Journal of Geological Resource and Engineering 6 (2018) 14-18
doi:10.17265/2328-2193/2018.01.002

Advances in Ordinary Kriging Using the Stampede and Bridges Supercomputers

Erin M. Hodgess and Kendra Mhoon

University of Houston-Downtown, 1 Main Street, Houston 77002, TX, USA

Abstract: A new approach for the implementation of variogram models and ordinary kriging using the R statistical language, in conjunction with Fortran, the MPI (Message Passing Interface), and the "pbdDMAT" package within R on the Bridges and Stampede Supercomputers will be described. This new technique has led to great improvements in timing as compared to those in R alone, or R with C and MPI. These improvements include processing and forecasting vectors of size 25,000 in an average time of 6 minutes on the Stampede Supercomputer and 2.5 minutes on the Bridges Supercomputer as compared to previous processing times of 3.5 hours.

Key words: Kriging, geostatistics, high performance computing.

1. Introduction

The theoretical underpinnings of kriging have been available for more than fifty years [1]. Yet due to kriging's computational intensity it has not been readily accessible until recently. Currently it can be found on the open source statistical programming language R [2]. However, the process in R can be quite slow with large vectors.

A combination of tools has been utilized to combat this slow processing issue with large vectors. Specifically, R, Fortran, the MPI (Message Passing Interface) wrappers to tie into Fortran (to render it a high performance language), and the pbdDMAT [3] package within R speedup ordinary kriging.

The following research begins with a description of the theoretical concept of ordinary kriging from geostatistics (Section 2), next, a description of the current kriging computational process in R (Section 3), followed by a simulation study of the new implementation process (Section 4), then a fitting of a function, based on longitude and latitude which produces a forecast and its variance, and finally constructs a map (Section 5). The map can be easily viewed on Google Earth via a laptop, smartphone, or iPad.

2. Definitions from Ordinary Kriging

2.1 Theory

We consider a response vector [4],

$$z = f(x_1, x_2) \qquad (1)$$

where, x_1 is longitude and x_2 is latitude.

Assuming first order stationarity,

$$E[Z(\mathrm{x})] = \mu \ \forall \mathrm{x} \in \Re \qquad (2)$$

Next, covariance must be considered. Therefore, assuming second order stationarity and the fact that one variable is measured against itself, autocovariance exists:

$$\begin{aligned} C[Z(x), Z(x+h)] &= E[\{Z(x)-\mu\}\{Z(x+h)-\mu\}] \\ C[Z(x), Z(x+h)] &= E[\{Z(x)\}\{Z(x+h)\}-\mu^2] \\ C[Z(x), Z(x+h)] &= C(h). \end{aligned} \qquad (3)$$

Then, examining autocorrelation:

$$\rho(h) = \frac{C(h)}{C(0)}. \qquad (4)$$

Finally, obtaining semivariances:

Corresponding author: Erin M.Hodgess, associate professor, research fields: geostatistics and time series.

$$\gamma(h) = C(0) - C(h). \quad (5)$$

The autocovariances will produce a positive semidefinite matrix. Yet, in order for the kriging equations to form a meaningful solution a positive definite matrix is necessary. Therefore, we model the semivariances as a function:

$$\gamma(h) = 0.5E[\{Z(x) - Z(x+h)\}^2] \quad (6)$$

The variogram function must:

- be monotonically increasing;
- be constant or have an asymptotic maximum (sill);
- have a nonnegative intercept (nugget).

Of particular concern are the following four models: the exponential, spherical, Gaussian, and the Stein version of the Matheron model. Fig. 1 displays a visual structure of the aforementioned models.

2.2 Sampling

Moving from the theoretical realm to the real world, sampling must be used to estimate the mean, covariances, and semivariances. Since covariance is a symmetric function, construction of the sample covariance and semivariance functions can be fashioned similar to that of a histogram. First, “select” a bin size, then obtain pairs of data, and calculate the average semivariances over the bins. This will provide an empirical histogram:

$$\overline{\gamma}(h) = \frac{1}{2m(h)} \sum_{i=1}^{m(h)} [z(x_i) - z(x_i + h)]^2 \quad (7)$$

Next, plot these values, then superimpose theoretical models selecting the best fitting model.

