FORECASTING AND PREDICTIVE MODELING

LECTURE 1

Lect. PhD. Oneț-Marian Zsuzsanna

Babeș - Bolyai University
Computer Science and Mathematics Faculty

2024 - 2025

Overview

- Course Organization
- What is forecasting?
- Example of a forecasting flow

Course Organization I

- Activities
  - Lecture: 2 hours / week
  - Seminar: 1 hour / week

Course Organization II

- Communication & Information
  - email: zsuzsanna.onet@ubbcluj.ro
  - Microsoft Teams chat
  - Personal webpage: www.cs.ubbcluj.ro/$\sim$ marianzsu

Course Organization III

- MS Teams of the course:
  - For students from **Data science for Industry and Society**:
    - Name: *2024-25 Forecasting and predictive modeling*
    - Code to join: **8axmoqe**
  - For students from **Cyber security**
    - Name: *2024-25 Business forecasting and predictive modeling*
    - Code to join: **ice6bbu**
- Lecture notes and other important documents will be uploaded here
- Announcements will be made here

Course Organization IV

- Grading and passing criteria
  - Will be discussed during the first seminar (and a document with the details will be uploaded to MS Teams).

Bibliography

- **R.J. Hyndman, G. Athanasopoulos Forecasting: Principles and Practice, OTexts, 3rd edition, 2018 (https://otexts.com/fpp3/)**
- P.J. Brockwell, R.A. Davis, *Introduction to Time Series and Forecasting*, Springer Verlag, 2nd edition, 2002.
- D.C. Montgomery, C.L. Jennings, M. Kulahci, *Introduction to Time Series Analysis and Forecasting*, Wiley, 2nd edition, 2015.
- M. Huber, D. Modlin, C. Wells. *Forecasting Using Model Studio in SAS Viya*, 2020
- Hadley Wickham, Mine Cetinkaya-Rundel, Garrett Grolemund, *R for Data Science*, O'Reilly, 2nd edition (https://r4ds.hadley.nz/)
- Other resources

Course objectives

- **General objective:** Introduce students in forecasting and predictive modelling

- **Specific objectives:**
  - present the field of forecasting and predictive modelling
  - familiarize students with the basic notions and concepts related to time series data and their analysis
  - offer the students the skills to develop different data analysis approaches for time series data
  - offer the students the skills to perform data analysis and forecasting in R

What is forecasting? I

- According to Wikipedia, *"Forecasting is the process of making predictions based on past and present data."*
- *"It is difficult to make predictions, especially about the future." - Neils Bohr (Danish physicist)*
- Prediction was one of the tasks that was targeted by supervised machine learning methods. What such methods are you familiar with?

What is forecasting? II

- In forecasting, in most cases, we work with *time series data*: it means that our instances (called in general observations), occur in successive order and were measured over some period of time.
- Our goal is to forecast (predict) the value of the same variable in the future (one or more time steps in advance). The number of future periods for which the forecast is made is called the *forecast horizon*.
- **Obs:** In this course we will consider *forecasting* and *predicting* to be synonyms.

Forecast horizon

- Forecasting problems are often classified in three categories, based on the length of the forecast horizon:
  - short-term (predict events only a few time periods - days, weeks, months - into the future)
  - medium-term (predict events 1-2 years in the future)
  - long-term (predict events many years in the future)
- Since in general historical data exhibits some inertia (meaning that the close future will likely be similar to the past), short- and middle-term forecasts are typically based on identifying, modeling and extrapolating patterns from historical data.
- This is why, we will focus on short-term and middle-term forecasting in this course.

Types of forecasting techniques

- *Qualitative forecasting*
  - used in situations where there is no historical data, for example when releasing a new product (ex. in 2012 Australia was the first country to pass legislation that required all cigarette packets to be dark green in colour)
  - in general it involves experts whose opinion is used for forecasting, thus it is subjective
- *Quantitative forecasting*
  - when there is historical data and we have reason to believe that some aspects of past patterns will repeat in the future, so we use that data and a forecasting model.
- In this course, we will focus on quantitative forecasting models.

What can be forecast? I

- A few examples of domains where forecasting is used:
  - *Business Operations management*: forecast product sales, demand for services, etc. in order to schedule production, determine staffing requirements, manage the supply chain, etc.
  - *Marketing*: forecast of sales as a function of advertising expenditure, forecast of sales after price changes, etc.
  - *Finance and risk management*: forecasting of stocks, currency exchange rates, returns of an investment, etc.
  - *Economy:* forecast of GDP, population growth, production, consumption, etc.
  - *Demography:* forecast of births, deaths, migration, etc. as a basis for different policies.

What can be forecast? II

- Some specific examples of possible forecasting tasks:

  1. daily electricity demand in 3 days time
  2. time of sunrise this day next year
  3. Google stock price tomorrow
  4. Google stock price in 6 months
  5. maximum temperature tomorrow
  6. exchange rate of RON/USD next week
  7. total sales of drugs in Romania next month
  8. timing of next Halley's comet appearance

- How would you rank the above tasks, based on difficulty? Which tasks seems the easiest and which seems to be the most difficult? Why?

