Session 8: Cross-Validation Method

Learning Objectives

After this lecture you should be able to

- Describe the concept and main steps of cross-validation
- Define common predictive performance metrics
- Perform k-fold cross-validation of an R-INLA model (Practical 5)

Motivation

Now we've made some models, how do we compare and select our models?

- We could just compare model posterior parameters or goodness-of-fit
- However these are not directly comparable between different models!
- So how can we judge the hackathon later?

Model Performance

Cross-validation

Cross-validation

Cross-validation is a methodology used to evaluate model prediction performance by:

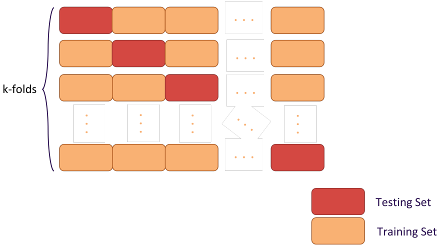
1. Splitting up the dataset into a training and a validation sample
2. Fitting the model using the training data
3. Evaluating the model against the validation set

This methodology allow us to evaluate how well our model reproduces the real-world quantities.

k-fold cross-validation

The most common method is the k-fold cross-validation.

Where the data is randomly partitioned into k non-overlaping subsets. The results from the different partitions are then combined to produce single estimations of the error.

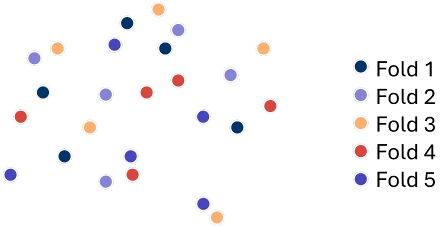


A typical value for k is 5 or 10, as these offer a good balance between reliable performance estimates and computational efficiency [Wikle et al., 2019]

Cross-Validation Method for Spatiotemporal Data

In a spatiotemporal framework, there are a number of ways to divide the data into these sets.

We will use randomised samples of the observed spatial locations.



Performance metrics

We can evaluate the predictive capability of competing spatiotemporal models using a number of different metrics.

For:

- m the total number of spatial locations
- n the total amount of time points
- y_{it} the observed data at i-th spatial location and t-th time.
- $\hat{y}_{it}$ the predicted value at i-th spatial location and t-th time.
- $\bar{y}$ the mean observed value

We define:

- Root Mean Squared Error: $RMSE = \sqrt{\frac{1}{mn}\sum_{i=1}^{m}\sum_{t=1}^{n}(\hat{y}_{it} - y_{it})^2}$
- Mean Absolute Error: $MAE = \frac{1}{m}\frac{1}{mn}\sum_{i=1}^{m}\sum_{t=1}^{n}|\hat{y}_{it} - y_{it}|$
- Bias: $Bias = \hat{y}_{it} - y_{it}$
- Simple correlation coefficients

References

Wikle, C. K., Zammit-Mangion, A., and Cressie, N. (2019).
Spatio-temporal statistics with R.
Chapman and Hall/CRC.