Outline of methods section for Emma's paper

April 25, 2023

Models considered

Empirical models are estimated for the variation of each of 14 geochemical quantities (each of which is represented generically by random variable Y) as a function of distance $d \in [0, 1800]$ km for five different models. Models are specified which explore the variation of Y with d in increasing complexity. The simplest model (C1C) assumes the existence of a single plume centre (at ***WHERE***), with respect to which d is defined for all three rifts; the variation of Y with d is assumed common to all rifts. Model C3C assumes the existence of three plume centres (at ***WHERE***); observations are allocated to the nearest plume centre, facilitating calculation of a single d for each observation; the variation of Y with d is assumed common to all rifts, regardless of plume allocation. Model C1D assumes one plume centre (like C1C) for calculation of d, but now the variation of Y with d is assumed to be different across rifts. Model C3D copies C3C for estimation of d, but variation of Y with d is assumed to be different across rifts. Finally, in model C3X we consider the presence of three plume centres, with different variation of Y with d for each combination of plume and rift.

Penalised B-splines

For each model, the variation of Y with d (possibly for a subset of the full sample) is described using a penalised B-spline (e.g. Eilers and Marx 1996, 2010), the characteristics of which are selected to provide optimal predictive performance. First, for a large index set of locations equally spaced on the domain of distance, we calculate a B-spline basis matrix $\boldsymbol{B}$ (e.g. de Boor 2001) for p equally-spaced cubic spline basis functions. Then the value of Y on the index set is given by the vector $\boldsymbol{B\beta}$, for spline coefficient vector $\boldsymbol{\beta}$ to be estimated. The value of p is specified to be sufficiently large to provide a good description of a highly variable Y. For a given data set, we penalise the difference between consecutive values in $\boldsymbol{\beta}$ using a roughness penalty, such that the penalised spline provides optimal predictive performance.

Estimating optimal spline roughness and predictive performance

For a sample of n_1 training data represented by the vectors of geochemical quantities and distances, $\boldsymbol{y}_1$ and $\boldsymbol{d}_1$, we first allocate each element of $\boldsymbol{d}_1$ to its nearest neighbour in the index set, and hence construct the appropriate spline basis matrix $\boldsymbol{B}_1$ for the sample. We then assume that $\boldsymbol{y}_1 = \boldsymbol{B}_1\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where the elements of $\boldsymbol{\epsilon}$ are independently and identically-distributed zero-mean Gaussian random variables. We penalise the roughness of β using a first-different penality $\lambda\boldsymbol{\beta}'\boldsymbol{P}\boldsymbol{\beta}$, where $\boldsymbol{P}$=$\boldsymbol{D}'\boldsymbol{D}$ and $\boldsymbol{D}$ is a first difference matrix (with elements $D_{ij} = -1$ if $i = j$; $= 1$ if $j = i + 1$; and $= 0$ otherwise; e.g. Jones et al. 2016). For a given choice of λ, we then find the optimal value of $\boldsymbol{\beta}$ by minimising lack of fit

$$\begin{aligned} \boldsymbol{\beta}^*(\lambda) &= \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\, (\boldsymbol{y}_1 - \boldsymbol{B}_1\boldsymbol{\beta})'(\boldsymbol{y}_1 - \boldsymbol{B}_1\boldsymbol{\beta})' + \lambda\boldsymbol{\beta}'\boldsymbol{P}\boldsymbol{\beta} \\ &= (\boldsymbol{B}_1'\boldsymbol{B}_1 + \lambda\boldsymbol{P})^{-1}\boldsymbol{B}_1'\boldsymbol{y}_1. \end{aligned}$$

We can evaluate the predictive performance of the resulting spline description using a tuning set of n_2 observations (independent of the training set) represented by vectors $\boldsymbol{y}_2$ and $\boldsymbol{d}_2$. We again start by finding the appropriate spline basis matrix $\boldsymbol{B}_2$ for this sample. Then we can calculate the predictive mean square error for the tuning sample

$$\mathrm{MSE}_2(\lambda) = \frac{1}{n_2}(\boldsymbol{y}_2 - \boldsymbol{B}_2\boldsymbol{\beta}^*(\lambda))'(\boldsymbol{y}_2 - \boldsymbol{B}_2\boldsymbol{\beta}^*(\lambda))$$

for each of a set of representative choices of values for λ. We can then select the optimal value of λ using

$$\lambda^* = \underset{\lambda}{\operatorname{argmin}}\, \mathrm{MSE}_2(\lambda).$$

The value $\mathrm{MSE}_2(\lambda^*)$ is a biased estimate of predictive performance, since the value of λ^* was tuned to minimise its value. We can obtain an unbiased estimate for the predictive performance of the spline model using a test set of n_3 observations (independent of the training and tuning sets) represented by vectors $\boldsymbol{y}_3$ and $\boldsymbol{d}_3$ (and corresponding spline basis matrix $\boldsymbol{B}_3$). Then the predictive performance is estimated using

$$\text{MSE} = \frac{1}{n_3}(\boldsymbol{y}_3 - \boldsymbol{B}_3\boldsymbol{\beta}^*(\lambda^*))'(\boldsymbol{y}_3 - \boldsymbol{B}_3\boldsymbol{\beta}^*(\lambda^*)).$$

Cross-validation and model comparison

We exploit cross-validation to evaluate MSE, by partitioning the full sample of data into $k > 2$ groups at random, withholding one group for tuning, another group for testing, retaining the remaining $k - 2$ groups for training. We then loop exhaustively over all possible combinations of choice of train, tune and test groups, evaluating overall predictive performance on the test data over all iterations, noting that each observation occurs exactly once in the test set. For models (C1D, C3D, C3X) requiring separate model fits to subsets of data, MSE is estimated using predictions from optimal predictive models for each subset. Further, we can repeat the analysis for different initial random partitioning of observations into k groups, to assess the sensitivity of overall predictive performance to this choice. We are careful to use the same cross-validation partitions to evaluate each of the five models, so that predictive performances can be compared fairly.

To quantify model performance over all 14 geochemical quantities, we define the overall standardised MSE

$$\text{SMSE} = \sum_{j=1}^{14} \frac{\text{MSE}_j}{s_j^2}$$

where MSE_j is the predictive performance for the j^{th} quantity, and s_j^2 is the sample estimate for the variance of that quantity.

Linear regression

For comparison, we also evaluate linear regression models for the variation of Y with d. In the current notation, these can be thought of as simple models with basis matrix $\boldsymbol{B} = [\mathbf{1} \quad \boldsymbol{d}]$, where $\mathbf{1}$ is a vector of appropriate length with each element $= 1$. $\boldsymbol{\beta}$ in this case is a 2-vector with elements corresponding to intercept and slope coefficients. Linear regression is approached using penalised B-spline models as the roughness coefficient $\lambda \rightarrow \infty$. That is, linear regression corresponds to a penalised B-spline model with very large λ. Therefore, a penalised B-spline model is guaranteed to perform at least as well as linear regression.

References

C. de Boor. *A practical guide to splines*. Springer-Verlag, New York (Applied Mathematical Sciences Revised Edition), 2001.

P. H. C. Eilers and B. D. Marx. Splines, knots and penalties. *Wiley Interscience Reviews: Computational Statistics*, 2:637–653, 2010.

P. H.C. Eilers and B. D. Marx. Flexible smoothing with b-splines and penalties. *Stat. Sci.*, 11:89–102, 1996.

M. Jones, D. Randell, K. Ewans, and P. Jonathan. Statistics of extreme ocean environments: non-stationary inference for directionality and other covariate effects. *Ocean Eng.*, 119:30–46, 2016.