STAT9005 CONTINUOUS ASSESSMENT 3 11TH APRIL 2025

You should work on this assignment and upload a report with its supporting R script and a video presentation on Canvas up to **Monday the 19th of May at 10pm**. This assignment is worth 40% (30% report + 10% presentation) of your final grade.

Format specifications:

Report

- The report should feature a **maximum of 15 pages**.
- You can either produce 1) a report in printable document format (PDF) with a supporting R script, or 2) an Rmarkdown file with both report and code configured to output a pdf file (notice the html output does not feature page counting). The file formats Canvas will accept are ".pdf", ".R" and ".Rmd".
- In the case your code is submitted separately as an R script, you should identify where each item is produced with appropriate comment lines in the R script.
- The report structure should follow those of the questions below. You are advised to extend the numbering within each section to each new item your report shows (figures, tables and so on).
- Write brief and objective sentences. Marks are assigned to the presentation and clarity of the text. It is important to number and describe each table and figure so they can be referred to clearly in the text.
- Likewise, a few marks are allocated to the readability of your script. Ensure the code is clean, contains only what is necessary to support your report and indicates where is what as you submit. Using RMarkdown may significantly help with this.
- You should also include a declaration at the beginning of your report that all work contained in the submission is your own, except for clearly referenced sources (libraries/packages written by others). Any collaboration and/or plagiarism will be penalised as per the MTU code of plagiarism and academic integrity.

Presentation

- The presentation should have a **minimum of 5 minutes and a maximum of 8 minutes**.
- All that is necessary is to record your voice over slides. It is not necessary to attach any video and/or animation.
- You should submit **only a video file with this presentation and not the slides**.
- To create a video with your presentation, you are advised to use PowerPoint or Google Slides since they allow partial editing of the audio in each slide. Alternatively, you can use ScreenPal.
- When prompted to choose a video resolution, you are suggested to pick 720p so that the video is sufficiently visible while keeping the file size not big.
- The preferred file format is ".mp4", which is what Canvas is configured to accept. If you absolutely need to submit a different file format, communicate your situation as soon as possible.
- Make sure to test your devices and software by recording test presentations beforehand. Leaving it to the last minute will create a risk for your submission deadline.

Please notice no submission past the deadline will be accepted. **This is an individual assignment. You are not allowed to collaborate on this assignment and plagiarism will be penalised as per MTU official academic guidelines.**

Time series analysis

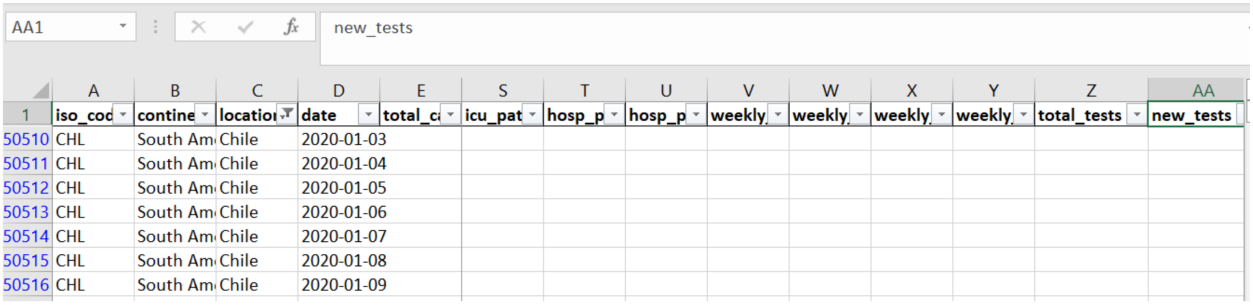
A dataset named "TS-covid-data_Full.xlsx" has collected information about several variables during the COVID-19 pandemic period for a number of countries all over the world. A sample of some of the variables are as follow:

Tests & positivity

On 23 June 2022, we stopped adding new datapoints to our COVID-19 testing dataset. You can read more at #2667.

Variable	Description
total_tests	Total tests for COVID-19
new_tests	New tests for COVID-19 (only calculated for consecutive days)
total_tests_per_thousand	Total tests for COVID-19 per 1,000 people
new_tests_per_thousand	New tests for COVID-19 per 1,000 people
new_tests_smoothed	New tests for COVID-19 (7-day smoothed). For countries that don't report testing data on a daily basis, we assume that testing changed equally on a daily basis over any periods in which no data was reported. This produces a complete series of daily figures, which is then averaged over a rolling 7-day window
new_tests_smoothed_per_thousand	New tests for COVID-19 (7-day smoothed) per 1,000 people
positive_rate	The share of COVID-19 tests that are positive, given as a rolling 7-day average (this is the inverse of tests_per_case)
tests_per_case	Tests conducted per new confirmed case of COVID-19, given as a rolling 7-day average (this is the inverse of positive_rate)

