Published in final edited form as:

J Stat Comput Simul. 2019 ; 89(11): 2121–2137. doi:10.1080/00949655.2019.1610884.

Threshold Knot Selection for Large-Scale Spatial Models With Applications to the *Deepwater Horizon* Disaster

Casey M. Jelsema,

Department of Biostatistics, School of Public Health, West Virginia University, Morgantown, West Virginia, 26505, USA

Richard K. Kwok,

Epidemiology Branch, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina, 27709, USA

Shyamal D. Peddada

Department of Biostatistics, School of Public Health, University of Pittsburgh, Pittsburgh, Pennsylvania, 15261, USA

Abstract

Large spatial datasets are typically modeled through a small set of knot locations; often these locations are specified by the investigator by arbitrary criteria. Existing methods of estimating the locations of knots assume their number is known *a priori*, or are otherwise computationally intensive. We develop a computationally efficient method of estimating both the location and number of knots for spatial mixed effects models. Our proposed algorithm, Threshold Knot Selection (TKS), estimates knot locations by identifying clusters of large residuals and placing a knot in the centroid of those clusters. We conduct a simulation study showing TKS in relation to several comparable methods of estimating knot locations. Our case study utilizes data of particulate matter concentrations collected during the course of the response and clean-up effort from the 2010 *Deepwater Horizon* oil spill in the Gulf of Mexico.

Keywords

fixed rank kriging; knot selection; reduced rank spatial model; spatial mixed effects

1. Introduction

On April 20, 2010, a catastrophic explosion occurred aboard the *Deepwater Horizon* oil drilling rig. The resulting oil spill over the ensuing three months became the largest maritime oil spill in the history of the United States (Ramseur, 2012). Toxins from the oil were released into the Gulf region, leading to concerns about potential health effects (Goldstein et al., 2011; McCauley, 2010).

Over 55,000 individuals participated in the response and clean-up effort. There has been concern raised about the potential short- and long-term human health effects of the oil spill and of the subsequent cleanup effort. In addition to exposure to crude oil, dispersants, and mixtures of the two within the environment, the clean-up operation also generated a large number of pollutants. One of these pollutants is $PM_{2.5}$, a combustion byproduct that has been linked to a number of negative morbidity and mortality outcomes (Hamra et al., 2014; Lee et al., 2014).

Several studies were launched to investigate the potential link between the oil spill and health effects, including efforts by the National Institute of Environmental Health Sciences (NIEHS) which initiated a study (the Gulf Long-term Follow-up Study (GuLF STUDY)) as well as the Deepwater Horizon Research Consortia to investigate potential adverse health effects from exposures experienced during the response and clean-up effort. (Huynh et al., 2014; Quick et al., 2014; DHRC, 2016). However, in order to accurately assess any potential health effects resulting from this oil spill, it is necessary to provide accurate estimates of exposures.

Two key issues arise when modeling the $PM_{2.5}$ data. First, they are spatially correlated. This means that observations located closely together tend to be more highly correlated than observations taken further apart. Secondly, there is a large volume of data. A map of the affected area is shown in Figure 1, along with a sample of $PM_{2.5}$ measurements between June 21 and June 27, 2010.

There are 9477 observations plotted in Figure 1, and in total over 100,000 measurements of $PM_{2.5}$ were taken. This presents a challenge: for a spatial process $\mathbf{Z}$ with n observations, the covariance matrix $cov(\mathbf{Z}) = \Sigma_{\mathbf{Z}}$ is of the order $n \times n$. Most spatial analysis requires the inversion of this covariance matrix, which has computational complexity of $\mathcal{O}(n^3)$. Traditional methods of modeling spatial processes are easily overwhelmed when the number of observed locations is too large. This is a “big data” problem in the context of geostatistical data. Multiple methods have been proposed for fitting reduced rank models of large spatial datasets in order to overcome the computational challenge, and a common feature in them is the use of knot locations selected across the spatial domain.

Cressie and Johannesson (2008) take an empirical Bayesian approach, using a mixed effects model where the random component is a latent reduced rank process defined over the knots, and its design matrix is a matrix of spatially varying basis functions. The full covariance matrix of the data may then be accessed analytically as a function of the basis functions and the covariance matrix of the latent process. Banerjee et al. (2008) take a Bayesian approach, modeling the spatial process through a lower dimensional Gaussian predictive process defined over the knots. These methods have been further developed, for instance Cressie et al. (2010) extend the spatial random effects model of Cressie and Johannesson (2008) to incorporate temporal dependencies, and Kang and Cressie (2011) consider Bayesian inference for similar models. Banerjee et al. (2010) developed a reduced rank model for multivariate responses.

The determination of knots, including both their number and locations, has been considered by several authors. Guhaniyogia et al. (2011) and Katzfuss (2013) each propose an adaptive scheme in which the knots are considered as a realization of a spatial point process (SPP). An attractive element of these two methods is that locations of knots are updated at every iteration of an MCMC algorithm, though as Bayesian methods, both can be somewhat computationally intensive.

There are also several non-Bayesian methods of estimating the knot locations. Crainiceanu et al. (2006) use the space-filling method of Nychka and Saltzman (1998) and more recently Ver Hoef and Jansen (2014) generated a grid of knots and used K-means clustering to select two sets of knots. In these methods the number of knots, r , is a fixed value, and selected based on the recommendation of Ruppert et al. (2003). While these methods help to ensure adequate coverage of the spatial domain, they are based on geometry only (the locations of observations), and ignore the value of the spatial process. Closer to the scope of this paper, Gelfand et al. (2011) describe an algorithm to sequentially estimate knot locations in the context of a Gaussian predictive process. However they also consider a fixed r , and select knots from a set of candidate locations until r knots is reached, where at each step the selected knot results in the largest drop in average predictive variance.

Motivated by the noted challenges and limitations of existing methods, in this paper we present a method which considers both the number and location of knots in a computationally efficient manner. We also consider the knot locations as arising from a SPP, in our work we obtain a realization of the SPP by generalizing the threshold sets described in Madrid et al. (2012). The proposed method is distribution-free and independent of any assumed spatial covariance function.

In Section 2 we describe a reduced rank spatial model and detail our knot-selection algorithm. The performance of the proposed algorithm is illustrated in Section 3 using a simulation study. In Section 4 we apply the proposed methodology to data from the *Deepwater Horizon* oil spill. We conclude by describing some challenges and open research problems in Section 5.

