \sim \mathsf{WN}\left(0, \sigma^2\right).$$

  Thus, $\{u_t\}$ is White Noise if for all $t \in \mathbb{T}$: $\mathsf{E}[u_t] = 0$, $\mathsf{E}\left[u_t^2\right] = \sigma^2 < \infty$, with u_t and u_{t-h} uncorrelated if $h \neq 0$, t and $(t-h) \in \mathbb{T}$.
- If the White Noise $\{u_t\}$ is Normally distributed, it is then called a **Gaussian White Noise:**

$$u_t \sim \mathsf{NID}\left(0, \sigma^2\right).$$

- A process $\{u_t\}$ whose components are independent and identically distributed is written IID:

$$u_t \sim \mathsf{IID}\left(\mu, \sigma^2\right).$$

Examples

1. Linear Deterministic Trend:

$$y_t = \alpha + \beta t + \epsilon_t, \quad \text{where} \quad \mathsf{E}\left[\epsilon_t\right] = 0.$$

2. Quarterly seasonal model

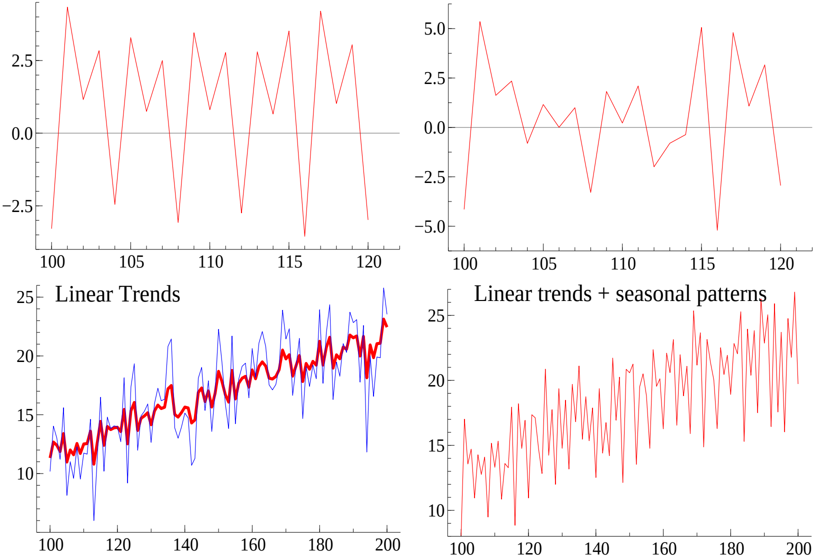
$$y_t = \alpha_1 1_{q1} + ... + \alpha_4 1_{q4} + \epsilon_t$$

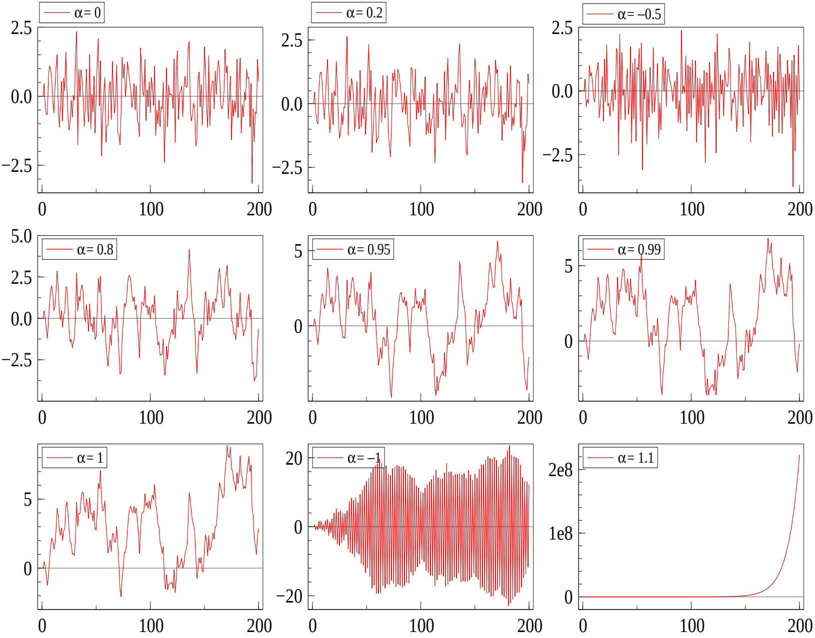
   where 1_{qj} takes value 1 in quarter j and zero elsewhere.

