Modelling groundwater contamination: A comparison of spatial and spatio-temporal methods

M McLean*, A Bowman, L Evers

School of Mathematics and Statistics, University of Glasgow
*m.mclean.1@research.gla.ac.uk

Introduction & Aim

Modelling contaminated groundwater can be difficult due to the impracticalities of obtaining samples from every monitoring well at each sampling period. In this work we compare spatial and spatio-temporal modelling techniques on groundwater data to determine whether the added computational complexity of the spatio-temporal methods is justified with their benefit of being able to carry forward information from previous sampling periods over the more efficient spatial methods.

Methods

Kriging aims to derive the best linear unbiased predictor at an unmeasured location $\boldsymbol{s}_0$ given observations $\boldsymbol{Z} = (Z(\boldsymbol{s}_1), ..., Z(\boldsymbol{s}_n))$.

$$\begin{pmatrix} Z(\boldsymbol{s}_0) \\ \boldsymbol{Z} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_z \\ \mu_z \mathbf{1} \end{pmatrix}, \begin{pmatrix} C_z(0, \boldsymbol{\theta}) & \boldsymbol{c}_z(\boldsymbol{s}_0, \boldsymbol{\theta})^\top \\ \boldsymbol{c}_z(\boldsymbol{s}_0, \boldsymbol{\theta}) & \boldsymbol{\Sigma}(\boldsymbol{\theta}) \end{pmatrix} \right)$$

Where, μ_z is the mean of the measured locations, $C_z()$ is a covariance function, $\boldsymbol{c}_z(\boldsymbol{s}_0, \boldsymbol{\theta}) = (C_z(\boldsymbol{s}_0 - \boldsymbol{s}_1; \boldsymbol{\theta}), \ldots, C_z(\boldsymbol{s}_0 - \boldsymbol{s}_n; \boldsymbol{\theta}))$ and $\boldsymbol{\theta}$ are the covariance model parameters. Then:

$$\mathbb{E}[Z(\boldsymbol{s}_0)|\boldsymbol{Z}] = \mu_z + \boldsymbol{c}_z(\boldsymbol{s}_0, \boldsymbol{\theta})^\top \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}(\boldsymbol{Z} - \mu_z \mathbf{1})$$

Covariance functions considered: Exponential & Matèrn.[2]

P-Splines[3] use a b-spline basis to build a model of the form:

$$\boldsymbol{Y} = \boldsymbol{B}\boldsymbol{\alpha} + \boldsymbol{\epsilon}$$

Where, $\boldsymbol{Y} \in \mathbb{R}^n$, $\boldsymbol{B} \in \mathbb{R}^{nxp}$ is a matrix of b-spline basis functions and $\boldsymbol{\alpha} \in \mathbb{R}^p$ are the basis coefficients. The objective function to be minimised with respect to $\boldsymbol{\alpha}$ is:

$$S(\boldsymbol{\alpha}) = (\boldsymbol{Y} - \boldsymbol{B}\boldsymbol{\alpha})^\top (\boldsymbol{Y} - \boldsymbol{B}\boldsymbol{\alpha}) + \lambda \boldsymbol{\alpha}^\top \boldsymbol{D}_d^\top \boldsymbol{D}_d \boldsymbol{\alpha}$$

Where, $\boldsymbol{D}_d \in \mathbb{R}^{pxp}$ is the d^{th} order difference matrix and λ is a smoothing parameter. The fitted values are then computed as:

$$\hat{\boldsymbol{Y}} = \boldsymbol{B}(\boldsymbol{B}^\top \boldsymbol{B} + \lambda \boldsymbol{D}_d^\top \boldsymbol{D}_d)^{-1} \boldsymbol{B}^\top \boldsymbol{Y}$$

Data

Two sampling designs were used to simulate contaminated groundwater data[1], across 167 sampling periods, from the well network in figure 1. Design 1 mimicked a real life design with incomplete sampling periods (1400 obs) whilst Design 2 was a full design with complete sampling periods (4843 obs). Equally spaced test data were also simulated for model validation.

