Spectral methods to approximate the likelihood for irregularly spaced spatial data[1]

Montserrat Fuentes

Mimeo Series 2568 - SUMMARY

Likelihood approaches for large irregularly spaced spatial datasets are often very difficult, if not infeasible, to use due to computational limitations. Even when we can assume normality, exact calculations of the likelihood for a Gaussian spatial process observed at n locations requires $O(n^3)$ operations. We present a version of Whittle's approximation to the Gaussian log likelihood for spatial regular lattices with missing values and for irregularly spaced datasets. This method requires $O(nlog_2n)$ operations and does not involve calculating determinants. Due to the edge effect the estimated covariance parameters using this approximated likelihood method are efficient only in one dimension. To remove this edge effect, we introduce data tapers. In spatial statistics, data tapers are often the tensor product of two one-dimensional tapers. However, we generally need more tapering for the corner observations. Thus, we introduce here a new spatial data taper, a circular taper, that gives more tapering to the corner observations. Therefore, with less overall tapering, we get the amount of smoothing that we need without losing so much information. We present simulations and theoretical results to show the benefits and the performance of the data taper and the spatial likelihood approximation method presented here for spatial irregularly spaced datasets and lattices with missing values.

[1]M. Fuentes is an Associate Professor at the Statistics Department, North Carolina State University (NCSU), Raleigh, NC 27695-8203, and a visiting scientist at the US Environmental Protection Agency (EPA). Tel.:(919) 515-1921, Fax: (919) 515-1169, E-mail: fuentes@stat.ncsu.edu.

Key words: covariance, Fourier transform, periodogram, spatial statistics, satellite data, tapering.

1 Introduction

Statisticians are frequently involved in the spatial analysis of huge datasets. In this situations, calculating the likelihood function, to estimate the covariance parameters or to obtain the predictive posterior density, is very difficult due to computational limitations. Even if we can assume normality, calculating the likelihood function involves $O(N^3)$ operations, where N is the number of observations.

However, if the observations are on a regular complete lattice, then it is possible to compute the likelihood function with fewer calculations using spectral methods (Whittle 1954, Guyon 1982, Dahlhaus and Künsch, 1987, Stein 1995, 1999). These spectral methods are based on the likelihood approximation proposed by Whittle (1954). Whittle's approximation is for Gaussian random fields observed on a regular lattice without missing observations. In practice, very often the data will be irregularly spaced or will not be on a complete regular lattice. Even for satellite data (Figure 1), the observations might actually be on a lattice but due to clouds or other phenomenon, there are generally missing values, so we could not use Whittle's approximation to the likelihood.

Spectral methods for irregular time series have been studied, e.g. by Parzen (1963), Bloomfield (2000), Neave (1970), Clinger and Van Ness (1976), and Priestly (1981, p. 585), in the context of estimating the periodogram of a time series. But, nothing has been done yet in terms of introducing spatial likelihood approximation methods using spectral tools for irregularly spaced spatial datasets. In a spatial setting is worth to mention the simple likelihood approximation introduced by Vecchia (1988), by partition the data into clusters, and assuming that the clusters are conditionally independent. Pardo-Igúzquiza et al. (1997) wrote a computer program for Vecchia's approximation method. However, there have not been efforts to evaluate the effect of Vecchia's approximation to the spatial Gaussian likelihood on the estimated parameters (Lark, 2002, Stein et al., 2004). Stein et al. (2004) adapted Vecchia's approach to approximate the restricted likelihood of a Gaussian process, and discussed the computational challenges to obtain the derivatives of the approximated likelihood and restricted likelihood functions. A similar clustering framework for likelihood approximation was presented by Caragea (2003), in which the clusters were assumed conditionally independent after conditioning on the cluster mean. For Markov fields, a coding method for parameter estimation was proposed by Besag (1974). Besag and Moran (1975) introduced an exact likelihood method for rectangular

autonormal processes.

We present here powerful spectral methods to handle lattice data with missing values and irregularly spaced datasets. We obtain a representation of the approximated likelihood function in the spectral domain to estimate covariance parameters. This approach can be used for estimation and also for Bayesian spatial prediction. We use our method to approximate the likelihood function in the predictive posterior density, and we study the impact of the likelihood approximation on the Bayesian inference made.

Spatial tapering is crucial in two and higher dimensional problems where there are a large number of edge observations. The relative edge effect in two dimensions is the reason why the Whittle approximation is not efficient (Guyon, 1982). We propose here a new data taper, a circular taper, that gives more tapering to the corner observations. Thus, with less overall tapering, we get the amount of smoothing that we need without losing so much information.

This paper is organized as follows. In Section 2 we introduce the notion of periodogram, spectral density and the Whittle's approximation to the Gaussian likelihood. In Section 3, we present an approach to approximate the likelihood for Gaussian lattice processes with missing values. In Section 4, we propose a method to approximate the Gaussian likelihood for irregularly spaced datasets. Section 5 describes a new data taper function for spatial data, that we called "rounded tapering". We finish with a discussion.

2 Spectral Domain

2.1 Spectral Representation of a Stationary Spatial Process

A random field Z in $\mathbb{R}^2$ is called weakly stationary, if it has finite second moments, its mean function is constant and it possesses an autocovariance function C, such that $C(\mathbf{x}-\mathbf{y}) = \text{cov}\{Z(\mathbf{x}), Z(\mathbf{y})\}$. If Z is weakly stationary random field with autocovariance C, then we can represent the process in the form of the following Fourier-Stieltjes integral:

$$Z(\mathbf{x}) = \int_{\mathbb{R}^2} \exp(i\mathbf{x}^T\boldsymbol{\omega}) dY(\boldsymbol{\omega}) \tag{1}$$

where Y are random functions with uncorrelated increments, (see Yaglom (1987), Cramér and Leadbetter (1967) for example). The representation of a stationary random process $Z(\mathbf{x})$, for $\mathbf{x} \in \mathbb{R}^2$, in the form of

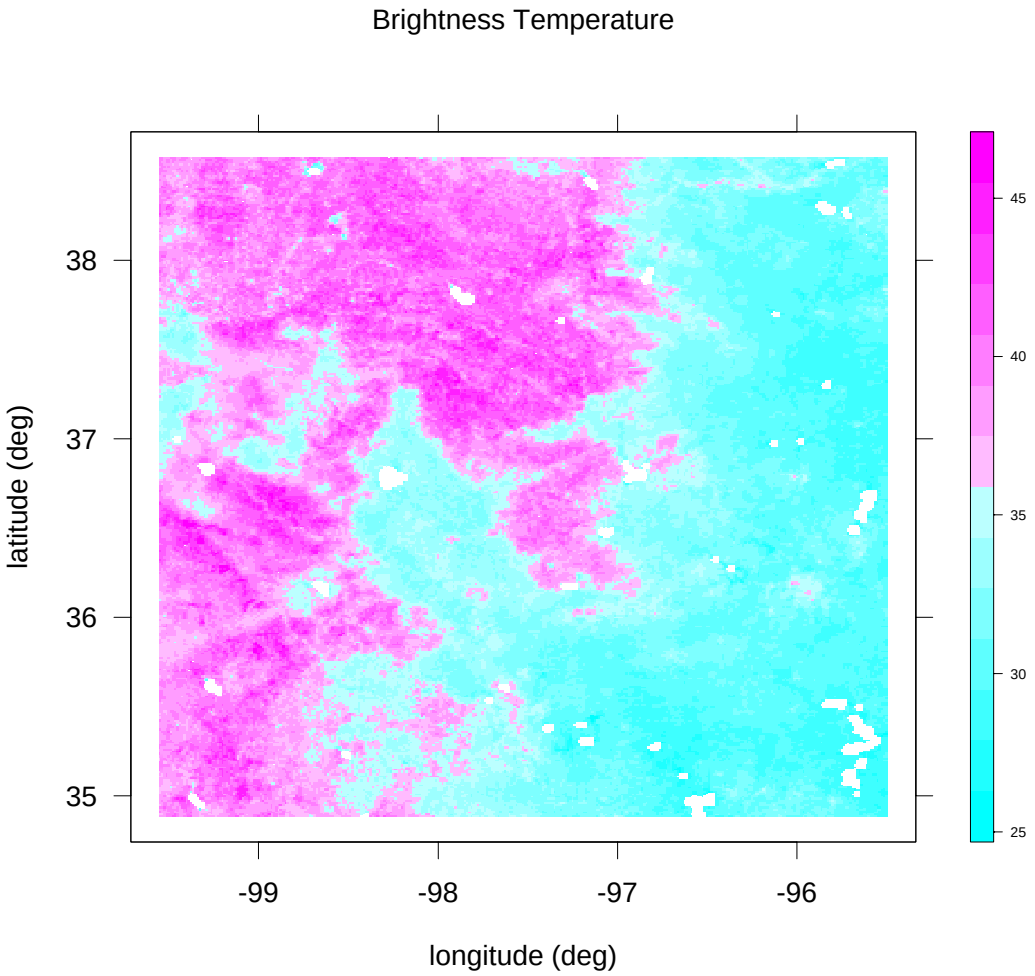

Figure 1: AVHRR data (1km×1km resolution), 140,000 pixels. The data represent satellite brightness temperature (°C) on June 19, 1996 in the Sourthern Great Plains. In white we have the missing values (due to clouds and lakes). We have 1% missing data.

the integral (1), is called the spectral representation of $Z(\mathbf{x})$, and $Y(\boldsymbol{\omega})$, with $\boldsymbol{\omega} \in \mathbb{R}^2$, is called the spectral process associated to Z. The spectral representation describes the harmonic analysis of a general stationary process $Z(\mathbf{x})$, i.e. its representation in a form of a superposition of harmonic oscillations. Using the spectral representation of Z and proceeding formally,

$$C(\mathbf{x}) = \int_{\mathbb{R}^2} \exp(i\mathbf{x}^T\boldsymbol{\omega}) F(\mathbf{d}\boldsymbol{\omega}) \tag{2}$$

where the function F is a nonnegative finite measure and it is called the spectral measure or spectrum for Z. The spectral measure F is the mean square value of the process Y,

$$E\{|dY(\boldsymbol{\omega})|^2\} = dF(\boldsymbol{\omega}).$$

If F has a density with respect to Lebesgue measure, this density is the spectral density, f, which is the Fourier transform of the autocovariance function:

$$f(\boldsymbol{\omega}) = \frac{1}{(2\pi)^2} \int_{\mathbb{R}^2} \exp(-i\mathbf{x}^T\boldsymbol{\omega}) C(\mathbf{x}) d\mathbf{x}.$$

Subject to the condition

$$\int_{\mathbb{R}^2} |C(\mathbf{x})| d\mathbf{x} < \infty$$

then F has a density, f.

By Bochner's Theorem, the function C is an autocovariance if and only if it can be represented as in (2), where F is a positive finite measure. Thus, the spatial structure of Z could be analyzed with a spectral approach or equivalently by estimating the autocovariance function.

We study now parametric models for the spectral density f. A class of practical variograms and autocovariance functions for continuous stationary processes Z can be obtained from the Matérn class (Matérn, 1960) of spectral densities

$$f(\boldsymbol{\omega}) = \phi(\alpha^2 + \|\boldsymbol{\omega}\|^2)^{\left(-\nu - \frac{d}{2}\right)} \tag{3}$$

with parameters $\nu > 0$, $\alpha > 0$ and $\phi > 0$, where d is the dimensionality of Z. Here, the vector of covariance parameters is $\theta = (\phi, \nu, \alpha)$. The parameter α^{-1} can be interpreted as the autocorrelation range. The parameter ν measures the degree of smoothness of the process Z, in that the higher the value of ν the

smoother Z would be, and ϕ is proportional to the the variance σ^2 times $\alpha^{2\nu}$. The corresponding covariance function for the Matérn class is given in (22) with a different parameterization. For further discussion about the Matérn class see Stein (1999, pp. 48-51).

