Journal of the Royal Statistical Society
SERIES C Applied Statistics

Appl. Statist. (2014)
63, *Part* 1, *pp.* 47–63

Flexible regression models over river networks

David O'Donnell, Alastair Rushworth, Adrian W. Bowman and E. Marian Scott

University of Glasgow, UK

and Mark Hallard

Scottish Environment Protection Agency, Stirling, UK

[Received July 2012. Final revision April 2013]

Summary. Many statistical models are available for spatial data but the vast majority of these assume that spatial separation can be measured by Euclidean distance. Data which are collected over river networks constitute a notable and commonly occurring exception, where distance must be measured along complex paths and, in addition, account must be taken of the relative flows of water into and out of confluences. Suitable models for this type of data have been constructed based on covariance functions. The aim of the paper is to place the focus on underlying spatial trends by adopting a regression formulation and using methods which allow smooth but flexible patterns. Specifically, kernel methods and penalized splines are investigated, with the latter proving more suitable from both computational and modelling perspectives. In addition to their use in a purely spatial setting, penalized splines also offer a convenient route to the construction of spatiotemporal models, where data are available over time as well as over space. Models which include main effects and spatiotemporal interactions, as well as seasonal terms and interactions, are constructed for data on nitrate pollution in the River Tweed. The results give valuable insight into the changes in water quality in both space and time.

Keywords: Flexible regression; Kernels; Network; Penalized splines; Smoothing; Spatial separation; Spatiotemporal models; Water quality

1. Introduction

Statistical models for data collected over a spatial region are widely available and heavily used in an enormous range of applications. However, the majority of these models assume that the spatial region of interest is a straightforward subset of $\mathbb{R}^2$ where Euclidean distance is the natural metric. One interesting example of spatial data which does not have these characteristics arises from measurements made over a network consisting of continuous, connected curved line segments. The sample space is intrinsically one dimensional, although embedded in two-dimensional space. River catchments are a particular, and commonly occurring, example of this. Fig. 1 illustrates both the network and a series of point sampling stations for the River Tweed, which spans the border between Scotland and England. (Note that the picture shows some apparently unconnected stream segments. This is simply because some small lochs and other types of water body are not shown.)

Address for correspondence: Adrian W. Bowman, School of Mathematics and Statistics, University of Glasgow, University Gardens, Glasgow, G12 8QQ, UK.
E-mail: adrian.bowman@glasgow.ac.uk

0035–9254/14/63047

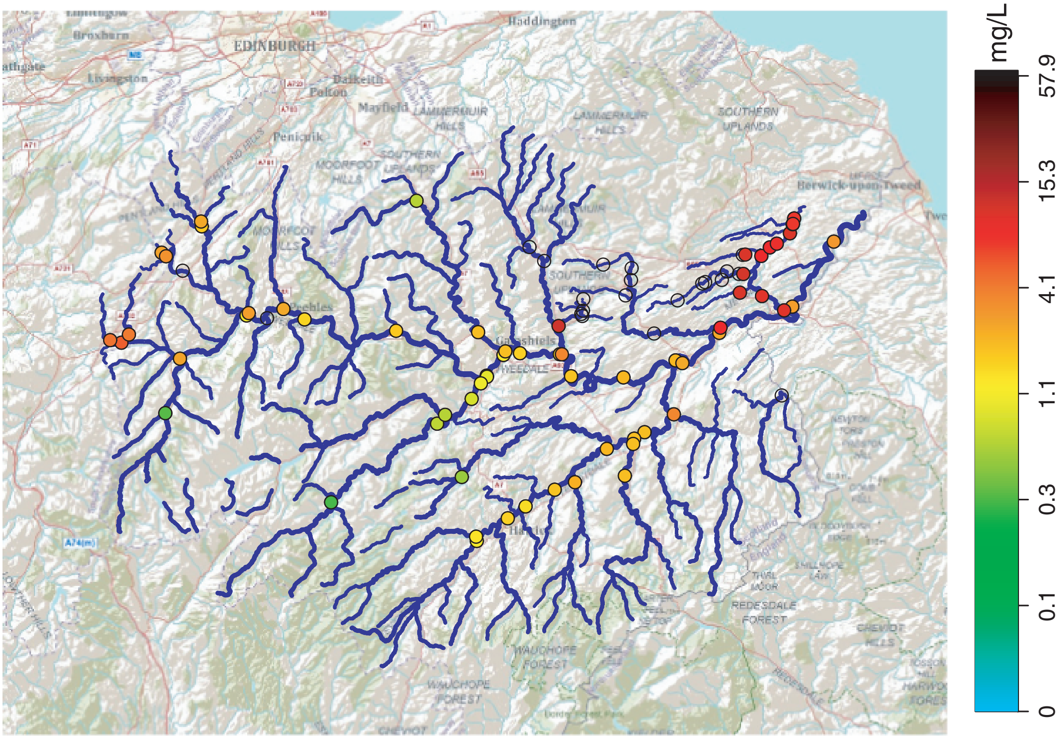


Fig. 1. River Tweed catchment, with sampling stations colour coded by nitrate level recorded in February 2004 (○, stations where no measurements were available at this time point): the scale of the map is approximately 110 km in each direction; the plot was produced by using the R package OpenStreetMap (Fellows, 2012) and the underlying map image was produced by Esri (www.esri.com) and its data providers

Models for this type of spatial data require different constructions. In particular, Euclidean distance needs to be replaced by 'stream distance', which was defined by Ver Hoef *et al.* (2006) as 'the shortest distance between two locations, where distance is only computed along the stream network'. This approach has been used in geostatistical models for stream networks for some time, e.g. by Cressie and Majure (1997) and Gardner *et al.* (2003). However, Ver Hoef *et al.* (2006) showed that substituting stream distance for Euclidean distance in standard geostatistical theory does not produce a valid spatial covariance model except when the exponential covariance structure is used. Ver Hoef *et al.* (2006) and Cressie *et al.* (2006) used moving average constructs to define a much broader class of valid spatial covariance models which use stream distance as well as other information, such as flow volume and the flow connectedness of locations. One of the defining properties of these models is that they assign a correlation of zero to pairs of locations which are not flow connected. Ver Hoef and Peterson (2010) developed the theory that had been set out in these earlier papers by defining both 'tail-up' and 'tail-down' moving average constructions, to allow for correlation between pairs of locations which are not flow connected. A variety of applications have subsequently been built on this theoretical structure; see Peterson and Ver Hoef (2010), Peterson and Urquhart (2006), Peterson *et al.* (2006) and Garreta *et al.* (2010) for examples.

Covariance functions, and the use of kriging for prediction at locations which have not been monitored, provide a very well-established approach to the construction of statistical models for spatial data. However, in some applications the principal focus is on the presence and nature of underlying trends, created by effects such as land use, geological patterns, dominant weather patterns or other influences with a strong systematic component which persists over repeated sampling of the same spatial region. Linear trends can be accommodated easily in covariance function models but in environmental settings trends often take the form of more flexible, non-parametric patterns. An attractive approach is then to place the emphasis on the direct modelling of these trends, using suitable forms of flexible regression, incorporating appropriate forms of spatial error where necessary. This line of thinking is also well established and expressed, for example, in the geoadditive models of Kammann and Wand (2003) and the more general semiparametric and additive modelling frameworks that have been described by Ruppert *et al.* (2003) and Wood (2006) among others. Bowman *et al.* (2009) described a model of this type for spatiotemporal data.