3. Autoregressive process of order 1, AR(1):

$$\begin{aligned} &y_t = \alpha y_{t-1} + \epsilon_t, \\ &\epsilon_t \sim \mathsf{WN}\left(0, \sigma^2\right) \text{ (white noise)} \end{aligned}$$

Seasonal Patterns




α= 0
α= 0.2
α= −0.5
α= 0.8
α= 0.95
α= 0.99
α= 1
α= −1
α= 1.1
2.5
0.0
−2.5
5.0
5
0
20
−20
2e8
1e8
100
200

Decomposition into

1. A trend $\mathcal{T}_t$: this is the element that is most persistent and determines the general orientation (upwards, downwards). The tend is *smoother*, less erratic, than the original series. This smoothness is often used to estimate the trend.
2. A seasonal component $\mathcal{S}_t$ that modifies in a regular and predictable fashion the "underlying" behavior.
3. An irregular element $\mathcal{I}_t$ with zero expectation: this element is the main focus of the course.

Example of quarterly series

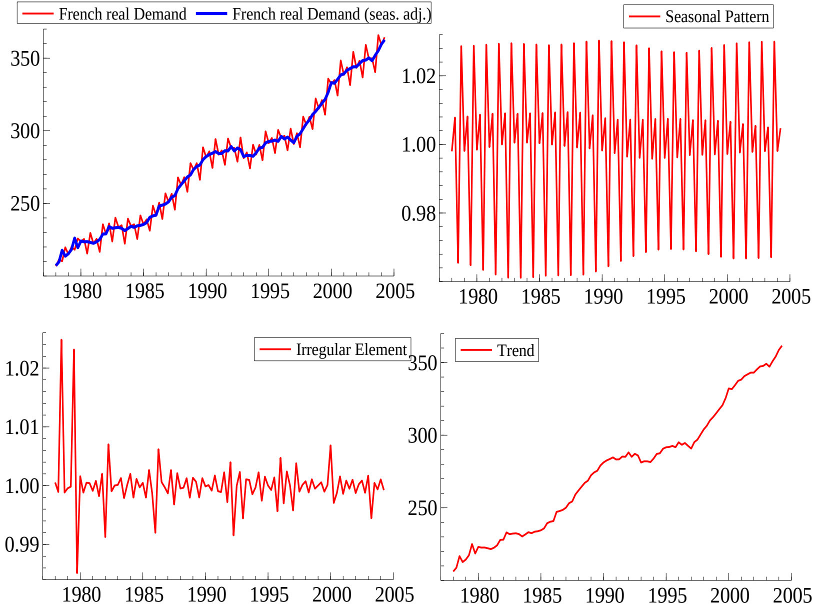
The quarterly time series generated by

$$y_t = \alpha + \beta t + \alpha_1 1_{q1} + ... + \alpha_4 1_{q4} + \epsilon_t, \quad où \quad \epsilon_t \sim \mathsf{WN}\left(0, \sigma^2\right)$$

can be decomposed as $y_t = \mathcal{T}_t + \mathcal{S}_t + \mathcal{I}_t$ with

$$\begin{aligned} \mathcal{T}_t &= \alpha + \beta t, \\ \mathcal{S}_t &= \alpha_1 1_{q1} + ... + \alpha_4 1_{q4}, \\ \mathcal{I}_t &= \epsilon_t. \end{aligned}$$

Here $\mathsf{E}[y_t] = \mathsf{E}[\mathcal{T}_t + \mathcal{S}_t + \mathcal{I}_t] = \mathcal{T}_t + \mathcal{S}_t$.
CAREFUL: here we must have $\alpha_1 + ... + \alpha_4 = 0$



Example of decomposition of French internal Demand using the X11 method.

