Spectral Analysis with Spatial Periodogram and Data Tapers

Hyon-Jung Kim, Montserrat Fuentes

Department of Statistics
North Carolina State University

The spatial periodogram is a nonparametric estimate of the spectral density, which is the Fourier Transform of the covariance function. It is a useful tool to explain the dependence structure of the underlying spatial process. Tapering (data filter) can be applied to the spatial data in order to reduce the bias of the periodogram. A data taper helps remove the edge effects even in high dimensional problems. However, the variance of the periodogram increases when the bias is reduced. We present a method to choose an appropriate smoothing parameter for data tapers and to get better estimates of the spectral density, by taking into account the trade-off between bias and variance of the tapered periodogram. An application to model and estimate the spatial structure for air pollutant concentration and fluxes is illustrated (using data provided by EPA).

KEYWORDS: periodogram, spectral density, data tapers, smoothing parameter, air pollutant concentration.

1 Introduction

Spectral analysis of stationary processes is particularly advantageous in the analysis of large data sets and in studying properties of multivariate processes. Geostatistical data are usually collected over a large region, and handling large data sets is often problematic for the commonly used techniques: inversion of a large covariance matrix to compute the likelihood function may not be possible or may require a long time in computation. The use of a Fast Fourier transform (FFT) algorithm for spectral densities can be a good resolution to these problems. However, FFT can be applied only to the regularly gridded data. Data observed on a irregular grid can be discretized for FFT when there are enough observations.

For a Gaussian stationary random field, $Z(\boldsymbol{x}),\ \boldsymbol{x} \in \mathbb{R}^{\mathbf{2}}$ with mean $\mathrm{E}(Z)$ and autocovariance function $C_Z(\boldsymbol{x})$, the spectral density at frequency $\boldsymbol{\lambda}$ is defined as

$$f_Z(\boldsymbol{\lambda}) = \frac{1}{(2\pi)^2} \int_{\mathbb{R}^{\mathbf{2}}} \exp(-i\boldsymbol{\lambda}^T \boldsymbol{x}) C_Z(\boldsymbol{x}) d\boldsymbol{x},$$

where

$$C_Z(\boldsymbol{x}) = \mathrm{cov}\{Z(\boldsymbol{x}+\boldsymbol{u}), Z(\boldsymbol{u})\}.$$

The spectral density function $f_Z(\cdot)$ is simply the Fourier transform of the autocovariance function $c_Z(\cdot)$. The covariance function can also be expressed in terms of the spectral density function as

$$C_Z(\boldsymbol{x}) = \int_{\mathbb{R}^2} \exp(i\boldsymbol{\lambda}^T\boldsymbol{x}) f_Z(\boldsymbol{\lambda}) d\boldsymbol{\lambda}.$$

Thus, the spectral density of a stationary process provides information equivalent to that obtained from its autocovariance.

The spatial periodogram, $I_N(\boldsymbol{\lambda}), N = (N_1, N_2)$, a non-parametric estimate of the spectral density, is a powerful tool for studying the properties of random fields observed on a lattice, $N_1 \times N_2$:

$$I_N(\boldsymbol{\lambda}) = \frac{1}{(2\pi)^2 N_1 N_2} | \sum_{\boldsymbol{s}\in D} Z(\boldsymbol{s}) e^{-i\boldsymbol{\lambda}^T\boldsymbol{s}}|^2,$$

where $\boldsymbol{s} \in D = \{0, ..., N_1 - 1\} \times \{0, ..., N_2 - 1\}$. It is the modulus-squared of a finite Fourier transform for the observed region of the process, introduced to search for hidden periodicities of processes. The periodogram itself is not a consistent estimator of the spectral density, but consistency can be achieved by applying linear smoothing filters to the periodogram. Smoothing the periodogram as is frequently done in time series does not remove large edge-effects (asymptotically) in two or more dimensions. In one-dimensional data, there is only one boundary point at each end, but the number of boundary points increases with the dimension, leading to more serious problems with edge effects. The sidelobes (subsidiary peaks) occurring on smoothing filters cause unnecessary large values of the periodogram ordinates for high frequencies and result in substantial bias. This phenomenon is called leakage. Instead of smoothing biased periodogram estimates, direct filtering of the data with a data taper before computing the periodogram can also provide a consistent estimate of the spectral density. The information lost through powerful frequencies by smoothing the periodogram can be better recovered by data tapers. Prewhitening is another method of filtering data to control this kind of bias from leakage.

