Forecasting and predictive analytics

Guillaume Chevillon (chevillon@essec.edu), Pierre Jacob (jacob@essec.edu)

Fall 2021

Contents

1 Introduction and overview

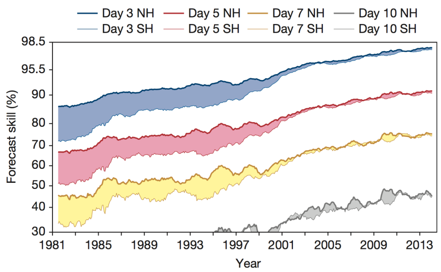
1.1 What is forecasting about?

A forecast is, according to the Cambridge dictionary, a prediction of what is likely to happen in the future. Let's start with a few examples.

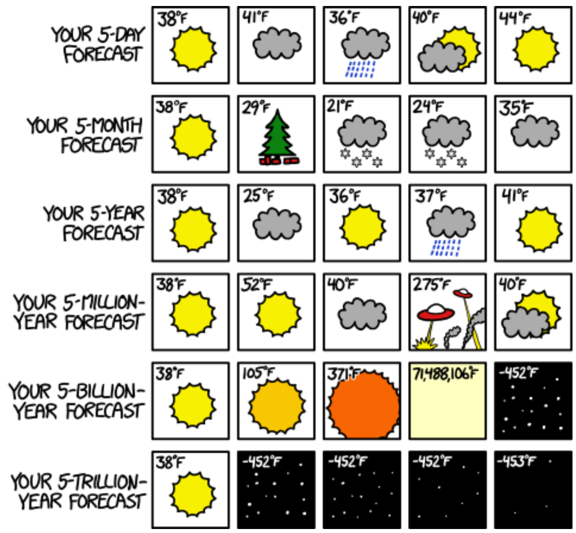
Sales Rob Hyndman and George Athanasopoulos' Forecasting book recalls these quotes.

> *I think there is a world market for maybe five computers.*
>
> (Chairman of IBM, 1943)

> *There is no reason anyone would want a computer in their home.*
>
> (President of DEC, 1977)



(a) Source: *The quiet revolution of numerical weather prediction*, P. Bauer, A. Thorpe, G. Brunet (2015)



(b) Source: xkcd, November 20, 2015

Depending on the industry, sale forecasting would be done over periods of months, quarters, or years. Sales forecast can be an essential tool in managing human resources, inventories, cash flows, and planning for future products. Sales forecast can influence a decision to invest in a marketing campaign which will itself alter future sales. Sales forecast can be done on the basis of data on past sales of comparable products, and surveys of potential clients. For instance, DSBA students participated over the last few years in a comparison of various forecasting models in the context of Direct Marketing Analytics (see Sarkar & De Bruyn, 2021)

Weather Weather forecasting is practised by... everyone but also by professionals with access to data from networks of weather stations and satellites. Figure 1(a) shows steady progress on "forecast skill" (what could this be?), for forecasting periods of 3 to 10 days, on both North and South hemispheres. It is worth wondering whether accurate weather forecasting is possible over much longer periods of time. Weather forecasting is critical to many decision-makers, for example in farming, transportation and tourism industries. However, weather forecasting does not (for now), lead to decisions that could directly affect the weather, although see "cloud seeding". Figure 1(b) highlights that weather forecasting becomes hard and then simple again as the forecasting period increases.

Elections and sport competitions Nate Silver, writer of "The Signal and the Noise: Why Most Predictions Fail - but Some Don't", is somewhat of a star among data analysts working on the prediction of election outcomes. He previously lived from playing online poker, and then from the analysis of baseball data. Sports analytics itself is now a large and expanding industry, popularized by the 2011 movie "Moneyball". Forecasting in these settings is commonly associated with gambling, particularly with the emergence of online "prediction markets", but it is also of immense value for the teams. Forecasting periods can range from the upcoming game to the entire season, or even to a player's prospective career. Data could include past performance of all teams or players, and even the complete trajectories of all players during all games of the past seasons. Going back to politics, the potential influence of forecasts onto the outcomes is often denounced. For example the mere availability of forecasts might affect voter turnout. To mitigate this effect, in France the law prevents polls to be published within two days of an election.

