Biometrics 000, 000–000 DOI: 000
000 0000

Finite area smoothing with generalized distance splines

David L. Miller[1*], Simon N. Wood[2]

[1]Department of Natural Resources Science, University of Rhode Island, Kingston, Rhode Island 02881, USA

[2]Dept of Mathematical Sciences, University of Bath, Claverton Down, Bath BA2 7AY, UK

*email: dave@ninepointeightone.net

Summary:

Most conventional spatial smoothers smooth with respect to the Euclidean distance between observations, even though this distance may not be a meaningful measure of spatial proximity, especially when boundary features are present. When domains have complicated boundaries leakage (the inappropriate linking of parts of the domain which are separated by physical barriers) can occur. To overcome this problem, we develop a method of smoothing with respect to generalized distances, such as within domain distances. We obtain the generalized distances between our points and then use multidimensional scaling to find a configuration of our observations in a Euclidean space of 2 or more dimensions, such that the Euclidian distances between points in that space closely approximate the generalized distances between the points. Smoothing is performed over this new point configuration, using a conventional smoother. To mitigate the problems associated with smoothing in high dimensions we use a generalization of thin plate spline smoothers proposed by Duchon (1977). This general method for smoothing with respect to generalized distances improves on the performance of previous within domain distance spatial smoothers, and often provides a more natural model than the soap film approach of Wood et al. (2008). The smoothers are of the linear basis with quadratic penalty type easily incorporated into a range of statistical models.

Key words: Generalized additive model; finite area smoothing; multidimensional scaling; spatial modelling; splines.

1. Introduction

In ecology one would often like to create a smooth map of some noisy response as a function of geographical coordinates. In such cases, care must be taken to account for the structure of the domain which is being modelled. If the domain is bounded (*finite area smoothing*) then problems can occur when the smoother does not respect the boundary shape appropriately, especially when the shape of the boundary is complex. This complexity may manifest itself as some peninsula-like feature(s) in the domain with notably different observation values on either side of the feature. Features such as peninsulae give rise to a phenomenon known as *leakage*. The top two panels of Figure 1 shows an example of leakage (taken from Wood, Bravington and Hedley, 2008) where the high values in the upper half of the domain (top panel) leak across the gap to the lower values below and vice versa (second panel). The phenomenon is problematic since it causes the fitted surface to be mis-estimated; this can then lead to incorrect inference, in particular the mis-identification of "hot spots" in biological populations.

[Figure 1 about here.]

The problem of leakage arises because spatial smoothers consider data which is close to eahc other to be similar, but in almost all cases distance between data locations is measured using straight line (Euclidean) distance. This approach is flawed in cases in which straight-line distance is not a meaningful measure of proximity. For example, since whales do not travel on land, the meaningful distance between sightings of two whales on either side of the Antarctic peninsula is not the straight line distance across the peninsula, but the shortest path between them that stays entirely in open water. This issue is ubiquitous in spatial ecology. Natural and man-made barriers carve up the landscape (and seascape), partitioning biological populations; spatial models should take this into account.

In this article we propose a general method for smoothing, based on generalized distances between points. We apply this to produce a finite area smoother, based on the *within-area distances* between points in the domain of interest. The general approach uses multidimensional scaling (MDS; e.g. Chatfield and Collins, 1980, Chapter 10) to associate a location in a $\mathcal{D}$ dimensional Euclidian space (*p-space*) with each original data point. The Euclidian distances between points in p-space then approximate the original generalized distance between the points. Smoothing is then performed with respect to locations in p-space. However, reasonable approximation of the generalized distances by the Euclidian distances in p-space can require $\mathcal{D}$ to be greater than the 2-4 dimensions in which conventional multidimensional smoothers work well. For this reason we revisit the general class of smoothers proposed in Duchon's (1977) thin plate spline paper, selecting a smoother that behaves well with increasing dimension. Note that when applied to the finite area problem our generalized distance smoother can be viewed as an extension of Wang and Ranalli (2007), albeit somewhat better founded (which we argue below).

The smoother proposed here has the attractive property of being representable using a linear basis expansion with an associate quadratic penalty. Such basis-penalty smoothers have a dual interpretation as Gaussian random fields (eg Rue and Held, 2005), and are appealing because of the ease with which they can be incorporated as components of other models (see e.g. Ruppert, Wand and Carroll, 2003 or Wood, 2006 for overviews). Before presenting our proposed method in detail we now briefly review spline type spatial smoothers, and previous approaches to the finite area smoothing problem.

1.1 *Spline smoothing for spatial data*

In the simplest case, we wish to find an f which is a smooth function of spatial coordinates, x_1 and x_2. We model f using a basis function expansion:

$$f(x_1, x_2) = \sum_{k=1}^{K} \beta_k b_k(x_1, x_2), \tag{1}$$

where the β_ks are coefficients to be estimated and the b_ks are flexible (known) basis functions, such as thin plate spline basis functions or tensor products of B-splines.

If K is made large enough to avoid substantial model mis-specification bias, then the estimates of f are almost certain to over-fit any data to which they are fitted. For this reason it is usual to associate a measure $J(f)$ of function wigginess with f, and to use this to penalize over fit during model estimation. For example, consider the simple generalized linear model

$$y_i \sim \text{EF}(\mu_i, \phi), \qquad g(\mu_i) = \eta_i = f(x_{1i}, x_{2i})$$

where EF denotes an exponential family distribution with mean μ_i and scale parameter ϕ, while g is a known link function and η_i is known as the 'linear predictor' of y_i. Letting $l(\boldsymbol{\beta})$ be the log likelihood then estimation of $\boldsymbol{\beta}$ is by maximization of

$$l(\boldsymbol{\beta}) - \lambda/2 J(f)$$

where λ is a tune-able smoothing parameter, used to control the wiggliness of the estimate of f. λ is typically estimated by GCV or marginal likelihood maximization (see e.g. Wood, 2011). A popular $J(f)$ in the spatial context is the 2 dimensional thin plate spline penalty

$$J(f) = \int \left(\frac{\partial^2 f}{\partial x_1^2}\right)^2 + 2\left(\frac{\partial^2 f}{\partial x_1 \partial x_2}\right)^2 + \left(\frac{\partial^2 f}{\partial x_2^2}\right)^2 dx_1 dx_2$$

which can conveniently be written as a quadratic form in $\boldsymbol{\beta}$.

Within this framework it is straightforward to allow η_i to depend on multiple smooth functions of various predictor variables, as well as on conventional parametric terms that are linear in any unknown parameters (e.g. Wood, 2006, based on Hastie and Tibshirani, 1990).

