Data Science And Statistical Modelling In Space And Time

Assessment 2, 2023

This assessment consists of three questions:

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

The total number of marks available is **160**, worth **80%** of the overall mark for MTHM505, and is split 55/50/55 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, 21st April.

You should submit a single pdf at the BART 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.

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

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

1. Spatial modelling [55 marks]

You have been given a dataset with 220 measurements of total monthly precipitation in the Netherlands in September 2019. This dataset was downloaded from the Copernicus Climate Data Store [1]. In the file `netherlands.csv`, each row contains a station name, longitude and latitude of the observation station, and total precipitation for the month in millimetres.

a) Plot the data, produce numerical summaries, and comment on any spatial relationships seen in the data. You might find it helpful to convert the data into a geodata object using `as.geodata()` *[5 marks]*

b) Select 3 of the observed locations at random, report the chosen stations (name, longitude, latitude, precipitation), and remove these from the dataset used for training models. Label these 3 locations as A, B and C. You'll need these later. *[1 mark]*

c) Calculate and plot the sample variogram of the data. Justify whether you need to set a maximum distance prior to fitting a model, and comment on whether you need to include a nugget. *[4 marks]*

d) Fit a spatial model using the variogram. You may want to try several 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]*

e) Repeat part d), but instead fit a Gaussian Process model using maximum likelihood. *[12 marks]*

f) Use your chosen variogram and maximum likelihood models to predict precipitation at locations A, B and C. Compare the predictions between the models, and to the true values. *[3 marks]*

g) Using the model fitted by maximum likelihood, produce plots of the mean and variance over a 0.05 degree grid covering the input data. *[3 marks]*

h) Fit a Bayesian model using discrete priors. Fit a model with and without a nugget, and compare your posterior distributions to each other and to your earlier parameter estimates. For each model, produce predictions for locations A, B and C. *[15 marks]*

Hints:

- When setting priors for the Bayesian model, you may use your results from earlier parts of the question to set sensible ranges for the discrete priors.
- 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.

Reference:

[1] Copernicus Climate Change Service, Climate Data Store, (2021): Global land surface atmospheric variables from 1755 to 2020 from comprehensive in-situ observations. Copernicus Climate Change Service (C3S) Climate Data Store (CDS). DOI: 10.24381/cds.cf5f3bac (Accessed on 23-MAR-2023)

2. Time series modelling [50 marks]

In this question, we are going to be modelling the strength of the Atlantic Meridional Overturning Circulation (AMOC) ocean current, measured at 26.5°N. Data from the RAPID AMOC monitoring project is funded by the Natural Environment Research Council and are freely available from www.rapid.ac.uk/rapidmoc. The file `AMOCdata.csv` contains daily measurements of the AMOC strength between January 2010 and December 2020, measured in Sverdrups (Sv), where 1 Sv = 1 million cubic metres per second ($1Sv = 10^6 m^3 s^{-1}$).

a) Average the data to quarterly means, plot the data and comment on any patterns/trends observed. You might find it helpful to convert the averaged data to a time series object using `ts()` *[4 marks]*

b) Fit an appropriate ARMA or ARIMA model (without a seasonal component), and use this to forecast the AMOC strength for the next 4 quarters. You may want to fit multiple models and select the best, justifying clearly why your chosen model is appropriate. *[12 marks]*

c) Repeat b), but with a DLM instead. Include both a trend and a seasonal component, clearly describing any modelling decisions you've made. *[12 marks]*

d) Compare the forecasts of parts b) and c). *[3 marks]*

e) Return to the original data, and calculate monthly averages instead. Find an appropriate 1) ARMA/ARIMA/SARIMA model and 2) a DLM for this monthly dataset, and use each to predict the AMOC strength for the next 12 months. *[16 marks]*

f) Compare the results of e) to your quarterly forecasts. *[3 marks]*

3. General modelling [55 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.

a. Provide spatial and time series plots of the dataset, and comment on trends seen in maximum daily temperature in California in 2012. *[5 marks]*

b. Fit a spatial Gaussian process model using maximum likelihood to predict the maximum temperature in Napa and DeathValley on November 13th 2012. *[20 marks]*

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

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

For the 2nd forecast period, you may simply refit the same models used for the 1st set of forecasts. *[22 marks]*

d. Compare your various predictions for the maximum temperature in Napa and DeathValley on November 13th, decide which model is best and discuss whether this is what you would have expected. Identify how prediction could be improved. *[8 marks]*