If Z is observed only at uniformly spaced spatial locations Δ units apart, the spectrum of observations of the sample sequence $Z(\Delta\mathbf{x})$, for $\mathbf{x} \in \mathbb{Z}^2$, is concentrated within the finite frequency band $-\pi/\Delta \leq \boldsymbol{\omega} < \pi/\Delta$ (aliasing phenomenon).

The spectral density f_Δ of the process on the lattice can be written in terms of the spectral density f of the continuous process Z as

$$f_\Delta(\boldsymbol{\omega}) = \sum_{Q \in \mathbb{Z}^2} f\left(\boldsymbol{\omega} + \frac{2\pi Q}{\Delta}\right). \tag{4}$$

for $\boldsymbol{\omega} \in \Pi_\Delta^2 = [-\pi/\Delta, \pi/\Delta]^2$.

2.2 Periodogram

We estimate the spectral density of a lattice process, observed in a grid $(n_1 \times n_2)$, with the the periodogram,

$$I_N(\boldsymbol{\omega}) = (2\pi)^{-2}(n_1 n_2)^{-1} \left| \sum_{s_1=1}^{n_1} \sum_{s_2=1}^{n_2} Z(\mathbf{s}) \exp\{-i\mathbf{s}^T\boldsymbol{\omega}\} \right|^2. \tag{5}$$

We compute (5) for $\boldsymbol{\omega}$ in the set of Fourier frequencies $2\pi\mathbf{f}/\mathbf{n}$ where $\mathbf{f}/\mathbf{n} = \left(\frac{f_1}{n_1}, \frac{f_2}{n_2}\right)$, and $\mathbf{f} \in J_N$, for

$$J_N = \{\lfloor -(n_1-1)/2 \rfloor, \ldots, n_1 - \lfloor n_1/2 \rfloor\} \times \{\lfloor -(n_2-1)/2 \rfloor, \ldots, n_2 - \lfloor n_2/2 \rfloor\}. \tag{6}$$

We define a spectral window,

$$W(\boldsymbol{\omega}) = \prod_{j=1}^{2} \frac{\sin^2\left(\frac{n_j\omega_j}{2}\right)}{\sin^2\left(\frac{\omega_j}{2}\right)}$$

for $\boldsymbol{\omega} = (\omega_1, \omega_2) = 2\pi\mathbf{f}/\mathbf{n}$ and $\mathbf{f} \in J_N \backslash \{0\}$, we have

$$E(I_N(\boldsymbol{\omega})) = (2\pi N)^{-2} \int_{(-\pi,\pi]^2} f_\Delta(\boldsymbol{\theta}) W(\boldsymbol{\theta} - \boldsymbol{\omega}) \mathbf{d}\boldsymbol{\theta},$$

where $f_\Delta(\boldsymbol{\omega})$ is the spectral density of the process on the integer lattice. As n_1 and n_2 increase, W places more mass near the origin, $(0,0)$, so that if the spectral density $f_\Delta()$ is smooth in a neighborhood of $\boldsymbol{\omega}$, then $I_N(\boldsymbol{\omega})$ will be approximately an unbiased estimate for $f_\Delta(\boldsymbol{\omega})$. Figure 2 shows the spectral window, W, in the vertical direction ($\omega_1 = 0$), we can clearly see the main lobe around the origin and some smaller side lobes

at other frequencies. These side lobes can lead to substantial bias in $I_N(\boldsymbol{\omega})$ as an estimator of $f_\Delta(\boldsymbol{\omega})$ since they allow the value of $f_\Delta()$ at frequencies far from $\boldsymbol{\omega}$ to contribute to the expected value. This phenomenon is called *leakage.*

The leakage can also extend to distant frequencies. For example, far away frequencies with relatively high power compared to neighboring frequencies can be completely submerged by leakage from distant frequencies with much higher power. If the side lobes of W were substantially smaller, we could reduce this source of bias for the periodogram considerably. Tapering is a technique that effectively reduces the side lobes associated with the spectral window.

Thus, we use a data taper to prevent the leakage from far away frequencies that could have quite a lot of power. We form the product $h(\mathbf{s})Z(\mathbf{s},t)$ for each value of $\mathbf{s} = (s_1, s_2)$, where $\{h(\mathbf{s})\}$ is a suitable sequence of real-valued constants called a *data taper.* The traditional tapers used for two dimensional data are the tensor product of two one-dimensional data tapers:

$$h_M(\mathbf{j}) = h_1(j_1)h_2(j_2),$$

where $\mathbf{j} = (j_1, j_2)$, $1 \leq j_1 \leq n_1 \quad \text{and} \quad 1 \leq j_2 \leq n_2$.

For instance, $h_1()$ could be a m-cosine taper (Dahlhaus and Künsch, 1987), where $1 \leq m < \frac{n_1}{2}$,

$$h_1(j_1) = \begin{cases} \frac{1}{2}\{1 - \cos(\frac{\pi(j_1 - 1/2)}{m})\} & 1 \leq j_1 \leq m \\ 1 & m+1 \leq j_1 \leq n_1 - m \\ h_1(n_1 - j_1 + 1) & n_1 - m + 1 \leq j_1 \leq n_1. \end{cases} \tag{7}$$

We define $h_2()$ in a similar way, and we form the product of the two data tapers to obtain, $h_M()$, the multiplicative data taper for two dimensional data.

Every time we do tapering we lose information. A multiplicative data taper defines a rectangle inside the grid (in Figure 3 the black rectangle represents the border of the grid and we define a blue rectangle inside), and gives small weight (close to 0) to the observations in the border of the grid (the black rectangle), weight 1 to the observations inside the (blue) rectangle, and a weight, $h_M()$, that goes smoothly from 1 to 0 to the observations in between the rectangle and the border of the grid (i.e in between the blue and the black rectangles). Since the goal here is to reduce the bias due to leakage without throwing away information

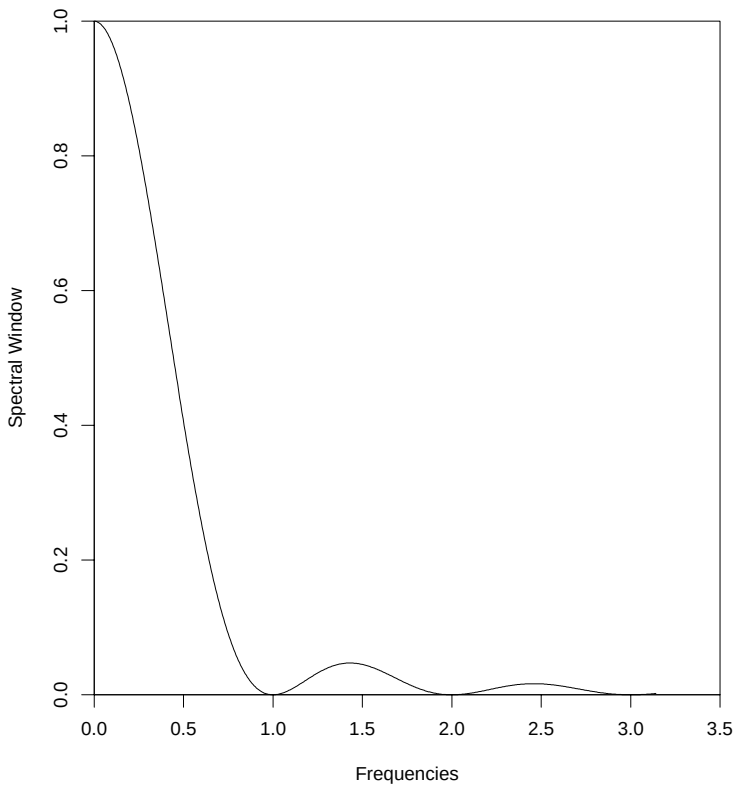

Figure 2: Spectral window along the vertical axis. The horizontal axis shows the frequencies, while the vertical axis shows the spectral window along the vertical axis for the periodogram (without tapering), for $n_2 = 500$.

unnecessarily. We will present in this paper (Section 5) a new class of data tapers that will overperformed the classic multiplicative taper, by not losing so much information and efficiently reducing the bias.

2.3 Likelihood function

For large datasets calculating the determinants that we have in the likelihood function can be often infeasible. Spectral methods could be used to approximate the likelihood and obtain the maximum likelihood estimates (MLE) of the covariance parameters: $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_r)$.

Spectral methods to approximate the spatial likelihood have been used by Whittle 1954, Guyon 1982, Dahlhaus and Küsch, 1987, and Stein 1995, 1999, among others. These spectral methods are based on Whittle's (1954) approximation to the Gaussian negative log likelihood:

$$\frac{N}{(2\pi)^2} \sum \log f(\boldsymbol{\omega}) + I_N(\boldsymbol{\omega}) f(\boldsymbol{\omega})^{-1} \tag{8}$$

where the sum is evaluated at the Fourier frequencies, I_N is the periodogram and f is the spectral density of the lattice process. The approximated likelihood can be calculated very efficiently by using the fast Fourier transform. This approximation requires only $O(Nlog_2N)$ operations. Simulation studies conducted by the author seem to indicate that N needs to be at least 100 to get good estimated MLE parameters using Whittle's approximation.

The asymptotic covariance matrix of the MLE estimates of $\theta_1, \ldots, \theta_r$ is

$$\left\{ \frac{2}{N} \left[\frac{1}{4\pi^2} \int_{[-\pi,\pi]} \int_{[-\pi,\pi]} \frac{\delta \log f(\boldsymbol{\omega}_1)}{\delta \theta_j} \frac{\delta \log f(\boldsymbol{\omega}_2)}{\delta \theta_k} d\boldsymbol{\omega}_1 d\boldsymbol{\omega}_2 \right]^{-1} \right\}_{jk} \tag{9}$$

this is much easier to compute that the inverse of the Fisher information matrix.

Guyon (1982) proved that when the periodogram is used to approximate the spectral density in the Whittle likelihood function, the periodogram bias contributes a non-negligible component of the mean squared error (mse) of the parameter estimates for 2-dimensional processes, and for 3-dimensions this bias dominates the mse. Thus, the MLE parameters of the covariance function based on the Whittle likelihood are only efficient in one dimension, but not in two and higher dimensional problems. Though, they are consistent. Guyon demonstrated that this problem can be solved by using a different version of the periodogram, an "unbiased peridogram", which is the discrete Fourier transform of an unbiased version of the sample covariance.

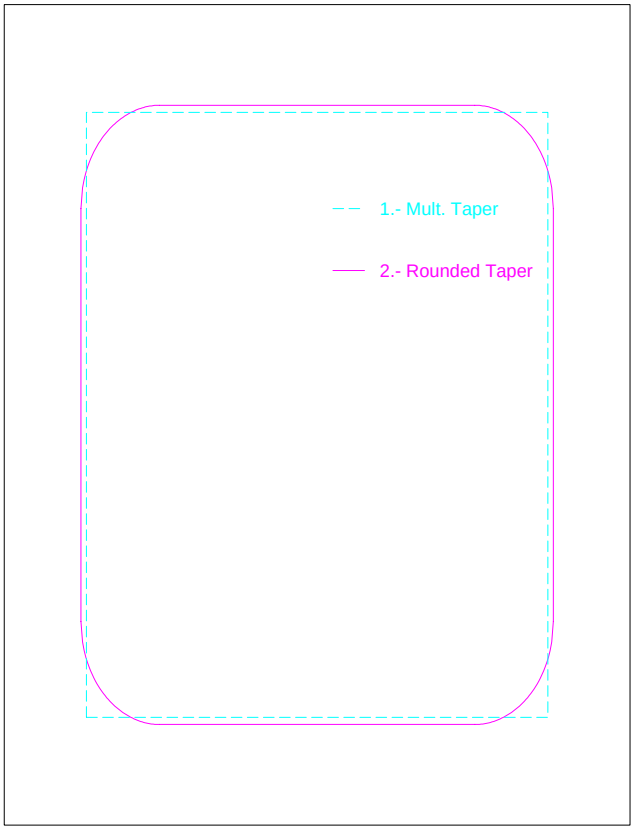

Figure 3: Multiplicative taper and rounded taper. The area of the rectangle is the same as the area of the rounded region. Line 1: Multiplicative Data Taper, gives weight 1 to the observations inside the rectangular region, and a weight $h_M()$ that goes smoothly from 1 to zero to the observations outside the rectangular region. Line 2: Rounded Data Taper, gives weight 1 to the observations inside the rounded region, and a weight $h_R()$ that goes smoothly from 1 to zero to the observations outside the rounded region.