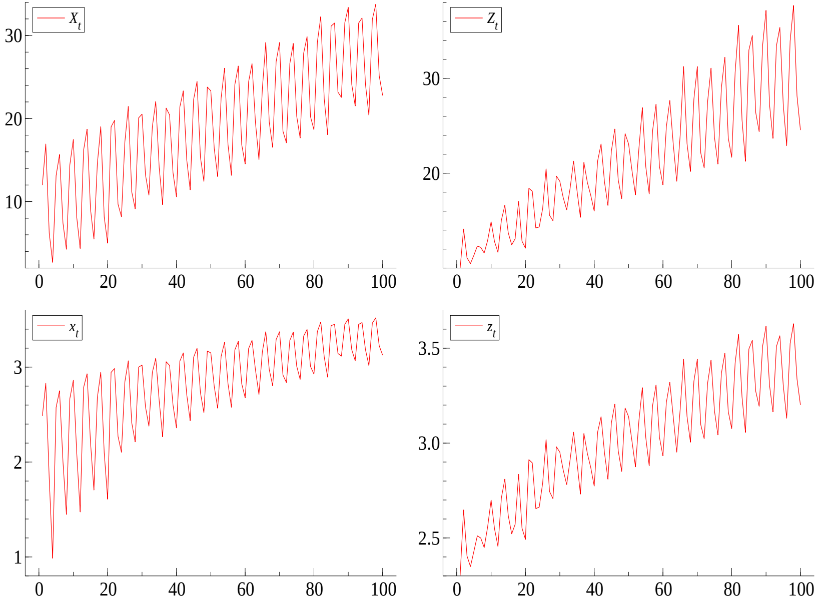
Types of adjustment

There exist two main ways to decompose time series: additively (A) or multiplicatively (M).

$$\text{A: } Y_t = \mathcal{T}_t + \mathcal{S}_t + \mathcal{I}_t$$
$$\text{M: } Y_t = \mathcal{T}_t \times \mathcal{S}_t \times \mathcal{I}_t$$

The difference between these two representations lies in that the trend and seasonal component can be expressed as units of Y_t or percent. There are numerous manners to filter the trend and seasonal variations of a series and the results also differ according to the order of extraction ($\mathcal{T}$ then $\mathcal{S}$ or reversely).

Examples of seasonality and trends



Trend Extraction, first case: no seasonality

Two main methods consist in using either a regression or a filtration.

- **Regression on a deterministic trend.** The simplest method consists in using a **polynomial function of time**:

$$\mathcal{T}_t = f(t) = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + ...\alpha_n t^n \tag{1}$$

  the (maximal) order n is chosen by visual inspection. We then regress y_t on $\left(1, t, t^2, t^3, ..., t^n\right)$. One should be careful in the choice of n : it should be too large. For any sample of T observations, there exist a polynomial of order $T-1$ that passes through all of $y_1, y_2, ..., y_{T-1}, y_T$. Unfortunately there is no reason why $f(T+1) = \widehat{y}_{T+1}$ should be any close to y_{T+1}. Hence (1) does not *describe* y_t, it only *reproduces it* and captures no salient characteristic usable in forecasting. Instead of a polynomial, any function of t can be used.

Moving Average filters

Filters are functions that produce an *output* signal (the filtered variable) from any *input*. Some can extract trends.

- Consider the centered moving average of Y_t of order p, written as $\mathsf{ma}_p^c\,(t)$, and defined, according to whether p is even or odd as:
  - odd $p = (2q+1)$:

$$\mathsf{ma}_{2q+1}^c\,(t) : W_t = \frac{1}{2q+1}\sum_{j=-q}^{q} Y_{t-j}, \quad q \in \mathbb{N}^*.$$

  - even $p = 2q$:

$$\mathsf{ma}_{2q}^c\,(t) : W_t = \frac{1}{2q}\left[\frac{1}{2}Y_{t-q} + \sum_{j=-(q-1)}^{q-1} Y_{t-j} + \frac{1}{2}Y_{t+q}\right], \quad q \in \mathbb{N}^*$$

Moving Average filters (cont'ed)

- The uncentered moving average of Y_t of order p, written as $\mathsf{ma}_p^u(t)$ is defined as