In this paper, we propose to use the tensor product of two one-dimensional tapering functions for spatial data.We also present a method to choose an appropriate smoothing parameter for data tapers and to get better estimates of the spectral density. A good choice of the amount of smoothing, taking into consideration the tradeoff between the bias and the variance of the tapered periodogram estimates, can be useful in practice in order to explain the underlying spatial structure of a process. Data tapers put relatively less weight on the boundary points and is highly effective in removing edge-effects, even in higher dimensional problems. Moreover, in the fixed-domain asymptotics where the number of observations in a fixed study area increases, it has been shown that using the periodogram of the raw data without any data tapers applied can yield highly misleading results (Stein, 1995).

In the time-space domain, empirical variogram estimates are most commonly used to estimate the correlation structure of a process. When a parametric variogram model is fit to empirical variogram estimates, frequently used techniques such as non-linear least squares or restricted maximum likelihood (REML) approaches generally do not take into account any correlation between estimated variogram values. The same data points are used to estimate the variogram at different lags, and the resulting variogram estimates are more correlated than the observations of the underlying process. There is not much known about the correlation structure of variogram estimates, and ignoring such correlation can mislead data analyses. On the other hand, in the spectral domain the periodogram values at any set of Fourier frequencies are asymptotically uncorrelated. Techniques such as the nonlinear least squares method can be naturally applied to those independent estimates. In this case, nonparametric smoothing of the tapered periodogram estimates might give better results than nonparametric variogram estimates. The asymptotic independence of the periodogram estimates is one of the big advantages of the spectral analysis. Comparison of the results obtained by estimating the spectral density in the frequency domain to those by estimating the covariance with parametric (or non-parametric) techniques which ignore the covariance structure in time-space domain will also be worth investigating as a further study.

In Section 2, we introduce the spatial periodogram which is a general extension of the traditional one-dimensional periodogram in time series analysis. In Section 3, the properties of tapering functions are briefly described and we present a method of estimating the spectral density (FT of the covariance function) using a two-dimensional data taper and the spatial periodogram in Section 4. Lastly, an application of our method to model the spatial structure of Nitric acid (HNO_3) concentrations is illustrated in Section 5.

2 Spatial Periodogram

The spatial periodogram for higher dimensional data is a natural extension of the periodogram traditionally used in time series. The periodogram reflects periodic behavior of a time series and is a useful tool to analyze stochastic processes. Smooth series are characterized by the periodogram which have most of their power at low frequencies whereas quickly oscillating series are characterized by the periodogram which have most of their power at high frequencies. We first examine the expected value of the spatial periodogram.

Proposition *1. Assume that Z is a stationary Gaussian random field defined on a grid $N_1 \times N_2$ with $\mathrm{E}[Z] = 0$ and $\sum_{\boldsymbol{u}} |c_Z(\boldsymbol{u})| < \infty$. Then, the expectation of the spatial periodogram on two-dimensional data is given by*

$$\begin{aligned}\mathrm{E}[I_N(\boldsymbol{\lambda})] &= (4\pi^2 N_1 N_2)^{-1} \\ &\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \left[\frac{\sin N_1(\lambda_1-\alpha_1)/2}{\sin(\lambda_1-\alpha_1)/2}\right]^2 \left[\frac{\sin N_2(\lambda_2-\alpha_2)/2}{\sin(\lambda_2-\alpha_2)/2}\right]^2 f_Z(\alpha_1,\alpha_2)d\alpha_1 d\alpha_2.\end{aligned}$$

where $f_Z(\cdot)$ is the spectral density for Z and $N = (N_1, N_2)$.
Thus, for the zero-mean processes, the expected value of the spatial periodogram is not $f_Z(\boldsymbol{\lambda})$, but a weighted integral of $f_Z(\boldsymbol{\lambda})$. It is asymptotically unbiased: $\lim_{min(N_1,N_2)\to\infty} E[I_N(\boldsymbol{\lambda})] = f_Z(\boldsymbol{\lambda})$.

