Project work for the teaching *Time Series Analysis*

Master in *Data Science and Statistical Learning*, A. A. 2021/2022

Teacher: Alessandro Magrini (alessandro.magrini@unifi.it)

The student must retrieve **two distinct time series with at least 50 measurements** from any publicly available source.

- **The first time series must show no seasonality**, for instance it may have annual frequency, having been deseasonalised or being a financial time series.
- **The second time series must** have frequency less than one year and **be characterized by a seasonal pattern**.

The source of the time series must not necessarily be an institute or an institution, the important is that the student is able to provide a clear reference for the data, that can also be a research paper. Ideally, the time series must contain no missing values but, if there are few of them, it is possible to perform an imputation making use of any method.

For each time series, produce a report of at most 10 pages with font 14 *Times New Roman* addressing the issues below. **All the procedures employed must be described clearly and each choice must be motivated with reference to the statistical theory**. The report may include formulas, tables and figures but no code or reference to the software employed. **Everything that reproduces the results included in the report** (data, R code, Excel files, etc) **must be presented separately as a zip file**.

Issues to be addressed

1. Describe the data, explaining which phenomenon they represent and providing details on the measurements (time span, frequency, scale).
2. Display the time series and discuss its appearance, emphasizing the presence of trends, breaks, cycles and seasonal patterns.
3. Find an **ARIMA model** and an **ARMA model with pure deterministic trend** that fit adequately to the time series. To do this, make use of a combination of graphical checks, significance tests and automated search. Some guidelines:
    - among deterministic trends, consider a polynomial one with no more than three degree;
    - for ARIMA models, evaluate whether a drift is necessary;
    - use a seasonal model for the second time series;
    - try to apply no more than one non-seasonal and one seasonal difference;

    - make use of transformations (logarithm or power functions) and/or reduce the window of observation (just in case of structural breaks) to improve the adequacy of the model.

4. Provide a formal definition and a description of the two models obtained, emphasizing the different theoretical properties.
5. Estimate the forecast accuracy of the two models at different horizons through cross-validation, provide a comparison, and discuss the results.
6. Display forecasts of future values of the time series based on the two models, and discuss the difference between them in light of their theoretical properties.

Suggested data sources

- ISTAT https://dati.istat.it/
- EUROSTAT https://ec.europa.eu/eurostat/data/database
- OECD https://data.oecd.org/
- The World Bank https://data.worldbank.org/
- Faostat https://www.fao.org/faostat/en
- Yahoo Finance https://it.finance.yahoo.com
- Kaggle https://www.kaggle.com/datasets

Suggested R functions

- Package ‘tseries’: `adf.test`, `kpss.test`
- Package ‘forecast’: `Acf`, `Pacf`, `Box.test`, `Arima`, `auto.arima`, `BoxCox.lambda`, `forecast`, `tsCV`