The aim of the present paper is to develop methods of flexible regression for data over a network. In common with all spatial models, a regression approach allows estimates to be constructed over an entire spatial region from point located data, but it also provides a framework within which spatial, temporal and other covariate effects can be treated simultaneously. Smoothing techniques form the basis of flexible regression methods and these have been applied to a variety of data structures. However, the published literature shows very little evidence of their use in a network setting. The challenge is to devise methods that are built on the concept of 'borrowing strength' locally, while respecting the specific topology of a network and the additional complications of directionality and size of flow. A key issue in addressing these issues is how to deal with confluence points, where different branches of the network combine. It is shown below that successful treatment of these issues leads to significant improvements over more standard smoothing techniques in this setting. In particular, the estimators exhibit features, such as sharp changes which are often expected at confluence points, but which cannot easily be reproduced by more standard approaches.

Monitoring systems which are designed to collect data spatially also commonly record data over time. In fact, in many applications, the detection of changes over time is equally important as the identification of spatial pattern and this has motivated a large body of research in spatiotemporal modelling. However, very little of this work is directed at a network setting. This provides an opportunity for a successful network flexible regression approach to be extended into the spatiotemporal setting, where spatial, temporal and interaction terms can all be identified informatively.

Different approaches to flexible regression over a network, including local fitting and penalized methods, are discussed in Section 3. A spatiotemporal model, including main effects and interactions, is constructed in Section 4 where a correlated error structure is also considered. Visualization of the complex nature of the interactions is also discussed here. Throughout the paper, the methods and models are applied to data from the River Tweed. Some final discussion is given in Section 5.

2. The Tweed data

The data that will be used in the analysis are supplied by the Scottish Environment Protection Agency (SEPA) and refer to the catchment of the River Tweed. The SEPA is responsible for the routine collection and evaluation of water quality data from Scotland's lochs, rivers and estuaries. The importance of this is underlined by European Union directives such as the Nitrates Directive, which was adopted in 1991, and the Water Framework Directive, which was adopted in 2002, which set targets in terms of water quality and ecological status. European Union members have committed to meet these targets and the collection and analysis of monitoring data are therefore essential.

Data on the Tweed catchment are available from January 1987 to August 2011 for 83 monitoring stations on the river. The timings of the observations are irregular, with frequencies which vary across the stations but are around one per month on average. Fig. 1 indicates the highly dendritic nature of the Tweed network and illustrates the monitoring stations by plotting nitrate measurements recorded in February 2004. Stations which were not monitored at this particular time point are shown as open circles.

Various chemical and biological measurements are available, but the most important is nitrate concentration. Diffuse pollutants such as sewage effluent and runoff from fertilizers are among the largest contributors to nitrate levels, which makes these an appropriate choice for the Tweed, which is surrounded by mainly arable land. Two different types of measurement are available, namely nitrate N and total oxidized nitrate TON, both measured in milligrams per litre. TON is the sum of nitrate and nitrite levels but, since the latter tend to be very small, TON and N are essentially equivalent and this will be assumed in the analysis. To improve normality and to stabilize the variance, nitrate levels will be analysed on the log-scale.

The analysis that is discussed in later sections of the paper will highlight the importance of measuring water flow. This is available at only a limited number of locations on the river and it would be impractical to measure this more widely. The SEPA therefore uses a hydrological model to estimate flow for each of the 298 separate stream segments. This is an adequate representation because, in later analysis, only relative flow across the stream segments is required. The widths of the stream segments that are shown in Fig. 1 are used to reflect these relative flow volumes. In cases where there is concern about the quality of flow information, an alternative is to follow Ver Hoef *et al.* (2006) by using a surrogate such as 'stream order', which indexes each stream segment by its location in the hierarchy of tributaries to the main river.

The Tweed catchment includes areas of outstanding natural resource which it is of high importance to monitor and preserve. However, the catchment also has considerable diversity of land use, including hill country, farmland and populated areas. The pattern of pollution is therefore likely to evolve differently over time across these land types and this motivates the need for a spatiotemporal description of pollution levels.

3. Network smoothing

There is a wide variety of approaches to the construction of smooth, flexible regression models. Wood (2006) provides a very useful starting point for the large associated literature, with Hastie and Tibshirani (1990), Bowman and Azzalini (1997), Schimek (2000) and Ruppert *et al.* (2003) providing earlier helpful overviews. Of the broad concepts involved, a common approach involves the use of local fitting through kernel functions or other forms of weighting schemes. Another involves the use of basis functions, often in conjunction with a penalty function to control smoothness. Both of these approaches are developed below for network data.

3.1. Kernel functions

When data $\{(y_i, x_i), i = 1, \ldots, n\}$ are to be modelled by a flexible regression $y_i = m(x_i) + \varepsilon_i$, where m denotes a smooth function and the ε_i denote error terms, a very simple approach to the estimation of m at any point of interest x is to compute a local average in the form

$$\hat{m}(x) = \frac{\sum_i w(x_i - x; h) y_i}{\sum_i w(x_i - x; h)},$$

where the weight function w decreases with distance from zero, at a rate that is determined by the parameter h, and so controls the degree of influence of each observation on the estimate. A normal density function with standard deviation h is a convenient choice for w. It turns out that fitting a local linear regression rather than a local mean has better theoretical properties, as described by Fan and Gijbels (1996) for example. However, where the data are sparse, as occurs in some areas of the Tweed catchment, this can cause stability problems, so the simple local mean is considered here.

One attraction of this approach is that the generality of the underlying idea allows it to be modified to suit data of very different types. In the network setting, these modifications follow the patterns of the spatial models for networks that were outlined in Section 1. If the locations of the observed values are denoted by s_i, and the location of the point of estimation by s, then the appropriate covariate values x_i and x now refer to 'river distance', using the river mouth as a natural origin. Secondly, the weights are non-zero only where there is a flow path between s_i and s. Thirdly, additional weights should be used to reflect the volume of water flowing in different sections of the river. This is exactly the approach that was proposed by Ver Hoef *et al.* (2006) but employed in the context of a regression model rather than spatial prediction through kriging. The flow weighting is derived from a 'tail-up' model, using the terminology of Ver Hoef and Peterson (2010), which assumes that points which are not flow connected are uncorrelated. Specifically, the additional weights are expressed in the function

$$\delta_i(x) = \begin{cases} \prod_{k \in B_{s,s_i}} \sqrt{\omega_k} & \text{if } s \text{ and } s_i \text{ are flow connected,} \\ 0 & \text{if } s \text{ and } s_i \text{ are not flow connected} \end{cases}$$

where B_{s,s_i} is the set of all stream segments between and including the locations s and s_i. Here a stream segment refers to a stretch of water between two neighbouring confluence points. The quantity ω_k denotes the proportion of flow contributed by water stretch k to its subsequent confluence. It contributes on the square-root scale to stabilize variance across the contributing stream segments at confluence points, as discussed by Ver Hoef *et al.* (2006).

An estimate of the mean value at s, located at a distance x from the mouth of the river, is then available as

$$\hat{m}(x) = \frac{\sum_i w(x_i - x; h)\, \delta_i(x) y_i}{\sum_i w(x_i - x; h)\, \delta_i(x)} = \sum_i v_i y_i,$$

where v_i denotes the combined effect of the weighting schemes, incorporating river distance, flow connectedness and flow values. A vector of estimated values can then be written as Sy, where the rows of the smoothing matrix S contain the weights that are applied to the data vector y to construct an estimate at a particular location. Where the estimation points are set to the observed locations, the trace of this 'hat matrix' has the interpretation of an approximate degrees of freedom, as discussed by Hastie and Tibshirani (1990) and many others. Further details of this approach are described in O'Donnell (2012).