3 Data Tapers

Data tapering was introduced in nonparametric time series analysis by Tukey (1967). It is a smoothing technique for raw data in the space domain, as opposed to smoothing the periodogram in the frequency domain. The most obvious advantage of tapering observed values prior to computing their Fourier transform is a better convergence of periodogram estimates to the spectral density.

We define a one-dimensional taper $g(u),\ u \in [0,1]$, with smoothness parameter ρ, by

$$g(u) = \begin{cases} w(2u/\rho) & (0 \le u < \frac{1}{2}\rho) \\ 1 & (\frac{1}{2}\rho \le u \le \frac{1}{2}) \\ g(1-u) & (\frac{1}{2} < u \le 1). \end{cases}$$

The typical shape of a tapering function, $g(u)$ where $u \equiv t/T$ involves a steady increase from 0 when $t = 0$ to a maximum of 1 for t between 0 and T, followed by a steady decrease to 0 as $|t|$ increases to T. A wide variety of tapering functions have been proposed, including the Tukey-Hanning taper with $w(u) = \frac{1}{2}\{1 - \cos(u\pi)\}$.

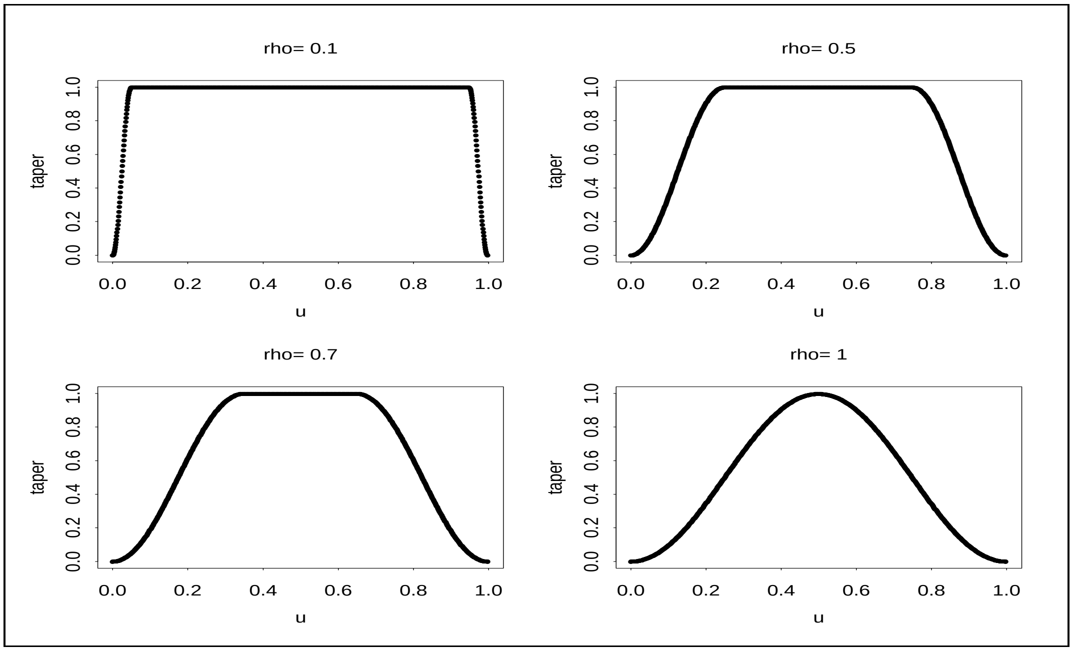

Figure 1: One dimensional tapering function with different values of a smoothing parameter ρ.