Stock markets Many analysts try hard to predict the returns of assets on stock markets to inform decisions to buy or sell, while institutions try to anticipate sources of risks, signs of upcoming crises. In this setting, all actors are forecasters, and their actions directly define the behavior of the values being predicted. The forecasting period involved could be very short, milliseconds in the case of high-frequency trading, or hours, days for most actors. Data that can inform forecast include historical records but also various types of additional sources of information, such as business and international news.

Disease outbreaks The covid-19 pandemic has made everyone aware of both the importance and the difficulty in predicting the spread of infectious diseases. Forecasters try to anticipate the next steps using information on the cases, on the social dynamics of infected individuals, and various other sources such as mobility data from mobile phones. We have all been affected by the decisions made in the light of such forecasts, or in some cases in the absence of precise forecasts.

Some might recall that in 2009, France was well prepared for an H1N1 pandemic: large amounts were invested in antiviral drugs and masks, that turned out to mostly unnecessary as the outbreak stopped quicker than expected. Good news! But the minister for health Roselyne Bachelot was then heavily criticized, as her decisions were viewed as leading to financial waste.

A nuanced story about forecasting and epidemiology is that of Google Flu Trends, a web service operated between 2008 and 2015. The idea was that search queries on Google could help tracking influenza (flu), for example because people might seek information on the web when they start to develop symptoms. This could enable live information on the spread of influenza, whereas national centres in charge of surveillance operate with delayed data: typically, symptomatic patients report to their doctors, who report to local authorities, who then communicate to the national level, the whole process taking days or weeks. However, after an initial performance that was promising, Google Flu Trends forecast completely missed a spectacular outbreaks of influenza in parts of the United States in the 2012-2013 season. After inspection, it appeared that Google Flu's algorithm relied overly on seasonal searches entirely unrelated to flu, such as "high school basketball". These searches were temporally associated with the flu over a few years and thus had some "predictive power"; see Figure 1. However, when flu came slightly "off season", these searches lost all their predictive power. See "What We Can Learn From the Epic Failure of Google Flu Trends", by David Lazer and Ryan Kennedy. Despite the failure of that first attempt, web searches have the potential to help in flu surveillance, as argued in "Accurate estimation of influenza epidemics using Google search data via ARGO", by Shihao Yang, Mauricio Santillana and Sam Kou, 2015.