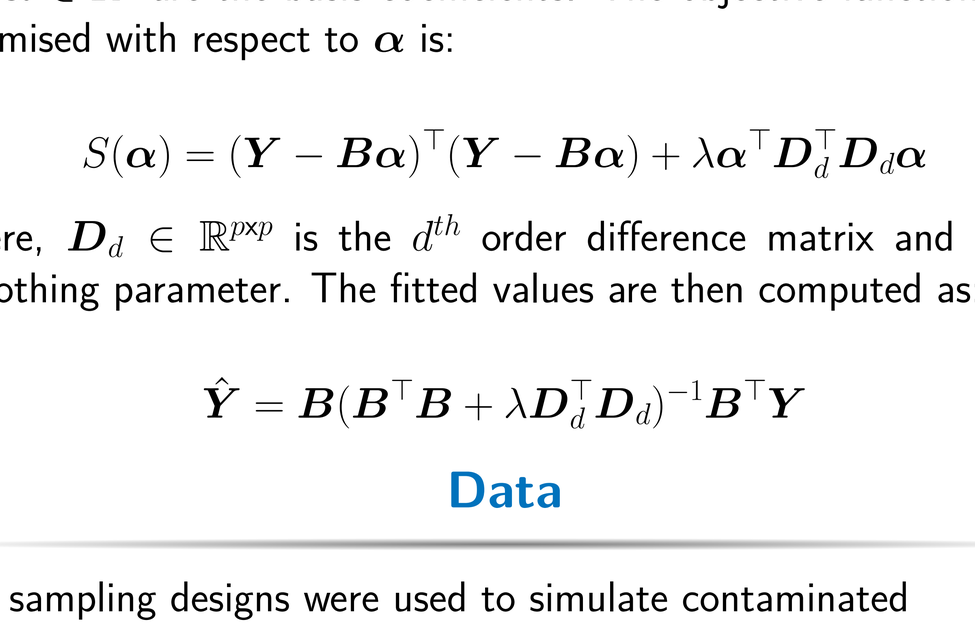

Figure 1: Well network, red wells are those sampled at time 167 under sampling scenario 1 (14 wells)

[1] Bowman, A. W., Evers, L., Jones, W. R., Molinari, D. A., Spence, M. J. (2013). Efficient and automatic methods for flexible regression on spatiotemporal data, with applications to groundwater monitoring. *Cornel University Library*, arXiv:1310.7815 . [2] Diggle, P., Ribeiro, P. J.,(2007). Model-based Geostatistics. *Springer series in Statistics*. [3] Eilers, P. H. C., Marx, B. D.,(1996). Flexible smoothing with b-splines and penalties. *Statistical Science*

Results

Predicted surfaces for sampling design 1:

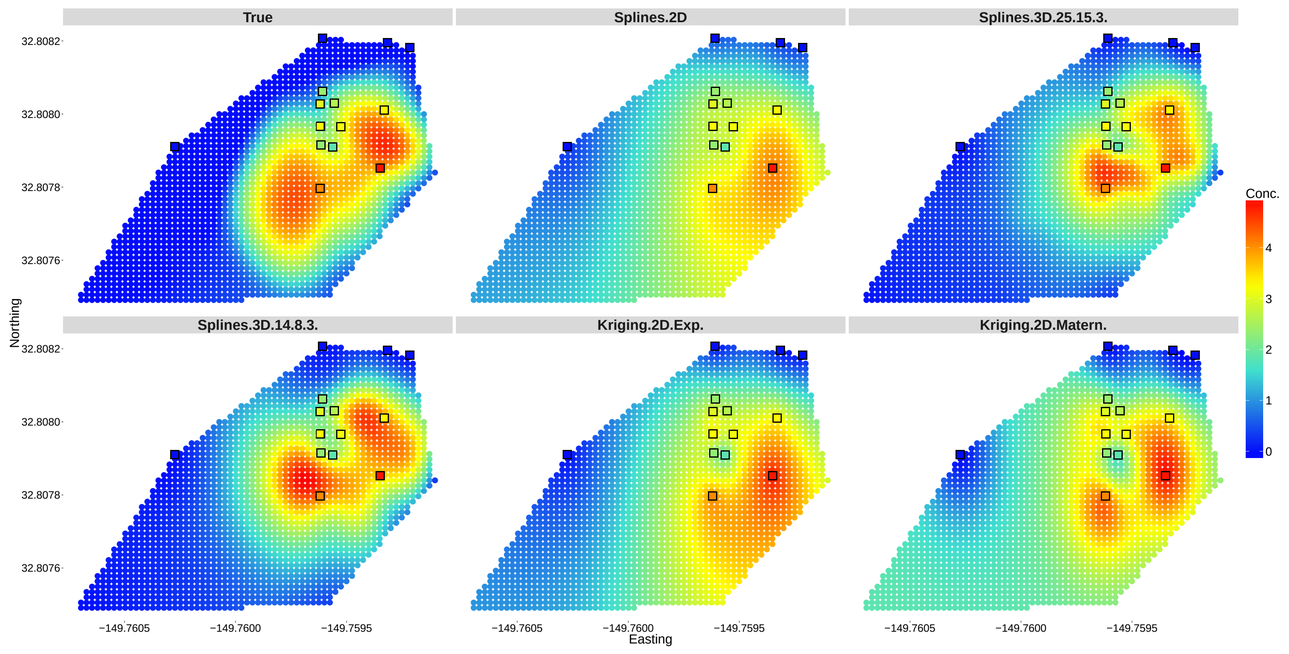

Figure 2: Predicted surfaces for test data at time 167 under the realistic design (sampling scenario 1). With the true surface obtained from the test data in the top left.