3.2. Penalized splines

An alternative approach to the estimation of smooth functions is to use a set of basis functions,

$\phi_j(x)$, $j=1,\ldots,p$, as the components in a regression model, representing the estimate in the form $\hat{m}(x)=\Sigma_j \beta_j \phi_j(x)$, with coefficients β_j. This has the advantage that the estimate can be expressed in proper functional form simply through the specification of the β_js and that the dimensionality of the estimation problem can be kept low, independent of the size of the data set. A convenient choice of basis functions is B-splines, whose construction from polynomial pieces gives them many attractive computational properties, as described by de Boor (1978). The approach to smoothing which is known as P-splines, which was proposed by Eilers and Marx (1996) and is now used by many other researchers, uses a rich B-spline basis but imposes a roughness penalty on the coefficients β_j.

The construction of a basis set over a network faces the difficulty of combining the basis components in a suitable manner at the confluence points. This is possible, but slightly awkward, with the usual pattern of smooth overlapping functions. An attractive alternative is to divide the network into a large number of small pieces within which the function m is likely to change very little. In the river setting, this arises naturally through the identification by the SEPA of 'stream units' corresponding to short water stretches which are judged to be relatively constant in terms of environmental conditions. There may be several stream units within each stream segment (the stretch of water between two adjacent confluence points). As the sizes of the stream units are very small compared with the network as a whole, a regression model can be constructed in a piecewise constant manner through a set of mean values β_j, $j=1,\ldots,p$, that are associated with the p stream units which make up the network. The estimate at any point of interest s is then simply the estimated value of β_j, where j indexes the stream unit in which s lies. If the number of stream units is large then the loss in resolution through the approximation of m in this piecewise constant manner is very small. This is, in fact, equivalent to the use of B-splines of order 0.

There are likely to be considerably more stream units than observed values and so the estimation process is ill defined. However, a penalty approach immediately overcomes this difficulty and also provides a means of controlling the smoothness that is exhibited by the estimate. The 'smoothness' of β-values corresponding to adjacent stream units j and k, with no intervening confluence, can be measured by $(\beta_j-\beta_k)^2$. Where a confluence point is involved, as illustrated in Fig. 2, the measure of smoothness needs to reflect the relative levels of flow in the contributing

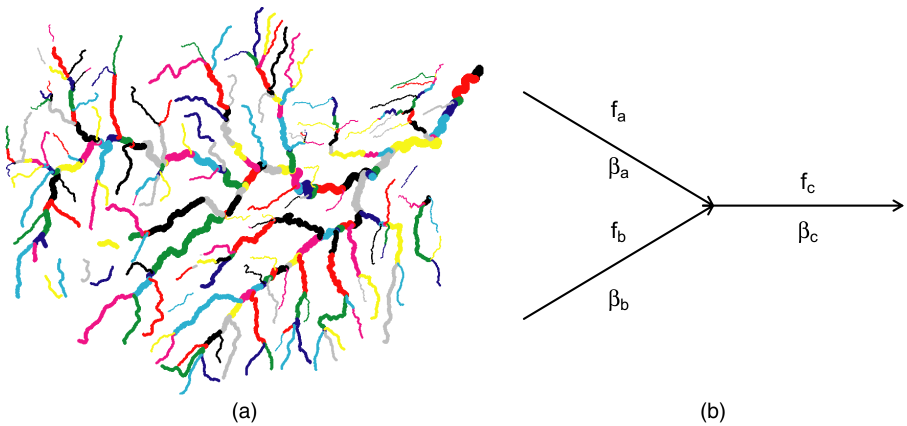


Fig. 2. (a) Decomposition of the river network into a large number of small stream units by using different colours and (b) a schematic representation of a confluence, with model parameters (β_a, β_b), flows (f_a, f_b) and the corresponding outgoing versions (β_c, f_c)

streams a and b. If the flows are denoted by f_{a}, f_{b} and f_{c}, then we expect $f_{\mathrm{c}} = f_{\mathrm{a}} + f_{\mathrm{b}}$ and the mixing of pollutants to be controlled by the relative flows of the inputs, $\omega_{\mathrm{a}} = f_{\mathrm{a}}/f_{\mathrm{c}}$ and $\omega_{\mathrm{b}} = f_{\mathrm{b}}/f_{\mathrm{c}}$. Following the principle of mass balance, the combined pollution input $\omega_{\mathrm{a}}\beta_{\mathrm{a}} + \omega_{\mathrm{b}}\beta_{\mathrm{b}}$ and the output β_{c} are identical if $\omega_{\mathrm{a}}(\beta_{\mathrm{a}} - \beta_{\mathrm{c}}) + \omega_{\mathrm{b}}(\beta_{\mathrm{b}} - \beta_{\mathrm{c}}) = 0$. Smoothness across the confluence can therefore be achieved through the penalty

$$\lambda\{\omega_{\mathrm{a}}^2(\beta_{\mathrm{a}} - \beta_{\mathrm{c}})^2 + \omega_{\mathrm{b}}^2(\beta_{\mathrm{b}} - \beta_{\mathrm{c}})^2\}, \tag{1}$$

where λ controls smoothness through the weight that is attached to the penalty. This has the attractive form of combining penalties for smoothness across each flow path of the confluence, with weights determined by the relative flow volumes.

The form of this model does not depend directly on any measure of spatial location s or stream distance x from the river mouth. Instead, dependence is modelled through a mixing process on the pollutant concentrations, expressed through the difference penalty. The use of the first-order penalty is a natural way of describing mixing, not only by imposing smoothness within stream segments but also across confluences, where it is reasonable to assume that outputs, conditioned on their immediate upstream neighbours, are independent of all further upstream neighbours.

A P-spline model can be formulated as $y = B\beta + \varepsilon$, where y, β and ε denote the vectors of responses, parameters and errors respectively, whereas the design matrix B is simply an $n \times p$ indicator matrix whose ith row has the value 1 in the column corresponding to the stream unit of y_i and 0s elsewhere. Following Eilers and Marx (1996), the model is fitted by minimizing the penalized sum of squares

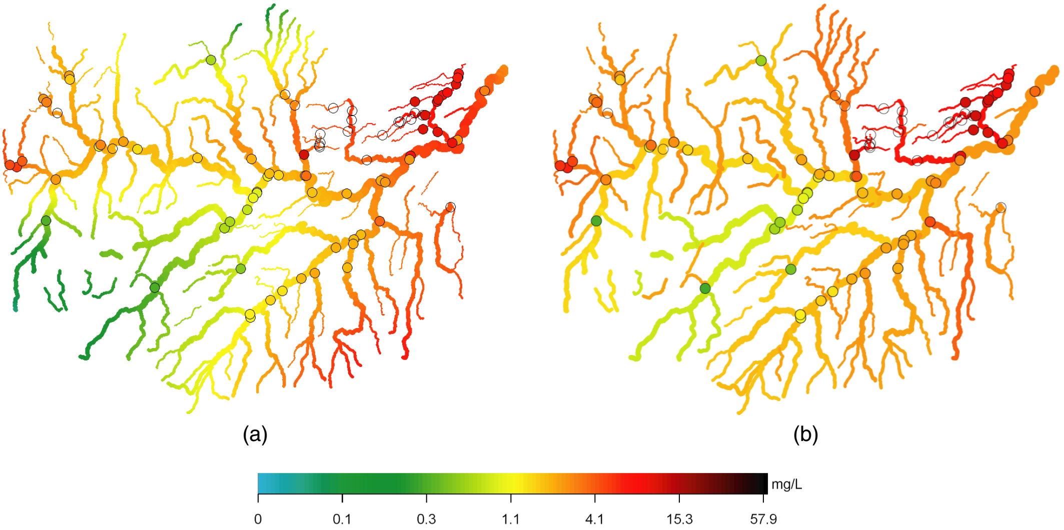
$$(y - B\beta)^{\mathrm{T}}(y - B\beta) + \lambda\beta^{\mathrm{T}}D^{\mathrm{T}}D\beta$$

with respect to β. Here D is the matrix which generates the differences between βs from adjacent stream units, weighted by flow where this is an intervening confluence point, as in the components of penalty (1). The penalty parameter λ controls the degree of smoothing. The solution to this least squares problem is easily shown to be $\hat{\beta} = (B^{\mathrm{T}}B + \lambda D^{\mathrm{T}}D)^{-1}B^{\mathrm{T}}y$. The linear form of this expression again allows an approximate degrees-of-freedom value to be computed as the trace of the hat matrix.