$$\mathsf{ma}_p^u(t) : Z_t = \frac{1}{p}\sum_{j=0}^{p-1} Y_{t-j}.$$

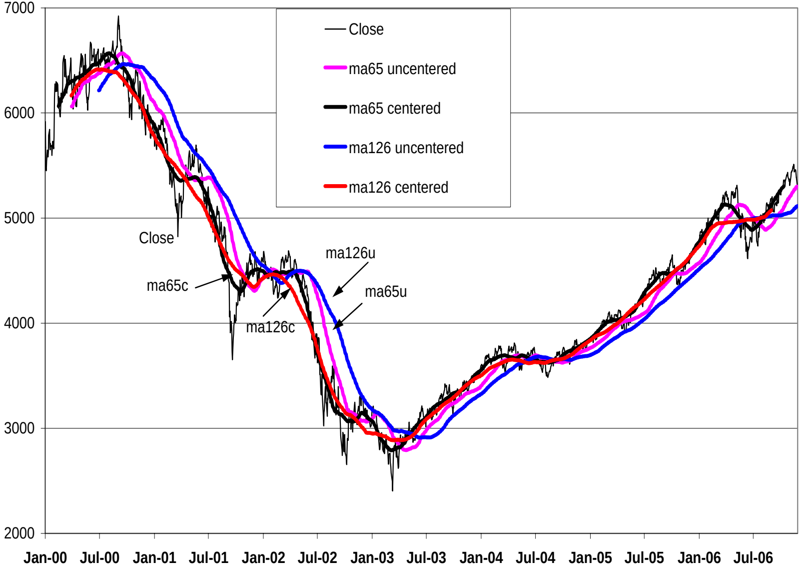
- Then , for $Y_t = \mathcal{T}_t + \mathcal{I}_t$, and p odd, ma_p^c is:

$$W_t = \frac{1}{2q+1}\sum_{j=-q}^{q} \mathcal{T}_{t-j} + \frac{1}{2q+1}\sum_{j=-q}^{q} \mathcal{I}_{t-j}$$

the centered $\frac{1}{2q+1}\sum_{j=-q}^{q} \mathcal{T}_{t-j} \approx \mathcal{T}_t$,. By the LLN $\frac{1}{2q+1}\sum_{j=-q}^{q} \mathcal{I}_{t-j} \approx 0$ and

$$W_t \approx \mathcal{T}_t$$

$q \uparrow$ implies $\frac{1}{2q+1}\sum_{j=-q}^{q} \mathcal{I}_{t-j} \to 0$, but the trend is more strongly "smoothed" and can give a poor estimator of $\mathcal{T}_t$.



7000
6000
5000
4000
3000
2000
Jan-00
Jul-00
Jan-01
Jul-01
Jan-02
Jul-02
Jan-03
Jul-03
Jan-04
Jul-04
Jan-05
Jul-05
Jan-06
Jul-06
Close
ma65 uncentered
ma65 centered
ma126 uncentered
ma126 centered
Close
ma65c
ma126c
ma126u
ma65u

Example

Centered Moving Average of order 3 on $Y_t = \mathcal{T}_t + \mathcal{I}_t$ with $\mathcal{T}_t = \alpha_0 + \alpha_1 t + \alpha_2 t^2$ and $\mathcal{I}_t = \epsilon_t$, where $\epsilon_t \sim \mathsf{WN}\left(0, \sigma^2\right)$.

Example

1. Moving average
   $W_t = \frac{1}{3}\left(Y_{t-1} + Y_t + Y_{t+1}\right) = \frac{1}{3}\sum_{j=-1}^{1} \mathcal{T}_{t-j} + \frac{1}{3}\sum_{j=-1}^{1} \mathcal{I}_{t-j}$.
2. For the trend, this becomes

$$\frac{1}{3}\sum_{j=-1}^{1} \mathcal{T}_{t-j} = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + \frac{2}{3}\alpha_2$$

3. Moving average of $\mathcal{I}_t$: $\frac{1}{3}\sum_{j=-1}^{1} \mathcal{I}_{t-j} = \frac{\epsilon_{t+1}+\epsilon_t+\epsilon_{t-1}}{3}$. The LLN yields, when $N \to \infty$, $\frac{1}{2N+1}\sum_{j=-N}^{N} \epsilon_{t-j} \to \mathsf{E}\left[\epsilon_t\right] = 0$ since all ϵ_t i.i.d.
4. $W_t = \left(\alpha_0 + \frac{2}{3}\alpha_2\right) + \alpha_1 t + \alpha_2 t^2$ is overestimated for the trend if $\alpha_2 > 0$ since $W_t - \mathcal{T}_t = \frac{2}{3}\alpha_2$.

