

Satellite Data: Massive but Sparse

Noel Cressie*

Tao Shi

The Ohio State University, Columbus, USA

Gardar Johannesson

Lawrence Livermore National Laboratory, Livermore CA, USA

Satellites and the instruments they carry have allowed the study of the earth and its environment as a system. The datasets are massive and they challenge data mining techniques to be scalable and adaptive to spatial heterogeneities at global and local levels. More recent instruments give higher-resolution observations but, as a consequence, the data are spatially sparse. For example, NASA's Multiangle Imaging SpectroRadiometer (MISR) provides level 2 data at $1.1km$ by $1.1km$ spatial resolution. Figure 1 shows a global map of the locations where MISR has successful retrieval of Aerosol Optical Depth on one particular day (February 1, 2005). Each white strip in the plot represents data retrieval from one MISR orbit, which is a narrow swath that MISR covers when it flies from north to south. By aggregating over time (e.g., several days), the data are somewhat less sparse, but one is still left with a highly inhomogeneous density of observation locations on the globe.

These level 2 data can be converted to level 3 data at a much lower resolution by averaging those observations falling in the lower-resolution pixels. The data are not nearly as massive, but they are still sparse in different regions of the globe. For example, the data retrieval shown in Figure 1 is just one of the 28 days used to produce a level-3 monthly product at the lower resolution. Figure 2 shows a standard MISR level-3 monthly, Aerosol Optical Depth product. The plot is a 0.5 degree by 0.5 degree global map of the averaged Aerosol Optical Depth values, where the averages are taken pixel-by-pixel over all level 2 data successfully retrieved in the pixel in the given month. Even for this low spatial resolution and monthly averaged level 3 data, there are pixels with no data (black pixels on the map) that still cover the poles and leave holes over other parts of the mid latitudes (e.g., over South America). These missing data create great difficulties for monthly global data comparison and other related studies. Our goal is to fill in the missing data and to de-noise the existing data, at level 3, in a statistically optimal way. We also wish to produce a measure

**Address for correspondence:* Noel Cressie, Department of Statistics, The Ohio State University, 1958 Neil Avenue, Columbus OH 43210-1247 (ncressie@stat.ohio-state.edu)

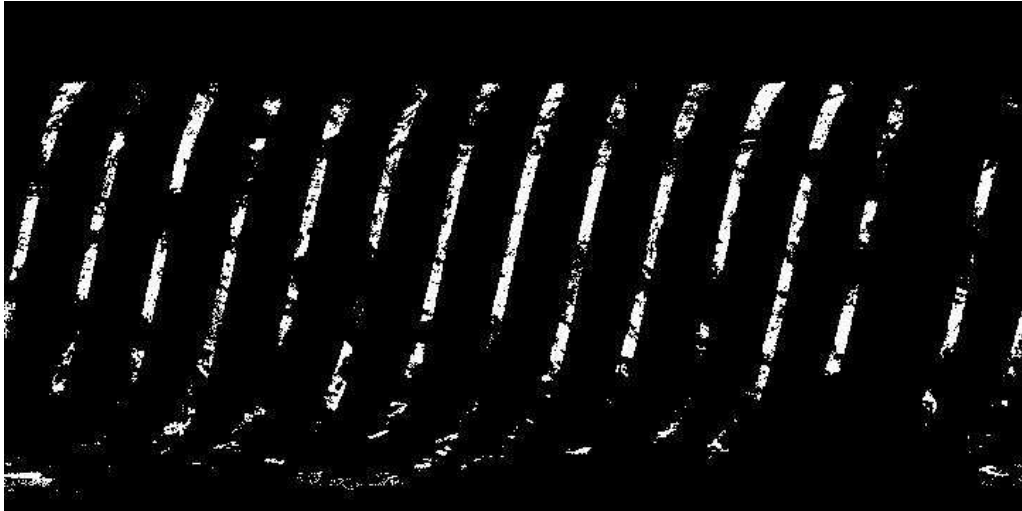


Figure 1: Coverage of MISR Aerosol Optical Depth, level 2 product, on February 1, 2005

of uncertainty associated with the completed dataset. We propose a method called kriging to accomplish these goals. The massive, but sparse, MISR Aerosol Optical Depth data will be used for testing our proposed methods.

Kriging, or spatial best linear unbiased prediction (spatial BLUP), has become very popular in the earth and environmental sciences, where it is sometimes known as optimum interpolation. With its internal quantification of spatial variability through the covariance function (or variogram), kriging methodology is able to produce maps of optimal predictions and associated prediction standard errors from incomplete and noisy spatial data (e.g., Cressie, 1993, Ch. 3). Solving the kriging equations directly involves inversion of an $n \times n$ variance-covariance matrix Σ , where n data may require $O(n^3)$ computations to obtain Σ^{-1} . Under these circumstances, straightforward kriging based on massive data is impossible.

It has been realized for some time that even a spatial dataset on the order of several thousand can result in computational difficulties. When datasets are on the order of tens of thousands to hundreds of thousands (e.g., the MISR Aerosol Optical Depth dataset has up to $720 \times 360 = 259,200$ level 3 data), kriging breaks down and *ad hoc* local kriging neighborhoods are typically used to solve the kriging equations. One avenue of recent research has been to approximate the kriging equations (Nychka et al., 1996; Nychka, 2000; Nychka, Wikle, and Royle, 2002; Furrer, Genton, and Nychka, 2005). Suggestions include giving an equivalent representation in terms of orthogonal bases and truncating the bases, doing covariance tapering, using approximate iterative methods such as conjugate-gradient, or replacing the data locations with a smaller set of space-filling locations. Kammann and Wand (2003) take up this latter idea when fitting a class of spatial models they call geoaddivitive models.