Dahlhaus and Künsch (1987) demonstrated that tapering also solves this problem.

3 Incomplete lattices

In this Section we introduce spectral methods to approximate the likelihood for spatial processes observed on incomplete lattices. In our approach, as a first step we fill-in with zeros the values of the process at the locations in this grid where we have no data (e.g. Marcotte, 1996), to then efficiently calculate the periodogram using FFT. We study the asymptotic properties of the estimated periodogram and the potential impact of this approximation on the likelihood approximation, prediction and inference made for spatial data.

Consider a random field Y observed on a rectangle $P_N = \{1, \ldots, n_1\} \times \{1, \ldots, n_2\}$ of sample size $N = n_1 n_2$. We write $Y(\mathbf{x}) = g(\mathbf{x})Z(\mathbf{x})$, where Z is the spatial lattice process under study with spectral density f_Z. We assume Z is a weakly stationary real-valued Gaussian process having mean zero, sumable covariance, and finite moments, but Z is not directly observed. Rather we observe, Y, an amplitude modulated version of Z for the observations on the grid, where g is defined as:

$$g(\mathbf{x}_j) = \begin{cases} 0 & \text{if } Z(\mathbf{x}_j) \text{ is missing at location } \mathbf{x}_j \\ 1 & \text{if } Z(\mathbf{x}_j) \text{ is observed at location } \mathbf{x}_j. \end{cases} \tag{10}$$

As an example, consider the simulated image in Figure 4, the circles represent the locations where we have missing values. The process Z of interest is not observed at those locations. Our filter function g would be zero at those locations and 1 everywhere else.

We propose the following estimate of the spectral density of Z,

$$\tilde{I}_Z(\boldsymbol{\omega}) = \frac{1}{H_2(\mathbf{0})} |\sum_{i=1}^{N} (Y(\mathbf{x}_i) - g(\mathbf{x}_i)\tilde{Z}) \exp\{-i\omega \mathbf{x}_i\}|^2$$

where $H_j(\boldsymbol{\lambda}) = 2\pi \sum_{i=1}^{N} g^j(\mathbf{x}_i) e^{i\boldsymbol{\lambda}^T \mathbf{x}_i}$, then $H_2(\mathbf{0}) = 2\pi \sum_{i=1}^{N} g(\mathbf{x}_i)^2$, and

$$\tilde{Z} = \left(\sum_{i=1}^{N} Y(\mathbf{x}_i) \right) / \left(\sum_{i=1}^{N} g(\mathbf{x}_i) \right).$$

If $g(\mathbf{x}_i) = 1$ for all $\mathbf{x_i}$ in P_N, then $Y \equiv Z$ on the lattice, and $\tilde{I}_N$ reduces to the standard definition of the periodogram.

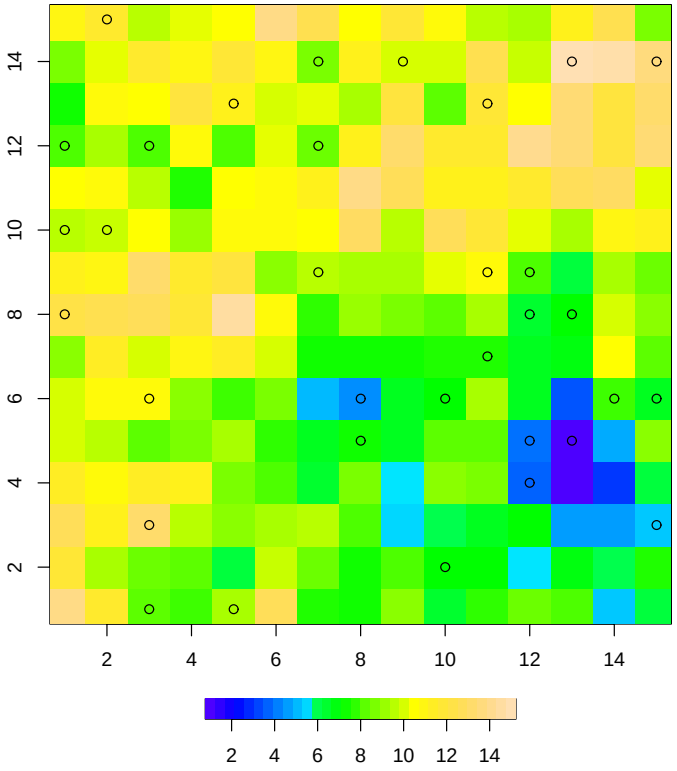

Figure 4: Simulated image (15×15), with 15% of missing values. Exponential covariance.

Let us study the asymptotic properties of this estimate of f_Z, as $N \to \infty$ (increasing domain asymptotics). The expected value of $\tilde{I}_Z$ is:

$$E[\tilde{I}_Z(\boldsymbol{\omega})] = \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} f_Z(\boldsymbol{\omega}-\boldsymbol{\phi})|H_1(\boldsymbol{\phi})|^2 d\boldsymbol{\phi}. \tag{11}$$

Thus, $E[\tilde{I}_Z(\boldsymbol{\omega})]$ is a weighted integral of $f_Z(\boldsymbol{\omega})$. Asymptotically,

$$E[\tilde{I}_Z(\boldsymbol{\omega})] = f_Z(\boldsymbol{\omega}) + O(N^{-1}). \tag{12}$$

This result is obtained by using (3.12) in Brillinger (1970), since $\tilde{I}_Z(\boldsymbol{\omega})$ is the periodogram of a tapered version of Z.

Sharp changes in g make its Fourier transform and the squared modulus of its Fourier transform exhibit side lobes. Therefore a lot of scatter missing values create very large side lobes in (11), and even if asymptotically the bias is negligible by (12), for small samples this bias could have some impact.

We obtain now the asymptotic variance for $\tilde{I}_Z$,

$$\text{var}\{\tilde{I}_Z(\boldsymbol{\omega})\} = |H_2(\mathbf{0})|^{-2}\left\{H_2(\mathbf{0})^2 + H_2(2\boldsymbol{\omega})^2\right\} f_Z(\boldsymbol{\omega})^2 + O(N^{-1}). \tag{13}$$

This can be proven by applying Theorem 5.2.8 in Brillinger (1981) to $\tilde{I}_Z(\boldsymbol{\omega})$.

The quantity multiplying f_Z in the expression (13) for the asymptotic variance is greater than 1 when we have missing values, and it is 1 when there is no missing values. Thus, a large number of missing values would increase the variance of the estimated spectrum.

We use now the estimated spectrum $\tilde{I}_Z$ to approximate the likelihood of the spatial process Z. By proposition 1 in Guyon (1982) we have that our estimated likelihood,

$$L_Z = \frac{N}{(2\pi)^2}\sum_{\mathbf{j}\in J_N}\left\{\log f_Z\left(2\pi\mathbf{j}/\mathbf{n}\right) + \tilde{I}_Z\left(2\pi\mathbf{j}/\mathbf{n}\right)\left(f_Z\left(2\pi\mathbf{j}/\mathbf{n}\right)\right)^{-1}\right\},$$

(as $N \to \infty$) converges to the exact likelihood of Z

$$\mathcal{L}_Z = \frac{1}{2}\log|\Sigma_N| + Z^T\Sigma_N^{-1}Z,$$

we have

$$E_{P_{\boldsymbol{\theta}}}\sup_{k=0,1,2}\sup_{\boldsymbol{\theta}}|L_Z^{(k)} - \mathcal{L}_Z^{(k)}| = O(\max(n_1,n_2)). \tag{14}$$

L_Z requires only $O(Nlog_2N)$ operations.

3.1 Simulation study

We simulate a spatial lattice (15× 15), with 15% missing values. The location of the missing values are represented by circles in Figure 4. The simulated process of interest is a stationary Gaussian spatial process with an exponential covariance function C,

$$C(\mathbf{h}) = \sigma e^{-|\mathbf{h}|/\rho}$$

the sill parameter (σ) is 2, and the range (ρ) 3. We calculate $\tilde{I}_Z$, and obtain the approximated likelihood function. Figure 5 shows a contour-plot for the likelihood function of the range and sill parameters using the spectral approach introduced here (filling up with zeros the missing values). Figure 5 compares the full spectral likelihood for the complete lattice (without missing values) to the approximated likelihood for the incomplete lattice using the approach presented here. With the filling-up approach we tend to be a little bit over-optimistic for the sill parameter, this is more clear in the next Figure, (Figure 6), in which we see a faster decay of the likelihood function as we move away from the pseudo-MLE for the sill. But, overall the estimated pseudo-MLEs for the sill and range, are practically the same for the incomplete and complete datasets and right on target (Figures 5 and 6). With datasets that had more than 20% missing values these results this did not hold any longer.

We also study the impact of this approximation on the Bayesian inference made about the data. Figure 7 shows the predictive posterior distributions (ppd) at the locations where we have missing values, using the approach presented here to approximate the likelihood. The prior for the range is a uniform on the interval $[0, 20]$, and the prior for the sill is

$$P(\sigma^2) \propto 1/\sigma^2.$$

The true value is represented in Figure 7 by a star. The approximated likelihood method does not seem to have an impact on the estimated center of the ppd (this is more clear in Figures 8 and 9). However, as we can appreciate in Figure 8, our method seems to underestimate slightly the spread of the ppd. The obtained ppd based on the approximated likelihood method introduced here has variance of .50 at location (row=3,column=1) in our image, while the variance of the ppd based on the exact likelihood function is .62. In Figure 9 we have the true values of the simulated spatial process at the locations where we have

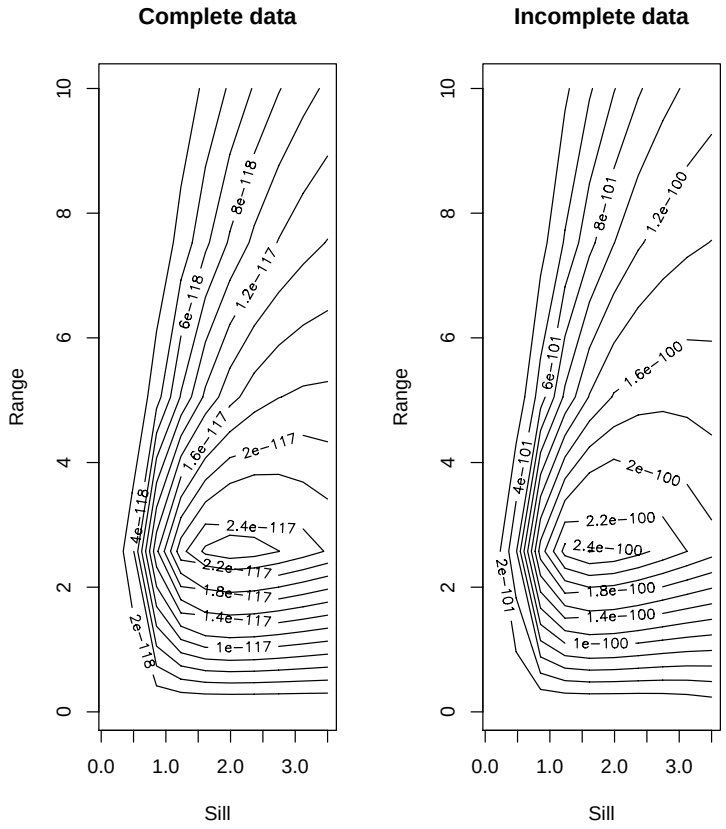

Figure 5: Contour plot for the likelihood. Left: using the complete dataset and the Whittle approximation. Right: using the approach presented here for incomplete datasets (with 15% missing values). Truth: Range is 3, Sill is 2.

the circles, versus the mean of the ppd using the exact likelihood and using the approximated likelihood for incomplete lattices introduced here. We can not appreciate a difference in the center (mean) of the ppd.

