MS4S09 - 2020/2021 Assessment 2

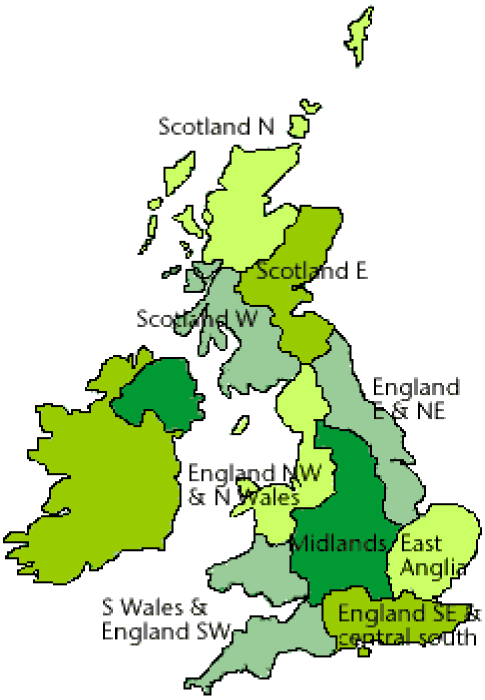
Deadline: Tuesday 2nd March 2021, 9:00 PM

This assessment is worth 60% of your overall mark for this module.

You have to use R to perform your statistical analyses and a report summarising your output and results should be produced. Your findings should be interpreted and valid conclusions drawn.

Only ONE file, which includes the assessment cover sheet and report, is to be submitted to Turnitin before the deadline, with contents converted to a PDF format.

On the Met Office website https://www.metoffice.gov.uk/research/climate/maps-and-data/uk-and-regional-series you can find weather data for 10 different districts in the UK.



The districts to consider, accordingly to the map above, are:

- Northern Ireland
- Scotland N
- Scotland E
- Scotland W
- England E & NE
- England NW & N Wales
- Midlands
- East Anglia
- England SW & S Wales
- England SE & central S

For each region you need to download the time series for the different parameters:

- Max temp
- Min temp
- Mean temp

Work out the following tasks by writing an appropriate R script. Your scripts should be parametric in order to achieve full marks: copying and pasting the same block of commands for each time series will result in a loss of marks. Use of the pipe '%>%' operator and writing functions is strongly advised to make your code more readable and easier to manage.

Although R includes many packages, the only packages that are really needed to solve this coursework are `magrittr` and `tseries` (for the `adf.test()` function). You can use additional packages to solve Task 1 (however the procedure seen in class can be easily adapted) and `ggplot2` if you want to make fancier plots (even if this was not discussed in class). Other packages are not allowed.

There are many ways of solving the following questions. It is strongly advised to avoid for loops, unless they are really necessary, as they usually make R scripts inefficient. It is suggested (but it is not mandatory) to use the list data structure. You may find the following references on R lists helpful:

- https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf chapter 6
- https://adv-r.hadley.nz/vectors-chap.html#lists

Task 1 – Getting the data (10%)

- Write an R script that downloads the data directly from the website for the 30 time series (3 time series for each of the 10 districts) using the "Year ordered statistics" option, and selecting the districts listed. Download up to December 2021.
- Create the 30 time-series objects in R to store the data you have downloaded. Remember to specify the appropriate starting point and frequency.

Task 2 – R programming (5%)

- Write an R script to identify the district and date (year and month) of the highest and the lowest max, min and mean temperature (six results in total).

Task 3 – Exploratory Data Analysis (25%)

- Carry out an EDA of the data you have downloaded. In order to complete your analysis, you may find it useful to answer (**but not only!**) the following questions:

  - Which district is the coldest/warmest? Describe used criteria.
  - Which district has the widest temperature range?
  - Are winters/summers getting colder/hotter?

Task 4 – Trend and Seasonality (30%)

For each district, consider the 3 time series: max temp, min temp and mean temp. Subset each of the 30 time series until December 2019.

- Estimate the trend of each time series using linear, quadratic and cubic regression. Compare your results and use appropriate plots and/or tables to confirm your observations.

