Data Science And Statistical Modelling In Space And Time

Assessment 2, 2025

This assessment consists of 3 questions:

1. Spatial modelling
2. Time series modelling
3. A combination of both

The total number of marks available is **100**, worth **80%** of the overall mark for MTHM505, and is split 35/35/30 between the 3 questions. Marks indicated for individual parts suggest the relative amount of detail required to answer questions.

The deadline for submission is 12 noon, 2nd May.

You should submit a single pdf at the ELE submission point containing all your solutions.

Commented `R` code (and any outcomes/plots) should be part of the answers, however only include `R` output that is helpful for answering the questions, and it should be clear from your answers which models you are fitting and why (i.e. don't **only** include code/plots), and ensure that plots are properly labelled and explained. Do not just submit code. The code should be there to help explain what you have done. We are looking for a good explanation in text for the models you have chosen and why.

You are expected to work independently - strict disciplinary action will be taken for any plagiarism. Late submissions will be penalised according to the University's late submission policy.

The data required in each question is found on the 'Assessment information and submission' tab on ELE.

All questions can be answered using models seen in the lectures and practicals. You may use any programming language or `R` package, however be careful that the code you are using is actually fitting the model that you think it is and answering the question - e.g., if the question says fit a 'Gaussian process with maximum likelihood', you won't get marks for fitting a different type of spatial model, or for using a different type of parameter estimation.

1. Sea Surface Temperature modelling [35 marks]

You have been given a dataset with 100 measurements of Sea Surface Temperature (SST) in the Kuroshio current off Japan from the first two days of January 1996. In the file `kuroshio100.csv`, each row contains the date/time of the observation, longitude and latitude of the observation station (a ship or buoy), the station id and the Sea Surface Temperature.

a) Plot and comment on the data. You might find it helpful to convert the data into a geodata object using `as.geodata()` *[3 marks]*

b) Select 5 points from the dataset at random, report the chosen stations (name, longitude, latitude, SST), and remove these from your training dataset for fitting models.

c) Using the sample variogram, comment on whether you need to set a maximum distance, and explain whether there should be a nugget in your model. Given this, fit a spatial model using the variogram/kriging approach. You may want to try different assumptions and see which one fits best. Clearly state what assumptions you are making about the trend/mean function, covariance function, and the nugget, and state *all* fitted model parameters. Validate your model. *[12 marks]*

d) Repeat part c), but instead fit a Gaussian Process model using maximum likelihood to estimate the parameters. *[9 marks]*

e) Now estimate the model parameters using a Bayesian approach with discrete priors. You may use your answer from d) to help set prior ranges. Include priors over the correlation length and nugget. Clearly state what modelling assumptions you are making, and compare the parameter estimates to those from part d). *[8 marks]*

f) Using your models from c), d) and e), predict precipitation at the 5 locations you removed. Compare your predictions. *[3 marks]*

Hint:

- The Bayesian approach can become extremely slow if you have multiple discrete priors and a large number of bins for each - it may be worth starting with a coarse discrete prior that allows you to fit the model relatively quickly, and then add more bins later if you have time.

2. The Atlantic Overturning Circulation [35 marks]

In this question, we are going to consider measurements of the Atlantic meridional overturning circulation (AMOC) at 26°N, a key component in the global ocean circulation and the climate system in general. The data are monthly values of the strength of the circulation starting in January 2017 and are stored in the variable MOCmean in the file MOC.Rdata. We are going to model the data and forecast ahead.

a) Plot the data and comment on any patterns/trends observed. *[3 marks]*

b) Fit appropriate ARMA **and** ARIMA models (both without seasonal components) to the dataset. Excludingthe last 8 months You may want to fit multiple models and select the best, justifying clearly why your chosen models are appropriate. (Do not use auto.arima!) *[12 marks]*

c) Average the dataset to quarterly means instead of monthly means, and find the most suitable ARMA **or** ARIMA **or** SARIMA model for this quarterly dataset (again excluding the last 8 months/2 quarters). *[8 marks]*

d) Fit a Dynamic Linear Model with a linear trend and a seasonal component, to both the original monthly dataset and the quarterly dataset from part c), again excluding the last 8 months/2 quarters). *[8 marks]*

e) Using your best models from each of b), c) and d), forecast the values of the overrturning anomaly for the last 2 quarters (8 months). Comment on your forecasts. *[4 marks]*

3. California daily temperatures [30 marks]

This question considers modelling maximum daily temperature in California. You have 2 files:

- `metadataCA.csv` containing longitude, latitude, elevation and place name for 11 sites in California.
- `MaxTempCalifornia.csv` containing maximum daily temperatures in degrees Celsius for each site from Jan 1, 2012 to Dec 30, 2012.

There are fewer individual parts in this question, but note that more marks are available for b) and c), and you should expect to carry out all the usual stages of modelling, e.g. making clear which model you are fitting and to which data, which assumptions you are making, etc. You should also perform appropriate validation checks for each model. Do not use auto.arima for fittingthe time series models.

a. Provide spatial and time series plots of the dataset, and comment on trends seen in maximum daily temperature in California in 2012. *[4 marks]*

b. Fit a spatial Gaussian process model using maximum likelihood to predict the maximum temperature in San Diego and Fresno on December 13th 2012. *[11 marks]*

c. Use time series modelling to produce forecasts of the maximum temperature in San Diego and Fresno for the following 2 sets of dates:

1) December 9th - December 13th
2) December 13th - December 17th

For the 2nd forecast period, you may simply refit the same models used for the 1st set of forecasts. *[11 marks]*

d. Compare your various predictions for the maximum temperature in San Diego and Fresno on December 13th, decide which model is best and discuss whether this is what you would have expected. Identify how prediction could be improved. *[4 marks]*