What can be forecast? III

- Some events can be forecast easily and accurately (ex. sunrise) others not at all (lotto numbers). Predictability of an event depends on:
  - how much data do we have
  - how well we understand the factors that contribute to it
  - how similar is the future to the past
  - how far in the future are we forecasting
  - whether the forecasts can affect the thing we are forecasting
- ex. short-term forecast of residential electricity demand can be highly accurate
- ex. forecasting currency exchange rates is hard, we only have data
- We do not need unchanging environment for forecasting. Every environment changes. We need to assume that *the way in which the environment changes will be the same*

Types of time series

- In most cases, the observations are numerical, although we can have *categorical time series* as well (for example: the data set on sleep of newborns)
- In most cases the observations are measured at equal time intervals (daily, weekly, monthly, quarterly, yearly, etc.), but there are time series with irregular time periods as well.
- Depending on how many variables we have, we can have:
  - univariate time series (only one time series variable)
  - bivariate time series (two connected times series variables)
  - multivariate time series (more than two connected time series variables)
- **During this course we will focus on univariate and multivariate time series with numerical data and measured at equal time intervals.**

Time series data

- When we have a time series, we need (at least) two information:
  - The time information (the year, quarter, month, etc.)
  - The data itself
- For example (US employment data)

	Month	Employed
	<mth>	<dbl>
1	1939 Jan	25338
2	1939 Feb	25447
3	1939 Mar	25833
4	1939 Apr	25801
5	1939 May	26113
6	1939 Jun	26485
7	1939 Jul	26481
8	1939 Aug	26848
9	1939 Sep	27468
10	1939 Oct	27830

Examples

- Data with measurements about penguins

	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
1	Adelie	Torgersen	39.1	18.7	181	3750	male	2007
2	Adelie	Torgersen	39.5	17.4	186	3800	female	2007
3	Adelie	Torgersen	40.3	18.0	195	3250	female	2007
4	Adelie	Torgersen	*NA*	*NA*	*NA*	*NA*	*NA*	2007
5	Adelie	Torgersen	36.7	19.3	193	3450	female	2007
6	Adelie	Torgersen	39.3	20.6	190	3650	male	2007
7	Adelie	Torgersen	38.9	17.8	181	3625	female	2007
8	Adelie	Torgersen	39.2	19.6	195	4675	male	2007
9	Adelie	Torgersen	34.1	18.1	193	3475	*NA*	2007
10	Adelie	Torgersen	42.0	20.2	190	4250	*NA*	2007
11	Adelie	Torgersen	37.8	17.1	186	3300	*NA*	2007

Examples

- Data with measurements about penguins

	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
1	Adelie	Torgersen	39.1	18.7	181	3750	male	2007
2	Adelie	Torgersen	39.5	17.4	186	3800	female	2007
3	Adelie	Torgersen	40.3	18.0	195	3250	female	2007
4	Adelie	Torgersen	*NA*	*NA*	*NA*	*NA*	*NA*	2007
5	Adelie	Torgersen	36.7	19.3	193	3450	female	2007
6	Adelie	Torgersen	39.3	20.6	190	3650	male	2007
7	Adelie	Torgersen	38.9	17.8	181	3625	female	2007
8	Adelie	Torgersen	39.2	19.6	195	4675	male	2007
9	Adelie	Torgersen	34.1	18.1	193	3475	*NA*	2007
10	Adelie	Torgersen	42.0	20.2	190	4250	*NA*	2007
11	Adelie	Torgersen	37.8	17.1	186	3300	*NA*	2007

- Even if it has an attribute for year, this is n ot a time series.

Examples

- Quarterly data with holiday overnight trips (in thousands) in Australia

	Quarter	Trips
1	1998 Q1	11.806038
2	1998 Q2	9.275662
3	1998 Q3	8.642489
4	1998 Q4	9.299524
5	1999 Q1	11.172027
6	1999 Q2	9.607613
7	1999 Q3	8.913887
8	1999 Q4	9.025688
9	2000 Q1	11.070866
10	2000 Q2	9.196262
11	2000 Q3	9.347506
12	2000 Q4	8.984014

Examples

- Quarterly data with holiday overnight trips (in thousands) in Australia

	Quarter	Trips
1	1998 Q1	11.806038
2	1998 Q2	9.275662
3	1998 Q3	8.642489
4	1998 Q4	9.299524
5	1999 Q1	11.172027
6	1999 Q2	9.607613
7	1999 Q3	8.913887
8	1999 Q4	9.025688
9	2000 Q1	11.070866
10	2000 Q2	9.196262
11	2000 Q3	9.347506
12	2000 Q4	8.984014

- Univariate time series with numerical data.

Examples

- Yearly data with Boston Marathon winners

	Event	Year	Champion	Country	Time
1	Men's open division	1897	John J. McDermott	United States	10510 secs
2	Men's open division	1898	Ronald J. MacDonald	Canada	9720 secs
3	Men's open division	1899	Lawrence Brignolia	United States	10478 secs
4	Men's open division	1900	John P. Caffery	Canada	9584 secs
5	Men's open division	1901	John P. Caffery	Canada	8963 secs
6	Men's open division	1902	Sammy A. Mellor	United States	9792 secs
7	Men's open division	1903	John C. Lorden	United States	9689 secs
8	Men's open division	1904	Michael Spring	United States	9484 secs
9	Men's open division	1905	Frederick Lorz	United States	9505 secs
10	Men's open division	1906	Timothy Ford	United States	9945 secs
11	Men's open division	1907	Thomas Longboat	Canada	8664 secs

Examples

- Yearly data with Boston Marathon winners

	Event	Year	Champion	Country	Time
1	Men's open division	1897	John J. McDermott	United States	10510 secs
2	Men's open division	1898	Ronald J. MacDonald	Canada	9720 secs
3	Men's open division	1899	Lawrence Brignolia	United States	10478 secs
4	Men's open division	1900	John P. Caffery	Canada	9584 secs
5	Men's open division	1901	John P. Caffery	Canada	8963 secs
6	Men's open division	1902	Sammy A. Mellor	United States	9792 secs
7	Men's open division	1903	John C. Lorden	United States	9689 secs
8	Men's open division	1904	Michael Spring	United States	9484 secs
9	Men's open division	1905	Frederick Lorz	United States	9505 secs
10	Men's open division	1906	Timothy Ford	United States	9945 secs
11	Men's open division	1907	Thomas Longboat	Canada	8664 secs

- Univariate time series with numerical data (the Time column), even if we have other attributes as well.

Examples

- Data about winning time of the 400m running at the Olympics.