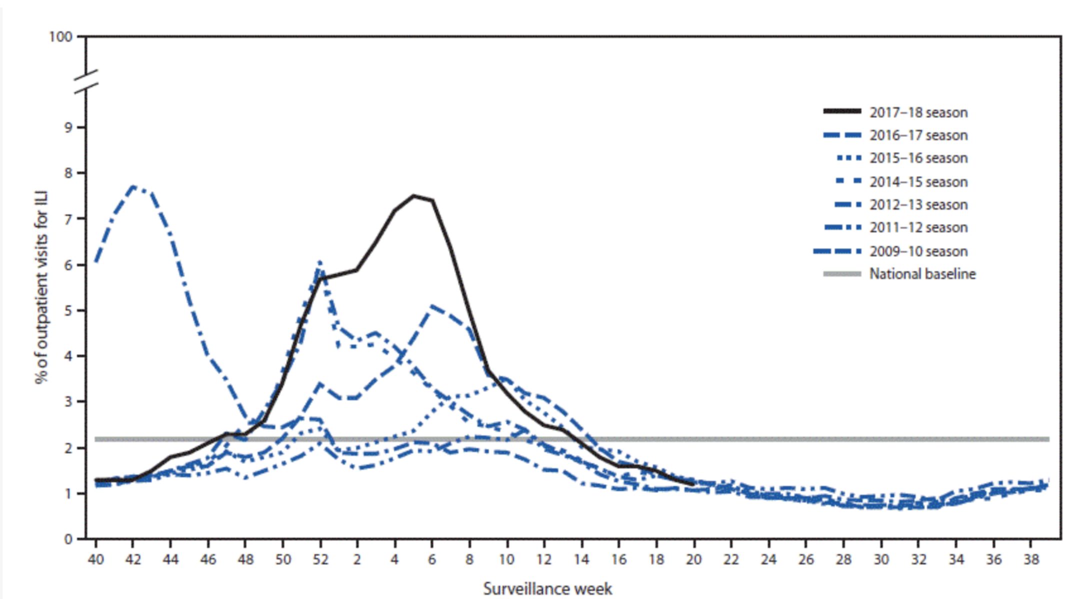


Figure 1: Week-by-week percentage of outpatient visits for influenza-like illnesses, displayed by season. Source: Centers for Disease Control and Prevention, https://www.cdc.gov/mmwr/volumes/67/wr/mm6722a4.htm.

Natural disasters Other natural disasters are the subjects of active surveillance and forecasts. In countries such as Japan or Chile, forecasting earthquakes is essential. Warning the population, even minutes in advance, can save many lives. In fact, a few seconds are enough to activate emergency breaks on fast trains and therefore to avoid derailment. So far no major earthquake has been predicted more than minutes in advance. There are debates among seismologists on the feasibility of the endeavor; see "Can We Predict Earthquakes At All?" by Sabrina Stierwalt in the Scientific

American, 2020.

In Summer 2021 dramatic mud floods have killed hundreds in Belgium and Germany, following days of heavy rain. In a Medium article “Floods of the century - Tale of a disaster foretold”, Laurent Eschenauer describes how the European Flood Awareness System provided accurate predictions, days before the floods, thanks to a sophisticated, data-heavy scientific machinery. They predicted a disaster of historic proportions, occurring with large probability, about four days in advance. Unfortunately, their alerts failed to be heard and processed quickly, and ultimately did not lead to safe evacuation plans for the population. A forecast is only useful if it is taken into account by decision-makers in a timely manner.

Can we forecast anything? In the book “Foundations”, by Isaac Asimov, published in 1951 but set in a distant future, a fictional “psychohistorian” analyzes thousands of years of historical data to predict events. He finds out, with great dismay but tragic certainty, that their current civilization will collapse in about 500 years, to be followed by a very long dark age. Although the collapse is unavoidable, actions can still be taken to reduce the length of the ensuing chaos. Outside of fiction few researchers aim at long-range socio-economic predictions, although some attempts are being made, for example under the name of “cliodynamics”.

The difficulty of the task of forecasting can come from different sources: ignorance about the dynamics at play, lack of information and data, or lack of predictability of the phenomenon of interest. Indeed for all practical purposes, many phenomena appear to be inherently random, in the sense that they can’t be fully determined in advance. We might be able to learn the odds (and that would really be in the best-case scenario) but must be ready to encounter situations where these odds leave many different possibilities as plausible. Expectations should be set accordingly. That being said, at least coin tosses can be perfectly predicted, see Figure 2.

1.2 Performance, uncertainty and interpretability

Forecasting is about predicting the future using available data and knowledge. The output of a forecasting exercise can take different forms. It could be a single number, considered to be the “most likely to occur”. This would be called a *point prediction* as it yields a single point. One would naturally assess the accuracy of such forecast through the absolute difference between the predicted point and the actual value, using “test data”. Clearly, higher accuracy is a plus.

Forecasting can also provide some notion of uncertainty associated with the predicted point. This can take the form of a “standard deviation” reported with the predicted value. We will also encounter prediction intervals. Such intervals aim at covering the future value at a nominal “coverage”. For a fixed coverage, one can measure the accuracy of intervals by their widths, which