4 Likelihood for irregularly spaced data

Assume Z is a continuous Gaussian spatial process of interest, observed at M irregularly spaced locations, and f_Z is the stationary spectral density of Z. We define a process Y at location $\mathbf{x}$ as the integral of Z in a block of area Δ^2 centered at $\mathbf{x}$,

$$Y(\mathbf{x}) = \Delta^{-2} \int h(\mathbf{x} - \mathbf{s}) Z(\mathbf{s}) d\mathbf{s} \tag{15}$$

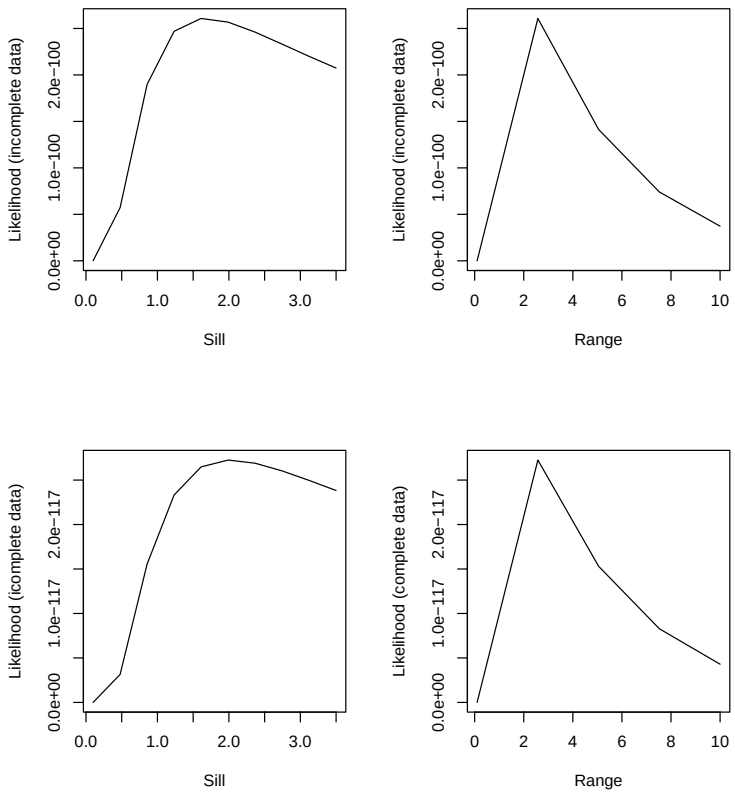

Figure 6: Profile likelihoods using Whittle approximation for the complete dataset on a regular lattice (bottom row) and using the approach presented here for irregular lattices (top row). Left: Sill parameter. Right: Range parameter.

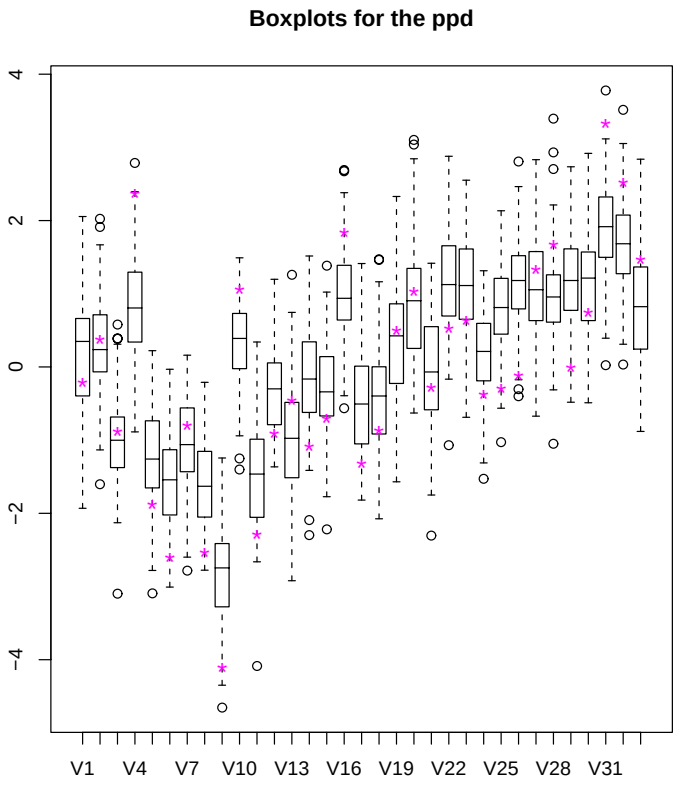

Figure 7: Boxplots for the predictive posterior distribution at the locations where we have missing data. The star represents the truth.

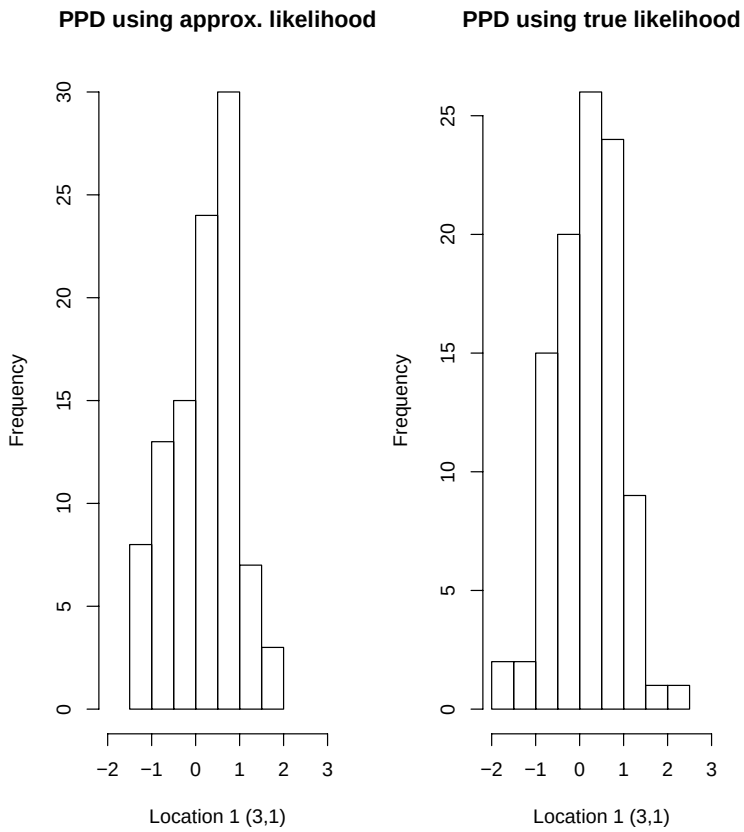

Figure 8: Predictive posterior distribution at location (3,1) for the irregular dataset with 15% missing values. Left: Using the approach presented here (mean:0.19, var:0.5, we are overoptimistic). Right: Using the exact likelihood (mean:0.17, var:0.62). Truth: -0.18.

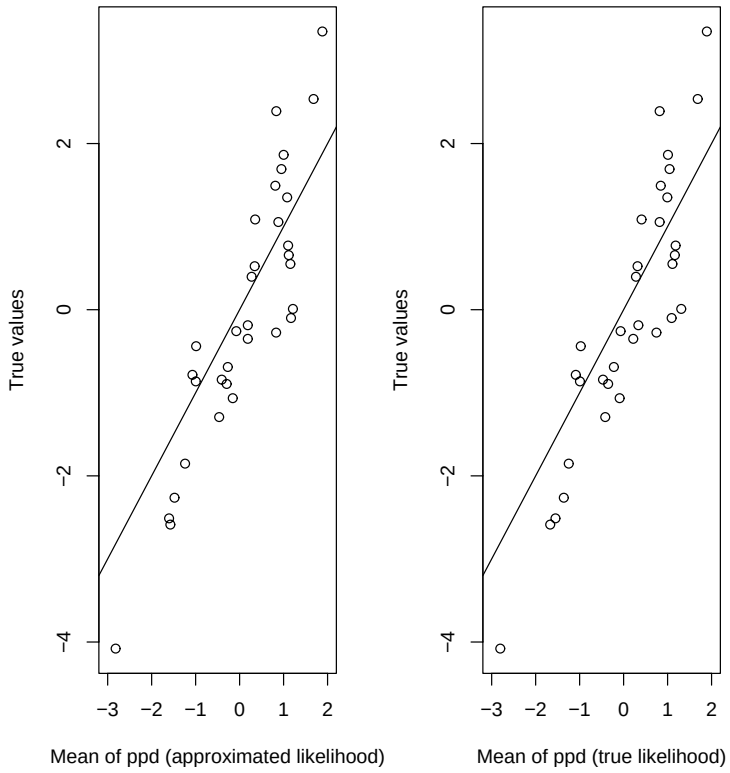

Figure 9: True values versus the mean of the predictive posterior distribution. Left: using the approximated likelihood method presented here for incomplete lattices. Right: using the exact likelihood.

where for $\mathbf{u} = (u_1, u_2)$ we have,

$$h(\mathbf{u}) = \begin{cases} 1 & \text{if } |u_1| < \Delta/2, |u_2| < \Delta/2 \\ 0 & \text{otherwise.} \end{cases}$$

Then, Y is also a stationary process with spectral density f_Y given by:

$$f_Y(\boldsymbol{\omega}) = \Delta^{-2}|\Gamma(\boldsymbol{\omega})|^2 f_Z(\boldsymbol{\omega}),$$

where $\boldsymbol{\omega} = (\omega_1, \omega_2)$ and $\Gamma(\boldsymbol{\omega}) = \int h(\mathbf{u}) e^{-i\boldsymbol{\omega}\mathbf{u}} = [2\sin(\Delta\omega_1/2)/\omega_1][2\sin(\Delta\omega_2/2)/\omega_2]$.

For small values of Δ, $f_Y(\boldsymbol{\omega})$ is approximately $f_Z(\boldsymbol{\omega})$, since we have:

$$\lim_{\Delta \to 0} \Delta^{-2}|\Gamma(\boldsymbol{\omega})|^2 = 1.$$

By (15), $Y(\mathbf{x})$ can be treated as a continuous spatial process defined for all $\mathbf{x} \in D$. But, here we consider the process Y only on a lattice $(n_1 \times n_2)$ of sample size $N = n_1 n_2$, i.e. the values of $\mathbf{x}$ in (15) are the centroids of the N grid cells in the lattice, having spacing Δ between neighboring sites (see Figure 10). Then, we have that the spectral density of the lattice process Y is,

$$f_{\Delta,Y}(\omega) = \sum_{Q \in \mathbb{Z}^2} |\Gamma(\omega + 2\pi Q/\Delta)|^2 f_Z(\omega + 2\pi Q/\Delta). \tag{16}$$

In practice, we truncate the sum in (16) after $2N$ terms, we include in the Appendix (A.1) the justification.