2. Methods

2.1. Reduced Rank Spatial Model (RSSM)

A spatial process Z observed at locations $\mathbf{s}_i$, $i = 1, \dots, n$ may be expressed as $Z(\mathbf{s}) = Y(\mathbf{s}) + \epsilon(\mathbf{s})$, where Y is a latent process and $\epsilon(\mathbf{s})$ is an independently and identically distributed (*iid*) zero-mean process which represents random errors (e.g., measurement error), and has variance σ^2 for all locations $\mathbf{s}$. For simplicity we stack these as vectors and matrices to obtain $\mathbf{Z} = \mathbf{Y} + \boldsymbol{\epsilon}$, so the $\mathbf{s}$ index is suppressed except when it is deemed necessary. Then $\mathbf{Y}$ is modeled with a mixed effects model, $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{S}\boldsymbol{\eta} + \boldsymbol{\delta}$. Here $\mathbf{X}$ is a matrix of covariates and $\boldsymbol{\beta}$ are the associated regression slopes. In this work we refer to the product $\mathbf{X}\boldsymbol{\beta}$ as the trend, while $\mathbf{S}\cdot$ is the spatial variation, and $\boldsymbol{\epsilon}$ and $\boldsymbol{\delta}$ are non-spatial variation. The $\boldsymbol{\delta}(\mathbf{s})$ process (stacked to obtain $\boldsymbol{\delta}$) is another *iid* zero-mean process with variance τ^2 for all locations $\mathbf{s}$. This is called the process error, and takes into account uncertainties arising from dimension

reduction. Often the variances σ^2 and τ^2 are non-identifiable, so instead the nugget effect $\nu^2 = \sigma^2 + \tau^2$ is estimated.

To accomplish dimension reduction, and thus computational feasibility for large datasets, one selects $r \ll n$ locations over the spatial domain to be knots. Then η is a reduced rank process over these knots with mean zero and covariance matrix $\mathbf{V}$, and $\mathbf{S}$ is an $n \times r$ matrix of basis functions which maps η to the full dimension of $\mathbf{Z}$. Gaussian distributions are typically assumed for η , ϵ , and δ (or conditional Gaussian distributions, e.g. on $\mathbf{Z}|\beta, \eta, \sigma^2$). The models described above can be rewritten as:

$$\mathbf{Z} = \mathbf{X}\beta + \mathbf{S}\eta + \delta + \epsilon. \quad (1)$$

Several options are available for estimating the model parameters $\mathbf{V}$ and ν^2 . Cressie and Johannesson (2008) use empirical Bayesian estimates, while Banerjee et al. (2008) and Kang and Cressie (2011) utilize fully Bayesian models. The reduced rank covariance matrix $\mathbf{V}$ may be further parameterized, for example by assuming some spatial covariance function. Katzfuss and Cressie (2011a) derived an Expectation-Maximization (EM) algorithm (Dempster et al., 1977) for reduced rank spatial models. We estimated $\mathbf{V}$ using this EM algorithm.

Any basis functions may be used for defining $\mathbf{S}$, including W-wavelets (Shi and Cressie, 2007; Kang and Cressie, 2011) and bisquare functions (Cressie and Johannesson, 2008; Katzfuss and Cressie, 2011b; Kang et al., 2012). In the present work we use a bisquare function to define $\mathbf{S}$,

$$\mathbf{S}_{i,j} = \begin{cases} \left(1 - \left(\frac{d_{ij}}{a_j}\right)^2\right)^2 & d_{ij} \leq a_j \\ 0 & d_{ij} > a_j \end{cases}, \quad (2)$$

where $d_{ij} = \|\mathbf{s}_i - \mathbf{u}_j\|$, $i = 1, \dots, n$, $j = 1, \dots, r$ is the distance between observed location $\mathbf{s}_i$ and knot location $\mathbf{u}_j$. The measure of distance to calculate d_{ij} should be appropriate to the data under consideration (e.g., Euclidean or geodesic distances may be appropriate in different contexts). The values a_j define the range of the knot $\mathbf{u}_j$. When knot locations are equidistant over the domain, this value is typically some multiple of the minimum distance between knot locations. Cressie and Johannesson (2008) use a constant range a for all knots of a given resolution. To accommodate knot locations that may not be equidistant, we let the ranges of the knots vary. Specifically, we define a_j as a multiple of the minimum distance from $\mathbf{u}_j$ to all the other knots (i.e. $a_j = \alpha \min \{\|\mathbf{u}_{-j} - \mathbf{u}_j\|\}$, where $\mathbf{u}_{-j}$ is the set of all knots except $\mathbf{u}_j$ and α is a scaling factor we take to be 1.5). This specification of basis functions allows the range for each knot to shrink or expand according to the intensity of knots in a region, which is conceptually relatable to the use of a variable bandwidth in kernel density estimation and kernel regression.

2.2. Threshold Knot Selection

The knots of a RRSM as described in Section 2.1 are nuisance parameters: they are required for fitting the model, but are not of interest themselves. Knot estimation requires balancing the competing goals of a finer spatial resolution and improved accuracy (more knots) against that of computational efficiency (fewer knots). Often the knots are selected on a regular grid (Banerjee et al., 2008; Sahu and Challenor, 2008), a modification of a regular grid (Diggle and Lophaven, 2006), or by some form of clustering algorithm (Crainiceanu et al., 2006). But in areas where the spatial process is smooth, fewer knots may be needed to adequately model the process, and Banerjee et al. (2010) note that in areas where data are sparse, extra knots may lead to inflated variability of predictions. This may be due to the fact that a relatively smaller number of observations are being used to estimate extra parameters. Gelfand et al. (2011) have noted that so long as an adequate number of knots is chosen, the spatial predictions are reasonably robust to the precise locations of knots. Consequently, our primary objective is an efficient data-informed method to estimate the number and locations of knots.