Such models are widely used in quantitative ecology, for example in the creation of density maps which can then be integrated over the domain to obtain an abundance estimate (see e.g. Hedley and Buckland, 2004; Williams et al, 2011) or as part of a larger model, taking into account nuisance spatial effects (e.g. Augustin et al, 2009).

2. Previous approaches to the problem of leakage

There 3 main types of existing approach to dealing with the finite area smoothing problem.

Partial differential equation methods

Ramsay (2002) exploited the link between smoothing with differential operator based penalties and partial differential equations to produce a smoother defined as the solution to a particular PDE problem defined only over a finite area. His FELSPINE method uses a finite element method to compute a smoother, based on the penalty

$$J(f) = \int_\Omega \left(\frac{\partial^2 f}{\partial x_1^2} + \frac{\partial^2 f}{\partial x_2^2} \right)^2 dx_1 dx_2 \tag{2}$$

Where Ω is the region of the x_1, x_2 plane of interest. Ramsay had to use the very strong boundary condition that contours of f meet the boundary of Ω at right angles, which leads to artefacts when the condition does no hold (see Wood et al, 2008). The computational method also makes it awkward to include such terms in larger models.

Wood et al (2008) use the physical analogy of a soap film to motivate an alternative which can be represented as a basis penalty smoother, and has better boundary behaviour. First consider the domain boundary to be made of wire, then dip this wire into a bucket of soapy water; a soap film with the same shape as the boundary will have then formed. If the wire lies in the spatial plane, the height of the soap film at a given point is the values of the smooth at that point. This film is then distorted smoothly toward each datum, while minimising the overall surface tension in the film. Mathematically the soap film consists of two sets of basis functions, one that is based entirely inside the domain (a set of interior knot locations

are specified) and one that is induced by the (known or estimated) boundary values. These functions are found by solving Poisson and Laplace's equations in two dimensions. The penalty associated with the former set is again (2).

The soap film approach has the basis-penalty form that is convenient for applied work and solves the boundary leakage problem, but basis setup is quite computationally expensive, and for many applications the approach is less natural than smoothing using within domain distances. A further problem with the soap film approach is that no distinction exists between 'open' boundaries (for example a boundary that is simply the limit of the region surveyed) and 'hard' boundaries (real physical barriers).

Within-area distances

Wang and Ranalli (2007) propose to replace straight-line distances with 'geodesic' distances in a smoother that is a sort of approximate thin plate spline. To calculate the geodesic distances, a graph is constructed in which each vertex is the location of an observation and is connected only to its k nearest neighbours. The within-area distances between each vertex pair is approximated using Floyd's algorithm (Floyd, 1962) to find the shortest path through the graph. This algorithm is cubic in the number of data, making the approach costly for large datasets. At large sample sizes the geodesic distances will tend towards 'within-area distance', i.e. the shortest path between two point that lies entirely within the domain of interest (Bernstein et al, 2000).

Wang and Ranalli use their geodesic distances in place of the usual Euclidian distances in the radial basis functions used to define a thin plate spline. They leave the basis for the null space of the thin plate spline penalty unchanged, so that some linkage across boundary features remains in the smoother. The principle difficulty in interpreting the results of their method is that it is unclear what their penalty term penalizes. The interpretational difficulty arises because Wang and Ranalli's expressions (3) and (9) involve the square roots of matrices

that are not positive semi-definite. In the case of their expression (3), which relates to a thin plate spline, this problem would be rectifiable if the spline coefficients had the usual thin plate spline linear constraints applied in order to force positive definiteness on the spline penalty. However in the case of (9), which defines their geodesic splines, there appears to be no sensible way to obtain positive semi-definiteness. This is a problem because matrix square roots in general only exist for positive semi-definite matrices plus some rather special cases not useful here (see e.g. Higham, 1987). It appears that for computational purposes Wang and Ranalli have used the generalization of a matrix square root given in appendix A.2.11 of Ruppert, Wand and Carroll (2003), but this square root lacks the basic properties that would allow Wang and Ranalli's (2) to be interpretable exactly as a (reparameterized) thin plate spline, or for it to be possible to work out what the penalty on their geodesic spline is actually penalizing.

The method appears to work well in simulations. However, the combination of an unmodified null space, the opacity of the penalty meaning and $O(n^3)$ computational cost are of some concern for practical work. For these reasons it seems worthwhile to try and come up with alternative ways of using the within-area distance idea, while avoiding these difficulties.

Domain warping

Paul Eilers (in a seminar at University of Munich in 2006) suggested conformally mapping the smoothing domain to a convex one via the Schwarz-Christoffel transformation (Driscoll and Trefethen, 2002). The idea is that smoothing can then be conducted on the convex domain, without leakage problems. The first author has extensively investigated such an approach (Miller, 2012, Chapter 3), but there are insurmountable difficulties associated with the extreme modification of inter-observation distances necessary to achieve domain convexity, which cause artefacts that are significantly more problematic than the leakage effects that the method seeks to avoid.

The methods proposed in the next section can be viewed as an attempt to put within-area distance methods on a more interpretable foundation by using an extension of the notion of domain warping.

3. The generalized distance smoothing model

We assume that we want to model a response y_i which is dependent on covariates via a linear predictor η_i. Our model is then

$$\eta_i = \alpha_i + f(\mathbf{d}_i)$$

where α_i may depend linearly on further model coefficients (or may simply be zero). f is a smooth function, dependent on $\mathbf{d}_i$, a vector of generalized distances between the i^{th} observation and either i) the other observations, or ii) some set of 'reference points'.

We complete the model by setting

$$f(\mathbf{d}_i) = f_{\mathcal{D}}\{\mathbf{x}(\mathbf{d}_i)\}$$

where $\mathbf{x}(\mathbf{d})$ is the location of the point with distance vector $\mathbf{d}$ in the $\mathcal{D}$ dimensional Euclidean space determined by multi-dimensional scaling applied to either i) the matrix for the complete data set or ii) the distance matrix for the set 'reference points' referred to above. $f_{\mathcal{D}}$ is a $\mathcal{D}$ dimensional *Duchon spline* (Duchon 1977), a generalization of the familiar thin plate spline.

The key idea here is that we smooth over a Euclidean space in which the Euclidean inter-observation distances are approximately equal to the original generalized distances. That is $\|\mathbf{x}(\mathbf{d}_i) - \mathbf{x}(\mathbf{d}_j)\| \approx d_{ij}$ when d_{ij} is the generalized distance between points i and j ($\|\cdot\|$ is the Euclidean norm). The choice of $\mathcal{D}$ determines the accuracy of the distance approximation. This can either be part of model specification, in which case $\mathcal{D}$ is chosen to achieve some specified level of approximation accuracy, or more pragmatically, can be chosen to optimize estimated prediction error (e.g. to optimize GCV).