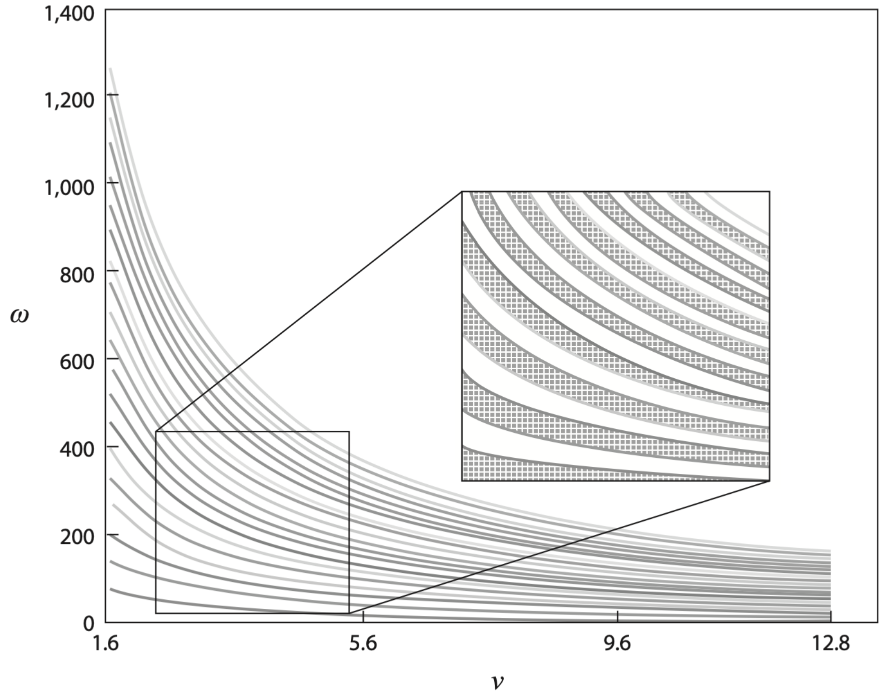


Figure 1.5. The hyperbolas separating heads from tails in part of phase space. Initial conditions leading to heads are hatched, tails are left white, and ω is measured in s^{-1}.

Figure 2: Source: Brian Skyrms & Persi Diaconis, "Ten Great Ideas about Chance", 2017. This corresponds to an experiment where a coin was sent upward at vertical speed ν and rotational speed ω in a controlled experiment where the outcome is deterministic and only function of the initial conditions (ν, ω).

we want small.

Beyond prediction intervals, one can aim for an entire probability distribution for the future value of interest. Using a predictive distribution, various questions can be considered: for example, what is the probability that the future value will exceed a certain level? This couldn't be obtained from a prediction interval. Predictive distributions can be represented using overlaid "ribbons" that report various quantiles of the distribution of the quantity of interest at future times. Figure 3 shows such "fan chart". It represents a sequence of marginal distributions at subsequent times in the future, and not a true "joint" distribution of trajectories. Again, accuracy of predictive distributions is of critical importance for the forecaster, and the course will cover the concept of "scoring rules" that is used to measure it. The following quote from Alan Greenspan, former Chairman of the Federal Reserve Board, advocates for probabilistic forecasting in macroeconomics:

Given our inevitably incomplete knowledge about key structural aspects of an ever-changing economy and the sometimes asymmetric costs or benefits of particular outcomes, a central bank needs to consider not only the most likely future path for the economy but also the distribution of possible outcomes about that path. The decision makers then need to reach a judgment about the probabilities, costs, and benefits of the various possible outcomes under alternative choices for policy.

