A multiscale method for spatial imputation applied to environmental data

Robert G. Aykroyd[1], Stuart Barber[1], Phil Northing[2], Alistair Murray[2] and Samuel J. Peck[1*]

[1] Department of Statistics, University of Leeds
[2] The Food and Environmental Research Agency, York

1 Introduction

In a spatio-temporal example from agriculture, concerning crop monitoring and pest control, it is required to fit surfaces describing covariates in a geographical region. We describe a method for imputing, or interpolating, values observed from functions on an irregularly spaced data grid using a Voronoi-based lifting scheme. The lifting scheme is a generalisation of wavelet decompositions used to obtain multiresolution analyses on irregular grids without restrictions on the number and spacing of data points (see Jansen *et al.* (2009), Sweldens (1997)). This allows us to identify activity with localisation in scale and position and also, under certain assumptions, gives us a method of minimising distortion due to noise.

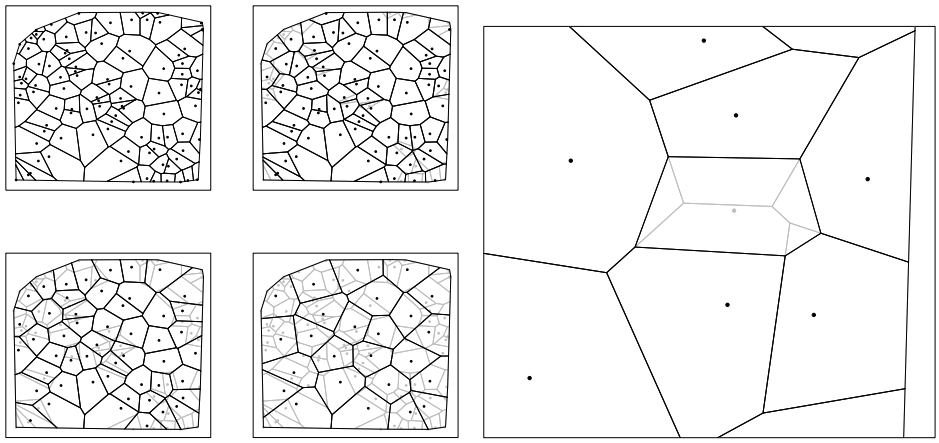
2 Method

Suppose we have some data, X_F, collected on an irregular two-dimensional grid, and for each data point, x_i, a corresponding function value, f_i, is observed with noise. We assume the model

$$f_i = g_i + \epsilon_i, \quad i = 1, \ldots, n,$$

where g_i are true function values, and $\epsilon_i \sim N(0, \sigma^2)$. We wish to make estimates of the function value at points where we have no observations, which we call X_M. To make these estimates, we first perform a lifting transformation on the combined grid, $X = (X_F, X_M)$. We place a mixture prior on the lifting coefficients and estimate the hyperparameters by an empirical Bayes approach similar to that used by Johnstone and Silverman (2005) in a wavelet context. We then estimate the lifting coefficients corresponding to grid points X_M utilising the sparsity property of the lifting transformation. The missing function values are then obtained by inverting the lifting transformation. We also obtain estimated posterior distributions for the missing function values, which are mixtures of the posterior coefficient distributions. It is possible to simulate directly from these mixture distributions; confidence intervals and other summaries can be derived from such samples. An advantage of this method over that propsed by Heaton and Silverman (2008) is that it does not require MCMC, hence providing good computational efficiency.

Figure 1 shows an illustration of a typical lifting decomposition. More specifically, figure 1(a) shows how fine-scale details (polygons) are removed early in the algorithm to leave polygons of approximately equal size nearer to the end of the algorithm. Figure 1(b) shows how the tesselation is updated when a point is removed. We predict the value of the removed point from its neighbours by using an average weighted by the proportion of its polygon allocated to those neighbours when it is removed.



(a) (b)

Figure 1: Progression of lifting decomposition. Panel (a): graphs showing an initial Voronoi tesselation (top-left) and updated versions at intermediate stages (in order, top-right, bottom-left, bottom-right). Panel (b): diagram showing how tesselation is updated when a point is removed – the grey point here is the removed point. This is a zoomed version of figure 1(a).

3 Results

We show an example of this method on a UK farm-based pest count data set which contains both densely and sparsely observed regions, ideal for the use of such multiscale methods. We build a two-dimensional surface of pest infestation levels by imputing on to a regular grid.

(a) (b)

Figure 2: Real data example: Pest infestation in UK with confidence intervals. Plot 2(a): Observation locations with intensity of infestation. Plot 2(b): Posterior mean surface using our lifting imputation method.

(a) (b)

Figure 3: Real data example: Pest infestation in UK with confidence intervals. Plots 3(a) and 3(b): Lower and upper bounds of 95% credible intervals for estimates derived from our method.

In figure 2, plot 2(a) shows the observed data with the level of infestation shown by the intensity and size of the dot – white is no infestation, black is heavy infestation. The other panel shows the imputed surface obtained using our method, as a portion of a 50×50 uniform grid. We see that the method picks up the high intensity region in the east, as well as the regions of very low intensity. Also, in areas where there is no data, the estimates are almost constant. Figure 3 shows the surfaces formed by the boundaries of 95% credible intervals derived from the posterior distributions.

Acknowledgements We thank both the EPSRC and the Food and Environment Research Agency (Fera; formerly Central Science Laboratory) for their funding of this work, and also Fera for providing data.

References

Heaton, T.J. and Silverman, B.W. (2008) A wavelet- or lifting-scheme-based imputation method. *Journal of the Royal Statistical Society: Series B*, **70**(3), 567–587

Jansen, M., Nason, G.P. and Silverman, B.W. (2009) Multiscale methods for data on graphs and irregular multidimensional situations. *Journal of the Royal Statistical Society: Series B*, **71**(1), 97–125

Johnstone, I.M. and Silverman, B.W. (2005) EbayesThresh: R Programs for Empirical Bayes Thresholding. *Journal of Statistical Software*, **12**(8), 1–38

Sweldens, W. (1997) The lifting scheme: A construction of second generation wavelets. *SIAM Journal on Mathematical Analysis*, **29**(2), 511–546