	Year	men	women
1	1968	43.80	52.00
2	1972	44.66	51.08
3	1976	44.26	49.29
4	1980	44.60	48.88
5	1984	44.27	48.83
6	1988	43.87	48.65
7	1992	43.50	48.83
8	1996	43.49	48.25
9	2000	43.84	49.11
10	2004	44.00	49.41
11	2008	43.75	49.62
12	2012	43.94	49.55

Examples

- Data about winning time of the 400m running at the Olympics.

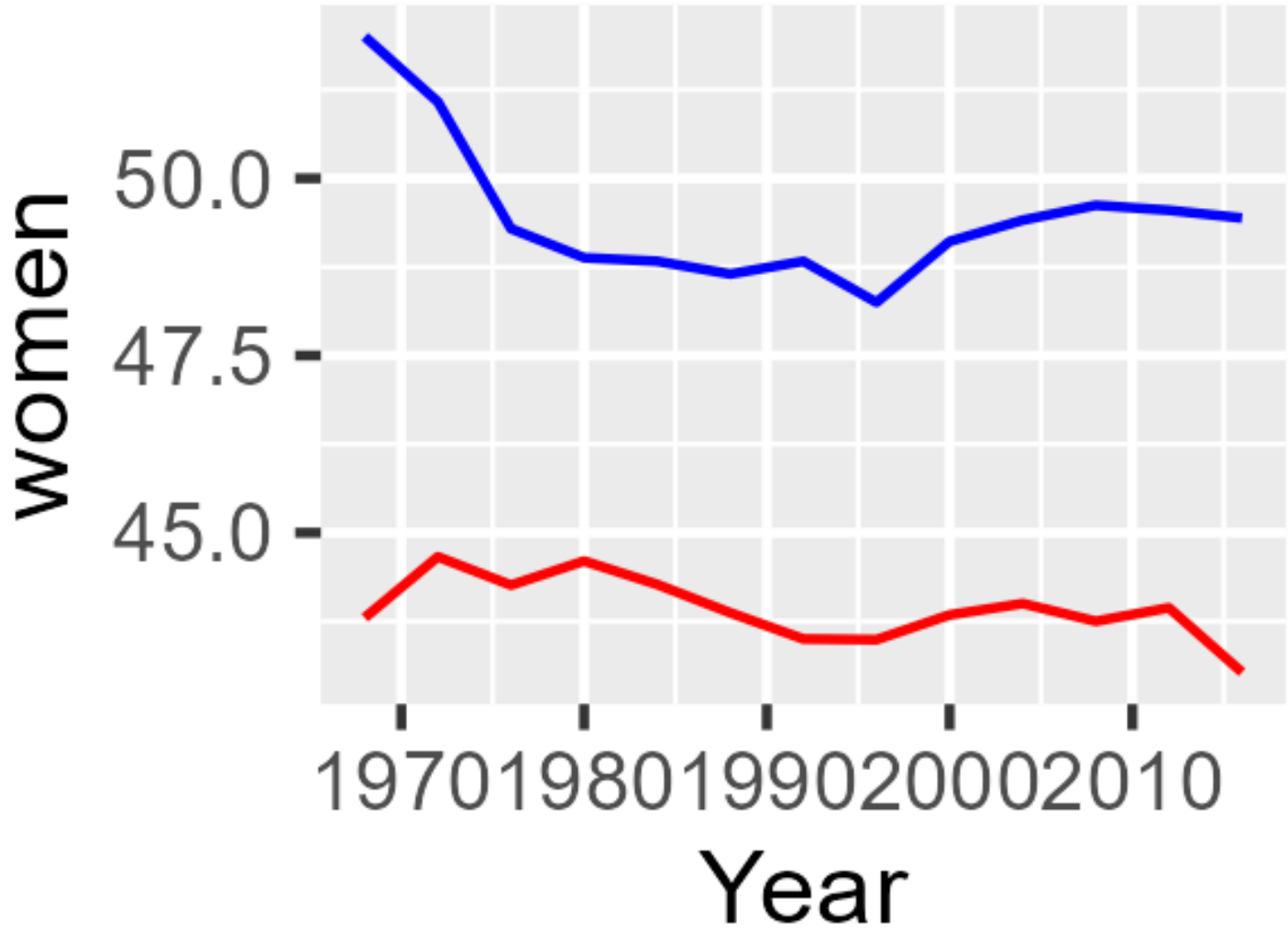
	Year	men	women
1	1968	43.80	52.00
2	1972	44.66	51.08
3	1976	44.26	49.29
4	1980	44.60	48.88
5	1984	44.27	48.83
6	1988	43.87	48.65
7	1992	43.50	48.83
8	1996	43.49	48.25
9	2000	43.84	49.11
10	2004	44.00	49.41
11	2008	43.75	49.62
12	2012	43.94	49.55

- This can be viewed as univariate or bivariate data.