In the case of finite area smoothing, the elements of $\mathbf{d}_i$ are 'within-area' distances between

points, that is to say the shortest path between two points, such that the path lies entirely within the domain of interest. We will refer to the original 2 dimensional data co-ordinates as being elements of the 'o-space' while $\mathcal{D}$ dimensional co-ordinates in the MDS projection will be referred to as elements of the 'p-space'. Web Appendix B gives and algorithm for calculating within-area distances for simple polygons.

These smoothers will be henceforth referred to as MDSDS (Multi-Dimensionally Scaled Duchon Splines), and the next three subsections provide the details for the MDS, smoothing and $\mathcal{D}$ selection steps.

3.1 *MDS as a transformation of space*

In this section we consider the construction of the mapping $\mathbf{x}(\mathbf{d})$ by multidimensional scaling (MDS; Gower, 1966). We start the process by choosing a representative set of locations of size n_s within the domain of interest (i.e. in o-space). This set might be all the locations at which we have observations, but in the case of finite area smoothing we would usually choose a set of locations spread uniformly over the region of interest in order to ensure that all the important geographic features in o-space will be represented in p-space.

The generalized distances between all pairs of points in the representative set are then computed, in order to obtain a matrix $\mathbf{D}$ such that D_{ij} is the squared generalized distance between points i and j. MDS then finds a configuration of points in $\mathcal{D}$ dimensions Euclidean space such that the Euclidean distances between the points approximate the original generalized distances. The recipe for achieving this is straightforward. Defining $\mathbf{H} = \mathbf{I} - \mathbf{1}\mathbf{1}^T/n_s$ we can obtain the double centred version of D, $\mathbf{S} = -\mathbf{HDH}/2$, which is then eigen-decomposed

$$\mathbf{S} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T.$$

The rows of $\mathbf{X} = \mathbf{U}\mathbf{\Lambda}^{1/2}$ then give locations in an n_s dimensional Euclidean space, such that the interpoint Euclidean distances in that space match the original generalized distances. In practice, dimension reduction is required, so we retain only the first $\mathcal{D}$ columns of $\mathbf{X}$ to

create our p-space, accepting that the interpoint distances in this $\mathcal{D}$ dimensional space will only approximate the original generalized distances. See e.g. Chatfield and Collins (1980, Chapter 10) for more details.

We then require some means for finding the location in p-space of a point in o-space that was not in the original representative set used in $\mathbf{D}$. Let $\mathbf{d}$ be the n_s vector of squared generalized distances (in o-space) between this new point and the points in the original representative set. Gower (1968) gives the following interpolation formula for the location of the new point in p-space

$$\mathbf{x} = -\mathbf{\Lambda}^{-1/2}\mathbf{U}^T\mathbf{d}'$$

where $d'_i = d_i - S_{ii}$. Again, $\mathbf{x}$ would usually be truncated, retaining only its first $\mathcal{D}$ components.

So, MDS combined with Gower's interpolation formula provide a means for constructing and computing with $\mathbf{x}(\mathbf{d})$. We now turn to the construction of a suitable smoother in p-space.

3.2 *Smoothing with Duchon splines*

In order for our smoother to have a convenient basis-penalty form, we need to smooth in p-space using a basis-penalty smoother. A thin plate spline (TPS) is the obvious choice for smoothing arbitrarily scattered data where the Euclidean distance between points determines similarity, but there is a technical problem. To achieve a smooth f requires $2m > \mathcal{D}$ where m is the order of differentiation in the TPS penalty, but the dimension of the space of unpenalized functions in a TPS basis is $M = \begin{pmatrix} m+\mathcal{D}-1 \\ \mathcal{D} \end{pmatrix}$. As Figure 2 shows the minimum possible M increases rapidly with $\mathcal{D}$, leading to the danger of substantial undersmoothing as $\mathcal{D}$ increases.

[Figure 2 about here.]

To combat this problem we propose to use a smoother from the larger class of functions considered in Duchon's (1977) paper introducing thin plate splines, which will allow us to obtain a smoother for which $M = \mathcal{D} + 1$. This larger class has been almost entirely ignored in the statistical literature, so we provide a brief summary here.

The difference between general Duchon splines and thin plate splines is in the smoothing penalty used. To understand the difference it helps to start with the general TPS penalty

$$J_{m,\mathcal{D}} = \int_{\mathbb{R}^d} \sum_{\nu_1+\cdots+\nu_d=m} \frac{m!}{\nu_1!\ldots\nu_d!} \left(\frac{\partial^m f(\mathbf{x})}{\partial x_1^{\nu_1}\ldots\partial x_d^{\nu_d}} \right)^2 \mathrm{d}x_1\ldots\mathrm{d}x_d. \tag{3}$$

By Plancherel's theorem (e.g. Vretblad, 2003, p. 180), if we take the Fourier transform, $\mathfrak{F}$, of the derivatives in (3) then the penalty can be re-expressed as

$$J_{m,\mathcal{D}} = \int_{\mathbb{R}^d} \sum_{\nu_1+\cdots+\nu_d=m} \frac{m!}{\nu_1!\ldots\nu_d!} \left(\mathfrak{F}\frac{\partial^m f}{\partial x_1^{\nu_1}\ldots\partial x_d^{\nu_d}}(\boldsymbol{\tau}) \right)^2 \mathrm{d}\boldsymbol{\tau}, \tag{4}$$

where $\boldsymbol{\tau}$ are now frequencies, rather than locations. Duchon then considers weighting the Fourier transform of the derivatives by some power of frequency, effectively increasing the penalization of high frequency components in the spatial derivatives if the power is positive. The resulting penalty is

$$\breve{J}_{m,\mathcal{D}} = \int_{\mathbb{R}^d} \|\boldsymbol{\tau}\|^{2s} \sum_{\nu_1+\cdots+\nu_d=m} \frac{m!}{\nu_1!\ldots\nu_d!} \left(\mathfrak{F}\frac{\partial^m f}{\partial x_1^{\nu_1}\ldots\partial x_d^{\nu_d}}(\boldsymbol{\tau}) \right)^2 \mathrm{d}\boldsymbol{\tau}. \tag{5}$$

where $-\mathcal{D} < 2s < \mathcal{D}$ and the restriction $m + s > \mathcal{D}/2$ is applied to ensure continuity of the splines that result from use of this penalty.