Fig. 3 shows the effects of smoothing on the average nitrate measurements from February 2004. Fig. 3(b) shows the effects of P-spline smoothing with 12 degrees of freedom. The ability to view the spatial structure over the whole network is a significant benefit, in comparison with plots of the original point-based observations shown in Fig. 1. In contrast with the use of standard two-dimensional Euclidean smoothing, which is shown in Fig. 3(a), proper recognition of the network structure gives appropriate weighting to observations from stream segments of different size and allows quite sharp changes in estimated level. For example, the relatively high concentrations of pollution that are exhibited by some of the tributaries in the northern periphery of the network are not immediately inherited by the larger and relatively unpolluted streams into which they flow, as a result of dilution effects. This behaviour reflects what we believe to be happening in the river but it cannot be captured by methods which ignore the special structures of a network.

3.3. Comparison of the two approaches

The use of kernel functions rather than P-splines produces an estimate which is qualitatively similar to Fig. 3(b). This confirms the general view that the particular form of construction of smoothing techniques is relatively unimportant—it is the choice of degree of smoothing which matters.


(a)

(b)

Fig. 3. Smoothing based on (a) Euclidean distance and (b) flow-weighted distance by using data from February 2004 and with 12 degrees of freedom

However, a comparison of these two approaches to smoothing over a network indicates that the P-spline method has some advantages. It characterizes the levels of pollution over the entire network through a set of parameters, one for every stream unit, which makes it particularly easy and efficient to identify estimated values at any location of interest, simply by identifying the stream unit in which it lies. In contrast, local mean estimation requires a new calculation for each new estimation point. The P-spline estimate can handle regions where the data are sparsely located, because the penalty can bridge the gap between data points, whereas local mean estimation can run into difficulties with very small weights. The P-spline construction is based simply on the relationship between values and flows in neighbouring stream units, so there is no need to define connectedness across the whole network simultaneously or to consider the cumulative effect of flow across distances which span several confluences. There is some potential loss of information with the use of stream units in the P-spline approach because the lengths of the stream units are not used. However, this loss of information is likely to be very small, precisely because the stream units have been defined as homogeneous stretches of water.

In view of these issues, further modelling work will use the P-spline approach. From a computational point of view, the kernel and P-spline methods are similar in their demands. However, within the context of larger models involving time, seasonal and possibly other covariates, to be discussed below, the direct nature of the estimation process that is involved in P-splines leads to much greater computational efficiency. This considerably strengthens the case for the use of P-splines for network smoothing.

With any method of flexible regression, the degree of smoothing, which expresses the complexity of the model, is an important choice. There is a very large literature on how this might be done automatically, using general principles such as cross-validation or an information criterion such as the Akaike information criterion AIC. Hurvich *et al.* (1998) proposed a version of AIC where the optimal value of λ is chosen to minimize

$$\mathrm{AICc} = \log(\hat{\sigma}^2) + 1 + \frac{2 + 2\,\mathrm{dof}}{n - \mathrm{dof} - 2}$$

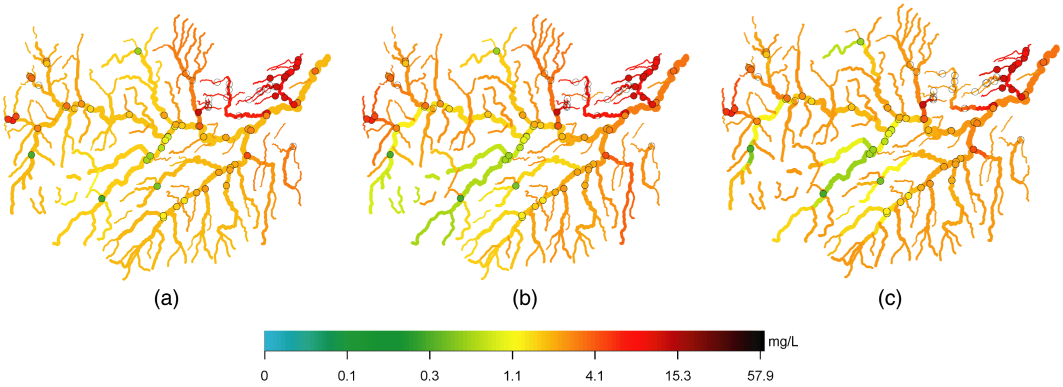


Fig. 4. Network smoothing with (a) 6, (b) 19 (optimal) and (c) 48 degrees of freedom, using the data from February 2004

where dof denotes the approximate degrees of freedom associated with λ. This imposes an additional penalty for large degrees of freedom and is designed to avoid the undersmoothing that might result in such cases when criteria such as generalized cross-validation or standard AIC are used.

However, there can also be merit in considering the effects of different levels of smoothing, in a sensitivity or multiresolution analysis. Expressing flexibility through degrees of freedom offers a very convenient scale on which to explore this. Fig. 4 shows the network estimates based on very small (6), AICc-optimal (19) and much larger (48) degrees of freedom. Low degrees of freedom allow very little movement away from the overall average whereas large degrees of freedom simply track the observed data, as shown for example by the stream segment in the central north part of the network where a single observation determines the estimated level. In contrast, the optimal value achieves a good balance between fit to the data and the identification of smooth trend.

4. Spatiotemporal models for networks

If the aim is to estimate the levels of pollution over a network at a single time point, then there is little to choose between kriging and flexible regression. Indeed, from a single realization of a spatial process it is not feasible to separate persistent trend from transient spatial variation. However, where observations are also available over time, the nature of spatial and temporal effects and their potential interactions become of considerable interest and, in contrast with approaches based on covariance functions, flexible regression methods can be extended to this setting relatively straightforwardly.

There are, of course, very many current examples of spatiotemporal models. Cressie and Wikle (2011) described a wide variety of modelling tools, and many researchers have described applications in areas such as air pollution (Guttorp *et al.*, 1994; Shaddick and Wakefield, 2002), rainfall (Brown *et al.*, 2001), snow water (Huang and Cressie, 1996) and fish population size (Reyjol *et al.*, 2005). There are also many examples of river data being modelled through space and time, as described for example by Cressie and Majure (1997), Clement and Thas (2007), Akita *et al.* (2007) and Thorp *et al.* (2006). However, these examples do not consider the essential network features of river distance, flow connectedness and flow weighting. Only a very small number of references consider a river network structure for data through space and time. Money *et al.* (2009) considered the use of the tail-up structure in space–time models, using a Bayesian–maximum entropy method of fitting, but the sparsity of flow-connected pairs in their data set makes the use of this information infeasible. Gardner and McGlynn (2009) also used the tail-up model for nitrate data from the Rocky Mountains, but the analysis is primarily based on a spatial analysis for each of a small set of time points over a period of approximately 1 year, to highlight seasonal changes.