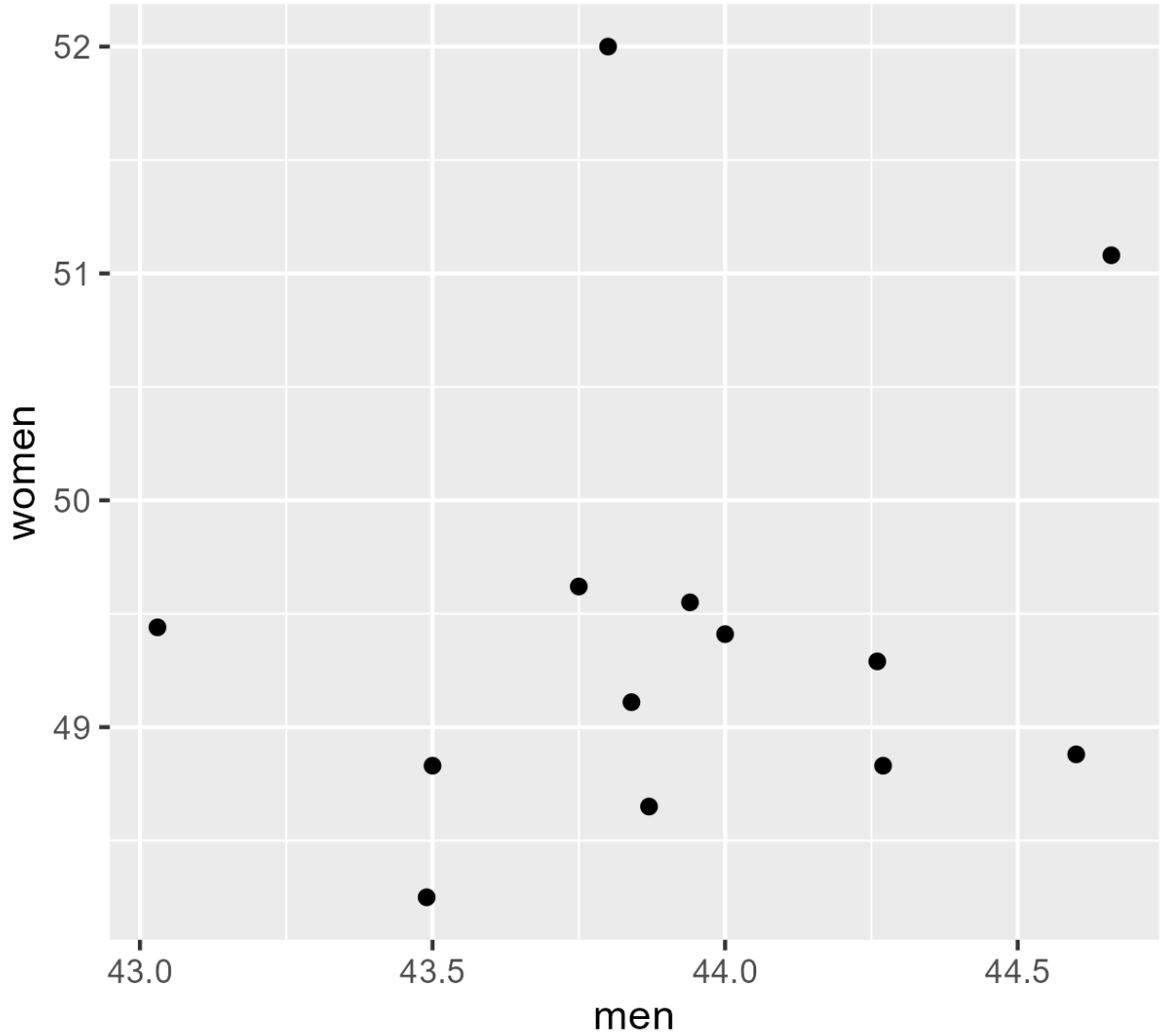
- How can we decide if something is univariate or multivariate data?

- How can we decide if something is univariate or multivariate data?
  - Ask an expert in the domain whether it makes sense to consider them connected.
  - Plot the data
  - Compute correlation (this might not always be a good idea! - more details later) - 0.19 for this example.

Olympic running - line plot



Olympic running - scatter plot



Caption

Examples

- Data on the monthly sales of different types of wine

	Fortified	Drywhite	Sweetwhite	Red	Rose	Sparkling	Total	Month	Year
1	2585	1954	85	464	112	1686	15136	January	1980
2	3368	2302	89	675	118	1591	16733	February	1980
3	3210	3054	109	703	129	2304	20016	March	1980
4	3111	2414	95	887	99	1712	17708	April	1980
5	3756	2226	91	1139	116	1471	18019	May	1980
6	4216	2725	95	1077	168	1377	19227	June	1980
7	5225	2589	96	1318	118	1966	22893	July	1980
8	4426	3470	128	1260	129	2453	23739	August	1980
9	3932	2400	124	1120	205	1984	21133	September	1980
10	3816	3180	111	963	147	2596	22591	October	1980
11	3661	4009	178	996	150	4087	26786	November	1980