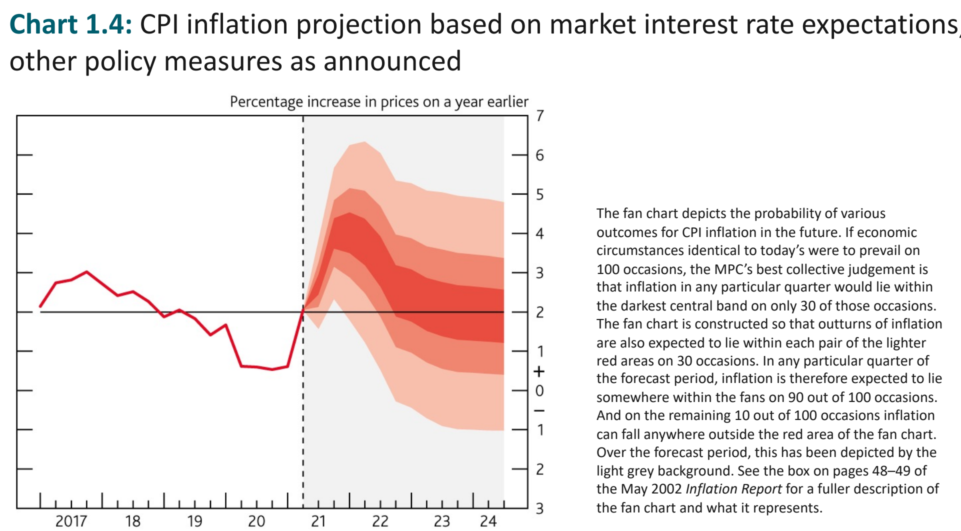


Figure 3: "Fan chart" inflation forecasts from the August 2021 Monetary Policy Bulletin of the Bank of England. See also "fanplot: An R Package for Visualising Sequential Distributions" by Guy J. Abel.

Beyond accuracy and a faithful representation of the uncertainty associated with a forecast, another criterion to consider is transparency. Decision-makers would pay attention to a forecast mostly if they can understand where it came from, what drove the prediction. Although it certainly depends on the setting, predictions that cannot be explained and justified will rarely be very useful for high-stake decisions. There are various good reasons to trade a bit of pure predictive performance for some interpretability. Readings on this include Todd Tomalak's "Communicating Forecasts to the C-Suite: A Six-Step Survival Guide", as well as Cynthia Rudin's "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead".

1.3 Time series data

We next describe the type of data that will be used to form predictions.

Time series are data measured over different points in time. The terms *observations*, *data*, *samples* and *measurements* will be used interchangeably. Strictly speaking, all data are collected over different points in time. Time series refer specifically to observations for which the time element is considered important.

Time element Consider some textbook examples of statistical analysis: testing if a coin is fair from n coin flips; estimating the difference between the means of some quantity measured in two groups; classifying emails as spam or not. In all these cases, there is certainly a time associated with each observation. However these might be considered irrelevant for the statistical question of interest, thus these examples can be handled without resorting to time series analysis.

In numerous situations the time element is indeed crucial, and thus the data are treated as time series. Figure 4 shows some examples. Consider the top-left one: Johnson & Johnson quarterly earnings per share between 1960 and 1981. Based on the plot, it is clear that the data show a specific structure of variations over time. It would sound absurd to ignore the times associated with these measurements if we wanted to predict future values. The top-right plot shows global mean land-ocean temperature deviations from the 1951-1980 average. The bottom-left shows Google search volumes for the term "chocolate" (from Google Trends). The bottom-right shows double total transect counts of red kangaroos (i.e. counts performed twice in a specific region), performed at irregular time intervals ranging from two to six months from 1973 to 1984, in New South Wales; see Caughley, G., N. Shepherd, and J. Short (1987), *Kangaroos, their ecology and management in the sheep rangelands of Australia.*

Beyond univariate time series The counts of red kangaroos form a bivariate time series. In general we can observe time series of any dimensions, and the observations can be structured along other features. For example, "panel data" refer to observations that are made for a large number of individuals over some (usually short) time span. Spatio-temporal data refer to observations that are associated with a time and a location. Likewise we might be interested in predicting more than one series into the future.

Raw and transformed data As in all parts of data analysis, the quality of the data is of critical importance for forecasting, and putting the available data in the most convenient format is somewhat of an art, sometimes called "feature engineering". Practitioners typically apply various transformations and adjustments to the raw series in order to find a representation that is meaningful and would lead to reliable forecast. Common transformations include logarithms, averaging through a moving window, differencing, adjusting prices for inflation, adjusting population-level aggregates by population sizes, converting units, standardizing, etc.