Detailed information about all the variables in the dataset can be found in the following link:
covid-19-data/public/data at master · owid/covid-19-data · GitHub



	A	B	C	D	E	S	T	U	V	W	X	Y	Z	AA
1	iso_cod	contine	location	date	total_ca	icu_pat	hosp_p	hosp_p	weekly	weekly	weekly	weekly	total_tests	new_tests
50510	CHL	South Am	Chile	2020-01-03										
50511	CHL	South Am	Chile	2020-01-04										
50512	CHL	South Am	Chile	2020-01-05										
50513	CHL	South Am	Chile	2020-01-06										
50514	CHL	South Am	Chile	2020-01-07										
50515	CHL	South Am	Chile	2020-01-08										
50516	CHL	South Am	Chile	2020-01-09										

The analysis conducted in this project will focus on the **daily** observations of the time series variable indicating the **"number of tests performed"** during the COVID-19 period. For example, in the picture below we are selecting the variable indicating the number of **"tests performed"** during the COVID-19 period in Chile from the data file (column 'AA' entitled `new_tests`)).

A dashboard previewing the time series pattern can be found in the following link:
COVID-19 Data Explorer - Our World in Data

Using that website, you can preview the time series for the different variables and countries. You should **choose a dataset from any country you are interested to analyse in this project**.

To perform the tests, you need to analyse the time series of the **number of tests performed** in a given day for the chosen country during the period of study using R. You should use the libraries available in R which you consider the most appropriate for this study.

To carry on the analysis, consider the following steps:

1 Data cleaning

Missing values of the variable 'number of tests performed in a given day' need to be replaced using a method of your choice (e.g. by interpolation of previous and posterior observations, average of the period/dataset and so on). *Hint: the seasonality observed in the data can also be used for this.*

Using **the number of tests performed every day in a chosen country**, create two new time series aggregating values for the weekly and monthly number of tests performed in the given period. Comment on the different time series (daily, weekly and monthly).

2 Preliminary analysis

1. Carry out a preliminary descriptive analysis on the data. Use summary statistics and plots to describe the dataset (using the 3 different datasets – daily, weekly and monthly).
2. Decompose the dataset and analyse the different components of the series. State what type of model between additive and multiplicative is more appropriate to describe this series. Support your conclusions by using appropriate plots and summaries.
3. Analyse the importance of each component (if any) in the time series and indicate if periodicity could be a problem in the time series. Discuss the possibility of multiple seasonalities in the datasets.
4. Considering only the daily dataset, identify three periods for different waves of COVID and generate a new subset of time series for the period of each wave (create 3 subsets of your choice from the original time series). Carry out a descriptive analysis on these different waves and discuss on the differences observed.

3 Time series modelling

Considering only time series with daily indexes (the base dataset on the number of tests for chosen country and the ones generated for the three different waves),

3.1 Exponential smoothing model

Implement an exponential smoothing model (SES, Holt's, Holt-Winter's) through the ETS framework following what you consider appropriate for describing the data. Explain why do you think this is the best option.

Indicate the best model options for the trend and seasonality components (additive and multiplicative). Explain which of the options fits the time series better. Support your claims with metrics and/or plots. Is this conclusion consistent with what you expected in the preliminary analysis?

3.2 Stationarity

Indicate whether the four time series are stationary or if they are not, which transformation is necessary to make them stationary. Use formal tests and graphical tools you consider appropriate to support your conclusion.

Use correlograms for analysing the transformed time series. Based on the correlograms, what type of ARIMA/SARIMA model do you expect to best describe these timeseries? Explain your answer.

3.3 ARIMA model

Chose two parameterisations you think could be good candidates for this dataset and analyse the residuals. Use plots and formal tests to compare the models. Refer to what you claimed in Section 3.2.

Finally, analyse the ARIMA produced automatically and discuss whether you think this is a good option for this dataset. Do you think that any of your models would be a better option?

4 Forecasting

Here, you should forecast 4 weeks for the four daily time series using the ETS and ARIMA models. That is, consider models fitted for the data excluding the period to be forecast.

4.1 ETS

Carry out the predictions between two different models (in terms of the trend and seasonality components) and compare the results between the forecasting from the two models. Comment on the differences observed to the model fitted in Section 3.1.

4.2 ARIMA

Carry out the predictions using the automatic parameterisation and also one of the models you proposed during your analysis in Section 3.3. Compare their results.

4.3 Comparison between models

Compare the forecasting obtained from the most adequate ETS model with the most adequate ARIMA model and discuss their performance.

4.4 Different country

Choose a second country to repeat your modelling analysis for the daily dataset of the full period (only the daily dataset and for the full length of the COVID period). Find the best ARIMA model for this new country. Contrast the performance of this ARIMA with the models you studied here in Section 4. Discuss whether any of the previous models is adequate for the different country.