Madrid et al. (2012) describe properties of the SPP arising from threshold exceedance sets. For a spatial process Z , the exceedance set corresponding to a threshold u is defined by $A = \{s_j: Z(s_j) > u\}$. Using single-linkage clustering, the elements of A may be placed into g clusters $C_1, C_2, \dots, C_g$ based on a bandwidth h , such that $\min \|a_{k,p} - a_{l,q}\| > h$, for all $p \neq q$, where $a_{k,p} \in C_p$ for $k = 1, \dots, |C_p|$ and $|C_p|$ denotes the cardinality of cluster C_p . The point $a_{k,p}$ thus represents the k^{th} element of the p^{th} cluster C_p (and similarly for $a_{l,q}$). Each cluster is then represented by the centroid of its locations (the mean along each coordinate), giving rise to a SPP. By defining a sequence of thresholds and indexing the exceedance sets by A_u , a family of SPPs is defined.

We use this family of SPPs to define the knot locations for a RRSM. However, there are several challenges which need to be addressed before one may do so. First, for T thresholds $u_1 < u_2 < \dots < u_T$, the exceedance sets will be increasing in cardinality: all elements of the exceedance set A_{u_t} will also be members of set $A_{u_{t+1}}$. Hence, the centroids of the clusters will tend to exhibit over-clustering and fail to adequately cover the spatial domain. To address this, we use interval sets instead of exceedance sets. That is, $A_{u_t} = \{s_j: u_t < Z(s_j) < u_{t+1}\}$. These interval sets represent non-overlapping regions, and knot locations derived from them more adequately cover the spatial domain.

Secondly, the purpose of the knots is to model the random effects component of equation (1). In the context of a RRSM, this term corresponds to spatially dependent variation about the mean trend. The natural choice therefore is to define exceedance sets based on the detail residuals $\mathbf{R} = \mathbf{Y} - \mathbf{X}\hat{\beta}_{ols}$, where $\hat{\beta}_{ols}$ is the ordinary least squares (OLS) estimate of β . We first place knots where the magnitude of $\mathbf{R}$ is largest. Once we have placed knots in the “peaks” and “valleys”, we simply need to populate the spatial domain with an adequate number and spread of knots. It should be noted that OLS estimate of β may differ from the generalized least squares (GLS) estimates that are typically used during kriging. During analysis, the user may wish to verify that OLS and GLS estimates of the regression parameter are similar.

Specifying the thresholds as quantiles of $\text{abs}(\mathbf{R})$ allows one to determine the lower bound at which they wish to supplement the trend with knots to capture spatial variation. In our implementations we specify the thresholds $u_1, \dots, u_T$ as equally spaced, decreasing quantiles of $\text{abs}(\mathbf{R})$, from the 98th down to the 50th percentile (starting at the 98th because we let $u_0 = \max(\text{abs}(\mathbf{R}))$). These thresholds are naturally all positive, so we find the intervals sets, separately, for both $\mathbf{R}$ and $-\mathbf{R}$, in order to capture both the peaks and the valleys of the residual process.

We note that the choice of percentiles used is arbitrary, but reasonable. The purpose of this method is to place knots in regions where the detailed residuals had large magnitude. Hence, some upper percentiles should be selected to ensure that the main “peaks” and “valleys” are identified. In the presence of spatial dependence and with large sample sizes, small changes to the particular percentiles are unlikely to have much impact on the results.

The proposed Threshold Knot Selection (TKS) can be summarized in the following steps.

- Step 1** Define T thresholds $u_1, \dots, u_T$ as decreasing percentiles of $\text{abs}(\mathbf{R})$.
- Step 2** For each $t = 1, \dots, T$ define the threshold interval sets (TIS) as $A_{u_t} = \{\mathbf{s}_i: u_t - \mathbf{R} < u_{t-1}\}$, where $u_0 = \max(\text{abs}(\mathbf{R}))$.
- Step 3** For each TIS A_{u_t} define clusters $C_1, C_2, \dots, C_{g_t}$ which satisfy $\min \|a_{k,p} - a_{l,q}\| > h$, for all $p \neq q$, where $a_{k,p} \in C_p$ for $k = 1, \dots, |C_p|$.
- Step 4** Define a set of g_t candidate knots $\mathbf{u}_t$ by the centroids of the clusters $C_1, C_2, \dots, C_{g_t}$.
- Step 5** Repeat steps 2-4 for $-\mathbf{R}$.
- Step 6** Discard any candidate knots from $\mathbf{u}_t$ if $\min \|\mathbf{u}_{t,i} - \mathbf{u}_{t',j}\| < h$, for $t' = 1, \dots, t-1$. I.e., discard any candidate knots placed too close to knots selected from previous thresholds.

The value of T may impact the estimation of knot locations: larger values of T correspond to a larger number of knots being selected. We conducted a simulation study to investigate the effect of T , and found that $T = 15$ thresholds tended to be sufficient for the datasets and simulations studies considered here. Increasing T beyond this did not provide much improvement in either the accuracy or the uncertainty of the predictions. See Section 3 for the details and results of this simulation study.

Note that any region in which there are no knots will implicitly have the spatial random effects constrained to zero (i.e., the $\mathbf{S}\boldsymbol{\eta}$ term in Equation (1) will have no effect). This will result in prediction uncertainties (e.g., mean square prediction error, MSPE) that are too small when making predictions in regions far from the knot locations. Consequently, when determining knot locations it is important to ensure that the spatial domain is adequately covered. If the observed locations cover approximately the whole domain, then TKS as described above will typically accomplish this. However there may be cases when TKS does not fully cover the spatial domain, particularly if there are large gaps in the domain where no data were observed. In such cases, it may be necessary to use some method of placing knot locations in the “empty” areas of the domain. We recommend augmenting TKS with a grid of knot locations selected by the following criteria.

First obtain a set of knot locations, denoted $\mathbf{u}$, using TKS, and define $\bar{d}_{\mathbf{u}}$, the average distance from a knot to the next two closest knots. Then create a grid of knot locations $\mathbf{u}^*$ over the entire spatial domain such that the distance between adjacent grid points is $\bar{d}_{\mathbf{u}}$. For each knot $\mathbf{u}_i$, remove any knot $\mathbf{u}_j^*$ which satisfies $\|\mathbf{u}_i - \mathbf{u}_j^*\| < \bar{d}_{\mathbf{u}} / 2$. The $\bar{d}_{\mathbf{u}} / 2$ criteria is used to define the approximate neighborhood of $\mathbf{u}_j^*$, and hence if a TKS-selected knot $\mathbf{u}_i$ is nearby, we give precedence to $\mathbf{u}_i$. We refer to this hybrid approach as *augmented* Threshold Knot Selection, or aTKS.