Duchon shows that the function minimising (5) while interpolating or smoothing data at locations $\mathbf{x}_i$ has the form

$$\hat{f}(\mathbf{x}) = \sum_i^n \gamma_i K_{2m+2s-\mathcal{D}}(\|\mathbf{x}-\mathbf{x}_i\|) + \sum_j^M \alpha_j \phi_j(\mathbf{x})$$

where the $\phi_j(\mathbf{x})$ form a basis for the polynomials of order $< m$, $\mathbf{x}_i$ is the ith observation location and γ_i and α_j are coefficients to be estimated, subject to the M linear constraints

$$\sum_i^n \gamma_i \phi_j(\mathbf{x}_i) = 0. \tag{6}$$

The other basis functions are given by

$$K_d(t) = \begin{cases} (-1)^{(d+1)/2}|t|^d & d \text{ odd} \\ (-1)^{d/2}|t|^d \log|t| & d \text{ even} \end{cases}$$

Finally, given the linear constraints (6),

$$\breve{J}_{m,\mathcal{D}} = \delta^T \mathbf{K} \delta$$

where $K_{ij} = K_{2m+2s-\mathcal{D}}(\|\mathbf{x}_i - \mathbf{x}_j\|)$. Notice that $s = 0$ gives a conventional TPS.

As an example of using this basis and penalty, we note that estimating f to minimize

$$\sum_i^n (y_i - f(\mathbf{x}_i))^2 + \lambda \breve{J}_{m,\mathcal{D}}$$

now reduces to the straightforward optimization problem of finding $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\alpha}}$ to minimize

$$\|\mathbf{y} - \mathbf{K}\boldsymbol{\gamma} - \mathbf{A}\boldsymbol{\alpha}\|^2 + \lambda \boldsymbol{\gamma}^T \mathbf{K} \boldsymbol{\gamma} \text{ s.t. } \mathbf{A}^T \boldsymbol{\gamma} = \mathbf{0}$$

where $A_{ij} = \phi_j(\mathbf{x}_i)$. Notice that this problem has exactly the same structure as the TPS problem, so exactly the same approach to computation can be taken. More importantly optimal rank reduced versions of these Duchon splines can be produced using the methods given in Wood (2003) for the TPS. Wood (2003) uses an eigen approximation to the full spline thereby avoiding the difficult problem of $\mathcal{D}$ dimensional knot placement that complicates other approaches to reduced rank splines, so we use this approach in what follows.

We are now in a position to produce a spline suitable for smoothing in p-space. Specifically we choose a (reduced rank) Duchon spline with $m = 2$ and $s = \mathcal{D}/2 - 1$, which will give us a smooth f for which $M = \mathcal{D} + 1$ (i.e. the unpenalized component of f grows only linearly with $\mathcal{D}$).

3.3 *Selecting* $\mathcal{D}$

An obvious question is whether we actually need $\mathcal{D}$ to be larger than 2 in practice. Figure 3 provides an illustration that in general we do. It shows what happens when within-area distances over a 2 dimensional domain with a peninsulae are used to obtain a 2D p-space. Some of the points in the resulting p-space configuration become quite severely squashed

together. In fact truncating the projection to two dimensions can sometimes cause the basic ordering of the points to be lost, making the task of the smoother impossible. Further investigation showed that increasing the dimensionality of p-space maintains the ordering of the points, however the number of dimensions required varied according to the shape of the domain.

[Figure 3 about here.]

Having accepted the need for $\mathcal{D} > 2$, we need some means for choosing $\mathcal{D}$. The two obvious strategies might be characterised as 'model driven' and 'data driven'.

In the model driven approach we specify the accuracy with which the inter-point distances in p-space must match the original generalized distances as part of model specification, and $\mathcal{D}$ is then increased during fitting until this specification is met. In the data driven approach then $\mathcal{D}$ is chosen in order to optimize some prediction error criterion, such as GCV. Figure 6 shows the relationship between $\mathcal{D}$ and GCV score for the Aral sea data analysed below.

Notice that selecting $\mathcal{D}$ is typically a small part of the computational burden, since the MDS and smoothing are usually cheap relative to the computation of distances (at least in the finite area smoothing case).

4. Examples

To illustrate the utility of the model two simulation studies are shown, followed by examples using real data. All concentrate on the finite area smoothing problem. In each case MDSDS was compared with thin plate splines as described in Wood (2003) (which do not account for the boundary), geodedic low-rank thin plate splines (GLTPS) and the soap film smoother (which both do account for the boundary). The GLTPS model was as described in Wang and Ranalli (2007), but with the within-area distances calculated as described in Web Appendix B (i.e. the same as for MDSDS); knots were placed using the `cover.design` method in the

package `fields` (again, as in Wang and Ranalli (2007)). In all cases smoothing parameters were selected by GCV. The R packages `mgcv` (available from CRAN), `soap` (available from `http://www.maths.bath.ac.uk/~sw283/simon/software.html`) and `msg` (available from `https://github.com/dill/msg`) were used to fit the models. Code for fitting the GLTPS is available at `https://github.com/dill/gltps`

In all the cases below the basis size specified refers to the maximum basis size allowed, since the penalty will reduce the complexity of the smoother, we simply need to specify an upper bound on the basis size.

4.1 *Ramsay's horseshoe*

The horseshoe shape shown in the top panel of Figure 1 is an obvious benchmark for techniques that wish to combat leakage. Although perhaps unrealistic (and bordering on pathological), any new method that works well on the horseshoe should have a good chance of working well in more realistic situations. A simulation experiment was run with the same setup as in Wood et al. (2008): 200 replicates were generated at each of three error levels (standard normal noise multiplied by 0.1, 1 and 10) with sample size 600. A thin plate regression spline, with basis size 100 and a soap film smoother with 32 interior knots and a 40 knot cyclic spline was used to estimate the boundary. For the MDSDS model, the basis size was set to 100 and a 20 by 20 initial grid was used for the MDS projection (see Web Appendix A), MDS projection dimension was selected by GCV in the range of 2 and the number of dimensions that explained 95% of the variation in the distance matrix of the initial grid. For the GLTPS 40 knot locations were selected as in Wang and Ranalli (2007). For each realisation the mean squared error (MSE) was calculated between the true function and a prediction grid of 720 points.

[Figure 4 about here.]

As can be seen in Figure 4, the thin plate regression spline has rather poor performance in MSE terms while MDSDS, the soap film smoother and GLTPS perform significantly better. MDSDS outperforms the soap film smoother and GLTPS at lower noise levels, becoming indistinguishable at the highest noise level. The median number of dimensions selected for the MDS projection using GCV was 3 (max. 14, min. 2). Looking more qualitatively at the bottom three plots in Figure 4, the predictions do not show any evidence of leakage.