Three principal variables need to be accommodated in a spatiotemporal model. One is space expressed through the river network locations s_i, the second is time t_i measured on a scale of years to express long-term trends, and the third is time within the year, z_i, to express the seasonal changes which are very often exhibited in environmental measurements. Additive models are natural tools to consider, as they provide a framework within which flexible regression can be extended to a wide variety of data structures. Hastie and Tibshirani (1990) provided an early synthesis of this approach which has encouraged wide use of these models in many application areas. Wood (2006) provided a modern overview which markedly extends the range of the tools that are available.

In the present setting, a very simple additive model is

$$y_i = \mu + m_s(s_i) + m_t(t_i) + m_z(z_i) + \varepsilon_i \tag{2}$$

where the three functions m_s, m_t and m_z describe spatial, temporal and seasonal trends and ε_i denotes error terms assumed to have an $N(0, \sigma^2)$ distribution marginally. If each of the trend functions is estimated by B-splines then, following the derivation in Section 3.2 above, they can be represented as $B_s\beta_s$, $B_t\beta_t$ and $B_z\beta_z$ where the columns of the design matrices evaluate each basis function at the observed values of the relevant covariate. B-splines of order 0 can be used for the spatial network, whereas cubic B-splines would be a good choice for the temporal and seasonal effects, as these are defined over more standard sample spaces. The full model can be represented as $y = B\beta + \varepsilon$, where B combines the columns of the individual design matrices, with an initial column of 1s, and β is the vector of combined parameters.

It remains to construct suitable penalty terms to induce smoothness on the estimates of the trend functions. First-order differences were the natural choice for the spatial network parameters, as described in Section 3.2 and computed through a difference matrix. For cubic B-splines, second-order differencing of the parameter vector is the more standard choice. The smoothness penalty can then be expressed as $\beta^{\mathrm{T}} P \beta$, where the matrix P has block diagonal form which combines the individual penalties as $(0, \lambda_s D_s^{\mathrm{T}} D_s, \lambda_t D_t^{\mathrm{T}} D_t, \lambda_z D_z^{\mathrm{T}} D_z)$. Cyclical behaviour in the seasonal term can be induced by requiring the coefficients of the first r basis functions to be identical with the last r basis functions. The penalty $\Sigma_{k=1}^{r} (\beta_{z,k} - \beta_{z,p+1-k})^2$ achieves this, with $r = 3$ for cubic splines. This can be adopted in the definition of D_z.

In the presence of an overall mean parameter μ in model (2), the identifiability of each additive component can be achieved by the addition of a ridge penalty, as described by Eilers and Marx (2002). This corresponds to a penalty of the form $\beta^{\mathrm{T}} Q \beta$, where Q is a diagonal matrix constructed from the vector $(0, \nu_s \mathbf{1}_s, \nu_t \mathbf{1}_t, \nu_z \mathbf{1}_z)$, with the ridge parameters denoted by ν_s, ν_t and ν_z and with $\mathbf{1}_a$ denoting a vector of 1s whose length is determined by the number of basis functions in the term denoted by a. The fitted model can then be expressed through the parameter estimates $\hat{\beta} = (B^{\mathrm{T}} B + P + Q)^{-1} B^{\mathrm{T}} y$. Denoting this as Hy, standard errors for $\hat{\beta}$, and so for fitted values, are available from the diagonal elements of HH^{T}, multiplied by an estimate of the error variance which is constructed as $\hat{\sigma}^2 = \mathrm{RSS}/(n - \mathrm{dof})$ where RSS denotes the residual sum of squares and dof the approximate degrees of freedom for the model. A penalized spline approach to (generalized) additive modelling was described by Marx and Eilers (1998), and many subsequent researchers including Wood (2006), where further details are available.

The additive model (2) is a natural starting point but it is implausible that the spatial pattern of pollution will change in exactly the same way over time, or throughout the year, at every location. It is therefore more appealing to consider an interaction model of the form

$$y_i = \mu + m_s(s_i) + m_t(t_i) + m_z(z_i) + m_{s,t}(s_i, t_i) + m_{s,z}(s_i, z_i) + m_{t,z}(t_i, z_i) + \varepsilon_i, \qquad (3)$$

where the functions $m_{s,t}$ and $m_{s,z}$ encapsulate the adjustments that are required to capture how the time trend and seasonal effects vary over the river network. The term $m_{t,z}$ allows an adjustment to the overall seasonal component, allowing different patterns in different years. The interaction terms can also be conveniently represented in spline basis form, this time by using a basis that is formed by all possible products of the spline basis functions on each separate variable. More precisely, we can write $m_{s,t} = \Sigma_{j,k} \beta_{jk} \phi_{s,j} \phi_{t,k}$, where $\phi_{s,j}$ and $\phi_{t,k}$ denote B-spline functions for space and time. Since the spatial basis is constructed from B-splines of order 0, this has the simple interpretation that the parameters associated with each stream unit are now allowed to evolve smoothly over time. Corresponding structures and interpretations can be adopted for the space–season and time–season interaction terms. In matrix notation, the model matrix is

$$B = (\mathbf{1} \quad B_s \quad B_t \quad B_z \quad B_s \Box B_t \quad B_s \Box B_z \quad B_t \Box B_z)$$

where '□' is the row-wise tensor product defined as $A \Box Y = (A \otimes \mathbf{1}') \odot (\mathbf{1}' \otimes Y)$ and '$\odot$' denotes the Hadamard (elementwise) product; see Eilers *et al.* (2006).

Smoothness in the model terms is induced by applying appropriate penalties, and corresponding penalty parameters λ_i, for each term. In the case of main effects, these are constructed through the difference matrices that were described above. Penalties for the interaction terms can be constructed by considering the coefficients $\{\beta_{jk}\}$ in matrix form and applying smoothness penalties to both the rows and the columns. For example, space–time smoothness is induced by applying a first-order network penalty to the columns of the matrix $\{\beta_{jk}\}$ and a second-order difference penalty over the rows. As described above, identifiability is ensured, and ill conditioning avoided, by adding a ridge penalty for each term in the model, expressed in a diagonal matrix Q. Other constraints or penalties exist that could have achieved a similar effect, e.g. constraints that force the mean value of each component to be 0. However, the straightforward specification of the ridge penalty and the subsequent retention of sparseness of model objects make this a convenient choice, as will be discussed in Section 4.1. Each of the λ_i is estimated by a short search procedure to find values that minimize the corrected AIC as defined in Section 3.3.

Fig. 5 shows the results of fitting this interaction model to the Tweed data, using AICc to select all the penalty parameters. Fig. 5(a) together with Figs 5(c) and 5(d) show estimates of the main effects for space, year and day of the year. This highlights that areas of high pollution are present in the tributaries to the north-east of the River Tweed. Across the years the overall levels of pollution are relatively stable, but with some indication of a slight decreasing trend. The overall seasonal effect is strong, as expected, with a gentle decrease from February to August and a sharper rise at the end of the calendar year. The shaded bands in Fig. 5 correspond to 2 standard errors on either side of the estimates, under an independence assumption, and these indicate high precision, as a result of the substantial size of the data set. Fig. 5(b) shows the estimate of interaction between year and season. The values of the adjustments that are plotted here are small, indicating that the change in seasonal pattern over the years is modest. Figs 5(e) and 5(f) show fitted values at four specific spatial locations, along with a comparison of a simple main effects model (the broken curve) and the interaction model (the full curve). This shows clear improvement at some sites as a result of fitting the interaction terms.