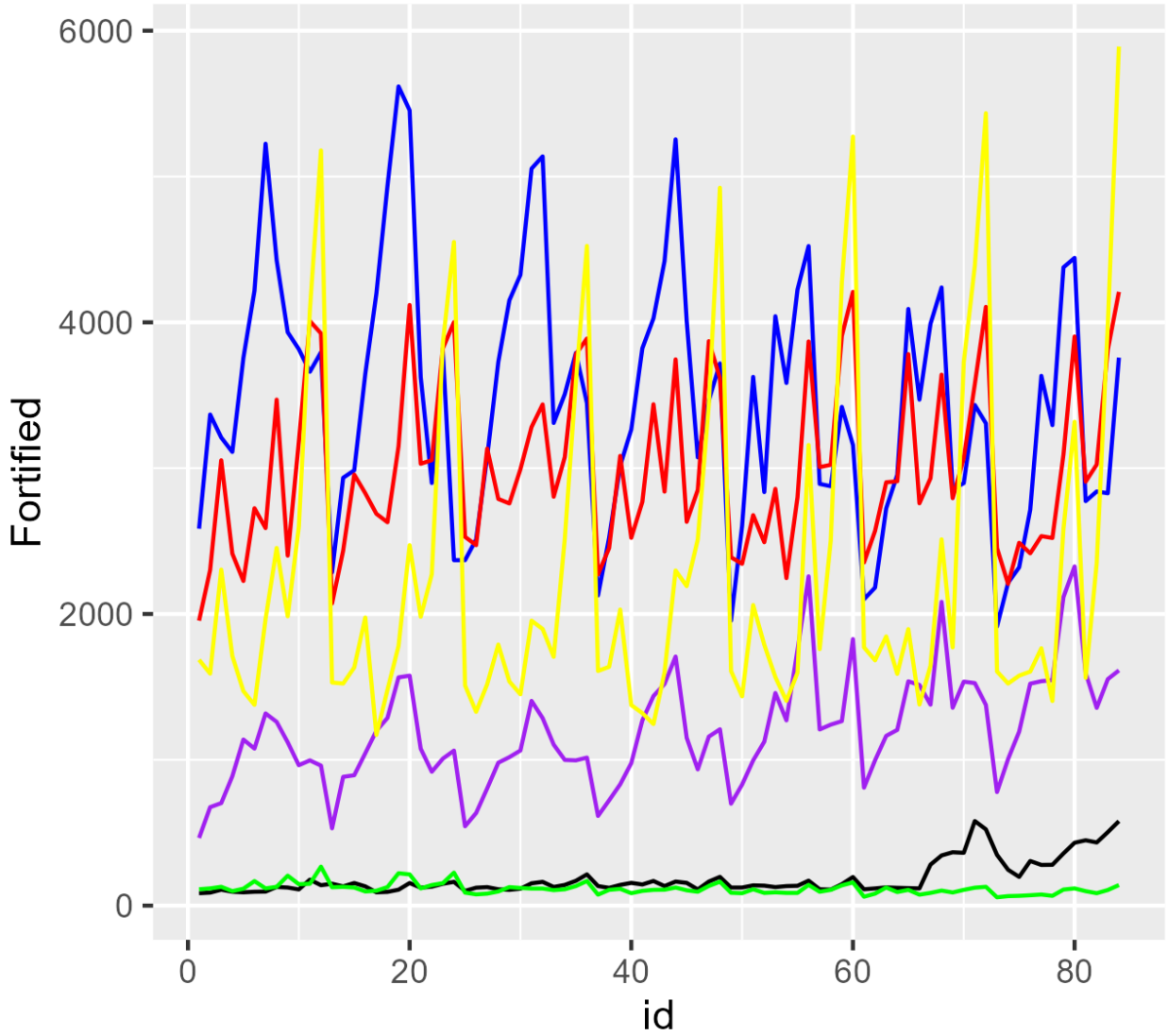
Examples

- Data on the monthly sales of different types of wine

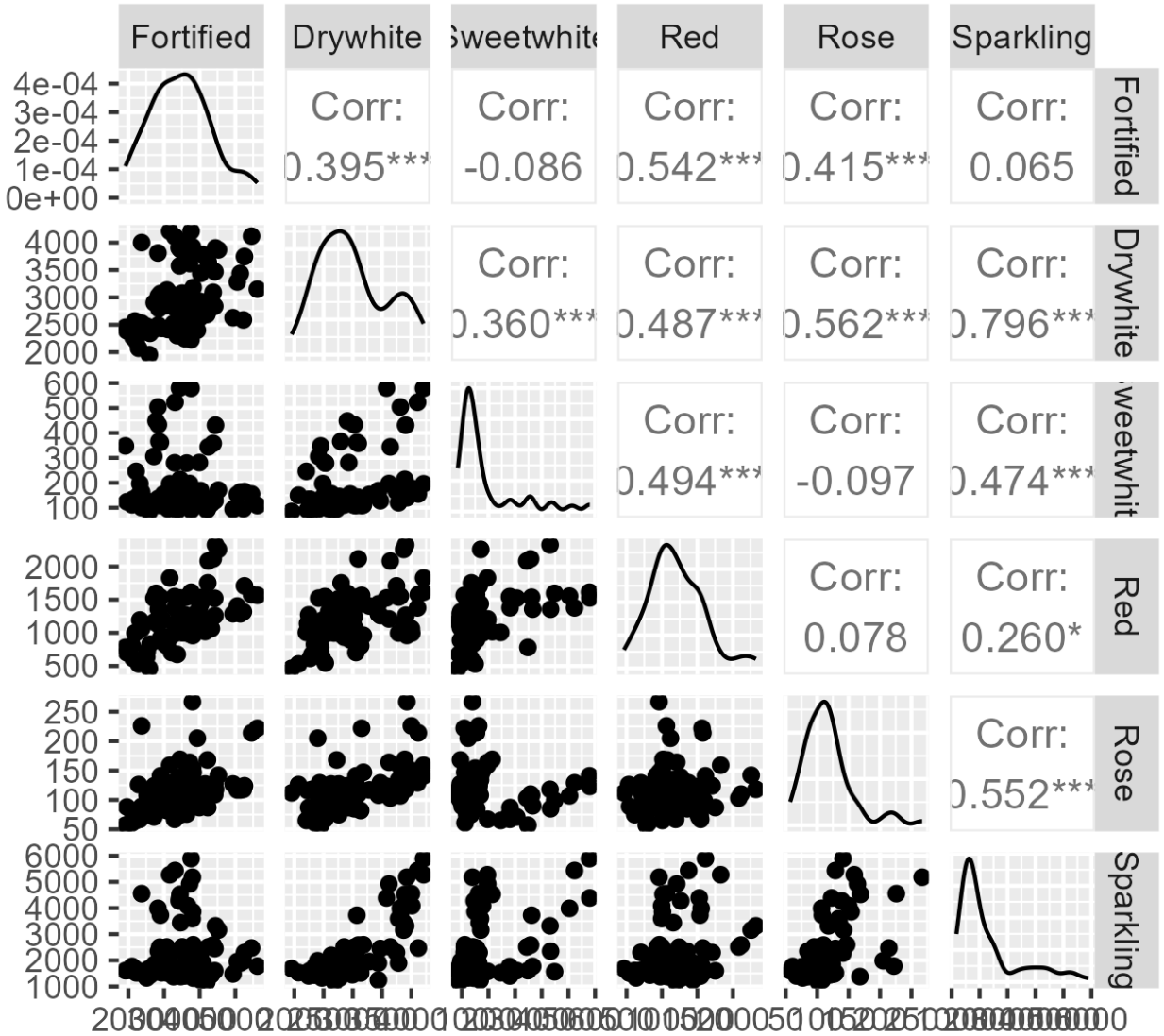
	Fortified	Drywhite	Sweetwhite	Red	Rose	Sparkling	Total	Month	Year
1	2585	1954	85	464	112	1686	15136	January	1980
2	3368	2302	89	675	118	1591	16733	February	1980
3	3210	3054	109	703	129	2304	20016	March	1980
4	3111	2414	95	887	99	1712	17708	April	1980
5	3756	2226	91	1139	116	1471	18019	May	1980
6	4216	2725	95	1077	168	1377	19227	June	1980
7	5225	2589	96	1318	118	1966	22893	July	1980
8	4426	3470	128	1260	129	2453	23739	August	1980
9	3932	2400	124	1120	205	1984	21133	September	1980
10	3816	3180	111	963	147	2596	22591	October	1980
11	3661	4009	178	996	150	4087	26786	November	1980