The idea is to apply Whittle likelihood to $f_{\Delta,Y}$, written in terms of f_Z. Therefore, we can obtain the MLE for the covariance/spectral density parameters of Z by writing the likelihood of the process Y. It might help the reader to interpret this key idea in the spatial domain rather than the spectral domain.

- Basic idea: interpretation in the spatial domain.

  The covariance for the block averages (the lattice process Y) is defined as

$$\text{cov}(Y(\mathbf{x}_{j_1}), Y(\mathbf{x}_{j_2})) =$$

$$\Delta^{-4} \int_{B_{j_1}} \int_{B_{j_2}} \text{cov}(Z(\mathbf{u}), Z(\mathbf{v})) d\mathbf{u} d\mathbf{v} =$$

$$\Delta^{-4} \int_{B_{j_1}} \int_{B_{j_2}} C_{\boldsymbol{\theta}}(\mathbf{u} - \mathbf{v}) d\mathbf{u} d\mathbf{v}$$

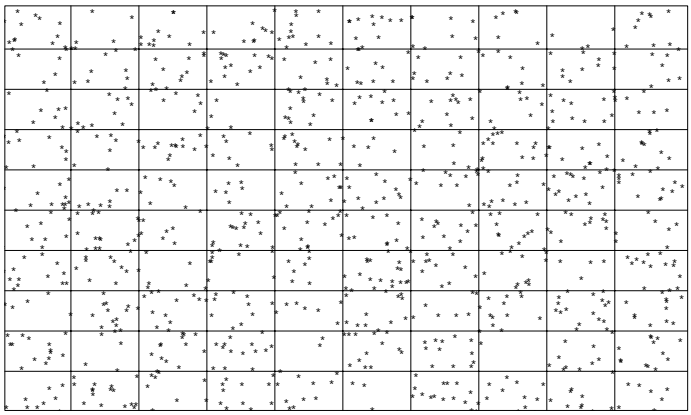

Figure 10: Simulation: Irregularly spaced observations. We grid the observations in a 10×10 grid, with an average of 10 observations per grid.

where $C_{\boldsymbol{\theta}}(\mathbf{u}-\mathbf{v})$ is the covariance for the continuous underlying process Z, and $\boldsymbol{\theta}$ are the covariance parameters. The continuous process Z is defined in terms of a pointwise covariance $C_{\boldsymbol{\theta}}(\mathbf{h})$, but we then use the previous expression to derive the covariances of the block averages $Y(\mathbf{x}_i)$, $i = 1, \ldots, N$, in terms of the pointwise covariance $C_{\boldsymbol{\theta}}$ This is then used to define a likelihood function for the parameters of the covariance function for the process Z in terms of the likelihood function of $Y(\mathbf{x}_1), Y(\mathbf{x}_2), \ldots, Y(\mathbf{x}_N)$.

To calculate the likelihood of Y, we first need to estimate $f_{\Delta,Y}$. With that purpose in mind we define Y_N as,

$$Y_N(\mathbf{x}) = 1/n_{\mathbf{x}} \sum_{\mathbf{s}_i \in J_{\mathbf{x}}} h(\mathbf{s}_i - \mathbf{x})\mathbf{Z}(\mathbf{s_i}), \tag{17}$$

where for $\mathbf{x} = (x_1, x_2)$,

$$J_{\mathbf{x}} = \{\mathbf{s} = (s_1, s_2), |x_1 - s_1| < \Delta/2, |x_2 - s_2| < \Delta/2\}, \tag{18}$$

and the cardinal of this set is $|J_{\mathbf{x}}| = n_{\mathbf{x}}$. For locations $\mathbf{x}$, such that $n_{\mathbf{x}} = 0$, the value of $Y_N(\mathbf{x})$ is not known.

As the observations become more dense, the covariance of Y_N converges to the covariance of Y, see Appendix (A.1). However, the approximation of Y_N to Y works worse in grid cells with very few observations. Thus, we apply a data taper to Y_N, that gives less weight to grid cells with less observations. We define

$g_1(\mathbf{x}) = n_{\mathbf{x}}/n$, with n the mean of the $n_{\mathbf{x}}$ values. This g_1 function plays a similar role to the g weight function (10) in the incomplete grid scenario.

We define $I_{g_1 Y_N}(\boldsymbol{\omega})$ the periodogram for the tappered process $g_1(\mathbf{x})Y_N(\mathbf{x})$,

$$I_{g_1 Y_N}(\boldsymbol{\omega}) = |H_2^*(\mathbf{0})|^{-1} \left| \sum_{s_1=1}^{n_1} \sum_{s_2=1}^{n_2} g_1(\mathbf{s}) Y_N(\mathbf{s}) \exp\{-i\mathbf{s}^T \boldsymbol{\omega}\} \right|^2, \tag{19}$$

where $H_k^*(\boldsymbol{\lambda}) = 2\pi \sum_{j=1}^{N} g_1^k(\mathbf{x}_j) e^{i\boldsymbol{\lambda}^T \mathbf{x}_j}$.

The peridogram $I_{g_1 Y_N}(\boldsymbol{\omega})$ is an asymptotically unbiased estimate of $f_{\Delta,Y}$. The bias is of order $O(N^{-1}) + O(\bar{n}^{-1})$, where $\bar{n}$ is the average of the n_x^2 values ($\bar{n} = \sum_{j=1}^{N} n_{\mathbf{x_j}}^2/N$):

$$E[I_{g_1 Y_N}(\boldsymbol{\omega})] = f_{\Delta,Y}(\boldsymbol{\omega}) + O(N^{-1}) + O(\bar{n}^{-1}),$$

the proof of this result is included in the Appendix (Theorem 1).

Thus, as long as $O(N/\bar{n}) \leq O(1)$ we have that (see Appendix, Theorem 2)

$$L_Y = \frac{N}{(2\pi)^2} \sum_{\mathbf{j} \in J_N} \left\{ \log f_{\Delta,Y}\left(2\pi \mathbf{j}/\mathbf{n}\right) + I_{g_1 Y_N}\left(2\pi \mathbf{j}/\mathbf{n}\right) \left(f_{\Delta,Y}\left(2\pi \mathbf{j}/\mathbf{n}\right)\right)^{-1} \right\}, \tag{20}$$

converges to $\mathcal{L}_Y = \frac{1}{2} \log |\Sigma_N| + Y^T \Sigma_N^{-1} Y$ (exact likelihood for Y), the order of convergence (in the sense of (14)) is $N^{1/2}$.

If M is the total number of observations of the process Z, the calculation of L_Y requires $O(N log_2 N + M)$ operations rather than $O(M^3)$ for the exact likelihood of Z. We choose $N \leq M^{2/3}$ (with the equality only when there are not many empty cells) to satisfy $O(N/\bar{n}) \leq O(1)$. If we have $N = M^{2/3}$, then the number of operations to obtain the likelihood function is $O(M^{2/3} log_2 M)$.

4.1 Simulation

We simulate 1000 observations of a Gaussian spatial process with a stationary exponential covariance (range=.25 and sill =1). We grid the observations in a 10× 10 lattice, and we obtain an average of 10 observations per grid. Figure 10 shows the grid and Figure 11 the empirical semivariogram for the gridded process and for the original data. The gridded process clearly shows a smaller sill (less variance) and larger range. We want to emphasize the fact that in our approach we do not estimate the parameters of the block covariance for the gridded process, what we do is to write this block covariance (or the corresponding spectrum) in terms of the covariance parameters of the continuous underlying process Z, and we estimate the

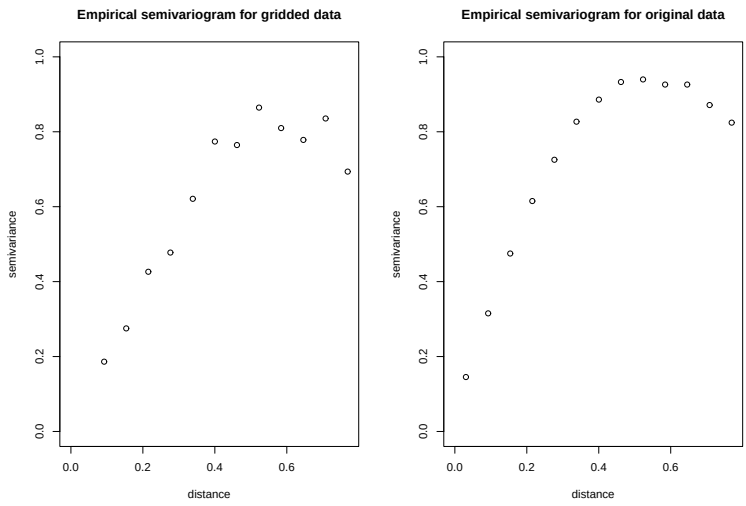

Figure 11: Empirical semivariogram for original data and gridded data. Truth: Exponential covariance, nugget=0, range =.25, sill =1.

parameters of the point-covariance of Z. Figure 12 shows the true semivariogram for Z and the empirical semivariogram with a confidence envelope, obtained simulating 1000 versions of the process Z, the upper and lower limits (dashed-lines) show the maximum and minimum values of the empirical semivariograms for the 1000 simulated Gaussian processes with an exponential covariance (range=.25 and sill=1). The variogram with the true MLE parameters (green line) is showed practically on top of the true variogram. Our approach gives the red dotted-line in Figure 12, just slightly below the true semivariogram.

4.2 Modified version of the approximated likelihood function

The approach presented in this Section for irregularly spaced datasets performs well when there is no nugget effect (measurement error). But, we could improve the estimation of the nugget and also the smoothness parameter (that explains the degree of differentiability of Z, see Stein (1999)) by adding to the likelihood function of Y information about the behavior of the process $Z(\mathbf{s}_i)$ within grid cells.

Thus, we randomly choose m blocks (no more than 10%-15% of the blocks) and treat them as if $n_{\mathbf{x}_i} = 0$ (i.e. we give them weight zero). We do not use the information from these m blocks in L_Y, the log-likelihood for Y.

Then, we add to the log-likelihood, L_Y, the log-likelihood of each one of the m blocks (treating the blocks

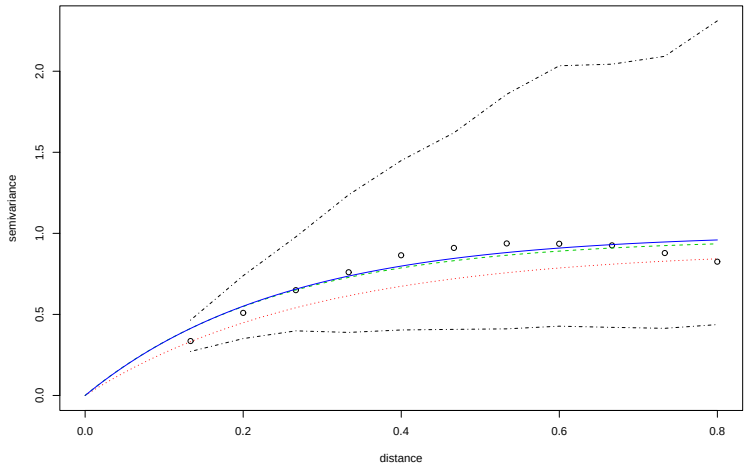

Figure 12: True semivariogram (solid line in blue, sill 1 and range .25) with an envelope, exact MLE (green line). Estimated MLE with our spectral method (red dotted-line).

as independent):

$$L_Y + \sum_{j=1}^{m} \frac{1}{2} \log |\Sigma_j| + \mathbf{Z}_j^T \Sigma_j^{-1} \mathbf{Z}_j, \tag{21}$$

where $\mathbf{Z}_j$ is a vector with the $n_{\mathbf{x}_j}$ observations within block j, and Σ_j is the covariance within the block written in terms of $C_{\boldsymbol{\theta}}(\mathbf{h})$ (covariance of Z). The calculation of (21) is very fast, since the blocks are small, approximately of order $M^{1/3}$.