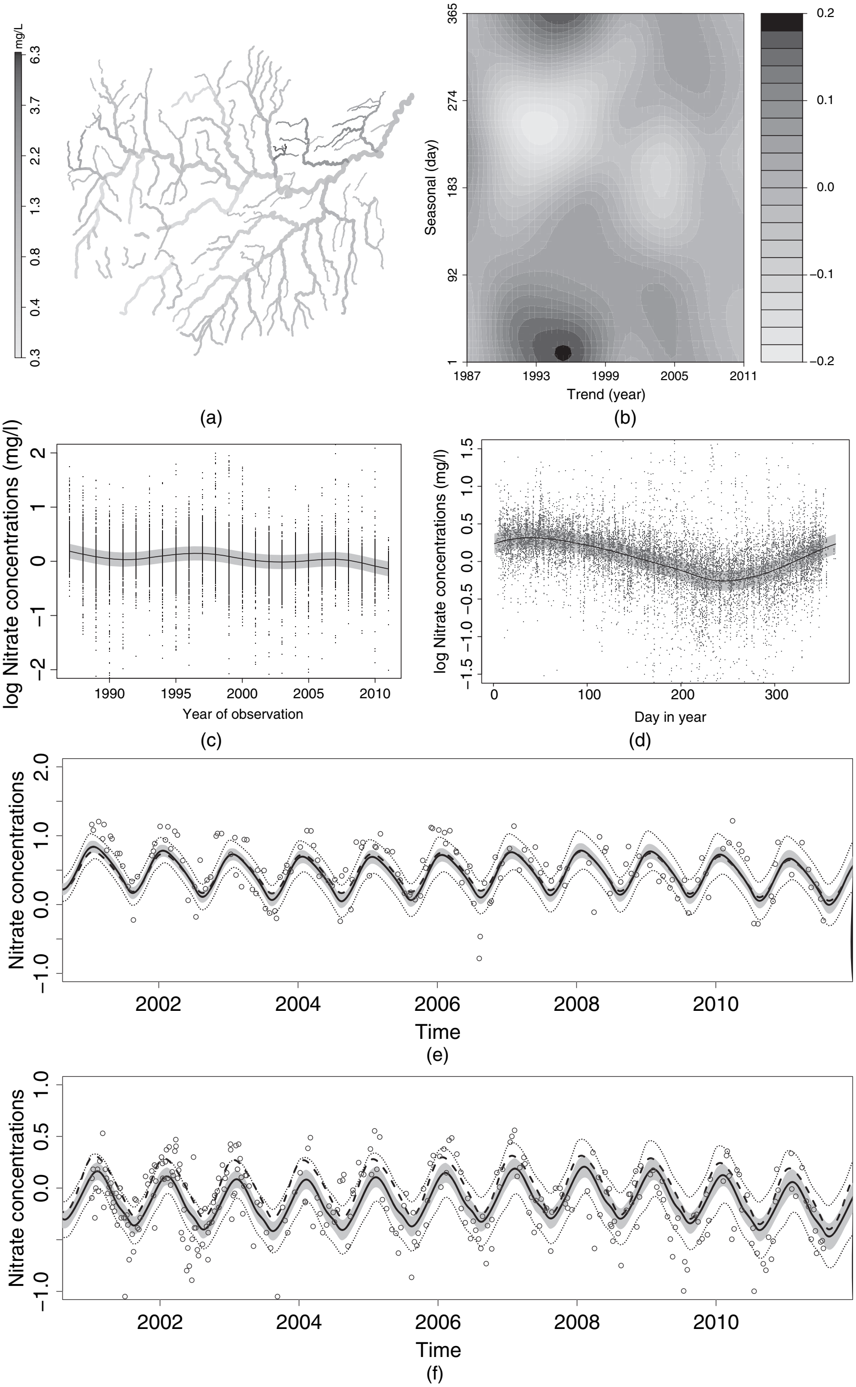


Fig. 5. (a)–(d) Main effects of space, year and day of the year, plus the interaction of these last two terms, and (e), (f) fitted values at two specific spatial locations, namely Gala Water Foot and Norham Gauging Station respectively, including a comparison of the simple main effects model (– – –) and the interaction model (———) in each case: · · · · · ·, 2 standard errors under the fitted covariance model; ▨, 2 standard errors under the independence model

4.1. Computational details

Tensor product spline smooths such as those specified in Section 3 rely on many basis parameters to represent each bivariate interaction and may therefore be intensive to fit. In the present case, the model matrix B is mostly composed of terms involving B_s, which is an $n \times p_s$ matrix where p_s is the size of the network partition. B_s contains exactly n non-zero entries. If V denotes the $n \times q$ matrix evaluating the bases for the other terms in the model, then $B_s \square V$ is at least $100(1-1/p_s)\%$ sparse and much more so if V is sparse. Sparse matrix algorithms can be used to decrease storage requirements and vastly increase performance. For example, the B that was fitted to the Tweed data was 12628×5960 where $p_s = 298$ and was 99.8% sparse.

The starting point in solving systems of sparse matrices is to define all of the model components in a compressed format. The `spam` or `Matrix` packages in R (R Development Core Team, 2011) perform this and allow the fitting of models with large amounts of data or complex structures. All model summaries and standard errors can then be easily and efficiently calculated by using a sequence of operations on the Cholesky factor L of $B^{\mathrm{T}}B + P + Q$ to obtain the quadratic form of interest, where $B^{\mathrm{T}}B + P + Q = LL^{\mathrm{T}}$. For example, the standard errors of the fitted values $\mathrm{se}(\hat{y}_i)$ are the diagonal elements of $B^{\mathrm{T}}L^{-\mathrm{T}}L^{-1}BB^{\mathrm{T}}L^{-\mathrm{T}}L^{-1}B$ and can be calculated efficiently by solving first $AL = B$, then $CL^{\mathrm{T}} = A$ and then summing the elementwise product of CB^{T} with itself.

4.2. Residual correlation

Having established a spatial interaction model for the River Tweed nitrate data that is appropriate for its network structure, it remains to check the assumption of independence that is made of the residuals. Evidence of residual temporal correlation at short time lags is to be expected, particularly as the model accounts for trends over longer time periods. Under the assumption of independent errors, all standard error estimates are likely to be underestimated when the underlying error process is correlated, so they must be adjusted appropriately. As a conservative measure, it was decided to fit a separable spatiotemporal model to the errors so that

$$\hat{\Sigma}_{ij} = \mathrm{cov}(\varepsilon_i, \varepsilon_j) = \omega_{ij}\sigma^2 \exp\left(-\frac{d_{ij}}{\rho} - \frac{|t_i - t_j|}{\psi}\right),$$

where $\omega_{ij} = \Pi_{k \in N}\,\omega_k$ and k indexes the set of stream units that lie between i and j and on the same flow path as both. The spatial and temporal correlation in the error process is assumed to depend on $t_i - t_j$, the time lag, and d_{ij}, the network separation measured in numbers of stream units. The correlation model was fitted by weighted least squares. Plots of the pairwise residual products against the fitted (weighted) covariance function showed that the model provides an adequate description of the correlation structure.