- Multivariate time series with numerical data (probably there is a relation between the sales for different types of wines)

Wine - line plot



Wine - scatter plot



Examples

- Hourly data on energy consumption (energy demand), precipitation and temperature in New York

timeStamp	demand	precip	temp
2012-01-01 0:00:00	4937.5	0	46.13
2012-01-01 1:00:00	4752.1	0	45.89
2012-01-01 2:00:00	4542.6	0	45.04
2012-01-01 3:00:00	4357.7	0	45.03
2012-01-01 4:00:00	4275.5	0	42.61
2012-01-01 5:00:00	4274.7	0	39.02
2012-01-01 6:00:00	4324.9	0	38.78
2012-01-01 7:00:00	4350	0	42.74
2012-01-01 8:00:00	4480.9	0	38.9
2012-01-01 9:00:00	4664.2	0	44.67
2012-01-01 10:00:00	4847.5	0	47.43
2012-01-01 11:00:00	4981.9	0	49.49
2012-01-01 12:00:00	5081	0	50.77
2012-01-01 13:00:00	5137.2	0	50.57
2012-01-01 14:00:00	5142.6	0	49.94
2012-01-01 15:00:00	5165.1	0	49.85
2012-01-01 16:00:00	5351.1	0	47.39
2012-01-01 17:00:00	5664	0	48.83

Examples

- Hourly data on energy consumption (energy demand), precipitation and temperature in New York

timeStamp	demand	precip	temp
2012-01-01 0:00:00	4937.5	0	46.13
2012-01-01 1:00:00	4752.1	0	45.89
2012-01-01 2:00:00	4542.6	0	45.04
2012-01-01 3:00:00	4357.7	0	45.03
2012-01-01 4:00:00	4275.5	0	42.61
2012-01-01 5:00:00	4274.7	0	39.02
2012-01-01 6:00:00	4324.9	0	38.78
2012-01-01 7:00:00	4350	0	42.74
2012-01-01 8:00:00	4480.9	0	38.9
2012-01-01 9:00:00	4664.2	0	44.67
2012-01-01 10:00:00	4847.5	0	47.43
2012-01-01 11:00:00	4981.9	0	49.49
2012-01-01 12:00:00	5081	0	50.77
2012-01-01 13:00:00	5137.2	0	50.57
2012-01-01 14:00:00	5142.6	0	49.94
2012-01-01 15:00:00	5165.1	0	49.85
2012-01-01 16:00:00	5351.1	0	47.39
2012-01-01 17:00:00	5664	0	48.83

- Multivariate time series with numerical data.

Examples

- 300 records of EEG sleep state scores of newborn infants, measured every minute, classified in six categories:
  - quiet sleep, trace alternant (qt)
  - quiet sleep, high voltage (qh)
  - transitional sleep (tr)
  - active sleep, low voltage (al)
  - active sleep, high voltage (ah)
  - awake (aw)
- The data itself: [ah, ah, ah, ah, ah, ah, ah, ah, tr, ah, tr, ah, ah, qh, qt, qt, qt, qt, qt, tr, ...]

Examples

- 300 records of EEG sleep state scores of newborn infants, measured every minute, classified in six categories:
  - quiet sleep, trace alternant (qt)
  - quiet sleep, high voltage (qh)
  - transitional sleep (tr)
  - active sleep, low voltage (al)
  - active sleep, high voltage (ah)
  - awake (aw)
- The data itself: [ah, ah, ah, ah, ah, ah, ah, ah, tr, ah, tr, ah, ah, qh, qt, qt, qt, qt, qt, tr, ...]
- This is a categorical time series.

Categorical data

- Other possible scenarios where we might have categorical time series could be:
  - Weather data (daily, hourly), with values like: Sunny, Cloudy, Rainy, etc.
  - Sentiment analysis over time for a product / service with values like: Positive, Negative, and Neutral.
  - Stock market information, with the change in the price of a given stock, with values like: Increase, Decrease.
- Nevertheless, in these cases we could argue that numerical values might be more useful than the categorical ones.

Examples

- Data about earthquakes all over the world (taken from: https://github.com/saiedmighani/earthquake_time_series_LSTM)