4.2 *Peninsulae domain*

The results from the modified Ramsay horseshoe are encouraging. However the domain is not particularly realistic. To further explore the performance of MDSDS a more realistic domain was used. The domain, which attempts to mimic a coastline, is shown in the left panel of Figure 3.

Simulations were run at a series of noise levels 0.35, 0.9 and 1.55 equating to signal-to-noise ratios of 0.50, 0.75 and 0.95, respectively. The soap film smoother used 109 internal knots and 60 for the cyclic boundary smooth. The MDSDS models used an initial grid of 120 by 126 points, the basis size was 140. The thin plate regression spline basis size was also 140. For the GLTPS, 80 knots were selected using the space filling design, as above.

Figure 5 shows the boxplots of the log of the MSE per realisation for each model. In the low noise cases, a paired Wilcoxon signed rank test showed that the soap film smoother and MDSDS were not significantly different at the 0.05 level. In all cases MDSDS was significantly better than both GLTPS and thin plate regression splines.

[Figure 5 about here.]

4.3 *Aral sea*

The Aral sea is located between Kazakhstan and Uzbekistan. It has been steadily shrinking since the 1960s when the Soviet government diverted the sea's two tributaries in order to

irrigate the surrounding desert. The NASA SeaWifs satellite collected data on chlorophyll levels in the Aral sea over a series of 8 day observation periods from 1998 to 2002 (Wood et al, 2008). The 496 data are averages of the 38^{th} observation period. Smooths were fitted to the spatial coordinates (Northings and Eastings; kilometres from a specified latitude and longitude) with the logarithm of chlorophyll concentration (modelled with a Gamma distribution) as the response.

The models that were fitted to the data were: a thin plate regression spline with basis size 70, MDSDS with a basis size of 70 (a 20 by 20 initial grid was used for the MDS projection), a GLTPS with 60 knots and soap film using 49 boundary knots and 74 internal knots. Using GCV for MDS projection dimension selection lead to a 5-dimensional projection. A plot of the relationship between projection dimension and GCV score can be seen in Figure 6; there is a clear minimum at 5 dimensions.

[Figure 6 about here.]

Predictions from the models over a grid of 496 points are shown in Figure 7. The fits are broadly similar across most of the domain. MDSDS, GLTPS and the soap film smoother do not show signs of leakage around (-50,-50), as the thin plate regression spline does.

[Figure 7 about here.]

5. Discussion

Our MDSDS approach appears to have competitive performance with existing methods, while providing a number of possible advantages. Relative to the soap film approach of Wood et al. (2008) the method has a more natural handling of open and closed boundaries, and is also often the more natural model, when the linkage between geographic areas is via movement of organisms. Relative to Wang and Ranalli (2007) our approach is somewhat more transparent in terms of what is being penalized when smoothing, and also uses a null

space basis that avoids leakage, unlike the Wang and Ranalli method for which the null space does not respect boundary features.

The MDSDS has an interesting link to Kriging with within-area distances. For example Løland and Høst (2003) used river network distances in the construction of a variogram, and overcame the problem of lack of positive definiteness (essentially the problem ignored in Wang and Ranalli, 2007) by using MDS and then constructing the variogram in the MDS configuration space. Jensen et al (2006) suggest using the proportion of variation explained or the Bayesian criterion of Oh and Raftery (2001) as possible metrics to perform projection dimension selection in the Krigging setting but do not fully address the issue, resorting to 2-dimensional projections. To date the Kriging literature in this area seems to have considered only stationary processes, and of course the Kriging approach does not make for the straightforward incorporation in general statistical models that our basis-penalty smoothers provide.

Further interesting work would involve considering more biologically motivated measures of distance. For example, distances based on the minimum energetic cost of moving between two locations. It is also of interest to investigate further the use of MDSDS for smoothing with respect to distances that have no geographic basis (the socio-economic similarity of parliamentary constituencies, or measures of genetic relatedness, for example), and we are investigating this at present.

Supplementary Materials

Web Appendices referenced in Sections 3 and 4 are available with this paper at the Biometrics website on Wiley Online Library.

Acknowledgements

David wishes to thank EPSRC for financial support during his PhD at the University of Bath.

References

Augustin, N.H., Musio, M., von Wilpert, K., Kublin, E., Wood, S.N. and Schumacher, M. (2009). Modeling spatiotemporal forest health monitoring data. *Journal of the American Statistical Association* **104**(487), 899–911.

Bernstein, M., de Silva, V., Langford, J.C., and Tenenbaum, J.B. (2000). Graph approximations to geodesics on embedded manifolds. Technical report.

Boisvert, J.B., Manchuk, J.G. and Deutsch, C.V. (2009). Kriging in the presence of locally varying anisotropy using non-Euclidean distances. *Mathematical Geosciences* **41**, 585–601.

Chatfield, C. and Collins, A.J. (1980). *Introduction to multivariate analysis*. CRC Press.

Craven, P. and Wahba, G. (1979), Smoothing noisy data with spline functions. *Numerische Mathematik* **31**, 377–403.

Curriero, F. (2006). On the use of non-Euclidean distance measures in geostatistics. *Mathematical Geology* **38**(8), 907–926.

Diggle, P.J. and Ribeiro, P.J. (2007). *Model-based Geostatistics*. Springer.

Driscoll, T.A. and Trefethen, L.N. (2002). *Schwarz-Christoffel Mapping*. Cambridge University Press.

Floyd, R.W. (1962). Algorithm 97: Shortest path. *Communications of the ACM* **5**(6), 345.

Gentleman, R., Ding, B., Dudoit, S., and Ibrahim, J. (2005). Distance measures in DNA microarray data analysis. In R. Gentleman, V. Carey, W. Huber, R. Irizarry, and S. Dudoit (Eds.), *Bioinformatics and Computational Biology Solutions Using R and*

Bioconductor, pp. 189–208. Springer.

Girosi, F., Jones, M. and Poggio, T. (1995). Regularization theory and neural networks architectures. *Neural computation*, **7**, 219-269.

Gower, J.C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, **53**(3 and 4), 325–338.

Gower, J. C. (1968). Adding a point to vector diagrams in multivariate analysis. *Biometrika*, **55**(3), 582–585.

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

Hastie, T., Tibshirani, R. and Friedman, J. (2001). *Elements of Statistical Learning*. Springer.

Hedley, S.L. and Buckland, S.T. (2004) Spatial Models for Line Transect Sampling. *Journal of Agricultural, Biological, and Environmental Statistics* **9**(2), 181–199.

Higham, N.J. (1987) Computing Real Square Roots of a Real Matrix *Linear Algebra and its Applications* **88/89**, 405–430.