Predicted surfaces for sampling design 2:

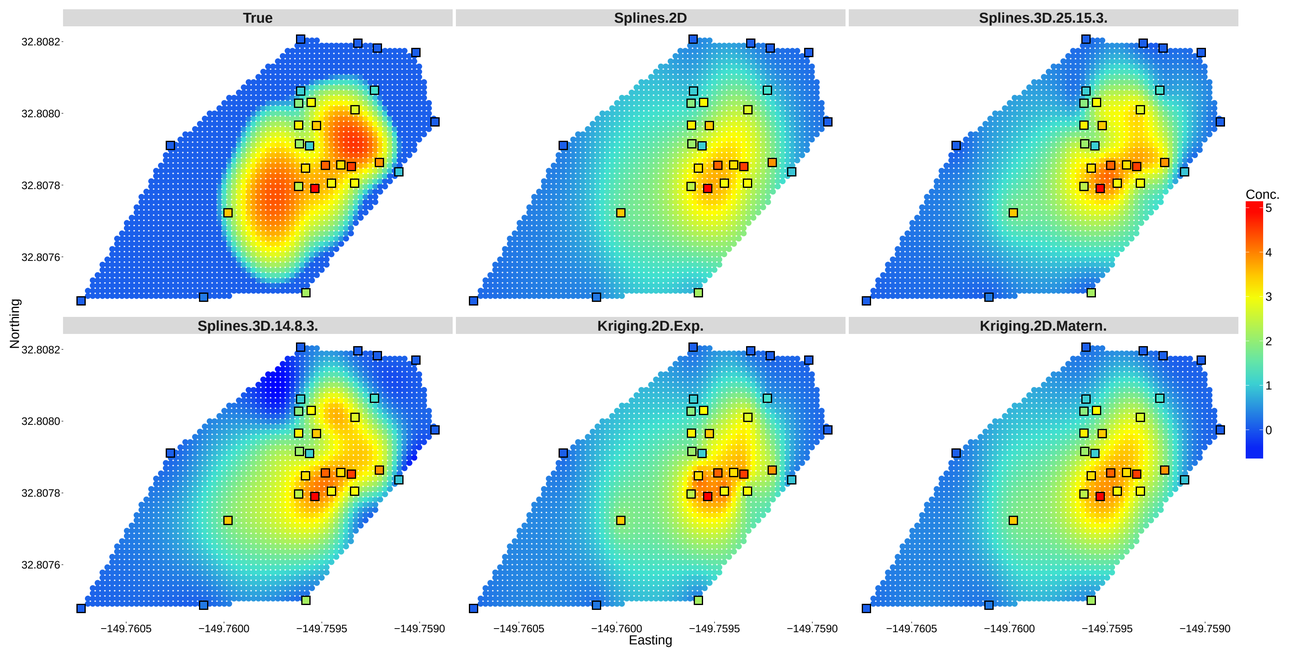

Figure 3: Predicted surfaces for test data at time 167 under the full design (sampling scenario 2). With the true surface obtained from the test data in the top left.

Figures 2 & 3 show that the 3D p-spline models best capture the true state for both designs, in line with the results in table 1.

Model	Design 1	Design 2
3D P-splines (25, 15, 3)	0.4876	0.4788
3D P-splines (14, 8, 3)	0.3730	0.6839
2D P-splines	1.0753	0.6682
Kriging - Matern	1.1667	0.6198
Kriging - Exponential	1.0670	0.6321

Table 1: Mean square prediction error (MSPE) values for the models predicting the test data at time 167. For 3D p-splines the numbers in brackets denote the number of basis functions used in each direction i.e. (easting, northing, time)

Conclusions

- The best method in terms of MSPE was 3D p-splines for both sampling designs.
- The spatial methods performed similarly for both designs with their predictions being better using the full design.
- However the spatial methods did not match the performance of the spatio-temporal methods for either design.