HEC Lausanne Forecasting II

Valérie Chavez-Demoulin {[valerie.chavez@unil.ch](mailto://valerie.chavez@unil.ch)}

Juraj Bodík {[Juraj.Bodik@unil.ch](mailto:Juraj.Bodik@unil.ch)}

Spring 2024

Projects Guidelines

Welcome to the Forecasting II course! This course evaluation will be based on two assignments. First accounts for 60% and the second for 40% of the total grade.

Common guidelines for the two projects:

* Group projects: form teams up to 5 people
* On the first page of the report, do not forget to write down your names, your student numbers and the date
* You are strongly advised to use R Markdown. In any case provide the code.

# Project 1 –Swiss tourism

**Introduction:**

In this project, you will have the opportunity to apply the forecasting techniques and methods learned throughout the course to a real-world scenario: predicting tourism in Switzerland. Tourism is a vital sector for Switzerland's economy, and accurate forecasting of tourist arrivals is crucial for effective planning and decision-making in the tourism industry.

**Project Objective:**

The main objective of this project is to predict

1. The total number of visitors to Vaud from October 2023 to December 2024.
2. Specifically, each group will be tasked with predicting as well as the number of visitors from a specific country in a different canton.

**Dataset Description:**

Your dataset comprises monthly data of arrivals in the hotel sector for open establishments, categorized by canton and visitors' country of residence. The dataset covers the period from 2005 to September 2023, providing a comprehensive historical perspective on tourism in Switzerland.

The data are available on Moodle. You may use auxiliary sources of information to improve your forecast, but you must acquire and (if necessary) forecast them yourselves. Some suggestions are:

* Weather conditions in Switzerland/canton[[1]](#footnote-1) (<https://gate.meteoswiss.ch/idaweb>)
* GDP of Switzerland and other countries
* Population growth
* …

**Project Guidelines:**

Your hand-in should consist of a PDF report, the code used (you may combine both with Rmarkdown if you wish) and an Excel or CSV file giving the forecasts.

Please follow the steps below for your analysis:

1. Exploration and Visualization: To better understand the data, you might want to visualize it first. By doing so, you will be able to discuss the features of your data. This part should not be longer than 2 pages.
2. Cleaning and Wrangling: Your data will probably require some preliminary transformations before being able to work on it. This includes, but is not limited to, dealing with missing values, detecting and treating outliers, creating new variables, etc.
3. Modeling: This part is about building on your knowledge of time series techniques to model your data. You can investigate various models but you should justify in your report your choices regarding these. Pay attention to the conditions that are needed to apply a specific model. Treat also carefully seasonality, outliers, colinearity, covariates, special events, etc. Remember the following steps:
   1. Aggregation choice for hierarchical time series
   2. Model building
   3. Model selection
4. Forecast and Validation: this part is about using the models selected in 3. to produce the most accurate forecast. You may try to compare different models using different scoring methods.
5. Limits and Discussion: Please discuss your results in an objective view, indicating potential improvements, issues and concerns, etc.

For this project, you will have to submit the following deliverables (60% of the total grade) before Tuesday May 21th 23h59:

* Scientific report
  + No more than 5 pages.
  + Evaluation - Clarity: you should explain every choice you made along the process of model building and forecasting.
* Forecast Table
  + Submit 2 tables of your forecast. One for Vaud (all countries of arrivals) and one for your group-specific Canton/Country of visitors. Each table must contain 4 columns containing:

1. Month (ranging from October 2023 to December 2024),
2. Forecast value
3. 95% Lower confidence interval
4. 95% Upper confidence interval

We will compare your forecast values with the actual values that we have (not yet public) for the most recent months. Most importantly is the methodology, not the actual results of your forecast.

|  |  |  |
| --- | --- | --- |
| **Groups:** | **Canton to predict** | **Origin of visitors** |
| Group 1 | Fribourg | Israel |
| Group 2 | Zürich | Philippine |
| Group 3 | Luzern | Japan |
| Group 4 | Jura | Germany |
| Group 5 | Genève | Philippine |
| Group 6 | St. Gallen | Austria (Österreich) |
| Group 7 | Bern | Singapore |
| Group 8 | Zürich | Japan |
| Group 9 | Ticino | Philippine |
| Group 10 | Bern | Taiwan (Chinese Taipei) |

# Project 2 - Open Data

For this project, you will choose the time series, define a question, and answer to it using forecasting methods.

1. Choose a time series. Your data has to be accessible by all times (i.e. open source). Here is a non-exhaustive list of sources you can download data from:
   * [CERN](http://opendata.cern.ch/)
   * [European Data Portal](https://www.europeandataportal.eu/en)
   * [Swiss Open Data](https://opendata.swiss/fr)
   * [UN Data](http://data.un.org/)
   * [World Bank Data](https://data.worldbank.org/)
   * [Google Dataset Search](https://datasetsearch.research.google.com/)
   * [Github Awesome Public Datasets](https://github.com/awesomedata/awesome-public-datasets)
   * [data.world](https://data.world/)
   * [Subreddit r/datasets](https://www.reddit.com/r/datasets/)
   * [Open Data Inception](https://opendatainception.io/)
   * [Corgis](https://corgis-edu.github.io/corgis/)
   * [Data is Plural: Google Sheet with 1000+ datasets](https://docs.google.com/spreadsheets/d/1wZhPLMCHKJvwOkP4juclhjFgqIY8fQFMemwKL2c64vk/edit#gid%3D0)

Your time series should have at least 500 observations, i.e. *n* ≥ 500, with regular time-stamps (daily, weekly, monthly, etc.).

1. Define the research or business question you want to investigate and use forecasting tools and methods to answer it.
2. Discuss your results.

For this project, you will have to submit the following deliverables:

* May 29: Business presentation (40% of the total grade)
  + 10 minutes + 5 minutes of questions
  + Format of your choice: PowerPoint, pdf, Markdown etc.
  + No report, but we require your R code (to prevent cheating and verify results) in .txt format.
  + Business presentation means it should be lively, accessible, with an emphasis on results and discussions. Imagine that we are CEOs of the company without strong knowledge in data science and you are presenting your results to us (as CEOs we do not know any words like ARIMA or cross-correlation etc, if you use them you have to explain them). You will be evaluated not only by your forecasting skills, but mostly by your ability to present them in an accessible manner.

1. Last year, these data were available only after providing a signature of us. If this is the case also this year, you can ask us for a signature. [↑](#footnote-ref-1)