As a final comment regarding selecting knot locations, Cressie and Johannesson (2008) strongly recommend that the basis functions used in a RRSM be multi-resolutional, in order to capture multiple scales of spatial dependence (e.g., long- and short-range dependence) (Nychka et al., 2002). Using the TKS algorithm described above, it is possible to select multiple resolution of knot locations. This can be achieved through separate runs of TKS (or aTKS) with different bandwidths h . We illustrate this in our data analysis in Section 4.

2.3. Selecting the Number of Knots

As in nonparametric regression or density estimation problems, the selection of the bandwidth is a component of this algorithm. If the bandwidth is too small, the knots selected will be too granular (too many knots, and increased computational demand). Conversely, too large a bandwidth will result in over-smoothing at best, or at worst in clustering together distinct regions of the domain and placing a knot between them, rather than within each. Unfortunately the spatial predictions cannot be expressed as a function of the bandwidth (or even as a function of the number of knots), and so an analytical expression for computing the bandwidth is not feasible. Instead, we develop a simple empirical procedure based on cross-validation to select the bandwidth.

First, set aside a set of m observations as a test set and use the remaining $n - m$ observations as a training set. Then, define a sequence of bandwidths and, for each bandwidth, use the training set to obtain the TKS knots. Once the knot locations have been determined, $\mathbf{S}$ is fully determined and one may fit the RRSM in equation (1) and obtain predictions $\hat{Z}_i$, $i = 1, \dots, m$. Multiple options exist to estimate parameters and make predictions. In our work, we use the framework described in Cressie and Johannesson (2008), where the prediction is the best linear unbiased predictor (BLUP) for a mixed effects model. After obtaining predictions, we evaluate an objective function such as $D(h) = (r_h^k / m) \sum_{i=1}^m (Z_i - \hat{Z}_i)^2$, where r_h^k is the number of knots that were selected for bandwidth h , raised to the power k . The selected bandwidth should be the one that minimizes $D(h)$. This procedure requires selecting a sequence of bandwidths to test, which can be a subjective choice. We suggest some fractions of the largest distance in the dataset. In our simulations, we found a successful approach to be selecting a set of bandwidths between 3% and 10% of the maximum distance of locations. The function $D(h)$ is simply a hold-out estimate of the mean squared error (MSE), penalized by a function of the number of knots selected for bandwidth h . One could substitute other measures functions instead of the MSE, such as the Continuous Ranked Probability Score (CRPS Wilks, 2006).

Based on the result of a simulation study (see Section 3, design 2) we set the penalty to be $k = 0.25$, which lead to good performance in the subsequent simulation comparing several methods (design 3). One of the goals of our simulations was to identify values for the tuning parameters which worked well in some contexts. However, a user may find it preferable to ‘tune’ TKS more fully for a specific analysis. This could be accomplished with a cross-validation approach on a hold-out sample of data, similar to how we assess the performance of TKS in the data analysis (see Section 4).

Our simulations suggest that this bandwidth selection method produces a stable estimation of knot locations in spite of using only one hold-out test set. As described previously, TKS initially places knots in regions where the magnitude of $\mathbf{R}$ is largest. These regions of large $\mathbf{R}$ are not markedly changed by holding out a relatively small proportion (e.g., 10%) of the data. Consequently, over many different random test sets, these regions of large $\mathbf{R}$ and the knots placed in them tend to remain stable. We conducted a simulation study to illustrate this stability of TKS over different random test sets. The simulations investigating the choice of penalty on r_h , and the stability of knot locations over many hold-out samples, are described in Section 3.

Because the bandwidth is only indirectly related (via the number of selected knot locations) to the prediction, the properties of this bandwidth selection rule cannot be studied analytically as in Härdle and Marron (1985). However, as demonstrated in the next section, the TKS algorithm, with the proposed bandwidth selection rule, performs well against comparable methods of choosing knot locations.

3. Simulation Studies

We investigate the properties and performance of TKS through simulation. First we describe how we simulate data, and then discuss results of the simulation studies.

3.1. Data Simulation

We simulate data according to the model in equation (1), we select $r_0 = 400$ initial knots on a uniform grid, and $n = 10,000$ locations randomly over a 100×100 square (though in design 3 we also investigate $n = 15,000$ and $n = 20,000$). The $\mathbf{X}$ matrix was defined as $[\mathbf{1}, \text{X-coord}, \text{Y-coord}]$, and $\boldsymbol{\beta} = (1, 2, 3)'$.

To create the reduced rank covariance $\mathbf{V}$, we first define $\mathbf{V}_0$ using the r_0 initial knots, where the $(i, j)^{th}$ element is determined from an exponential covariance function, given by $\mathbf{V}_0(i, j) = \exp \{-\|\mathbf{u}_i - \mathbf{u}_j\|/\Theta\}$. The parameter Θ controls the range of spatial dependence; small values lead to less spatial dependence, and larger values lead to greater spatial dependence. Then $\mathbf{V}$ is taken as a random realization from an Inverse-Wishart distribution with $\text{df} = r_0$ and $\mathbf{V}_0$ as the scale matrix. The bisquare function given in equation (2) was used to define $\mathbf{S}_0$, using the r_0 initial knots and setting $a_j = 1.5 d_k$, where $d_k = 5$ was the minimum distance between two knots. Exploring other forms of $\mathbf{V}_0$ may be of interest, though by drawing $\mathbf{V}$ from an Inverse-Wishart distribution, the covariance matrix is not constrained to be stationary or isotropic. For simplicity, we specify $\mathbf{V}$ in this manner for all of our simulations.

The white-noise errors δ and ϵ were simulated together from $N(0, \nu^2 \mathbf{I})$, where $\mathbf{I}$ is the identity matrix, and setting $\nu^2 = 0.1$, and η was sampled from $N_{r_o}(\mathbf{0}, \mathbf{V})$. Finally, $\mathbf{Z}$ was computed by $\mathbf{Z} = \mathbf{X}\beta + \mathbf{S}_0\eta + \delta + \epsilon$.