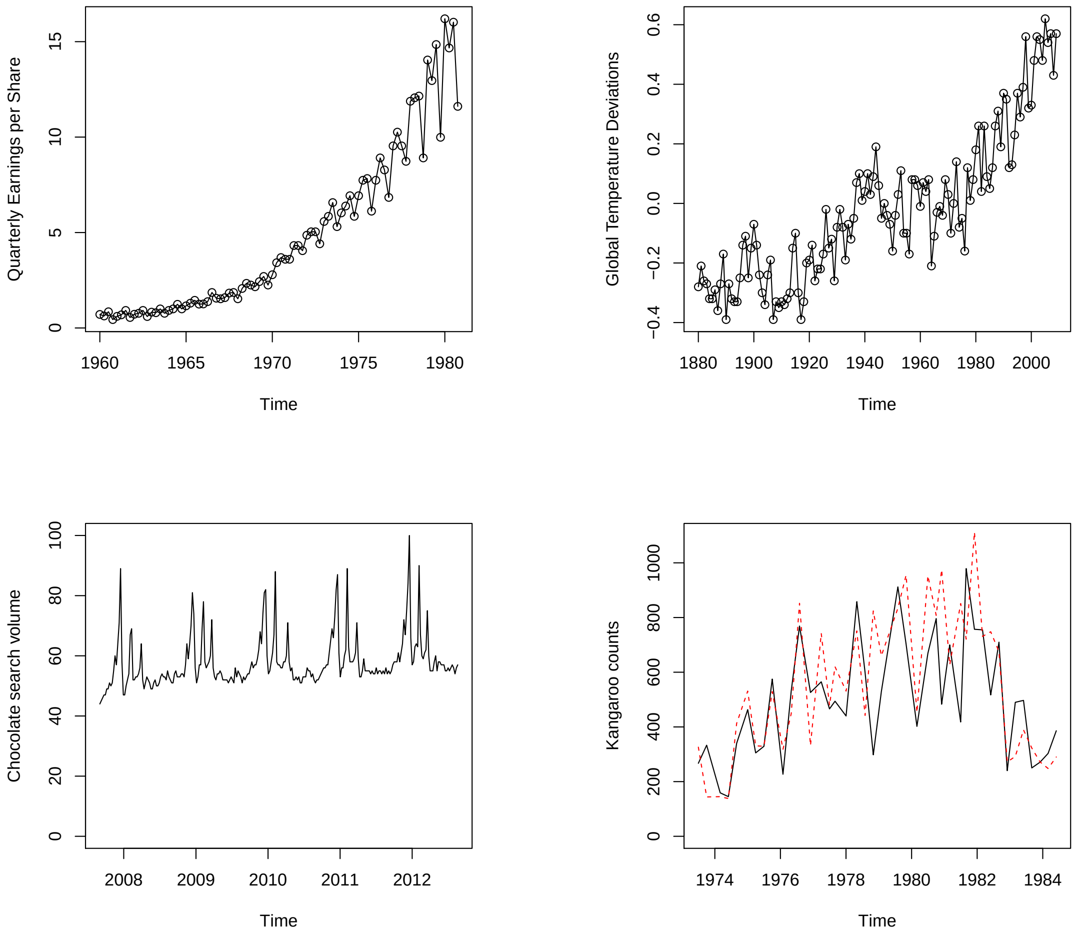


Figure 4: Top-left: Johnson & Johnson quarterly earnings per share. Top-right: global mean land-ocean temperature. Bottom-left: Google search volumes for the term “chocolate”. Bottom-right: double counts of red kangaroos in New South Wales (black line and dashed red line for both counts).

Features of time series The examples shown in Figure 4 present specific features. The top row exhibits upward *trends*. Without further knowledge on the phenomena under study, we would naturally predict future values by extending these trends. The top-left plot exhibits variations around the trend that amplify over time; on the other hand, in the top-right plot the magnitude of the variations seems stable over time. The bottom-left plot shows a slight upward trend, but more strikingly, a *seasonal* behavior: for each year we see the sames spikes around the same dates (perhaps Chritmas, Valentine's day and Easter?). The bottom-right plot does not show very clear features: there are no overall trend, and no obvious *seasonalities*. However, we see an initial upward trend followed by a final downward trend, which suggests the relevance of time series analysis for this series.

On top of situations where the time element is important, the concepts developed in time series analysis are routinely applied to other sequences of measurements. For instance, a biological sequence made of nucleotides (e.g. ATTCGGAA...) can be treated as a time series, where "time" would in fact refer to the index of each element in the sequence. In texts, letters and words constitute sequences of elements in which the order clearly matters. In the analysis of ice cores, the measurements made at different depths of the core can be analyzed as time series; there, depths and times are in fact related, since deeper measurements corresponds to older ice.