- The moving average can also be "weighted"" e.g.

$$\overline{Y}_t = \frac{0.5Y_{t-2} + Y_{t-1} + Y_t + Y_{t+1} + 0.5Y_{t+2}}{4} \tag{2}$$

$$\overline{Z}_t = \frac{Z_{t-2} + 2Z_{t-1} + 3Z_t + 2Z_{t+1} + Z_{t+2}}{9} \tag{3}$$

Expression (2) is used for seasonal adjustment (see below)..

Exponential smoother

Forecasting using moving averages, example of the exponential smoother

$$\widehat{y}_T = \alpha \sum_{i=0}^{T} (1-\alpha)^i \, y_{T-i}.$$

which can also be obtained as

$$\widehat{y}_0 = y_0 \quad \textit{and} \quad \widehat{y}_t = (1-\alpha)\,\widehat{y}_{t-1} + \alpha y_t$$

and used for forecasting, say $\widehat{y}_t$ could be a forecast of y_{t+1}. The issue with ES is that all future forecasts $\widehat{y}_{T+k|T}$, $k \geq 0$, are equal

$$\widehat{y}_{T+k|T} = \widehat{y}_T$$

Double exponential smoothing is used when forecasting with a local linear trend.

Holt and Holt-Winters filters

- The principle of Holt-Winters filters is based on double exponential smoothing with different coefficients for the level and the trend. Holt-Winters also adjusts for seasonality.
- The smoothed value is such that the trend is centered towards the observed value using the formula:

  $$\widetilde{Y}_{t+k} = A_t + B_t k$$

  where the parameters are themselves computed recursively with smoothing coefficients (α, β) that are chosen by the user:

  $$A_t = \alpha Y_t + (1-\alpha)(A_{t-1} + B_{t-1}) \tag{4}$$
  $$B_t = \beta (A_t - A_{t-1}) + (1-\beta) B_{t-1} \tag{5}$$

  One of the interest of Holt-Winters is that we do not need to save all the data for updating and forecasting, we only need the recent values.
- Extensions: exponential or damped trends.

Other methods

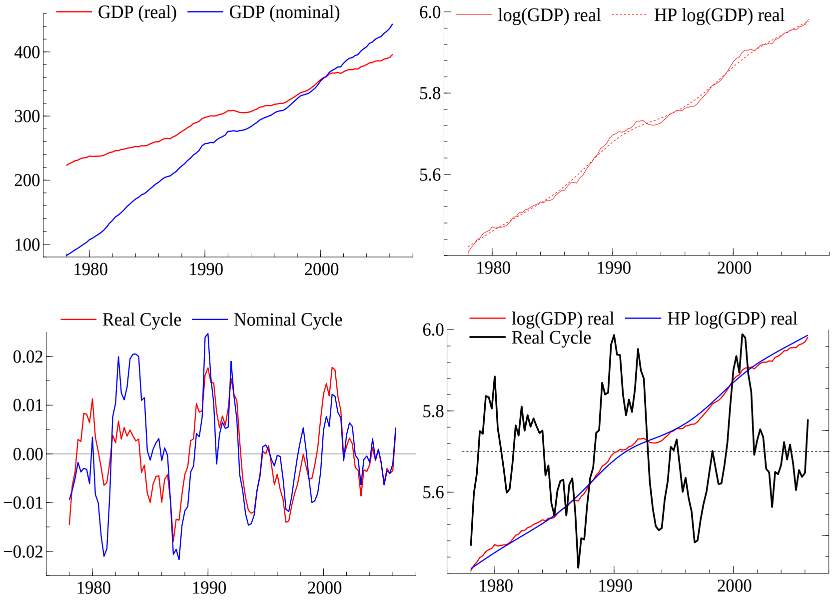
- Moving Averages can also correct for seasonality.
- Many methods to correct for seasonal patterns,
  - most common: x11 (or x12) (U.S. Census Bureau) seasonal patterns + number of working days
- Multiplicative method: very similar to additive
  - geometric means and ratios rather than arithmetic means and differences.
  - X_t is computed in the same way, then $r_t = Y_t / X_t$, seasonal indices are ω_i, adjusted as $\omega_m / \sqrt[12]{\omega_1 \omega_2 ... \omega_{12}}$ to get $\widehat{\mathcal{S}}(j)$
  - the product of the $\widehat{\mathcal{S}}(j)$ equals unity. The seasonally adjusted series is $Y_t / \mathcal{S}_t$.