Having obtained an estimate for $\hat{\Sigma}$, the standard errors for the fitted values were then adjusted by

$$\mathrm{se}(\hat{y}) = \surd \mathrm{var}(\hat{H}y) = \surd \mathrm{diag}(\hat{H}\hat{\Sigma}\hat{H}^{\mathrm{T}}),$$

where $\hat{H}$ is the projection or hat matrix given by $B(B^{\mathrm{T}}B + P + Q)^{-1}B^{\mathrm{T}}$. The estimated parameters in the correlation model were $\rho = 13.3$ and $\psi = 27.4$ which represent moderate residual temporal correlation and (after adjusting with weights) weak residual spatial correlation. These parameters refer to a spatial scale in miles relative to a catchment diameter of approximately 100 km, and a temporal scale in days, relative to a span of 26 years for the whole data set. The overall variance parameter σ^2 was estimated as 0.1554, which is very close to the estimate under an independence assumption (0.1442). The corresponding adjustments to standard errors are displayed in Fig. 5 as broken curves, from which it is clear that the increases in width over the independence model are not sufficiently large to lead to any substantive change in conclusions.

It would be possible to consider incorporating the correlation structure into the fitting process for the model. This would, however, considerably increase the complexity of the computations, particularly as sparsity would be compromised. The post-fitting adjustment approach combines computational efficiency with an effective first-order approximation to the correlation structure, which has been used to good effect in similar settings, as discussed by Giannitrapani *et al.* (2011).

4.3. Visualization

Simple spatial terms can be plotted in map or network form but interactions with spatial components are more problematic to view. Figs 5(e) and 5(f) show temporal effects at particular point locations. An alternative illustrated in Fig. 6 is to display the estimated spatial effects at different time points, here at three different months (January, May and October) in 2005. This helpfully focuses attention on the spatial areas where seasonal change is strongest. However, changes in colour alone can be difficult to assess, especially where those changes are modest. The plots shown in Fig. 6 represent the values over the network as 'nodes', plotted approximately in the geographical midpoint of each stream unit. In addition to colour code, each node has radius proportional to the estimated nitrate pollutant level (on the original rather than log-scale). This form of display is particularly effective at illustrating changes over time as small changes in size are more easily identifiable than small changes in colour.

A more satisfactory solution involves animation of the spatial pattern across time. This kind of effect can be achieved with graphical tools such as those provided by the `rpanel` package (Bowman *et al.*, 2007) for R (R Development Core Team, 2011). This allows the time setting for the spatial display to be controlled through a slider. In a similar manner, sliders can also be used to control the degrees of smoothing through interactive selection of values for the approximate degrees of freedom. Since visualizing and understanding spatiotemporal model fits is challenging from static printed plots, two animations of the fitted models are provided in the on-line supplementary material. These illustrate spatial and temporal variation in the fitted mean nitrate levels in both network and node form. The effect of the spatial penalty across

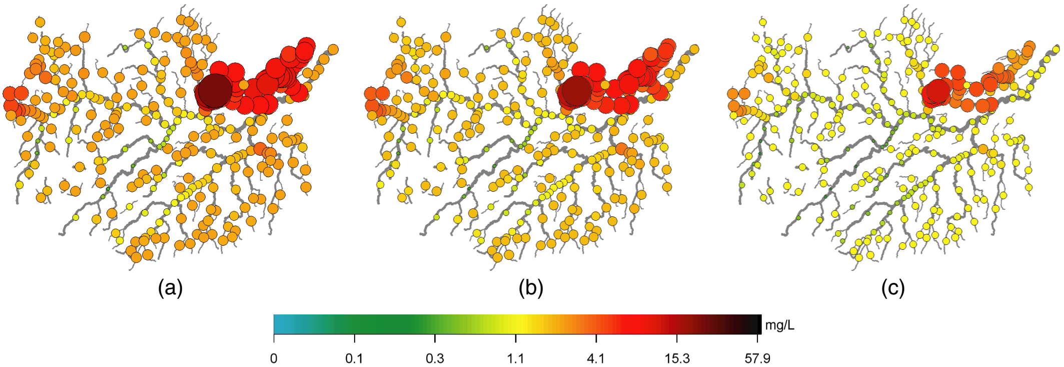


Fig. 6. Estimated spatial effects at (a) January, (b) May and (c) October 2005, indicated by colour and scaling of 'nodes' located at the stream units

neighbouring stream units is more evident in the first, whereas the degree of pollution and change through time is arguably better represented by the second.

5. Discussion

This paper has proposed flexible regression models which respect the unique spatial structure of data which arise from a directed network and which allow appropriate spatially varying coefficient models for capturing spatial change. In addition, the methods are capable of capturing the complex temporal changes which can occur in environmental pollutant concentrations. The results from the River Tweed data show evidence of a strong overall seasonal effect but only a small degree of overall change in nitrate levels over the 26-year period. Of greater interest is the insight that is gained from interaction terms, as these allow the inclusion of spatially varying effects which also respect the nature of the network. It is clear from Fig. 5 that local effects account for much of the variation that is seen at a specific site over time. This may support an argument in favour of the retention of a robust, well-designed long-term monitoring network which can detect changing local environmental pressures.

The penalized models that were described here use a discrete approximation to what is essentially a continuous spatial process which, in combination with the assumption of conditional independence between neighbouring sites, renders all model components sparse and computationally straightforward to store and manipulate. The sparseness property allows complex network models to be constructed and further covariates to be included without the computational limitations which can hamper other approaches. The issue of computational speed is also relevant to the method that is used to select smoothing parameters, especially as there are many examples of data sets with much larger numbers of stream segments than those considered for the Tweed. For example, it is popular to formulate P-splines as mixed models (Ruppert *et al.*, 2003) with smoothing parameters representing variance components which can be estimated by restricted maximum likelihood. We plan to investigate this approach but the computational issues, and the retention of sparsity, require careful construction.

The penalized model specification has an interpretation as a Bayesian hierarchical model in which the β_i are treated as random effects with a Gaussian Markov random-field prior and a fixed variance, controlling pairwise differences on neighbouring stream units. There is a particularly strong connection with conditional auto-regressive models, which are often used in applications such as disease mapping, where spatial dependence is defined in a similar manner through an adjacency metric. A Bayesian approach has attractive features, particularly because sampling from the full posterior distribution allows uncertainty associated with smoothing parameters to be integrated out. At present, smoothing parameters are selected through a grid search over a range of candidate values, which can be cumbersome for multi-dimensional smooths or many covariates. In addition, for Markov chain Monte Carlo updates, the conditional independence structure of the random effects β_i lends itself to the efficient block updating for Gaussian Markov random fields described by Fahrmeir and Lang (2001). This approach will be the focus of future research.

Attention has focused on the estimation of model terms and their standard errors, as these give clear and interpretable insight into the structure of the data. If more formal methods of model comparison are required, these can be implemented through approximate F-tests as described by Bowman *et al.* (2009) in the spatiotemporal setting.

Regardless of estimation procedure, average flow values are required for each stream unit in the network partition. Flow data that were used here were not observed but modelled and supplied by the SEPA. Observed flow data would allow a model to adapt to different flow settings over time, as observed by Cressie and O'Donnell (2010). However, here it is the flow ratios which are the crucial quantities.

The choice of spatial metric, namely river distance in the kernel approach or separation by a number of stream units in the P-splines approach, represents the belief that pollution changes in a slow and consistent way along stream segments. An argument might also be made for a metric with a Euclidean component. This could be valuable where the surrounding land is the source of pollution, rather than point sources, as land characteristics are naturally mapped on a Euclidean scale. It would be possible to use a combination (weighted average) of river distance and Euclidean distance, as in Cressie *et al.* (2006) for spatial prediction. O'Donnell (2012) has explored this in the context of kernel methods. Alternatively, land use information could be included as covariates where such data are available.

