

3. Title: Mapping Middle English Dialects

Proposer: Dr Geoff Nicholls

Description:

"A Linguistic Atlas of Late Mediaeval English" (see [1] for background) is a collection of data used by mediaevalists to 'locate' documents within England using dialect features of the writing. The data is a set of about 1000 scribal profiles. Each profile describes the dialect of a given manuscript (or part manuscript). For each of 238 words or parts of words, the profile gives the spelling variations used in the manuscript. A typical line might be

```
$daughter/npl DOUGHTEREN 3 DOUGHTERS 3 DOUGHTerEN 3 DOUGHTerES 3  
DOWGHTERS 3
```

for the plural we write 'daughters'. Manuscripts coming from similar locations use similar spellings. We can think of a scribal profile as a 238 component vector. At each point in space (i.e. location in mediaeval England) this vector takes some value representing spelling habits at that location, and the collection of vectors arranged in space are called a 'random field' - a realization of a stochastic spatial process. Now there are several problems. First, the 'training data' is 1000 or so vectors drawn from the random field. Only some of these vectors carry location data. The problem is to put the vectors back in the field so that the ones we haven't located somehow interpolate between the vectors we have located. If we have a distribution for the field then this is a well defined problem, from a Bayesian [2] point of view: draw a sample from an N-dimensional field distribution on a lattice. Now, remove some subset of the vectors from the lattice, recording the vector, but not its location. Can we put the removed-vectors back in the correct places in the lattice? We can use MCMC (Chapter 10 of [2]) to sample the posterior distribution for the vector locations. Both the idealized and the applied problems are of interest, and one may or may not suggest a solution to the other.

Prerequisite courses: BS2a (for the Bayesian inference and MCMC)

Data analysis/simulation project

[1] Benskin, M. (1991). *Regionalism in late medieval manuscripts and texts : essays celebrating the publication of A linguistic atlas of late mediaeval English*, Chapter The 'Fit'-Technique Explained, pp. 9-26. Cambridge : D.S. Brewer.

[2] O'Hagen, A and Forster, J (2004). 'Bayesian Inference' in Kendal's *Advanced Theory of Statistics*, Vol 2B, 2nd Ed. London: Arnold.