Hodrick-Prescott Filter

The Hodrick-Prescott filter is commonly used in trend/cycle decompositions, as with the GDP for instance. It consists in generating an output signal ($\mathcal{T}_t$) by minimizing a criterion that weights the proximity to the input and the smoothness (via the second order derivative,
$\Delta^2 \mathcal{T}_t = \Delta \left(\mathcal{T}_t - \mathcal{T}_{t-1} \right) = \mathcal{T}_t - 2\mathcal{T}_{t-1} + \mathcal{T}_{t-2}$); $\mathcal{T}_t$ is obtained by minimizing the following criterion:

$$\mathcal{T}_t = \underset{g(\cdot)}{\arg\min} \sum_t \left\{ \left(Y_t - g\left(t\right) \right)^2 + \lambda \left(\Delta^2 g\left(t\right) \right) \right\},$$

where λ is chosen according to the periodicity of the the variable ($\lambda = 14400$ for monthly data, 1600 for quarterly series, or $\lambda = 100$ for annual observations). The difference between the input and output is called cycle $C_t = Y_t - \mathcal{T}_t$. Note that the HP filter uses all the available data: it is very sensitive to the observations on the boundaries of the sample (first and last observations).



French Nominal and Real GDP. The HP filter is applied to the logarithms to get cycles in percent (0.02 indicates a quarterly rate of growth of 2% above the trend rate). The real cycle is computed using real GDP, nominal cycle using nominal GDP. Data source: INSEE.

Forecasting

Most sensible method consists in extrapolating from past behavior, hoping that they keep on. Some methods are better than others though!
Simplest method: use decomposition model presented above and prolong its behavior assuming that “unpredictable” variables are set to their expectations.

- For instance, assuming that we dispose of a sample of T observations that follow a white noise: $\epsilon_t \sim \mathsf{WN}\left(0, \sigma^2\right)$, if we wish to obtain forecasts of the future values of ϵ_{T+h} for the forecast horizons $h \geq 1$ from a forecast origin at T, which we write $\widehat{\epsilon}_{T+h|T}$, then we use as forecast of ϵ_{T+h} its conditional expectation

$$\widehat{\epsilon}_{T+h|T} = \mathsf{E}\left[\epsilon_{T+h}|\mathcal{I}_T\right]$$

where, since ϵ_{T+h} is uncorrelated with the elements in $\mathcal{I}_T$, $\mathsf{E}\left[\epsilon_{T+h}|\mathcal{I}_T\right] = \mathsf{E}\left[\epsilon_{T+h}\right] = 0$, and a confidence interval at probability 95% around $\widehat{\epsilon}_{T+h|T}$ is $\pm 1.96\sigma$ in the case of Gaussian white noise.

- If we wish to forecast, from T, and at horizon $h \geq 1$ a variable that is linearly trending, $y_t = \alpha + \beta t + \epsilon_t$, it is natural to choose

$$\widehat{y}_{T+h|T} = \mathsf{E}\left[\alpha + \beta\left(T+h\right) + \epsilon_t\right] = \alpha + \beta\left(T+h\right)$$

and we notice that

$$\widehat{y}_{T+h|T} = \widehat{y}_{T+h-1|T} + \beta$$

Forecasts can therefore be defined by iterations from a model:

$$y_t = f\left(t, y_{t-1}, y_{t-2}, ..., y_0; \Theta\right) + \epsilon_t \tag{6}$$

where Θ represents a vector of parameters. The forecast is the product of an estimation $\widehat{\Theta}$ and a model $f(\cdot)$ as

$$\widehat{y}_{T+1|T} = f\left(T+1, y_T, y_{T-1}, ..., y_0; \widehat{\Theta}\right)$$

where by iterations, where future values are replaced by their forecasts

$$\widehat{y}_{T+h|T} = f\left(T+h, \widehat{y}_{T+h-1|T}, \widehat{y}_{T+1|T}, y_T, ..., y_0; \widehat{\Theta}\right)$$