Løland, A. and Høst, G. (2003). Spatial covariance modelling in a complex coastal domain by multidimensional scaling. *Environmetrics* **14**(3), 307–321.

Jensen, O.P., Christman, M.C. and Miller, T.J. (2006). Landscape-based geostatistics: a case study of the distribution of blue crab in Chesapeake Bay. *Environmetrics* **17**(6), 605–621.

Mahalanobis, P.C. (1936). On the generalised distance in statistics. *Proceedings of the National Institute of Sciences of India* **2**(1), 49–55.

Marra, G. and Radice, R. (2010). Penalised regression splines: theory and application to medical research. *Statistical Methods in Medical Research*, 19, 107–125.

Miller, D.L. (2012) *On smooth models for complex domains and distances*. PhD thesis, University of Bath.

Oh, M-S and Raftery, A.E. (2001). Bayesian multidimensional scaling and choice of dimension. *Journal of the American Statistical Association*, **96**(455), 1031–1044.

Ramsay, T. (2002). Spline smoothing over difficult regions. *Journal of the Royal Statistical Society: Series B* **64**(2), 307–19.

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

Ruppert, D., Wand, M.P. and Carroll, R.J. (2009). Semiparametric regression during 2003–2007. *Electronic Journal of Statistics*, **3**, 1193–1256.

Shabenberger, O. and Gotway, C.A. (2005). *Statistical methods for spatial data analysis*. CRC Press.

de Silva, V. and Tenenbaum, J.B. (2004). Sparse multidimensional scaling using landmark points. Technical report, Stanford University.

Venables, W.N. and Ripley, B.D. (2002). *Modern applied statistics with S*. Springer.

Vretblad, A. (2003). *Fourier Analysis and Its Applications*. Springer.

Wang, H. and Ranalli, M.G. (2007). Low-rank smoothing splines on complicated domains. *Biometrics* **63**(1), 209–217

Williams, R., Hedley, S.L., Branch, T.A. and Bravington, M.V., Zerbini, A.N. and Findlay, K.P. (2011). Chilean Blue Whales as a Case Study to Illustrate Methods to Estimate Abundance and Evaluate Conservation Status of Rare Species. *Conservation Biology* **25**(3), 526–535.

Wood, S.N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society: Series B*, **65**(1) 95–114.

Wood, S.N. (2006). *Generalized Additive Models: An Introduction with R*. Chapman & Hall.

Wood, S.N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B*, **73**(1), 3–36.

Wood, S.N., Bravington, M.V. and Hedley, S.L. (2008). Soap film smoothing. *Journal of the*

Royal Statistical Society: Series B **70**(5), 931–55.

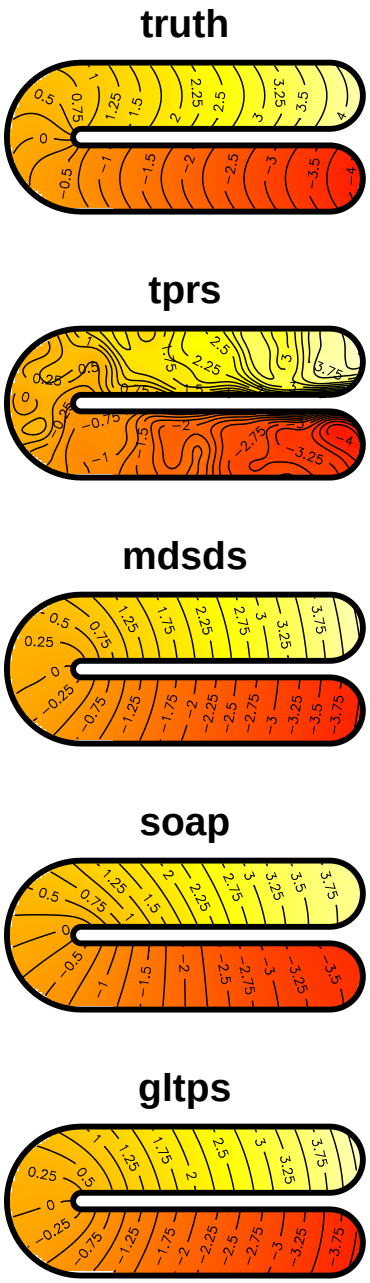

Figure 1. Top: the modified Ramsay horseshoe function from Wood et al (2008). Below: predictions from models using thin plate regression splines ("tprs"), MDSDS ("mdsds"), the soap film smoother ("soap") and geodesic low-rank thin plate splines ("gltps") when 600 points were sampled from the horseshoe and standard normal noise was added. Note that the predictions from the thin plate regression spline fit shows severe leakage. This figure appears in colour in the electronic version of this article.

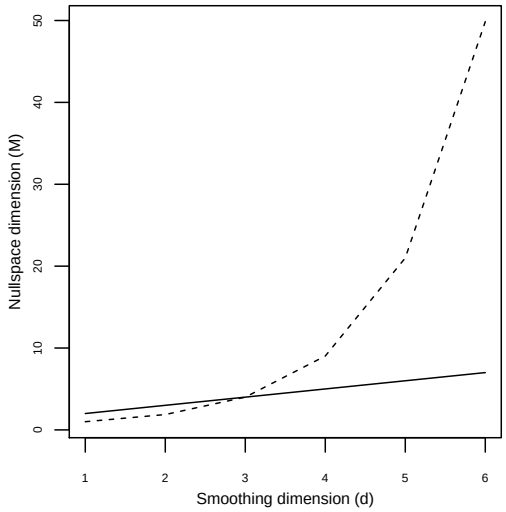

Figure 2. Relationship between smoothing dimension (d) and the nullspace dimension (M) when m (the derivative penalty order) is set to 2 for thin plate regression splines (dashed) and Duchon splines (solid). Note that as the nullspace dimension increases, the complexity of those functions in the nullspace increases too. For the thin plate splines a combination of the continuity condition that $2m > d$ and the form of M makes the size of the nullspace increase very quickly with smoothing dimension.