The tensor product of two one-dimensional tapering functions $h(\boldsymbol{s}/\boldsymbol{N}) = g_1(s_1/N_1)\ g_2(s_2/N_2)$ is commonly used as a two-dimensional data taper. In two dimensions, the tapering function takes on the form of a rectangle of 1's centered in the middle of the data grid, then decreases smoothly to 0 at the edges of the grid.

4 Tapered Spatial Periodogram

After the data are filtered with a data taper, the spatial periodogram for tapered data with the normalizing factor (the tapered spatial periodogram) is given as:

$$I_N^{(t)}(\boldsymbol{\lambda}) = ((2\pi)^2 \sum_{\boldsymbol{s}} h(\boldsymbol{s}/\boldsymbol{N})^2)^{-1} | \sum_{\boldsymbol{s}} e^{-i\boldsymbol{\lambda}^T \boldsymbol{s}} h(\boldsymbol{s}/\boldsymbol{N}) Z(\boldsymbol{s})|^2.$$

Choosing an appropriate value for the smoothing parameter $\boldsymbol{\rho} = (\rho_1, \rho_2)$ will enable us to summarize the spatial structure of a spatial random field of interest. Such a smoothing parameter can be obtained by minimizing the mean squared error (MSE) of the tapered spatial periodogram. The expectation and the covariance of the tapered spatial periodogram are first reexpressed in terms of $\boldsymbol{\rho}$ for MSE and we use the Lagrange multiplier for minimization by adding a constraint that $|\boldsymbol{\rho}|$ must be between 0 and 1.

Proposition *2. (Kim, Fuentes, 2000) Assume that Z is a stationary Gaussian random field defined on a grid $N_1 \times N_2$ with $\mathrm{E}[Z] = 0$ and $\sum_{\boldsymbol{u}} |c_Z(\boldsymbol{u})| < \infty$. Then the expectation of the spatial periodogram on the tapered 2-dimensional data, $I_Z^{(t)}(\boldsymbol{\lambda})$, is*

$$\mathrm{E}[I_N^{(t)}(\boldsymbol{\lambda})] = ((2\pi)^2 \sum h(\boldsymbol{s}/\boldsymbol{N})^2)^{-1} \int_{\mathbb{R}^2} |H(\boldsymbol{\tau})|^2 f_Z(\boldsymbol{\lambda} - \boldsymbol{\tau}) d\boldsymbol{\tau}$$

where $|H(\boldsymbol{\tau})|^2 = |H_1(\tau_1)|^2 |H_2(\tau_2)|^2$, $H_1(\tau_1)$ and $H_2(\tau_2)$ are the Fourier transforms of two different tapering functions, $g_1(s_1/N_1)$ and $g_2(s_2/N_2)$, respectively.

Assume that the same conditions in the above hold and assume further that the (finite) fourth order moments correspond to stationarity. The covariance of the spatial periodogram on the tapered 2-dimensional data is

$$\mathrm{Cov}[I_N^{(t)}(\boldsymbol{\lambda}), I_N^{(t)}(\boldsymbol{\mu})] = ((2\pi)^2 \sum h(\boldsymbol{s}/\boldsymbol{N})^2)^{-2}$$
$$\{\int_{\mathbb{R}^2} H(\boldsymbol{\lambda} - \boldsymbol{\tau}) H(-\boldsymbol{\lambda} - \boldsymbol{\alpha}) H(\boldsymbol{\mu} + \boldsymbol{\tau}) H(\boldsymbol{\alpha} - \boldsymbol{\mu}) f_Z(\boldsymbol{\tau}) f_Z(\boldsymbol{\alpha}) d\boldsymbol{\tau} d\boldsymbol{\alpha}$$
$$+ \int_{\mathbb{R}^2} H(\boldsymbol{\lambda} - \boldsymbol{\tau}) H(-\boldsymbol{\lambda} - \boldsymbol{\alpha}) H(\boldsymbol{\mu} + \boldsymbol{\alpha}) H(\boldsymbol{\tau} - \boldsymbol{\mu}) f_Z(\boldsymbol{\tau}) f_Z(\boldsymbol{\alpha}) d\boldsymbol{\tau} d\boldsymbol{\alpha}\}$$