In the simulations designs where we consider predictive performance (simulation designs 2 and 3), the spatial predictions are computed using the BLUP, as we used during the bandwidth selection, with parameters estimated through an expectation-maximization (EM) algorithm. Unless otherwise noted (simulation design 2), we use parameter settings as described previously in all simulations. Particularly, we (1) set the number of thresholds at $T = 15$, (2) defined the thresholds as equally spaced percentiles between the 98th and 50th percentile (so the levels were equally spaced, but not necessarily the thresholds themselves), (3) used the $k = 0.25$ penalty on the number of knots in the objective function, and (4) tested a sequence of bandwidths between 3% and 10% of the maximum possible distance (in our simulations, 141 units).

3.2. Simulation design 1: Stability of TKS

We stated in Section 2.3 that the TKS algorithm is stable in spite of using only one hold-out test set. To demonstrate this, we generated four datasets as described, and for each we ran 100 repetitions of TKS using different hold-out samples for each repetition. Figure 2 shows the results of this simulation. The detail residuals of the four datasets are plotted, and we overlay the knots from each of the 100 repetitions. These four datasets can be considered as random blocks: each dataset should be considered in isolation, comparisons between them are not meaningful. The concept of interest is to observe a sort of sampling distribution of the knot locations, and that over many different hold-out samples this distribution of knot locations is sufficiently stable (recall that the goal of TKS is to place knots where the detail residuals are large, and then to sufficiently fill out the spatial domain).

Visible in Figure 2 are areas showcasing the stability of the TKS knots. For example in dataset 3, at coordinates (93, 79), there is a cluster of very large detail residuals (e.g., a local peak), and all 100 repetitions placed a knot in nearly the exact same location. Other examples of these local peaks and valleys can be seen throughout these datasets, and the knots there tend to be placed at nearly coincident locations for all 100 repetitions.

We also see regions where the knot locations are not as stable, but these regions are where the detail residuals are close to 0. As discussed in Section 2.2, the location of the knots is less important in these regions because the large-scale trend has already produced reasonable predictions, and therefore the goal is primarily to adequately fill the domain with knot locations (which may then be augmented with a grid to ensure full spatial coverage). To summarize: in areas where $\mathbf{R}$ are large, the estimated knot locations are highly stable over different hold-out test sets, and in areas where the estimated knot locations shift between different hold-out sets, $\mathbf{R}$ are small and the increased variability is less important. Consequently, using a single hold-out sample to estimate the knot locations is adequate, rather than a more complex and computationally intensive cross-validation procedure.

3.3. Simulation design 2: Investigation of Penalty

In a second simulation study, our goal was to determine the optimal penalty to use on the number of knots in the objective function, and the number of thresholds, T , to define. In Section 2.3 we noted that we set $k = 0.25$ in the objective function $D(h)$. In this simulation we explored the use of other powers for this penalty, looking at a sequence of values $k = \{0, 0.25, 0.33, 0.50, 0.66, 1\}$ for each of $T = \{3, 5, 10, 15, 20, 25\}$. We first generated a dataset of size $n = 10000$, and held out $m = 500$ observations as a test set. To compare the levels of T and k , we computed several metrics over the test set, including the MSE, MSPE, and the CRPS; we also timed the process of selecting knots and making predictions. Small values are preferable for all of these measures. Because one should ensure full spatial coverage, the results shown consider the aTKS described earlier.

The results of simulation design 2 are shown in Figure 3. This figure considers only the CRPS, but results for the MSE and MSPE were similar. In particular, we call attention to the effect of the number of thresholds, T . Note that there is, as expected, a decreasing trend: as T increases, the predictions tend to improve. However, there is a decreasing return which can be observed. Approximately 15 thresholds appears to be a reasonable value, and therefore we set $T = 15$ for the rest of our simulations.

Figure 4 shows results of a subset of simulation design 2 where $T = 15$. The plotted lines are the empirical cumulative density function (ECDF) of the product of the CRPS and the computation time. This value measures a balance of accuracy and computation time, both of which should be minimized.

While Figure 4 considers only the case where $T = 15$, similar patterns were observed across the values of T . None of the penalties considered performed optimally in all cases, but $k = 0.25$ tended to be one of or the best performing penalty in most cases. While the penalty $k = 0.5$ is similar in this figure, the 0.25 penalty produced superior predictions when considering CRPS only. For this reason, we chose $k = 0.25$ in $D(h)$, and use this throughout our simulations and data example.

3.4. Simulation design 3: Comparison of Predictive Performance

Finally, we seek to evaluate TKS compared to other methods, such as the space-filling design (SF method) mentioned earlier, and placing knots on a regular grid (Grid method). One issue to address is that neither the Grid nor SF methods are designed to estimate the appropriate number of knot locations. Consequently, we equalize the number of knots between the three methods. For example, if TKS selected r knots, then we would select r locations using the SF method, and also create a grid of approximately r locations (because r may not yield a grid of equidistant locations). We take this approach because it has been noted (Gelfand et al., 2011) that the number of knots is more important than the precise locations, hence equalizing the number of knots seems to be the most fair manner by which to compare the methods. As in simulation design 2, we consider the augmented scheme, aTKS (the SF method was similarly augmented).

Other methods of knot placement were either unavailable, or used inherently more time-intensive methods (e.g., Bayesian methods), which was not the focus of the present research.

Furthermore, the two methods to which we compare are, like aTKS, independent of any assumed covariance function: The Grid method is not informed by the data at all, while the SF method is based on geometry only, the locations of observations.

We assess the impact of three factors on the performance of our method: the sample size, the degree of spatial dependence, and the distribution of locations. For the sample size, we consider three values $n = \{10000, 15000, 20000\}$; for the spatial dependence parameter, we consider the values $\Theta = \{0.001, 0.05, 5, 10, 20, 50, 100\}$. Three patterns of locations were tested, which we refer to as ‘uniform’, ‘radial’, and ‘spots’. The uniform pattern had locations approximately uniformly over the domain. In the radial pattern the density of locations decreased from dense to sparse as distance from a selected epicenter increased. Finally, the spots pattern contained five epicenters of highly dense locations, with sparse regions all around.

Each combination of n , Θ , and the pattern of locations was tested, totaling 63 parameter combinations. For each combination of parameters, we ran 100 replications. For each replication, we held out all observations in a region of the domain (the square bounded by the point (15, 15) and (35, 35)), and used the rest of the data to select the knots. For each method, we obtained spatial predictions and computed the MSE, MSPE, and CRPS over the held-out observations.