- Select a trend model for each time series using an appropriate criteria. Are the models selected all the same? If not is there a pattern depending on the region and/or the group (max, min and mean)?

- After removing the trend using the model selected in the previous step, use the output to estimate the seasonality of each time series employing averaging and sine-cosine models. Compare your results and use appropriate plots and/or tables to confirm your observations.

- Select a seasonal model for each time series using an appropriate criteria. Are the models selected all the same? If not is there a pattern depending on the region and/or the group (max, min and mean)?

- Estimate a combined model for trend and seasonality using the results of the previous steps. Call this model “final”.

- Estimate trend and seasonality using a combined quadratic and sin-cosine (of order 2) models. Call this model “test”.

Task 5 – ARMA and Forecasting (20%)

- Using the final and the test model estimated in the previous task, remove trend and seasonality from each of the 30 time series. You will now have 60 residuals time series.

- Fit the residuals with an appropriate ARMA model.
- Forecast the average max, min and mean temperature for each month of 2020. Remember that you also have to forecast the trend and seasonal components.
- Compare your forecasts with the actual values. You may find it useful to look at the following link https://otexts.com/fpp2/accuracy.html. Which model performs better?

Report (10%)

The report (R Markdown) must be well structured, well written and grammatically correct. Titles, heading and figures should be correct and labelled in a meaningful way and referenced accordingly. The report should be no more than 20 pages. The report should not include code (use the option '`echo = FALSE`'), but just the output of your commands (including graphs and tables) and your written interpretation of them. Not all plots have to be included: choose the ones that you consider valuable for your report.

Note: Send a copy of the Rmd file to filippo.cavallari@southwales.ac.uk before the deadline. The subject of the email must be: "MS4S09 CW2 Markdown".

Marking Guidelines

	80-100	**70-79**	**60-69**	**50-59**	**40-49**	**30-39**	**0-29**
	Exceptional First	**First**	**Upper 2nd**	**Lower 2nd**	**Third**	**Narrow Fail**	**Fail**
Analysis outline	Professional outline of analysis presented.	Detailed purpose of analysis provided.	Adequate outline of analysis provided.	Outline of analysis provided but with some flaws.	Simple outline of analysis provided, but lacking key detail.	Inadequate outline of analysis provided.	No outline of analysis provided.
Methods used and assumptions made	Sophisticated investigation of assumptions and description of methods used.	Comprehensive investigation of assumptions and description of methods used.	Adequate and correct investigation of assumptions and description of methods used.	Investigation of assumptions and description of methods used is provided but with some flaws.	Limited investigation of assumptions and description of methods used.	Inadequate investigation of assumptions and description of methods used.	No investigation into assumption and no description of methods used.
Statistical Modelling	Sophisticated time series techniques utilised.	Comprehensive time series techniques utilised	Adequate time series techniques utilised.	Time series techniques attempted but with some flaws.	Limited time series techniques utilised.	Inadequate time series techniques utilised.	No time series techniques utilised.
Key results and correctness of content	Unanticipated results and implementations presented. Appropriate, substantial, correct and sophisticated nature.	Comprehensive results and implementations, presented and employed well. Appropriate, substantial and correct.	Expected results and implementations presented. All appropriate, largely correct, with few flaws.	Not all expected results and implementations presented. All appropriate, largely correct, with few flaws	Few or simple results and implementations presented. Much appropriate material, but flawed.	Seriously flawed results or no implementation. Appropriate but seriously flawed material.	No results or implementation. Incorrect or inappropriate content.
Conclusions	Deep and critical understanding provided.	Thorough understanding shown.	Good understanding shown.	Key concepts generally understood.	Some evidence of understanding.	Little of superficial understanding shown.	No evidence of understanding.
Report	Like a publishable report, virtually error-free.	Like a publishable report with isolated minor errors.	Can be followed easily with very few errors.	Can be followed easily with some weaknesses.	Can be followed with difficulty.	Poor structure or containing significant errors.	Unstructured and with many errors.