The plots of Figure 4 show some time series $(y_1, \ldots, y_T)$, also denoted $y_{1:T}$, as a function of the time index, for instance $t = 1, \ldots, T$ or $t = 1960, \ldots, 1980$. These are called *traceplots*. Plotting the data should be among the first things you do when you set out on a forecasting adventure.

1.4 Statistical specificity of time series

The analysis of time series can be considered a sub-field of data analysis. Compared to usual settings such as linear regression, the distinguishing feature of time series lies in dependencies between observations. That is, successive measurements y_t and y_{t+1} are expected to be related to one another, more so than measurements that are distant in time. Thus, we cannot rely on the usual intuition that measurements constitute a representative sample of exchangeable individuals from an underlying, fixed population. More technically a typical assumption in regression is that of *independent and identically distributed* (i.i.d.) observations. This assumption allows the use of probability results such as the *law of large numbers*, or the *central limit theorem*, which are crucial in the justification of statistical methods such as maximum likelihood estimators.

Instead, a time series is a *single* trajectory of dependent measurements over time. From a statistical point of view, this is very challenging: in a sense, we have only *one observation* of the series. The vast literature on time series provides resolutions of this challenge, for example by finding some appropriate transformations of the data such that the resulting sample can be considered *stationary*. Although *stationary* is not the same as *i.i.d.*, it will allow us to think of an underlying, fixed distribution that we can learn about, and predict new realizations. This perspective will also help understanding how some powerful ideas such as the bootstrap and cross-validation can be applied to the setting of time series.

1.5 Models and stochastic processes

Forecasting relies largely on *models*. A model is an attempt at finding a plausible description for the data, from which we can hopefully extrapolate in order to predict future values. A model tries to encapsulate important features observed in the data, such as trends, seasonalities and the dependence structure between successive observations, while remaining as simple as possible. Simplicity in modelling is often key to avoid overfitting, computational issues, and to help with interpretability.

The first step in statistical modeling is conceptual: we view a time series $y_1, \ldots, y_T$, which up to now was only a collection of numbers, as a realization of a collection of random variables $Y_1, \ldots, Y_T$. We will hereafter denote random variables with uppercase letters, and realizations with lowercase letters. A collection of random variables is called a *stochastic process*. We will understand the time t that indexes the stochastic process (Y_t) to be an integer, either in $\mathbb{N}$ or $\mathbb{Z}$. We thus work in "discrete time", as opposed to continuous time.

As a simple collection of random variables, let us consider $(W_t)_{t=1}^T$, a collection of i.i.d. Normal variables, with mean 0 and variance σ^2. The variables $(W_t)_{t=1}^T$ constitute a stochastic process, called the *Gaussian white noise process*. To make things more interesting, we can add a deterministic trend to the noise. A first model for real-valued time series with a trend would be

$$\forall t \in \{1, \ldots, T\} \quad Y_t = \beta_1 + \beta_2 t + W_t. \tag{1.1}$$

Effectively, this model is a linear regression of Y_t on the covariate X_t defined as the time index: $X_t = t$. The model parameters are $(\beta_1, \beta_2, \sigma^2)$, respectively characterizing the intercept, the slope of the trend and the variance of the noise. We expect to be able to estimate these parameters with some accuracy, for example by maximizing the likelihood given enough measurements, and by plugging these parameter estimates we can come up with forecasts for y_{T+1}. The uncertainty about future values would come from the error associated with the estimation of $(\beta_1, \beta_2, \sigma^2)$, as well as the randomness of the noise terms (W_t), with variance equal to σ^2 for all t.

Another classic model for series with trends is the random walk model with drift, where

$$Y_1 = W_1 \quad \text{and} \quad \forall t \in \{2, \ldots, T\} \quad Y_t = \delta + Y_{t-1} + W_t. \tag{1.2}$$