Partial results of the simulation study are shown in Figure 5 which plots the median CRPS over the 100 replications at each combination of parameters. While not exhibiting uniformly better results, the aTKS method produced the smallest median CRPS in 38 of the 63 cases, frequently by a substantial margin. In the cases where TKS did not produce the smallest median CRPS, it was frequently beaten by a small margin. This indicates that for the simulation designs considered, aTKS is able to select a set of knots which produce equally good or superior predictions than either the Grid or SF methods, when considering both the accuracy and uncertainty of the prediction.

As opposed to aggregated results, we also considered the results of the three methods for each of the 6300 simulation runs (100 replications of 63 cases) on a more individual level. From this perspective, aTKS produced a smaller CRPS than the SD method in 62.2% of the runs, and a smaller CRPS than the Grid method in 60.6% of the runs. As with the aggregate results, this shows that while aTKS is not uniformly superior, it tends to produce better predictions.

In addition to the estimation of r , and improved accuracy, aTKS also offers some computational efficiency compared to the space-filling method. Table 1 (column 2) provides the median elapsed time for each combination of parameters, which shows that aTKS tended to be quicker computationally than the SF method. Again looking at the 6300 individual simulation runs, in 61.2% of them aTKS used less computation time than the SF method. Computation times include selecting the knot locations (for the aTKS and SF methods), estimating model parameters using an EM algorithm, and computing the BLUP predictions for the held-out locations.

The Grid method, as one should expect, used less computation time than aTKS, but there are two important points to raise in this regard. First, the Grid method does not estimate the number of knot locations, but simply utilized the value of r estimated by aTKS as if it took no time to determine (the same applies to the SF method). Secondly the Grid method is not informed by the data. On the other hand, aTKS does draw information from the data, which is reflected in the generally improved predictive performance of aTKS over the Grid method shown in this simulation.

4. Application to *Deepwater Horizon* Data

Over the course of the oil spill clean-up, CTEH collected 113,992 measurements of $PM_{2.5}$ along the Gulf Coast. The data were collected by mobile monitors which did not follow regular intervals in either time or space. Spatially this presents no difficulty, as geostatistics assumes that the spatial locations are continuous across some domain. As a proof of concept, we focus on a short interval, the week of June 21 and June 27, 2010. During this time, oil was still leaking from the well, and *in situ* burnings were taking place. Exploratory analysis revealed differences in $PM_{2.5}$ between days, but not over the course of a day. Therefore, we included an indicator for each day in the matrix of covariates (X). Other covariates were latitude and longitude coordinates, distance from the *Deepwater Horizon* oil rig; angle from the longitude of the oil rig (due north of the rig would have angle 0° , due east would have angle 90° , etc), to help account for atmospheric conditions such as prevailing wind patterns; and the interaction (product) between the distance and angle.

Distance was measured in kilometers using the haversine formula. If λ and ϕ denote the longitude and latitude, respectively (in radians), then for two locations $s_1 = (\lambda_1, \phi_1)$, $s_2 = (\lambda_2, \phi_2)$ the haversine great-circle distance is, $\text{distance} = 2 \times 6371 \times \text{atan2}(\sqrt{a}, \sqrt{1-a})$, where, $a = \sin^2\left(\frac{\phi_2 - \phi_1}{2}\right) + \cos(\phi_1) \times \cos(\phi_2) \times \sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)$. We set the thresholds in the same manner as in the simulations: 15 equally spaced percentiles of $\text{abs}\{\mathbf{R}\}$, from the 98th percentile down to the 50th percentile. The bandwidths tested were 13, 15, ..., 21 km for the first resolution, and 3, 7, ..., 11 km for the second resolution. Since the observed distribution of $PM_{2.5}$ was right-skewed, we analyzed the data after taking the natural logarithm of the measurements. As with the simulations, we used the BLUP to make spatial predictions, which in this case entailed log-normal kriging (Cressie, 1993).

In total there were 9477 measurements of $PM_{2.5}$ in the time interval we analyzed. For evaluation purposes, we randomly selected $M = 477$ observations to hold out as a test set, and use $n = 9000$ observations as a training set to estimate the knot locations. Due to the highly irregular pattern of observed locations, we obtained knot locations only using TKS and the SF method. We first ran aTKS to obtain two resolutions of knots of size $r_1 = 196$ and $r_2 = 432$. We then used the SF algorithm to obtain two resolutions of knots of approximately the same size, specifically $r_1 = 199$ and $r_2 = 449$.

In this illustration, TKS used more time to estimate knot locations than the SF method (20.7 minutes vs. 1.3 minutes). However, the aTKS method improved predictive performance relative to the SF design. For example, the MSE over the held-out observation was 135 for

aTKS, and 136 for the SF method. While this does not appear to be a dramatic improvement, aTKS also improve the prediction uncertainty. For example, the MSPE for aTKS was 14% smaller, on average, for aTKS than the SF method; and the CRPS for aTKS was 7.92 compared to 7.99 for the SF method. Note that aTKS attained these results in spite of the fact that slightly more knot locations were selected for the SF method.

PM_{2.5} can come from a number of sources in addition to the *in situ* burns and flaring of oil (e.g., exhaust from ships and vehicles participating in the clean-up, decontamination of vessels and beaches, etc.); future analyses could benefit by incorporating such information could result in better estimation of the trend, and therefore improve the predictions. Unfortunately, the CTEH data did not include any additional information by which to do so. Given the fact that this exercise was a proof of concept for the TKS method, we did not attempt to incorporate any additional data sources to be able to use these data in any analyses of potential health outcomes.

5. Summary

In this paper we have introduced the TKS method. This is an objective means of simultaneously estimating the number and locations of knots for a RRSM. The method is computationally efficient and, for a given r , generally improves the fit of a RRSM compared to a gridded knot method or the space-filling method. Other attractive features of the TKS algorithm are that its mechanics can be intuitively understood, and that it estimates knot locations while respecting the underlying smoothness of the observed data.