	time	type	mag	place	status	tsunami	sig	net	nst
1	1970-01-02 10:45:20	earthquake	3.14	24km S of Santa Barbara, CA	reviewed	0	152	ci	11.0
2	1970-01-02 21:47:53	earthquake	2.61	12km NE of Inyokern, CA	reviewed	0	105	ci	6.0
3	1970-01-03 02:51:58	earthquake	4.00	San Francisco Bay area, California	reviewed	0	246	ushis	20.5
4	1970-01-03 19:48:40	earthquake	3.16	6km NE of Banning, CA	reviewed	0	154	ci	9.0
5	1970-01-04 02:27:15	earthquake	2.74	8km N of Big Bear City, CA	reviewed	0	116	ci	9.0
6	1970-01-05 12:04:34	earthquake	3.04	7km SSW of Salton City, CA	reviewed	0	142	ci	4.0
7	1970-01-06 02:29:07	earthquake	3.92	24km ENE of Soledad, CA	reviewed	0	236	ci	13.0
8	1970-01-06 02:56:03	earthquake	3.05	18km WSW of Greenfield, CA	reviewed	0	143	ci	3.0
9	1970-01-06 22:18:10	earthquake	3.66	69km SW of Avila Beach, CA	reviewed	0	206	ci	5.0
10	1970-01-07 01:14:49	earthquake	3.31	72km E of Maneadero, B.C., MX	reviewed	0	169	ci	3.0
11	1970-01-07 03:25:26	earthquake	2.40	22km N of Ridgecrest, CA	reviewed	0	89	ci	6.0
12	1970-01-07 18:04:50	earthquake	3.07	45km WSW of Vandenberg Air Force Base, CA	reviewed	0	145	ci	5.0
13	1970-01-08 02:16:41	earthquake	2.75	8km NE of Lake Arrowhead, CA	reviewed	0	116	ci	13.0
14	1970-01-08 17:00:33	earthquake	3.70	98km W of Vandenberg Air Force Base, CA	reviewed	0	211	ci	19.0

Examples

- Data about earthquakes all over the world (taken from: https://github.com/saiedmighani/earthquake_time_series_LSTM)

	time	type	mag	place	status	tsunami	sig	net	nst
1	1970-01-02 10:45:20	earthquake	3.14	24km S of Santa Barbara, CA	reviewed	0	152	ci	11.0
2	1970-01-02 21:47:53	earthquake	2.61	12km NE of Inyokern, CA	reviewed	0	105	ci	6.0
3	1970-01-03 02:51:58	earthquake	4.00	San Francisco Bay area, California	reviewed	0	246	ushis	20.5
4	1970-01-03 19:48:40	earthquake	3.16	6km NE of Banning, CA	reviewed	0	154	ci	9.0
5	1970-01-04 02:27:15	earthquake	2.74	8km N of Big Bear City, CA	reviewed	0	116	ci	9.0
6	1970-01-05 12:04:34	earthquake	3.04	7km SSW of Salton City, CA	reviewed	0	142	ci	4.0
7	1970-01-06 02:29:07	earthquake	3.92	24km ENE of Soledad, CA	reviewed	0	236	ci	13.0
8	1970-01-06 02:56:03	earthquake	3.05	18km WSW of Greenfield, CA	reviewed	0	143	ci	3.0
9	1970-01-06 22:18:10	earthquake	3.66	69km SW of Avila Beach, CA	reviewed	0	206	ci	5.0
10	1970-01-07 01:14:49	earthquake	3.31	72km E of Maneadero, B.C., MX	reviewed	0	169	ci	3.0
11	1970-01-07 03:25:26	earthquake	2.40	22km N of Ridgecrest, CA	reviewed	0	89	ci	6.0
12	1970-01-07 18:04:50	earthquake	3.07	45km WSW of Vandenberg Air Force Base, CA	reviewed	0	145	ci	5.0
13	1970-01-08 02:16:41	earthquake	2.75	8km NE of Lake Arrowhead, CA	reviewed	0	116	ci	13.0
14	1970-01-08 17:00:33	earthquake	3.70	98km W of Vandenberg Air Force Base, CA	reviewed	0	211	ci	19.0

- Irregular time series

Time series data

- When we deal with time-series data, there are two possible tasks:
  - Time series analysis - try to understand data, find patterns in it
  - Time series forecasting - try to predict the future
- Time series analysis should always be the first step, even when we want to do forecasting.

Problem setting I

- Consider the task of forecasting the *hourly electricity demand* (ED). We might start in three directions:

Problem setting II

- **Explanatory model** (also called regression model) - build a model based on predictor variables (variables whose value influences the electricity demand). In this case the model would be

$$ED = f(current\ temperature, population, time\ of\ day, day\ of\ week, error)$$

Problem setting III

- **Time series model** - assume that future ED depends on past ED. The model would be:

$$ED_{t+1} = f(ED_t, ED_{t-1}, ED_{t-2}, ..., error)$$

Problem setting IV

- **Mixed model** - combines the features of the above two methods:

$$ED_{t+1} = f(ED_t,\ current\ temp.,\ population,\ time\ of\ day, day\ of\ week,\ error)$$

Basic steps of forecasting I

1. Problem definition
   - Requires understanding what needs to be forecast exactly, who will use the forecasts and for what purpose.
   - Some example of questions to be answered at this point are:
     - the form of the forecast (monthly, quarterly, etc.)
     - the forecast horizon (how many time steps in the future we want to forecast)
     - what error is acceptable in order to consider it a good forecast
     - if multiple products are involved then the level of the forecast is also important (do we forecast individual products, or families of similar products)

Basic steps of forecasting II

2. Gathering information
   - If possible, historical data is needed (although there are methods that can handle lack of historical data)
   - Other, relevant predictor variables can also be identified and collected.
   - Expertise of people who collect data and will use the forecast

Basic steps of forecasting III

3. Preliminary (exploratory) analysis
   - Plotting the data to see if there are patterns, outliers, relationships between variables, etc. (think about what we did with the olympic running times and wines data).
   - Get a *feel* for the data.

Basic steps of forecasting IV

4. Choosing and fitting models
   - Fitting a model to the (historical) data means estimating the unknown parameters of the data, based on the actual data we have.
   - While almost any model can be fit on any data, obviously not every model is equally good. One first check is to see how well the model fits the historical data (fitting error)
   - The best model depends on the data, the relationship between the variables and how forecasts are used. It might be a good idea to try several models.

Basic steps of forecasting V

5. Evaluating the forecasting model
   - Besides checking how the model performs on the data on which it was built, it should be evaluated on new, previously unused data (forecast error).
   - Finding the best relevant performance measure is also important.