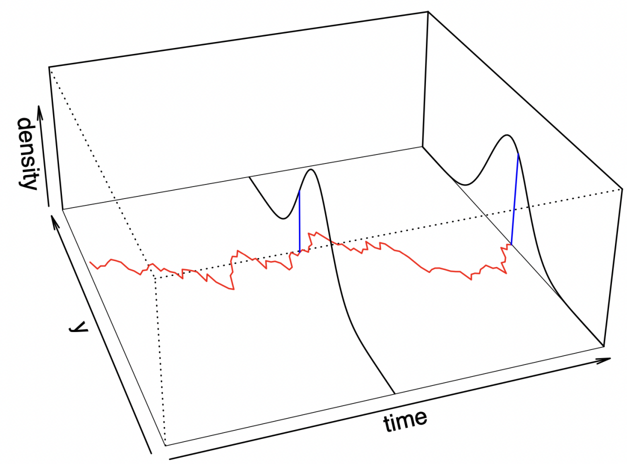


Figure 5: Trajectory of a stochastic process, along with curves representing two marginal distributions at different times.

.

The parameter δ is called the drift. We can compute $\mathbb{E}[Y_t] = \mathbb{E}[(t-1)\delta + \sum_{s=1}^{t} W_s] = (t-1)\delta$ by the linearity of expectations, and we can also compute the variance $\mathbb{V}[Y_t] = t\sigma^2$. Therefore the random walk with drift has a linear trend in expectation, with slope δ, and a variance that increases linearly with t, contrarily to the model in (1.1) where Y_t has a constant variance. Thus this model might be more adequate than the previous one for series that exhibit noise with increasing magnitude.

Finally, we introduce the "autoregressive model" (AR) which is defined by the recursion

$$Y_t = \delta + \rho Y_{t-1} + W_t. \tag{1.3}$$

Here ρ is a real value; we recover the random walk model if $\rho = 1$. However the properties of the process are much different if $|\rho| < 1$, as we will see in the next chapter. Figure 5 represents the trajectory of an AR, along with two of its marginal distributions.

Forecasting as will be seen in this course involves the introduction of models, estimation of their parameters using data, and forward extrapolation through the model over the forecasting period. The uncertainty associated with the forecast comes both from the estimation of parameters using available data, and the variability inherent to the model, represented by the noise terms.

1.6 Aims of the course

Through this course you will acquire a solid understanding of the principles of forecasting as understood in econometrics and statistics. You will be familiar with a range of methods, old and new, that are routinely used by businesses and institutions to make predictions about variables of interest, and take principled decisions, accounting for uncertainty and using available data and knowledge. You will learn how to come up with forecasts and how to assess and criticize the forecasts of others.

Perhaps more importantly, having understood the core principles of forecasting you will have the capacity to learn or even develop the tools that are best suited to your activities, and to be critical of forecasting tools.

On the conceptual level you will discover a rapidly evolving intellectual landscape, with roots at the intersection of econometrics and statistics, and connections to virtually every field of natural and social science.

This course is intended for master students with some familiarity with econometrics and statistical inference (basic probability, linear regression, bias-variance trade-offs, parameter estimation and likelihood-based inference, basic algorithms/machine learning, programming in R or python), but no previous experience with time series or stochastic processes.

1.7 References and reading suggestions

- "Economic Forecasting" by Graham Elliott & Allan Timmermann, a rich and a very readable resource on forecasting from an econometrics point of view.
- "Time Series Models for Business and Economic Forecasting", by P.H. Franses, D. van Dijk and A. Opschoor.
- "Forecasting: Principles and Practice", Rob J. Hyndman and George Athanasopoulos, with third edition available online here: `https://otexts.com/fpp3/`. A very practical and comprehensive resource written by some of the most renowned forecasters.
- Time Series Analysis and Its Applications (with R examples), by Robert H. Shumway and David S. Stoffer. It is an excellent entry point to time series analysis that does not assume much prerequisites. The companion R package `astsa` can be used to reproduce the figures of the book.
- The same authors have made available a free version, if you have difficulties finding the above book.
- Basic knowledge of probability and statistics are given e.g. in the first chapters of All of Statistics, by Larry Wasserman.

- Standard books on time series analysis include: Brockwell and Davis (1991), Tsay's Analysis of Financial Time Series (2010), Lütkepohl's New Introduction to Multiple Time Series Analysis (2005). Eric Zivot and Jiahui Wang also have a free book online with code, which covers nonlinear time series. All of these books are very relevant for the course, but again, by no means are they necessary.