Important to note is that the bandwidth selection described in Section 2.3 is a novel contribution in the selection of knot locations. Most of the existing methods of placing knot locations consider a fixed value for the number of knots. By incorporating a function like $D(h)$, we introduce a mechanism to select the number of knot locations based on the data. Though we did not investigate it here, one could use the same mechanism for the space-filling or grid designs. Hence, our contribution in this paper is two-fold: the TKS algorithm, and an objective function $D(h)$ to determine the number of knot locations.

As noted, the use of single-linkage clustering is a potential drawback of TKS. When there is no spatial correlation, TKS tended to select very few knots due to the noise in the detail residuals. Single-linkage clustering also causes TKS to not consider the shape of any peaks or valleys in the detail residuals. Better representation of the shape of clusters could lead to further improvements in the predictions. Though we note that TKS shows promise in spite of these potential drawbacks.

TKS is also a distribution-free method of estimating knot locations. In Section 3 we simulated from a multivariate Normal distribution, and made use of an EM algorithm to estimate the spatial covariance matrix, but one could easily use another estimate for the spatial covariance matrix that is robust or distribution-free, for example the method of moments estimate discussed in Cressie and Johannesson (2008).

The spatial mixed effects model, upon which TKS is based, is a general framework for modeling large-scale spatial data. While we used Empirical Bayesian methods for estimation

and prediction, TKS is not tied to this choice. Bayesian hierarchical models are a popular choice for modeling large-scale spatial data, and TKS can be used in this context. For example, the adaptive Bayesian model of Guhaniyogia et al. (2011) approximately doubles the computation time necessary. Significant computational time could be saved by using TKS to estimate the knot locations, and then using a non-adaptive Bayesian model to estimate parameters and make predictions. Given that the TKS algorithm has shown evidence of significant computational efficiencies while outperforming several existing methods of estimating knots, it should be considered when analyzing large-scale spatial datasets.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Software, Acknowledgments, and Disclosure

The Threshold Knot Selection algorithm has been implemented in the R programming language (R Core Team, 2017). The code for TKS is available at <https://github.com/jelsema/RRSM>. In particular, see the function TKS. Should the link deprecate, contact the first author for information regarding how to obtain the code.

The authors affirm that they have no financial interest or benefit resulting from direct applications of this work. This research was supported (in part) by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences (Z01 ES101744 and Z01 ES 102945) and the NIH Common Fund. In addition to the editorial staff of the Journal of Statistical Computation and Simulation and the peer reviewers, the authors thank Drs. Rajib Paul, Lawrence S. Engel, Patricia Stewart, and Mark Stenzel. All of these individuals provided suggestions which substantially improved the content and presentation of this work.

References

- Banerjee Sudipto, Gelfand Alan, Finley Andrew, and Sang Huiyan. Gaussian predictive process models for large spatial datasets. *Journal of the Royal Statistical Society: Series B*, 70(4):825–844, 2008.
- Banerjee Sudipto, Finley Andrew, Waldmann Patrik, and Ericsson Tore. Hierarchical spatial process models for multiple traits in large genetic trials. *Journal of the American Statistical Association*, 105(490):506–521, 2010. [PubMed: 20676229]
- Crainiceanu Ciprian, Diggle Peter, and Rowlingson Barry. Bivariate binomial spatial modelling Loa loa prevalence in tropical africa. Johns Hopkins University Dept of Biostatistics Working Papers, Working paper 103, 2006 URL <http://biostats.bepress.com/jhubiostat/paper103>.
- Cressie Noel. *Statistics for Spatial Data*. Wiley, New York, second edition, 1993.
- Cressie Noel and Johannesson Gardner. Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society: Series B*, 70(1):209–226, 2008.
- Cressie Noel, Shi Tao, and Kang Emily L.. Fixed rank filtering for spatio-temporal data. *Journal of Computational and Graphical Statistics*, 19(3):724–745, 2010.
- Dempster Arther P, Laird Nan M, and Rubin Donald B. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–38, 1977.
- DHRC. Deepwater Horizon Research Consortia. <http://www.niehs.nih.gov/research/supported/dert/programs/gulfconsortium/>, 2016 [Online; accessed: 2016-02-16].
- Diggle Peter and Lophaven Sored. Bayesian geostatistical design. *Journal of Scandanavian Journal of Statistics*, 33:53–64, 2006.
- Gelfand Alan, Banerjee Sudipto, and Finley Andrew. Spatial design for knot selection in knot-based dimension reduction models In Mateu Jorge and Miller Werner, editors, *Spatio-temporal Design: Advances in Efficient Data Acquisition*. Wiley, 2011.