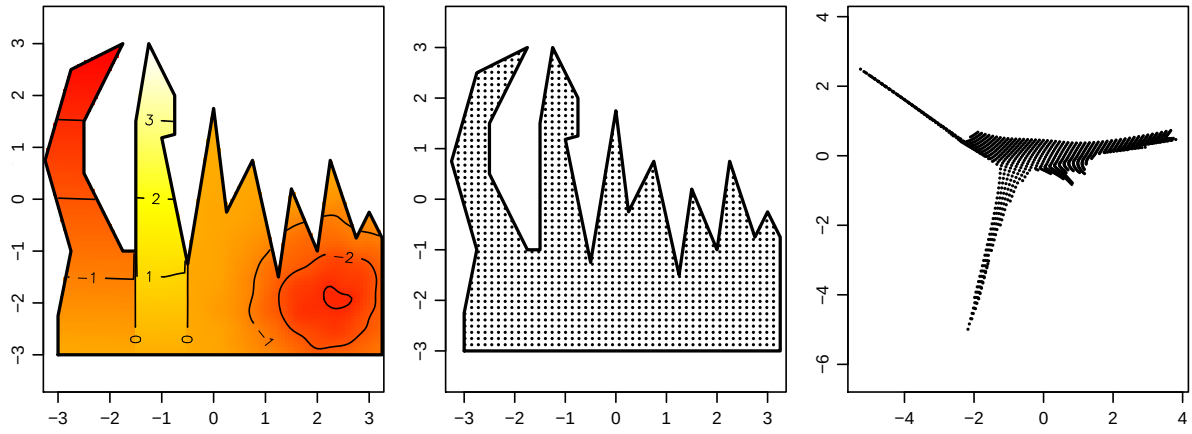

Figure 3. Left to right: Test function over the peninsulae domain, points in the domain and finally their projection into 2-dimensional p-space when within-area distances are used to calculate the distance matrix. The p-space plot shows that some squashing can happen in two dimensions. The large left peninsula and some of the smaller peninsulae have lost their 'width' and, in fact, points within them have lost ordering. This figure appears in colour in the electronic version of this article.

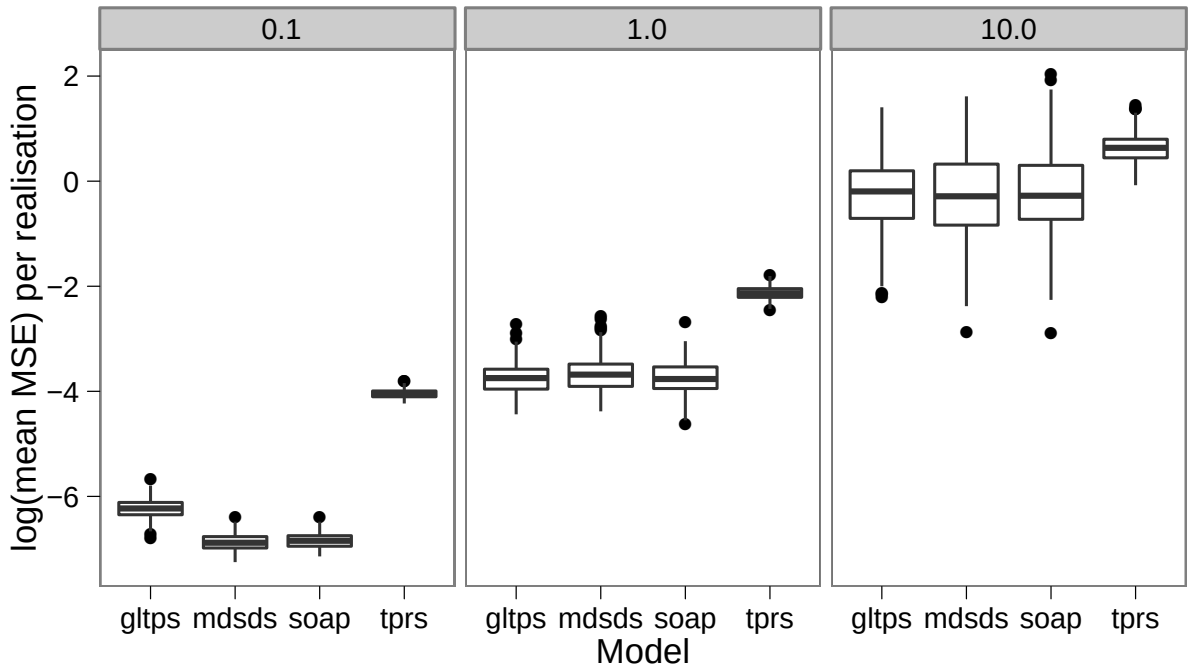

Figure 4. Top: boxplots of per-realisation log mean squared error at the three noise levels. Using a paired Wilcoxon signed two-sample test, the difference between the new approach ("mdsds") and the other models was significantly different for the two lower noise levels (at the 0.05 level). For the higher noise level, the three methods that accounted for the boundary ("gltps" and "soap") were not distinguishable from MDSDS. The thin plate regression spline ("tprs") was worse.

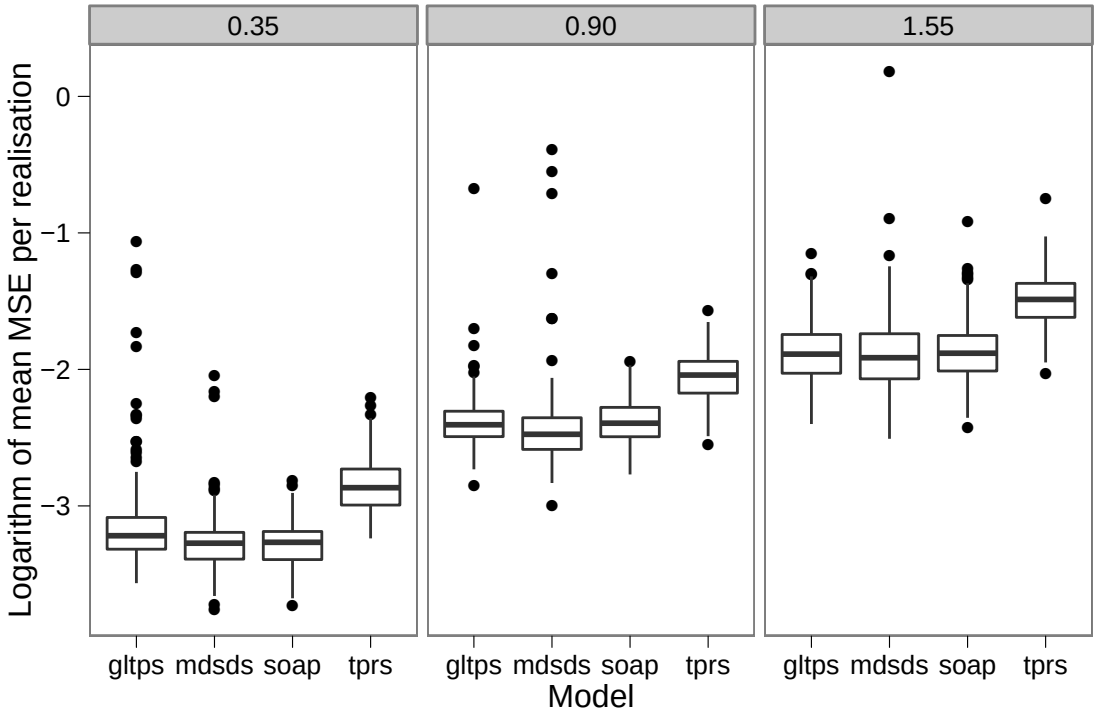

Figure 5. Boxplots of logarithm of mean MSE per realisation for the models tested on the peninsulae domain at three noise levels. At each noise level, the median mean MSE was lower for MDSDS than for the thin plate regression spline, soap film smoother or GLTPS. A paired Wilcoxon signed rank test showed that the only significant difference (at the 5% leve) occurred between MDSDS and the soap film smoother at the 0.35 noise level.