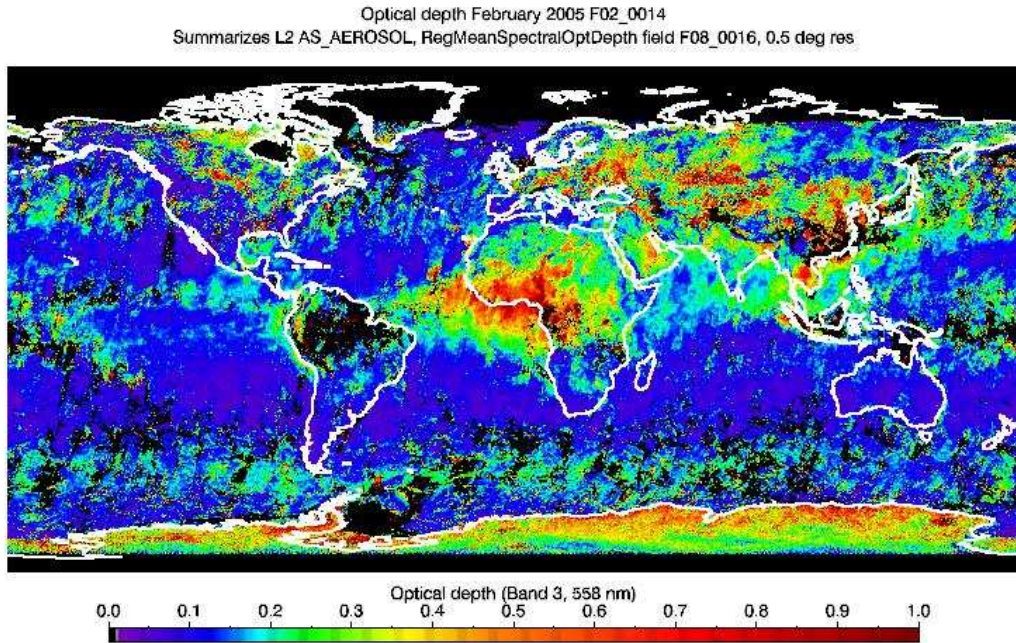


Figure 2: MISR Level 3 Monthly Aerosol Optical Depth product for February 2005

Another approach has been to choose classes of covariance functions for which kriging can be done exactly, even though the data are massive (e.g., Huang, Cressie, and Gabrosek, 2002; Johannesson and Cressie, 2004; Johannesson, Cressie, and Huang, 2007). In these papers, a multi-resolution spatial (and spatio-temporal) process is constructed explicitly so that (simple) kriging predictors can be computed extremely rapidly, with computational complexity linear in the size of the data. In the spatial case, Johannesson and Cressie (2004) achieved speed-ups on the order of 10^8 over direct kriging. They were able to compute optimal spatial predictors and their associated mean squared prediction errors in about 3 minutes for $n \simeq 160,000$. The advantage of having a spatial model that allows exact computations is that there is no concern about how close approximate kriging predictors and approximate mean squared prediction errors are to the corresponding theoretical values. For exact methods, an important question is then, “How flexible are the spatial covariance functions used for kriging?”

For the multi-resolution models referred to above, the implied spatial covariances are nonstationary and “blocky”. In this talk, we shall take a different approach to achieve orders-of-magnitude speed-ups for optimal spatial prediction, using covariance functions that are very flexible and can be chosen to be smooth or not, as determined by the type of spatial dependence expected.

We shall show that there is a very rich class of spatial covariances from which kriging of massive spatial data can be carried out exactly. Furthermore, since satellite data are global, any spatial dependencies in the data are almost certainly heterogeneous across the globe. Our intention in this talk is to address both problems (massiveness and heterogeneity) directly.

Point kriging and block kriging can be handled equally well within the class of covariance models we choose. This is important, because change of resolution is always of interest when analyzing and processing satellite data. Spatio-temporal BLUP is potentially useful for filtering and forecasting; some discussion of this will be given at the end of the talk.

References

- Cressie, N. (1993) *Statistics for Spatial Data, rev. edn.* New York: John Wiley & Sons.
- Furrer, R., Genton, M. G., and Nychka, D. (2005) Covariance tapering for interpolation of large spatial datasets, *Journal of Computational and Graphical Statistics*, submitted.
- Huang, H.-C., Cressie, N., and Gabrosek, J. (2002) Fast, resolution-consistent spatial prediction of global processes from satellite data. *Journal of Computational and Graphical Statistics*, **11**, 63-88.
- Johannesson, G. and Cressie, N. (2004) Variance-covariance modeling and estimation for multi-resolution spatial models, in *geoENV IV - Geostatistics for Environmental Applications*, X. Sanchez-Vila, J. Carrera, and J. Gomez-Hernandez (eds). Dordrecht: Kluwer Academic Publishers, 319-330.
- Johannesson, G., Cressie, N., and Huang, H.-C. (2007) Dynamic multi-resolution spatial models. *Environmental and Ecological Statistics*, forthcoming.
- Kammann, E. E. and Wand, M. P. (2003) Geoaddivitive models. *Applied Statistics*, **52**, 1-18.
- Nychka, D., Bailey, B., Ellner, S., Haaland, P., and O'Connell, M. (1996) *FUN-FITS: Data analysis and statistical tools for estimating functions*. North Carolina Institute of Statistics Mimeoseries, No. 2289, North Carolina State University, Raleigh, NC.
- Nychka, D. (2000) Spatial-process estimates as smoothers, in *Smoothing and Regression: Approaches, Computation, and Application*, M. G. Schimek (ed.). New York: John Wiley & Sons, 393-424.
- Nychka, D., Wikle, C., and Royle, J. A. (2002) Multiresolution models for non-stationary spatial covariance functions. *Statistical Modeling*, **2**, 315-331.