4.3 Simulation

We simulate 1000 observations of a Gaussian spatial process with a stationary Matérn covariance (nugget= .25, range=.25, smoothness parameter =3, and partial sill =1):

$$C(\mathbf{h}) = \sigma_0 I(\mathbf{h}) + \frac{\sigma_1}{2^{\nu-1}\Gamma(\nu)} (2\nu^{1/2}|\mathbf{h}|/\rho)^{\nu} \mathcal{K}_{\nu}(2\nu^{1/2}|\mathbf{h}|/\rho), \tag{22}$$

where $\mathcal{K}_{\nu_s}$ is a modified Bessel function and $\boldsymbol{\theta} = (\sigma_0, \nu, \sigma_1, \rho)$. $I(\mathbf{h})$ is an indicator function, it takes the value 1 when $\mathbf{h} = (0, 0)$, and it is zero otherwise. The nugget parameter is σ_0 (microscale variation). The parameter ρ measures how the correlation decays with distance; generally this parameter is called the *range*. The partial sill, σ_1, is the total variance of the process minus the nugget. The parameter ν measures the

degree of smoothness of the process Z. The higher the value of ν the smoother Z would be; e.g. when $\nu = \frac{1}{2}$, we get the exponential covariance function. In the limit as $\nu \to \infty$ we get the Gaussian covariance.

We grid the observations in a $10\times$ 10 lattice (as in Figure 10). Figure 13 presents the empirical semivariogram for the gridded process and for the original data. The gridded process clearly does not capture the microscale variation, and estimates the nugget as zero. This still remains a problem when we write this block covariance (or the corresponding spectrum) in terms of the point-covariance parameters of the continuous underlying process Z.

The following results show the improvement in the nugget estimation by adding to the likelihood of the gridded process, L_Y in (20), the information within blocks (using 10 blocks, randomly selected) using expression (21). In Table 1, the nugget is estimated as 0, using the gridded process, and it is estimated as .3 (with standard error .23) using expression (21). Regarding the smoothness parameter, which is always very difficult to estimate, the exact MLE is .7. We estimate this parameter as 5 using the information within 10 blocks. We should note that the spectral likelihood method for the gridded process (L_Y) seems to estimate well the range and partial sill parameters and also their standard errors (s.e.). The standard errors are obtained using expression (9). The s.e. for the range and smoothness parameters are underestimated using L_Y, but when we add the information in 10% of the blocks we estimate better not only these parameters but also the uncertainty about them.

Parameters:	Nugget	Partial Sill	Range	Smoothness
TRUTH	.25	1	.25	3
MLE (exact likelihood)	.24 (.2)	.9 (.2)	.5 (.25)	.7 (3.5)
MLE (Spectral gridded)	0 (.1)	.8 (.2)	.12 (.2)	1 (2.1)
MLE (Spectral combining)	.3 (.23)	.8 (.27)	.4 (.3)	5 (4.8)

Table 1. Estimated covariance parameters, the values in parenthesis are standard errors.

In the next Section we introduce a new data taper. Tappering does not help to improve the asymptotic order of approximation of the spectral likelihood to the exact likelihood. But, it does help to obtain better pseudo-MLE parameters, by reducing the bias in the empirical covariance and periodogram due to the edge effect.

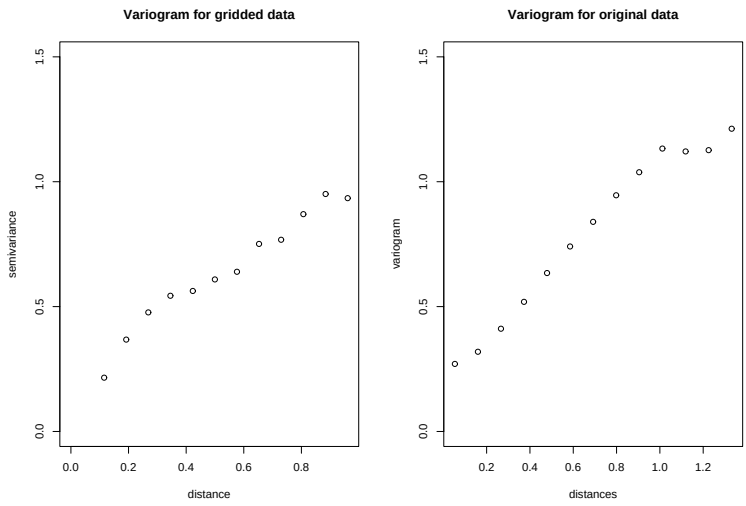

Figure 13: Empirical semivariogram. Truth: Matern covariance, smoothness=3, range=.25, partial sill=1, nugget=.25

5 New data taper

The periodogram, I_N, defined in (5) in terms of the process Z, it is also the discrete Fourier transform of the biased estimate of the covariance, the sample covariance $c_N(\mathbf{k}) = N^{-1} \sum (Z_{\mathbf{s}} - \bar{Z})(Z_{\mathbf{s}+\mathbf{k}} - \bar{Z})$ where N is the total number of observations and $\bar{Z}$ the sample mean, rather than the unbiased version (Guyon, 1982). No matter how large N becomes, the periodogram always involves the *tail* of the sample covariance, which is a poor estimate of the corresponding theoretical covariance, since in this region the sample covariance is based on just a small number of pairs of observations. One of the advantages of tapering is the reduction of the bias due to the boundary effect on the sample covariance. Edge effects are a serious problem in spatial statistics because the number of boundary points increases with the dimension. Thus, we generally need more tapering for the corner observations. A rounded taper gives more tapering to the corner observations, so with less overall tapering, we get the amount of smoothing that we need without losing so much information.

We propose here a "rounded taper". First, we introduce some notation and we define new coordinates $(r_1, r_2) \in \left[-\frac{n_1}{2}, \frac{n_1}{2}\right] \times \left[-\frac{n_2}{2}, \frac{n_2}{2}\right]$ in terms of $(s_1, s_2) \in [0, n_1] \times [0, n_2]$,

$$r_1 = |s_1 - (n_1 - 1)/2| \,,$$

$$r_2 = |s_2 - (n_2 - 1)/2|\,,$$

$$d = \sqrt{[r_1 - (n_1/2 - \epsilon)]^2 + [r_2 - (n_2/2 - \epsilon)]^2},$$

$$S = \{(r_1, r_2), \quad \text{s.t.} \quad r_1 > (n_1/2 - \epsilon) \quad \text{and} \quad r_2 > (n_2/2 - \epsilon)\}.$$

We give now a partition of the grid in terms of the new coordinates $(r_1, r_2) \in \left[-\frac{n_1}{2}, \frac{n_1}{2}\right] \times \left[-\frac{n_2}{2}, \frac{n_2}{2}\right]$ (see Figure 3), because centering the observations in this way makes it easier to define the rounded taper,

$$A = \{(r_1, r_2), \quad \text{s.t.} \quad r_1 \leq n_1/2 - \epsilon \quad \text{and} \quad r_2 > n_2/2 - \delta\}$$

$$B = \{(r_1, r_2), \quad \text{s.t.} \quad r_2 \leq n_2/2 - \epsilon \quad \text{and} \quad r_1 > n_1/2 - \delta\}$$

$$C = \{(r_1, r_2), \quad \text{s.t.} \quad r_1 \leq n_1/2 - \delta \quad \text{and} \quad r_2 < n_2/2 - \epsilon, \quad \text{or}$$
$$r_1 \leq n_1/2 - \epsilon \quad \text{and} \quad r_2 < n_2/2 - \delta\}$$

$$D = \{(r_1, r_2), \quad \text{s.t.} \quad (r_1, r_2) \in S \quad \text{and} \quad d \in [\epsilon - \delta, \epsilon]\}$$

$$E = \{(r_1, r_2), \quad \text{s.t.} \quad (r_1, r_2) \in S \quad \text{and} \quad d \geq \epsilon\}$$

$$F = \{(r_1, r_2), \quad \text{s.t.} \quad (r_1, r_2) \in S \quad \text{and} \quad d \leq \epsilon - \delta\}$$

Finally, we present the weight function, $h_R()$, that defines the rounded data taper,

$$h_R(r_1', r_2') = \begin{cases} \frac{1}{2}\left\{1 - \cos\left(\frac{\pi(n_2/2 - r_2)}{\delta}\right)\right\} & \text{for } (r_1, r_2) \in A, \\ \frac{1}{2}\left\{1 - \cos\left(\frac{\pi(n_1/2 - r_1)}{\delta}\right)\right\} & \text{for } (r_1, r_2) \in B, \\ 1 & \text{for } (r_1, r_2) \in C \text{ or } F, \\ \frac{1}{2}\left\{1 - \cos\left(\frac{\pi(\epsilon - d)}{\delta}\right)\right\} & \text{for } (r_1, r_2) \in D, \\ 0 & \text{for } (r_1, r_2) \in E. \end{cases} \tag{23}$$

The parameters δ and ϵ define the rounded data taper. The parameter $\delta \in [0, n/2]$, with $n = \min(n_1, n_2)$, plays the same role as m in the multiplicative taper (7), and the parameter $\epsilon \in [0, n/2]$ defines the rounded region (see Figure 14).

Figure 3, illustrates the difference between the two data tapers. The black rectangle represents the border of the site. The rounded data taper gives weight 1 to the observations inside the magenta curve, whereas a multiplicative data taper that gives the same amount of tapering to the corner observations would give

weight 1 only to the observations inside the blue rectangle. Therefore, we generally lose more information with a multiplicative data taper, if we equalize the amount of tapering in the corners.

Figure 15 shows the spectral windows on a decibel scale along the vertical axis ($\boldsymbol{\omega}_1 = 0$). The idea is to select a data taper so that $W()$ is uniformly small at frequencies far from 0. Note that the spectral windows corresponding to the two data tapers have significantly smaller side lobes than $W()$ with no taper. However, notice that the widths of the main lobes of these spectral windows are slightly larger: we have suppressed the side lobes in the windows at the expense of wider main lobes. The spectral window for the rounded taper, at almost all frequencies along the vertical axis, is much smaller than the spectral window for the multiplicative.

Figure 16 shows the spectral windows on a decibel scale along the diagonal ($\boldsymbol{\omega}_1 = \boldsymbol{\omega}_2$). We see in Figure 15 that the rounded taper has a slightly broader main lobe. Along the diagonal the side lobes for both tapers are much smaller than along the axes (compare Figures 15 and 16), and the differences between the two tapers are also smaller along the diagonal. Thus, even if the side lobes are slightly larger for the rounded data taper along the diagonal, overall the spectral window for the rounded data taper has better properties (smaller side lobes) than the spectral window for the multiplicative taper.

In summary, tapering is an operation that replaces the spectral window of the periodogram with one having better side lobe properties, which is a useful method for reducing the bias due to leakage in direct spectral estimators. A useful characterization of spectral density functions, f, is in terms of their *dynamic range,* which we define by the ratio

$$10 \log_{10} \left(\frac{\max_{\boldsymbol{\omega}} f(\boldsymbol{\omega})}{\min_{\boldsymbol{\omega}} f(\boldsymbol{\omega})} \right),$$

the dynamic range of a white noise process is 0. The bias in the periodogram for processes with high dynamic range can be attributed to the side lobes of the spectral window. Thus, tapering is particularly important for spectral densities with large dynamic range.