Acknowledgements

David O'Donnell and Alastair Rushworth gratefully acknowledge support from Engineering and Physical Sciences Research Council studentships (EP/P501423/1 and MCA06-2008/2009). Some of the research was conducted while Alastair Rushworth was seconded to the SEPA under a project that was funded by the Scottish Sensors Systems Centre. The work also benefited from Adrian Bowman's participation in the discussions on spatiotemporal models under project MTM2008-02901 from the Spanish Ministry of Science and Innovation. The comments of the Joint Editor and Associate Editor and two reviewers were very helpful in revising the paper.

References

Akita, Y., Carter, G. and Serre, M. L. (2007) Spatiotemporal nonattainment assessment of surface water tetrachloroethylene in New Jersey. *J. Environ. Qual.*, **36**, 508–520.

de Boor, C. (1978) *A Practical Guide to Splines*. New York: Springer.

Bowman, A. and Azzalini, A. (1997) *Applied Smoothing Techniques for Data Analysis*. Oxford: Oxford University Press.

Bowman, A., Crawford, E., Alexander, G. and Bowman, R. W. (2007) rpanel: simple interactive controls for r functions using the tcltk package. *J. Statist. Softwr.*, **17**, 1–18.

Bowman, A. W., Giannitrapani, M. and Scott, E. M. (2009) Spatiotemporal smoothing and sulphur dioxide trends over Europe. *Appl. Statist.*, **58**, 737–752.

Brown, P. E., Diggle, P. J., Lord, M. E. and Young, P. C. (2001) Space–time calibration of radar rainfall data. *Appl. Statist.*, **50**, 221–241.

Clement, L. and Thas, O. (2007) Spatio-temporal statistical models for river monitoring networks. *J. Agric. Biol. Environ. Statist.*, **12**, 161–176.

Cressie, N., Frey, J., Harch, B. and Smith, M. (2006) Spatial prediction on a river network. *J. Agric. Biol. Environ. Statist.*, **11**, 127–150.

Cressie, N. and Majure, J. J. (1997) Spatio-temporal statistical modeling of livestock waste in streams. *J. Agric. Biol. Environ. Statist.*, **2**, 24–47.

Cressie, N. and O'Donnell, D. (2010) Comment: Statistical dependence in stream networks. *J. Am. Statist. Ass.*, **105**, 18–21.

Cressie, N. and Wikle, C. K. (2011) *Statistics for Spatio-temporal Data*. New York: Wiley.

Eilers, P., Currie, I. and Durbán, M. (2006) Fast and compact smoothing on large multidimensional grids. *Computnl Statist. Data Anal.*, **50**, 61–76.

Eilers, P. and Marx, B. (1996) Flexible smoothing with b-splines and penalties. *Statist. Sci.*, **11**, 89–102.

Eilers, P. and Marx, B. (2002) Generalized linear additive smooth structures. *J. Computnl Graph. Statist.*, **11**, 758–783.

Fahrmeir, L. and Lang, S. (2001) Bayesian inference for generalized additive mixed models based on Markov random field priors. *Appl. Statist.*, **50**, 201–220.

Fan, J. and Gijbels, I. (1996) *Local Polynomial Modelling and Its Applications*. London: Chapman and Hall.

Fellows, I. (2012) OpenStreetMap: access to open street map raster images. *R Package Version 0.2*. (Available from `http://CRAN.R-project.org/package=OpenStreetMap`.)

Gardner, K. K. and McGlynn, B. L. (2009) Seasonality in spatial variability and influence of land use/land cover and watershed characteristics on stream water nitrate concentrations in a developing watershed in the rocky mountain west. *Wat. Resour. Res.*, **45**, article W08411.

Gardner, B., Sullivan, P. and Lembo, A. (2003) Predicting stream temperatures: geostatistical model comparison using alternative distance metrics. *Can. J. Fish. Aquat. Sci.*, **60**, 344–351.

Garreta, V., Monestiez, P. and Hoef, J. M. V. (2010) Spatial modelling and prediction on river networks: up model, down model or hybrid? *Environmetrics*, **21**, 439–456.

Giannitrapani, M., Bowman, A. and Scott, E. (2011) Additive models for correlated data with applications to air pollution monitoring. In *Statistical Methods for Trend Detection and Analysis in the Environmental Sciences* (eds R. Chandler and E. Scott), ch. 7, pp. 267–282. Chichester: Wiley.

Guttorp, P., Meiring, W. and Sampson, P. (1994) A space-time analysis of ground-level ozone data. *Environmetrics*, **5**, 241–254.

Hastie, T. and Tibshirani, R. (1990) *Generalized Additive Models*. London: Chapman and Hall.

Huang, H. and Cressie, N. (1996) Spatio-temporal prediction of snow water equivalent using the Kalman filter. *Computnl Statist. Data Anal.*, **22**, 159–175.

Hurvich, C. M., Simonoff, J. S. and Tsai, C.-L. (1998) Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *J. R. Statist. Soc.* B, **60**, 271–293.

Kammann, E. E. and Wand, M. P. (2003) Geoadditive models. *Appl. Statist.*, **52**, 1–18.

Marx, B. and Eilers, P. (1998) Direct generalized additive modeling with penalized likelihood. *Computnl Statist. Data Anal.*, **28**, 193–209.

Money, E., Carter, G. P. and Serre, M. L. (2009) Using river distances in the space/time estimation of dissolved oxygen along two impaired river networks in New Jersey. *Wat. Res.*, **43**, 1948–1958.

O'Donnell, D. (2012) Spatial prediction and spatio-temporal modelling on river networks. *PhD Thesis*. University of Glasgow, Glasgow.

Peterson, E. E., Merton, A. A., Theobald, D. M. and Urquhart, N. S. (2006) Patterns of spatial autocorrelation in stream water chemistry. *Environ. Monit. Assessmnt*, **121**, 571–596.

Peterson, E. E. and Urquhart, N. S. (2006) Predicting water quality impaired stream segments using landscape-scale data and a regional geostatistical model: a case study in Maryland. *Environ. Monit. Assessmnt*, **121**, 615–638.

Peterson, E. E. and Ver Hoef, J. M. (2010) A mixed-model moving-average approach to geostatistical modeling in stream networks. *Ecology*, **91**, 644–651.

R Development Core Team (2011) *R: a Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Reyjol, Y., Fischer, P., Lek, S., Rosch, R. and Eckmann, R. (2005) Studying the spatiotemporal variation of the littoral fish community in a large prealpine lake, using self-organizing mapping. *Can. J. Fish. Aquat. Sci.*, **62**, 2294–2302.

Ruppert, D., Wand, M. P. and Carroll, R. (2003) *Semiparametric Regression*. London: Cambridge University Press.

Schimek, M. G. (ed.) (2000) *Smoothing and Regression: Approaches, Computation, and Application.* New York: Wiley.

Shaddick, G. and Wakefield, J. (2002) Modelling daily multivariate pollutant data at multiple sites. *Appl. Statist.*, **51**, 351–372.

Thorp, J., Thoms, M. and Delong, M. (2006) The riverine ecosystem synthesis: biocomplexity in river networks across space and time. *Riv. Res. Applic.*, **22**, 123–147.

Ver Hoef, J. M. and Peterson, E. E. (2010) A moving average approach for spatial statistical models of stream networks. *J. Am. Statist. Ass.*, **105**, 6–18.

Ver Hoef, J. M., Peterson, E. and Theobald, D. (2006) Spatial statistical models that use flow and stream distance. *Environ. Ecol. Statist.*, **13**, 449–464.

Wood, S. (2006) *Generalized Additive Models: an Introduction with R*. London: Chapman and Hall–CRC.

Supporting information

Additional animations can be found in the on-line version of this article as supporting information.