- Goldstein Bernard D., Osofsky Howard J., and Lichtveld Maureen Y.. The gulf oil spill. *New England Journal of Medicine*, 364:1334–1348, 2011. [PubMed: 21470011]
- Guhaniyogia Rajarshi, Finley Andrew, Banerjee Sudipto, and Gelfand Alan. Adaptive gaussian predictive process models for large spatial datasets. *Environmetrics*, 22:997–1007, 2011. [PubMed: 22298952]
- Hamra Ghassan B., Guha Neela, Cohen Aaron, Laden Francine, Raaschou-Nielsen Ole, Samet Jonathan M., Vineis Paolo, Forastiere Francesco, Saldiva Paulo, Yorifuji Takashi, and Loomis Dana. Outdoor particulate matter exposure and lung cancer: A systematic review and meta-analysis. *Environmental Health Perspectives*, 122:906–911, 2014. [PubMed: 24911630]
- Härdle Wolfgang and Marron James Stephen. Optimal bandwidth selection in non-parametric regression function estimation. *The Annals of Statistics*, 13(4):1465–1481, 1985.
- Huynh Tran, Ramachandran Gurumurthy, Banerjee Sudipto, Monteiro Joao, Stenzel Mark, Sandler Dale P., Engel Lawrence S., Kwok Richard K., Blair Aaron, and Stewart Patricia A.. Comparison of methods for analyzing left-censored occupational exposure data. *Annals of Occupational Hygiene*, 58(9):1126–1142, 2014. [PubMed: 25261453]
- Kang Emily L. and Cressie Noel. Bayesian inference for the spatial random effects model. *Journal of the American Statistical Association*, 106(495):972–983, 2011.
- Kang Emily L., Cressie Noel, and Sain Stephan R.. Combining outputs from the North American Regional Climate Change Assessment Program by using a Bayesian hierarchical model. *Journal of the Royal Statistical Society C*, 61(2):291–313, 2012.
- Katzfuss Matthias. Bayesian nonstationary spatial modeling for very large datasets. *Environmetrics*, 24:189–200, 2013.
- Katzfuss Matthias and Cressie Noel. Tutorial on fixed rank kriging (FRK) of CO₂ data. Technical Report 858, Department of Statistics, The Ohio State University, Columbus, OH, 2011a.
- Katzfuss Matthias and Cressie Noel. Spatio-temporal smoothing and EM estimation for massive remote-sensing datasets. *Journal of Time Series Analysis*, 32:430–446, 2011b.
- Lee Byeong-Jae, Kim Bumseok, and Lee Kyuhong. Air pollution exposure and cardiovascular disease. *Toxicological Research*, 30(2):71–75, 2014. [PubMed: 25071915]
- Madrid AE, Angulo JM, and Mateu J. Spatial threshold exceedance analysis through marked point processes. *Environmetrics*, 23(1):108–118, 2012.
- McCauley Linda A.. Environments and health: Will the BP oil spill affect our health? *American Journal of Nursing*, 110(9):54–56, 2010.
- Nychka Douglas and Saltzman Nancy. Case studies in environmental statistics In *Lecture Notes in Statistics*, volume 132 New York: Springer-Verlag, 1998.
- Nychka Douglas, Wikle Christopher, and Royle J. Andrew. Multiresolution models for nonstationary spatial covariance functions. *Statistical Modeling*, 2(4):315–331, 2002.
- Quick Harrison, Groth Caroline, Banerjee Sudipto, Carlin Bradley P., Stenzel Mark R., Stewart Patricia A., Sandler Dale P., Engel Lawrence S., and Kwok Richard K.. Exploration of the use of Bayesian modeling of gradients for censored spatiotemporal data from the Deepwater Horizon oil spill. *Spatial Statistics*, 9: 166–179, 2014. [PubMed: 25599019]
- R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2017 URL <https://www.R-project.org/>.
- Ramseur Jonathan L.. Oil spills in US coastal waters: Background and governance. <https://www.fas.org/sgp/crs/misc/RL33705.pdf>, 1 2012 [Online; accessed: 2016-02-16].
- Ruppert David, Wand MP, and Carroll RJ. *Semiparametric Regression*. Cambridge University Press, 2003.
- Sahu Sujit and Challenor Peter. A space-time model for joint modeling of ocean temperature and salinity levels as measured by Argo floats. *Environmetrics*, 19: 509–528, 2008.
- Shi Tao and Cressie Noel. Global statistical analysis of MISR aerosol data: a massive data product from NASA's Terra satellite. *Environmetrics*, 18:665–680, 2007.
- Ver Hoef Jay M. and Jansen John K.. Estimating abundance from counts in large data sets of irregularly spaced plots using spatial basis functions. *Journal of Agricultural, Biological, and Environmental Statistics*, 20(1):1–27, 12 2014 URL 10.1007/s13253-014-0192-z.

Wilks DS. Statistical Methods in the Atmospheric Sciences. Academic Press, San Diego, USA, 2 edition, 2006.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 1.

Measurements of PM_{2.5} (in micrograms per cubic meter, $\mu\text{g}/\text{m}^3$) taken along the Gulf Coast between June 21, 2010 and June 27, 2010. For better visualization the latitudes and longitudes have been randomly ‘jittered’ by 0.03°. The triangle at -88.38 longitude, 28.74 latitude is the location of the *Deepwater Horizon* drilling rig.

Figure 2.

Four datasets simulated as described above, the values plotted are the detail residuals. Knot locations estimated using TKS from 100 repetitions using different hold-out test sets are plotted over the detail residuals as black dots. In the online version of this manuscript, the detail residuals are shown in color to better differentiate the simulated data and the estimated knot locations.

Figure 3.

Boxplot of continuous ranked probability score for each combination of T (number of thresholds) and k (penalty on selected number of knots).

Figure 4.

ECDF of the continuous ranked probability score, multiplied by computational time for each of a sequence of values of k in the objective function $D(h)$. Note that the penalty $k = 0.25$ has been plotted with a solid line, all others with dashed lines. In the online version of this manuscript, this figure has color to better distinguish the dashed lines.

Figure 5.

Results of simulation design 3, evaluating the predictive performance of several methods of knot placement. The y-axis denotes the CRPS, and the x-axis denotes the value of the spatial dependence parameter Θ . The solid lines with circles denote the aTKS method, dashed lines with squares denote the SF method, and the dot-dashed lines with triangles denote the Grid method. The values plotted are the median CRPS over 30 replications.

Table 1.

Results of simulation design 3, continuous ranked probability score and computational time between aTKS, the SF method, and the Grid method. Results shown are for $n = 15,000$. Full results are provided in the Supplementary Materials.

Pattern	θ	Median CRPS			Median Time (s)		
		aTKS	SF	Grid	aTKS	SF	Grid
Uniform	0.001	1.26	1.47	1.50	272	299	66
Uniform	0.05	1.14	1.46	1.54	288	317	70
Uniform	5	1.17	1.23	1.41	261	273	62
Uniform	10	1.12	1.22	1.30	225	225	55
Uniform	20	0.93	1.19	1.00	221	180	38
Uniform	50	0.72	0.80	0.67	196	138	26
Uniform	100	0.57	0.62	0.62	197	116	18
Spots	0.001	1.33	1.70	1.55	370	493	175
Spots	0.05	1.25	1.53	1.64	385	519	192
Spots	5	1.29	1.25	1.61	363	464	168
Spots	10	1.09	1.14	1.25	311	350	146
Spots	20	1.00	0.98	1.01	287	311	118
Spots	50	0.67	0.69	0.59	246	235	81
Spots	100	0.60	0.62	0.53	221	177	54
Radial	0.001	1.27	1.53	1.72	321	586	150
Radial	0.05	1.33	1.36	1.63	335	574	156
Radial	5	1.35	1.34	1.48	328	506	148
Radial	10	1.00	1.12	1.13	277	397	114
Radial	20	1.05	1.10	1.03	267	319	97
Radial	50	0.64	0.74	0.69	214	218	56
Radial	100	0.53	0.60	0.59	186	165	40