Forecast error

Define the **forecast error** (once it becomes available, i.e. at time $T+h$) the difference

$$\widehat{e}_{T+h|T} = y_{T+h} - \widehat{y}_{T+h|T}$$

whose properties depend on whether the model is adequate, i.e. if (6) is *well specified* (e.g. if ϵ_t is effectively white noise) then

$$\begin{aligned}\widehat{e}_{T+h|T} =& f\left(T+h, y_{T+h-1}, y_{T+h-2}, ..., y_0; \Theta\right) \\ &- f\left(T+h, \widehat{y}_{T+h-1|T}, \widehat{y}_{T+1|T}, y_T, ..., y_0; \widehat{\Theta}\right) + \epsilon_t,\end{aligned}$$

ideally

$$\mathsf{E}\left[\widehat{e}_{T+h|T}\right] = 0.$$

Owing to the (e.g. additive) decomposition of time series

$$Y_t = \mathcal{T}_t + \mathcal{S}_t + \mathcal{I}_t \tag{7}$$

we can generate the forecasts

$$\begin{aligned}\widehat{Y}_{T+h|T} &= E\left[\mathcal{T}_{T+h} + \mathcal{S}_{T+h} + \mathcal{I}_{T+h}\right] \\ &= \mathcal{T}_{T+h} + \mathcal{S}_{T+h}\end{aligned}$$

The forecast error is

$$\widehat{e}_{T+h|T} = Y_t - \widehat{Y}_{T+h|T}.$$

Forecast uncertainty

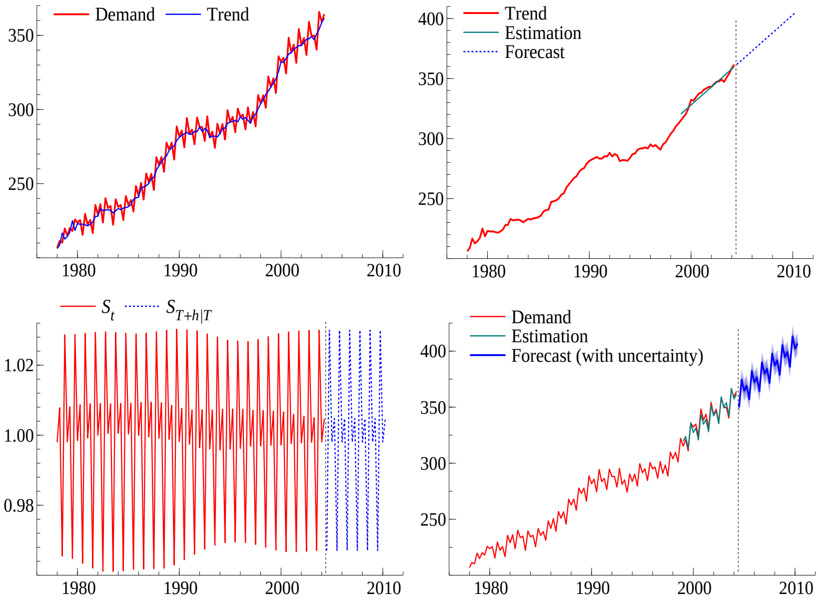
In practice, the uncertainty that surrounds the mechanism at work (the so called **Data Generating Process**) leads us to using (7) for Y_{T+h}, i.e.

$$\widehat{e}_{T+h|T} \approx \mathcal{I}_{T+h}$$

and the confidence interval at probability $(1-\alpha)$ about Y_{T+h} is

$$Y_{T+h} \underset{(1-\alpha)}{\in} \left[\widehat{Y}_{T+h|T} + t_{\alpha/2}\sqrt{\mathsf{Var}\left[\widehat{e}_{T+h|T}\right]}, \widehat{Y}_{T+h|T} + t_{1-\alpha/2}\sqrt{\mathsf{Var}\left[\widehat{e}_{T+h|T}\right]}\right]$$

where $t_{\alpha/2}$ is a quantile at the probability $\alpha/2$ from the Student distribution (T degrees of freedom) or from the Normal distribution if $T \geq 30$ and $\mathsf{Var}\left[\widehat{e}_{T+h|T}\right]$ is estimated in-sample (it is different from the variance of the residuals when $h \geq 2$). The uncertainty that surrounds the forecast is therefore related to the second moment of the forecast error (this is called **Mean-Square Forecast Error, MSFE or MSE**), and we can have an estimate thereof using the available sample.


Demand
Trend
350
300
250
1980
1990
2000
2010
Trend
Estimation
Forecast
400
350
300
250
1980
1990
2000
2010
S_t
$S_{T+h|T}$
1.02
1.00
0.98
1980
1990
2000
2010
Demand
Estimation
Forecast (with uncertainty)
400
350
300
250
1980
1990
2000
2010