Basic steps of forecasting VI

6. Deploying the forecasting model and monitoring the performance
   - Getting the model to be used by the customer
   - When you have a model, you can start creating forecasts and once you have the real data, its performance can be evaluated and should be monitored to see if the model still fits the data.

Our forecasting workflow

- During this course, we will focus on Steps 3, 4, and 5. We will consider as input data different data sets available online, and we will perform analysis and forecast on them.

- Let's take an example of what we will discuss during the semester, considering a data set called *global_economy*.

	Country	Code	Year	GDP	Growth	CPI	Imports	Exports	Population
1	Afghanistan	AFG	1960	537777811	NA	NA	7.024793	4.132233	8996351
2	Afghanistan	AFG	1961	548888896	NA	NA	8.097166	4.453443	9166764
3	Afghanistan	AFG	1962	546666678	NA	NA	9.349593	4.878051	9345868
4	Afghanistan	AFG	1963	751111191	NA	NA	16.863910	9.171601	9533954
5	Afghanistan	AFG	1964	800000044	NA	NA	18.055555	8.888893	9731361
6	Afghanistan	AFG	1965	1006666638	NA	NA	21.412803	11.258279	9938414
7	Afghanistan	AFG	1966	1399999967	NA	NA	18.571429	8.571429	10152331
8	Afghanistan	AFG	1967	1673333418	NA	NA	14.209827	6.772908	10372630
9	Afghanistan	AFG	1968	1373333367	NA	NA	15.210356	8.899677	10604346
10	Afghanistan	AFG	1969	1408888922	NA	NA	14.984227	10.094637	10854428
11	Afghanistan	AFG	1970	1748886596	NA	NA	11.944108	9.784002	11126123
12	Afghanistan	AFG	1971	1831108971	NA	NA	16.140795	10.922342	11417825
13	Afghanistan	AFG	1972	1595555476	NA	NA	18.105850	14.763231	11721940
14	Afghanistan	AFG	1973	1733333264	NA	NA	14.743591	12.948718	12027822
15	Afghanistan	AFG	1974	2155555498	NA	NA	14.845361	14.020619	12321541
16	Afghanistan	AFG	1975	2366666616	NA	NA	14.272301	12.676057	12590286
17	Afghanistan	AFG	1976	2555555567	NA	NA	14.869565	13.217391	12840299
18	Afghanistan	AFG	1977	2953333418	NA	NA	14.823175	11.662904	13067538
19	Afghanistan	AFG	1978	3300000109	NA	NA	13.872054	10.841750	13237734
20	Afghanistan	AFG	1979	3697940410	NA	NA	NA	NA	13306695
21	Afghanistan	AFG	1980	3641723322	NA	NA	NA	NA	13248370
22	Afghanistan	AFG	1981	3478787909	NA	NA	NA	NA	13053954
23	Afghanistan	AFG	1982	NA	NA	NA	NA	NA	12749645
24	Afghanistan	AFG	1983	NA	NA	NA	NA	NA	12389269

- This data set contains data about 263 countries, so it actually contains 263 time series. We will work with data from Sweden.

Our forecasting workflow

- Data preparation and understanding
  - It should always be the first step, you need to understand your data and prepare it for the subsequent analysis.
  - Some forecasting models have different requirements (for ex. no missing data), before applying one, we need to check if these requirements hold.

```
gdppc <- global_economy |>
    mutate(GDP_per_capita = GDP / Population)
```

Our forecasting workflow

- Plot the data
  - Plotting the data will help understand it better. There are several plots that we can use, here we will go with a simple version where the x axis contains the time, and the y axis contains the value of the observation. The values of two consecutive observations are linked by a line.

```
gdppc |>
  filter(Country == "Sweden") |>
  autoplot(GDP_per_capita) +
  labs(y = "$US", title = "GDP per capita for Sweden")
```

Our forecasting workflow

- Choose and define a time series forecasting model
  - We will talk about several models during the semester.
  - In this example, we will use a simple *time series linear model*, where the output parameter is the *GDP_per_capital*
- Train the selected model on data

```
fit <-
    gdppc |> model(trend_model = TSLM(GDP_per_capita ~ trend()))
```

Our forecasting workflow

- Check model performance
  - There are several ways to check the performance of a model

```
augment(fit)
fit |> gg_tsresiduals()
```

Our forecasting workflow

- Produce forecasts
  - Need to specify the number of observations to be forecast

```
fit |>
    forecast(h = 5) |>
    filter(Country == "Sweden") |>
    autoplot(gdppc) +
    labs(x = "Time", y = "GDP per capita", title = "Sweden's GDP per capita –
     forecast for the next 5 years")
```

Prediction interval

- When thinking about forecasting something, in general we think in terms of single values (single numbers). This is called a *point forecast*.
- Point forecasts are almost always wrong; you have a *forecast error*.
- It would be useful to have an estimate of how big that error is likely to be: this is called a *forecast interval* or *prediction interval*.
- Prediction intervals are accompanied by probabilities, for ex: 80% prediction interval (meaning that there is an 80% chance that the actual value will be in that interval).

Notations I

- We will use subscript t for denoting time: y_t is an observation at time t.
- The set of all observations (all the information we have) will be denoted by: $\mathcal{I}$
- We write $y_t|\mathcal{I}$ to denote the value of y_t given what we know in $\mathcal{I}$.
- $y_t|\mathcal{I}$ can take a set of values and each has its own probability; this is called the **forecast distribution**. In general we want to *forecast* the average value of this distribution, denoted by $\hat{y}_t$ (so $\hat{y}_t$ is the forecast of y_t)

Notations II

- Sometimes we want to express clearly what information we considered for computing the forecast.
- $\hat{y}_t|_{t-1}$ means that we have considered all the previous observations: $y_1, y_2, \ldots y_{t-1}$.
- $\hat{y}_{t+h}|_t$ means that we have considered all the previous observations: $y_1, y_2, \ldots y_{t-1}$ but we are forecasting the value y_{t+h} (h steps in advance)