A Taylor series expansion of the spectral density $f_Z(\boldsymbol{\lambda} - \boldsymbol{\tau})$ yields,

$$\mathrm{E}[I_N{}^{(t)}(\boldsymbol{\lambda})] - f_Z(\boldsymbol{\lambda}) \approx A_N(\boldsymbol{\rho}) f_Z''(\boldsymbol{\lambda}),$$

assuming $f_Z(\boldsymbol{\lambda})$ is twice differentiable.Thus, as $|\boldsymbol{\rho}|$ increases to 1, $A_N(\boldsymbol{\rho})$ goes to 0 and the bias decreases.

$$\mathrm{Cov}[I_N^{(t)}(\boldsymbol{\lambda}), I_N^{(t)}(\boldsymbol{\mu})] \approx f_Z(\boldsymbol{\lambda})^2 \; |H * H(\boldsymbol{\lambda} - \boldsymbol{\mu})|^2, \quad * : \text{convolution operator}.$$

Correlation of the tapered periodogram values is negligible at Fourier frequencies, but the variance increases as $|\boldsymbol{\rho}|$ increases. As we smooth more, the bias is made smaller but the covariance increases and We avoid loss of efficiency by making $\boldsymbol{\rho}$ depend on N.

5 Data Application

EPA provides data for atmospheric pollutant concentrations in regular grids in parts of the U.S. by running the regional scale air quality models, e.g. Models-3. Our main goal is to understand and quantify the spatial structure of nitric acid (HNO_3) concentrations measured hourly in the Eastern U.S. using the output of Models-3.

To apply the technique presented in this paper, we first chose a particular region (North Carolina) to be our main study domain where we can assume the stationarity for the concentration process. We first filter the data using the multiplicative (two-dimensional Tukey-Hanning) taper with $\rho_1 = \rho_2 = 0.1$ and the spatial periodogram is computed over the tapered data. Tapering HNO_3 data gives much reduced sidelobes for the (tapered) periodogram estimates compared to the periodogram of the raw data. This confirms that tapering will lead to a better convergence of the periodogram estimates to the spectral density.

6 Conclusion

Spectral methods provide powerful tools to analyze large spatial data. The spatial periodogram is a nonparametric estimate of the spectral density for a stationary spatial process Z, defined over $\mathbb{R}^2$. Tapering is a highly effective technique to remove edge-effects in high dimensional problems and we use data tapers to reduce the bias and variance of the periodogra In this paper, we have shown that the tapered spatial periodogram at a particular frequency is approximately unbiased and that the tapered periodogram values at distinct Fourier frequencies are asymptotically independent. For large values of the smoothing parameter in a tapering function, we obtain smaller bias of the tapered periodogram but larger variance. The loss of efficiency by smoothing can be avoided by making the amount of smoothing dependent on the sample size and the spacing between neighboring observations. By studying the asymptotic MSE of the tapered periodogram, it is shown that one can control the trade between the bias and the variance by finding an efficient normalizing factor for the tapered periodogram. In practice, to find an appropriate amount of smoothing, we propose to use a plug-in method which directly uses the data information.

7 References

Brillinger, D. R. (1981). *Time Series: Data Analysis and Theory.*
Cressie, N. (1993). *Statistics for Spatial Data.* John Wiley, New York.
Fuentes, M. (2000). A new high frequency kriging approach for nonstationary environmental processes, submitted to *Environmetrics.*
Guyon, X. (1982). Parameter estimation for a stationary process on a d-dimensional lattice. *Biometrika*, 69, 95-105.
Stein, M. L. (1999). *Interpolation of Spatial Data: some theory for kriging.* Springer-Verlag, New York.
Whittle, P. (1954). On stationary processes in the plane, *Biometrika*, 41, 434-449.