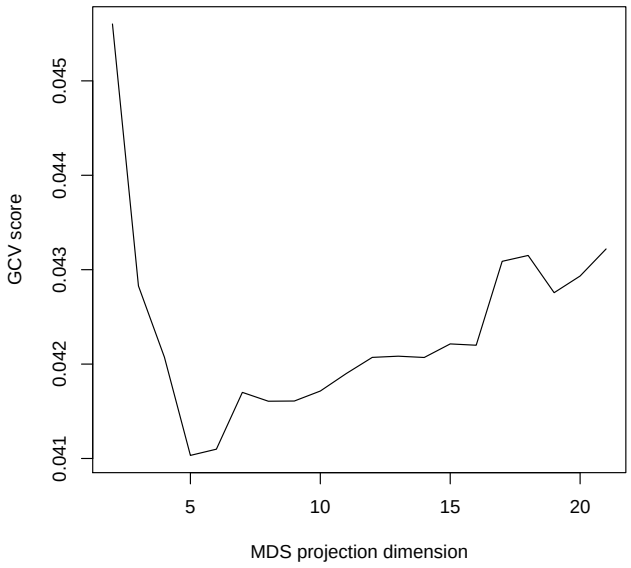

Figure 6. Plot of the relationship between GCV score and MDS projection dimension for the Aral sea data set. Here a clear minima at 5 dimensions can be seen, however there is no particular reason to believe that there will always be such a pronounced optima.

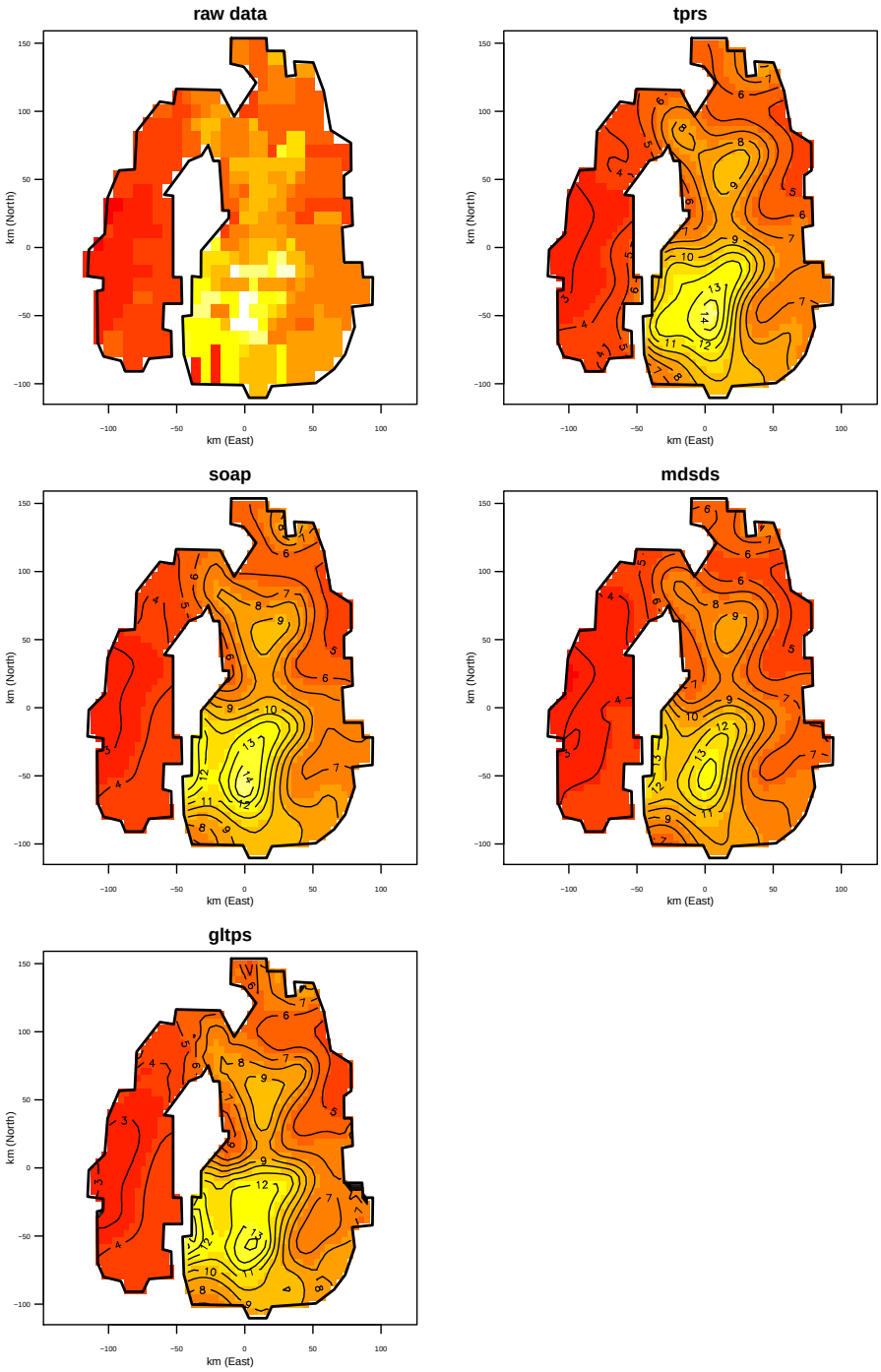

Figure 7. Aral sea analysis. Top to bottom, left to right. First row: raw data, predicted surface for thin plate regression spline. Second row: predicted surfaces for MDSDS and the soap film smoother. Bottom row: predicted surface for GLTPS. The latter three avoid the leakage seen in the $(-50, 50)$ region of the thin plate regression spline fit. This figure appears in colour in the electronic version of this article.