The asymptotic properties of the estimated covariance parameters (in terms of efficiency and consistency) using Whittle's likelihood and applying the data taper presented in this paper are the same as the ones obtained using a multiplicative taper (Guyon, 1982). However, we have seen in this section that with finite

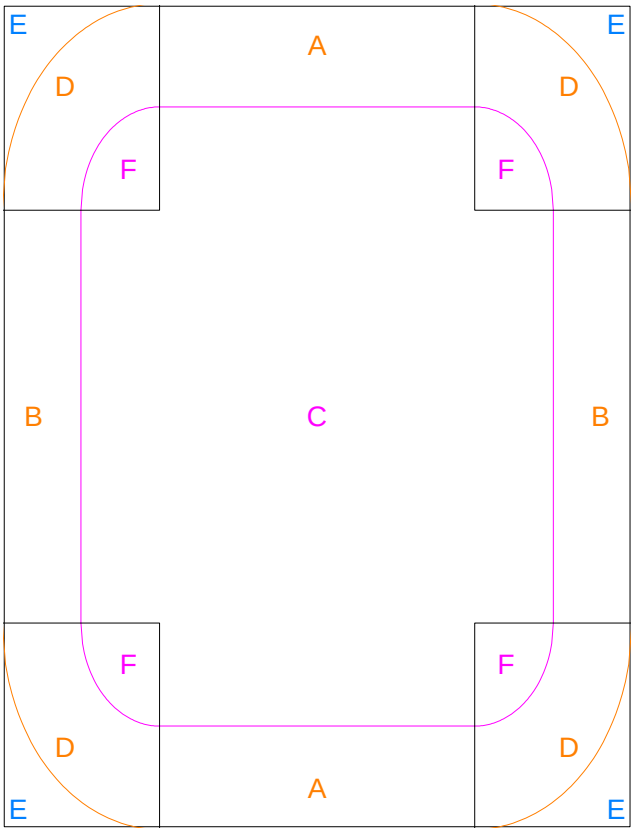

Figure 14: Rounded taper. The weight in the regions F and C is 1, the weight in the regions E is 0, and the weight in the regions A, B, and D is a cosine function $h_R()$ that goes smoothly from 1 to 0. The two parameters that define the rounded tapering are ϵ, and δ. The parameter ϵ is the radius of the circle that defines the regions D, and $\epsilon - \delta$ is the radius of the circle that defines the regions F.

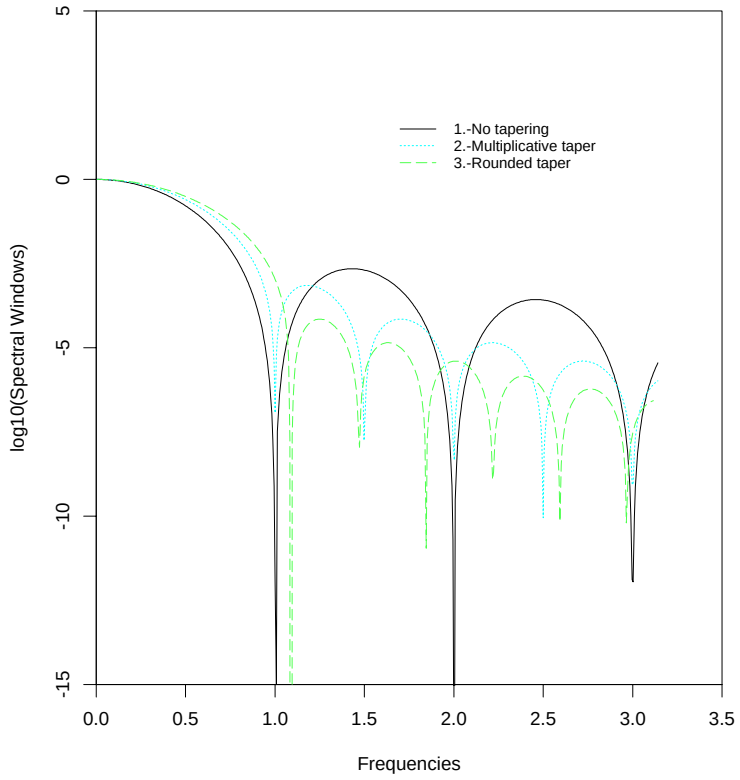

Figure 15: Spectral windows along the vertical axis. The horizontal axis shows the frequencies, while the vertical axis shows the spectral window along the vertical axis ($\omega_1 = 0$) on a decibel scale; for the periodogram (line 1), for the periodogram applying a *multiplicative* data taper (line 2), and for the periodogram applying a *rounded* data taper (line 3).

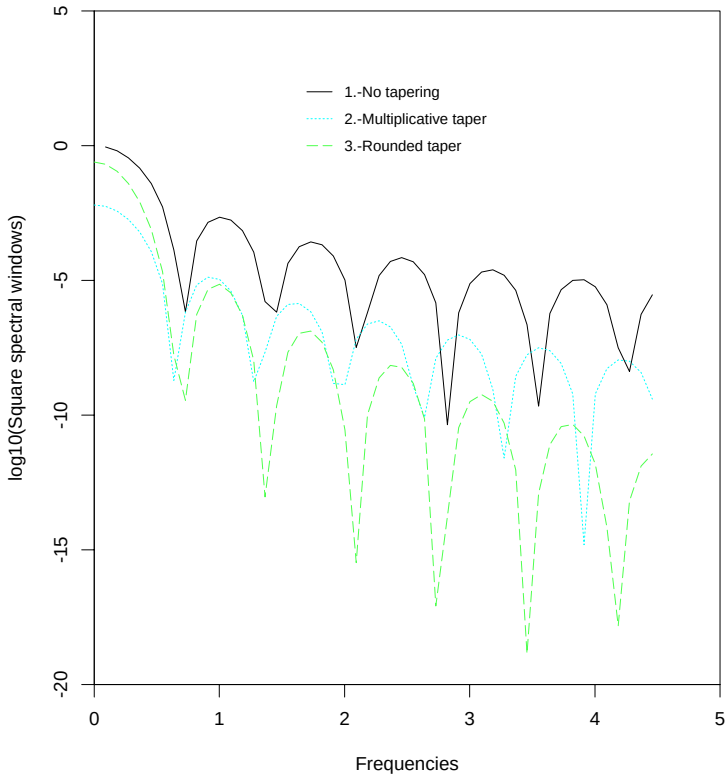

Figure 16: Spectral windows along the diagonal. The horizontal axis shows the frequencies, while the vertical axis shows the spectral window along the diagonal ($\omega_1 = \omega_2$) on a decibel scale; for the periodogram (line 1), for the periodogram applying a *multiplicative* data taper (line 2), and for the periodogram applying a *rounded* data taper (line 3).

samples the rounded taper helps to reduce the bias in the periodogram of the spectral likelihood function.

6 Discussion

In this paper we introduce likelihood approximation methods for lattice data with missing values and for irregularly spaced datasets. We use a spectral framework that offers enormous computational benefits. There are other alternative spectral approaches for irregularly sampled processes that are more computationally expensive:

- a spectral likelihood based on a periodogram for irregularly sampled processes obtained using generalized prolate spheroidal sequences (Bronez, 1988),
- the EM algorithm (Dempster, Laird, and Rubin 1977), which is a very well-known technique to find maximum likelihood estimates in parametric models with incomplete data. In the EM algorithm, we could first impute the values of the process at the locations in the grid where we have no data and then calculate the complete-data likelihood using spectral methods. We would need to iterate through these two steps.

The spectral likelihood approach presented here is attractive because of its simplicity and because is very computationally efficient and fast compared to any other known likelihood approximation method for spatial data that gives consistent estimates.

The weight function g introduced here to handle incomplete lattices, could have more sophisticated structures than the one used in this paper. For instance, instead of just taking values 1 and 0, g could go smoothly from 1 to 0 using a cosine function or a spline function to capture better the transition zones in the areas with missing values. This can be particularly helpful when we have missing values clustered together, e.g. clouds in AVHRR satellite data.

The new spatial data taper introduced in this paper, does not improve the asymptotic properties of the estimated covariance/spectrum parameters, but with finite samples it helps to reduce the periodogram bias due to the edge effect and therefore produces more reliable estimates. Any tapering technique is defined in terms of some parameters, and the problem arises as to how one should choose the values of those critical

parameters. If the objective is spatial prediction, cross validation is often used to determine the values of the taper parameters. The two parameters in the rounded data taper proposed here are ϵ and δ, and they can be chosen such that they minimize the relative mean squared error (rmse) of the estimated periodogram. The rmse is the mse of the periodogram divided by the squared of spectral density, f^2. The mean and the asymptotic variance of the periodogram for a tapered process are presented in Section 3, and can be used to approximate the mse. We suggest using a plug-in approach and replace f with the periodogram in the expression for the mse.

7 Appendix

A.1

Truncation of $f_{\Delta,Y}$:

Let us assume that for large frequencies (as $|\boldsymbol{\omega}| \to \infty$) the spectral density of a continuos spatial process Z satisfies:

$$f_Z(\boldsymbol{\omega}) \propto |\boldsymbol{\omega}|^{-\alpha}, \text{ with } \alpha > 2. \tag{24}$$

The spectral densities models generally used for continuous spatial process (i.e. Matérn) satisfy condition (24). Under this condition we need to prove that the residual term in the expression for $f_{\Delta,Y}$ given in (16), when we truncate the sum the sum after $2N$ terms, is negligible compared to $O(N^{-1})$, which is the bias of our estimated function of $f_{\Delta,Y}$ (Section 4).

The spectral density of the lattice process Y, $f_{\Delta,Y}(\boldsymbol{\omega})$ for $\boldsymbol{\omega} \in [-\pi/\Delta, \pi/\Delta]^2$, is defined in (16) in terms of f_Z. Here we study the order of the residual term R,

$$f_{\Delta,Y}(\boldsymbol{\omega}) = \sum_{Q \in N_Y} |\Gamma(\boldsymbol{\omega} + 2\pi Q/\Delta)|^2 f_Z(\boldsymbol{\omega} + 2\pi Q/\Delta) + R(\boldsymbol{\omega}, N_Y) \tag{25}$$

where $N_Y = \{(q_1, q_2) \in \mathbb{Z}^2; -n_1 < q_1 < n_1, -n_2 < q_2 < n_2\}$. We have that,

$$\begin{aligned} &\sum_{q_1=n_1}^{+\infty} \sum_{q_2=n_2}^{+\infty} |\Gamma(\boldsymbol{\omega} + 2\pi(q_1, q_2)/\Delta)|^2 f_Z(\boldsymbol{\omega} + 2\pi(q_1, q_2)/\Delta) \leq \\ &\qquad \int_{\pi/\Delta + 2\pi n_1/\Delta}^{+\infty} \int_{\pi/\Delta + 2\pi n_2/\Delta}^{+\infty} |\omega_1|^{-1} |\omega_2|^{-1} |\omega_1^2 + \omega_2^2|^{-\alpha/2} d\omega_1 d\omega_2 = O(N^{-\alpha/2}). \end{aligned} \tag{26}$$

Similarly, we have,

$$\sum_{q_1=-n_1}^{-\infty}\sum_{q_2=-n_2}^{-\infty}|\Gamma(\boldsymbol{\omega}+2\pi(q_1,q_2)/\Delta)|^2 f_Z(\boldsymbol{\omega}+2\pi(q_1,q_2)/\Delta)\leq \int_{-\infty}^{\pi/\Delta-2\pi n_1/\Delta}\int_{-\infty}^{\pi/\Delta-2\pi n_2/\Delta}|\omega_1|^{-1}|\omega_2|^{-1}|\omega_1^2+\omega_2^2|^{-\alpha/2}d\omega_1 d\omega_2 = O(N^{-\alpha/2}). \tag{27}$$

Therefore, the order of convergence to zero of the residual term in (25) is faster than $O(N^{-1})$, which is the bias of $I_{g_1 Y_N}$ (defined in (19)), and it is our estimate of $f_{\Delta,Y}$.

A.2

Theorem 1:

Consider a continuous weakly stationary spatial process Z observed at M locations in a domain D of interest. We define two lattice processes Y (as in (15)) and Y_N (as in (17)), both written in terms of the process Z, and defined on a lattice $n_1 \times n_2$ covering D, with spacing Δ between neighboring observations. We define $f_{\Delta,Y}$, the spectral density of the process Y. We propose $I_{g_1 Y_N}$, defined in (19), as an estimate of $f_{\Delta,Y}$. As $N \to \infty$ and $\bar{n} \to \infty$, where $\bar{n} = \frac{1}{N}\sum_i n_{\mathbf{x}_i}^2$ and $n_{\mathbf{x}_i}$ is the number of observations of the process Z in the grid cell i, we have,

$$E[I_{g_1 Y_N}(\boldsymbol{\omega})] = f_{\Delta,Y}(\boldsymbol{\omega}) + O(N^{-1}) + O(\bar{n}^{-1}).$$

Proof of Theorem 1:

$$I_{g_1 Y_N}(\boldsymbol{\omega}) = |H_2^*(\mathbf{0})|^{-1}\left|\sum_{s_1=1}^{n_1}\sum_{s_2=1}^{n_2} g_1(\mathbf{s})Y_N(\mathbf{s})\exp\{-i\mathbf{s}^T\boldsymbol{\omega}\}\right|^2.$$

First, we need to study the convergence of the second order moments of Y_N to the ones of the process Y as the observations become more dense, i.e. as each $n_{\mathbf{x}_i} \to \infty$. Assume $C_{\boldsymbol{\theta}}(\mathbf{h})$ is the stationary covariance of the process Z at a distance $\mathbf{h}$, with parameters $\boldsymbol{\theta}$. We have

$$\mathrm{cov}(Y_N(\mathbf{x}_1), Y_N(\mathbf{x}_2)) = \frac{1}{n_{\mathbf{x}_1} n_{\mathbf{x}_2}}\sum_{\mathbf{s}_i \in J_{\mathbf{x}_1}}\sum_{\mathbf{s}_j \in J_{\mathbf{x}_2}} C_{\boldsymbol{\theta}}(\mathbf{s}_i - \mathbf{s}_j),$$

where $J_{\mathbf{x}}$ is defined in (18), and

$$\mathrm{cov}(Y(\mathbf{x}_1), Y(\mathbf{x}_2)) = \Delta^{-2}\int\int h(\mathbf{u}-\mathbf{x}_1)h(\mathbf{v}-\mathbf{x}_2)C_\theta(\mathbf{u}-\mathbf{v})d\mathbf{u}d\mathbf{v} = \Delta^{-2}\int_{B_\Delta}\int_{B_\Delta} C_{\boldsymbol{\theta}}((\mathbf{x}_1-\mathbf{x}_2)-(\mathbf{u}-\mathbf{v}))d\mathbf{u}d\mathbf{v},$$

where $B_\Delta = [-\frac{1}{2}\Delta, \frac{1}{2}\Delta]^2$. Clearly the covariance of Y is stationary. By the expressions above for the covariance functions of Y_N and Y, we have that the covariance of Y_N between points $\mathbf{x}_1$ and $\mathbf{x}_2$, coverges to $\text{cov}(Y(\mathbf{x}_1), Y(\mathbf{x}_2))$ as $n_{\mathbf{x}_1} \to \infty$ and $n_{\mathbf{x}_2} \to \infty$. The order of convergence is $O(n_{\mathbf{x}_1}^{-1} n_{\mathbf{x}_2}^{-1})$.

Thus, it is straighforward to see that the order of converge of the covariance of the tapered process $g_1 Y_N$ to the covariance of $g_1 Y$ is $O(n^{-1}n^{-1})$, (n average of the $n_{\mathbf{x}_i}$ values). Then,

$$E[g_1(\mathbf{x}_1)Y_N(\mathbf{x}_1)g_1(\mathbf{x}_2)Y_N(\mathbf{x}_2)] = E[g_1(\mathbf{x}_1)Y(\mathbf{x}_1)g_1(\mathbf{x}_2)Y(\mathbf{x}_2)] + O(n^{-1}n^{-1}),$$

uniformly in $\mathbf{x}_1, \mathbf{x}_2$. This is a uniform convergence, because C is a uniformly bounded function, since we assume that the variance of Z is finite.

Thus, since

$$E[I_{g_1 Y_N}(\boldsymbol{\omega})] = |H_2^*(\mathbf{0})|^{-1} \left| \sum_{s_1=1}^{n_1} \sum_{s_2=1}^{n_2} g_1(\mathbf{s}) Y_N(\mathbf{s}) \exp\{-i\mathbf{s}^T\boldsymbol{\omega}\} \right|^2,$$

we have,

$$E[I_{g_1 Y_N}(\boldsymbol{\omega})] = E[I_{g_1 Y}(\boldsymbol{\omega})] + \epsilon_N(\boldsymbol{\omega}) \tag{28}$$

where $I_{g_1 Y}$ is the periodogram of the tapered process $g_1 Y$. As $N \to \infty$, $E[I_{g_1 Y}(\boldsymbol{\omega})]$ in the expression above converges uniformly to $f_{\Delta,Y}(\boldsymbol{\omega})$,

$$E[I_{g_1 Y}(\boldsymbol{\omega})] = f_{\Delta,Y}(\boldsymbol{\omega}) + O(N^{-1}),$$

and the residual term $\epsilon_N(\boldsymbol{\omega})$ in expression (28) is of the following order,

$$\epsilon_N(\boldsymbol{\omega}) \leq |H_2^*(\mathbf{0})|^{-1} \left| \sum_{s_1=1}^{n_1} \sum_{s_2=1}^{n_2} \frac{1}{n^2} \exp\{-i\mathbf{s}^T\boldsymbol{\omega}\} \right|^2 = O(\bar{n}^{-1}).$$

Therefore,

$$E[I_{g_1 Y_N}(\boldsymbol{\omega})] = f_{\Delta,Y}(\boldsymbol{\omega}) + O(N^{-1}) + O(\bar{n}^{-1}).$$

A.3

Let us assume that the spectral density of the lattice process Y (as in (15)), $f_{\Delta,Y}$ with parameters $\boldsymbol{\theta}$, satisfies the following conditions:

(a.1) $f_{\Delta,Y}(\boldsymbol{\omega})$ is rational with respect to $e^{i\boldsymbol{\omega}}$, without zeros or poles,

(a.2) the second derivative of $f_{\Delta,Y}$ in $\boldsymbol{\theta}$ is continuous in $\boldsymbol{\theta}$.

All classical spectral density models satisfy these two conditions.

Theorem 2:

Assume the order of convergence to zero of $\bar{n}$ as $N \to \infty$, is at least $O(N^{-1})$. This means, $O(N/\bar{n}) \leq O(1)$. Then, under conditions **(a.1)**-**(a.2)**, we have that,

$$L_Y = \frac{N}{(2\pi)^2} \sum_{\mathbf{j} \in J_N} \left\{ \log f_{\Delta,Y}\left(2\pi \mathbf{j}/\mathbf{n}\right) + I_{g_1 Y_N}\left(2\pi \mathbf{j}/\mathbf{n}\right) \left(f_{\Delta,Y}\left(2\pi \mathbf{j}/\mathbf{n}\right)\right)^{-1} \right\}, \tag{29}$$

converges to $\mathcal{L}_Y = \frac{1}{2}\log|\Sigma_N| + Y^T \Sigma_N^{-1} Y$ (exact likelihood function for the lattice process Y), and if n_1 and n_2 are of the same order, the rate of approximation (in the sense of (14)) is $N^{1/2}$.

Proof of Theorem 2:

By condition $O(N/\bar{n}) \leq O(1)$, the proposed peridogram function, $I_{g_1 Y_N}$, approximates the spectral density $f_{\Delta,Y}$ with a bias of the same order (Theorem 1) than if we use $I_{g_1 Y}$, the periodogram of a tapered version of Y. Then, by Proposition 1 in Guyon (1982). In which the convergence of the spectral Whittle likelihood function for a tapered process $g_1 Y$, to the exact likelihood of Y is proven. We obtain that (29) holds and the order is $N^{1/2}$.

Acknowledgements

This research was sponsored by a National Science Foundation grant DMS 0353029. The author would like to thank Michael Stein, at the University of Chicago, for providing invaluable suggestions for the taper function introduced in Section 5.

References

Besag, J.E. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *J. R. Statist. Soc. B* **36**, 76-86.

Bessag, J.E. and Moran P.A.P. (1975). On the estimation and testing of spatial interaction in Gaussian lattice processes. *Biometrika*, **62**, 555-562.

Bronez, T. P. (1988). Spectral estimation of irregularly sampled multidimensional processes by generalized prolate spheroidal sequences. *IEEE Transactions on Acoustics, Speech, and Signal Processing,* **36**, 1862-1873.

Caragea, P. (2003). Approximate Likelihoods for Spatial Processes. Ph.D. Dissertation at UNC. http://www.stat.unc.edu/postscript/rs/caragea.pdf

Clinger, W. and Van Ness, J. W. (1976). On unequally spaced time points in time series. *Ann. Statist.* **4**, 736-745.

Cramér, H. and Leadbetter, M. R. (1967). *Stationary and related stochastic processes. Sample function properties and their applications.* Wiley, New York.

Dahlhaus, R. and Küsch, H. (1987), Edge effects and efficient parameter estimation for stationary random fields. *Biometrika,* **74** 877-882.

Bloomfield (2000). *Fourier Analysis of Time Series.* Wiley, New York.

Brillinger, D. R. (1970).The frequency analysis of relations between stationary spatial series, *Proceedings of the Twelfth Biennial Seminar of the Canadian Mathematical Congress,* (ed. R. Pyke), Montreal, Canadian Math. Congress, 39-81.

Brillinger, D. R. (1981). *Time Series: Data Analysis and Theory.* Expanded edition. Holden-Day, Inc, San Francisco.

Guyon, X. (1982). Parameter estimation for a stationary process on a d-dimensional lattice. *Biometrika,* **69**, 95–105.

Marcotte, D. (1996). Fast variogram computation with FFT. *Computers and Geosciences,* **22**, 1175-1186.

Matérn, B. (1960). *Spatial variation.* Medded. Stat. Skogsforkskinst, 49, 5. Second ed. (1986), Lectures Notes in Statistics 36, New York: Springer.

Neave, H.R. (1970). Spectral analysis of a stationary time series using initially scarce data. *Biometrika*, **57**, 111-122.

Pardo-Iguzquiza, and Dowd (1997). AMLE3D: a computer program for the inference of spatial covariance parameters by approximate maximum likelihood estimation. *Comput. Geosci.*, **23**, 793-805.

Parzen, E. (1963). On spectral analysis with missing observations and amplitude modulation. *Sankhya,*

Ser. A, **25**, 383-392.

Priestley, M. B. (1981). *Spectral Analysis and Time Series.* Academic Press, London.

Stein, M. L. (1995). Fixed domain asymptotics for spatial periodograms. *Journal of the Americal Statistical Association,* **90**, 1277-1288.

Stein, Chi and Welty. (2004). Approximating likelihoods for large spatial data sets. *Journal of the Royal Statistical Society, Series B.* **66**, 275-296.

Stein, M. L. (1999). *Interpolation of Spatial Data: some theory for kriging.* Springer-Verlag, New York.

Vecchia (1988). Estimation and model identification for continuous spatial processes. *Journal of the Royal Statistical Society, Series B.* **50**, 297-312.

Whittle, P. (1954). On stationary processes in the plane, *Biometrika,***41**, 434-449.

Yaglom, A. M. (1987). *Correlation theory of stationary and related random functions.* Springer-Verlag, New York.