2.3 Kriging Calculations

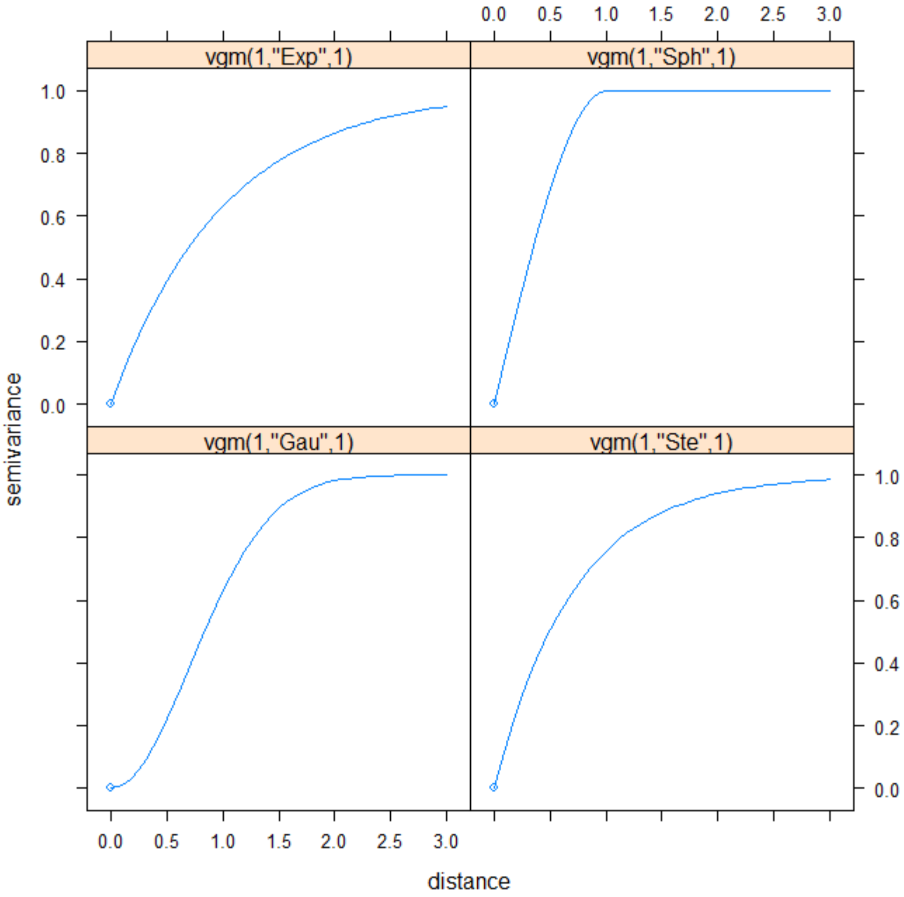

Fig. 1 The variogram of the Exponential, Speherical, Gaussian, and the Stein version of the Matheron Model.

The first step of the kriging calculations is to produce the spatial mean. The spatial mean is not the average of the observations, but rather,

$$\hat{\mu} = (1'\square C^{-1}\square 1)^{-1}\square(1'\square C^{-1}\square z) \qquad (8)$$

where C is the matrix of the semivariances, z is the original response values, and 1 is a vector of ones.

Thus the actual kriging prediction system for one point is

$$\hat{z} = \hat{\mu} + c_0 C^{-1}(z - \hat{\mu}1) \qquad (9)$$

Note, $\hat{z}$ is easily generalized for many points.

This system is a best linear unbiased predictor. Therefore, it should provide the optimum information for mapping purposes.

3. Discussion of Existing Tools

A famous data set in spatial statistics, the Meuse River (a river which runs through the Netherlands) data set was utilized in order to model the log of the zinc content near a river. The original data set only contains 155 points, for this analysis the points were resampled for *n* = 1,000, 2,000, 10,000, 20,000, and 25,000. For the existing tool as per [5, 6] the process begins by fitting an empirical variogram model, either by eyeballing, or automatically, via the automap package in R [7], running the kriging function and producing a forecast. Tests were run on a MacBook, running El Capitan, Version 10.11.6 (Table 1).

4. Implementation of the New Process

In order to improve the existing R tool, its steps were timed. The examination determined that the model selection and the kriging prediction were the steps in the process where bottlenecks seemed to occur. The first attempt at improvement was the Rmpi package [8], which ties MPI wrappers into C, for the model selection section, but that did not show great promise. Next, Fortran with MPI was implemented, and that seemed to improve model selection timing. Then, R, Fortran, and MPI were combined.

The analyses were performed on the Stampede and Bridges supercomputers. The Stampede supercomputer is located at the Texas Advanced Computing Center at the University of Texas. Stampede ranks Number 12 on the Top 500 list of supercomputers in the world. Most recently, access to Bridges was obtained, which has accelerators in the Fortran language. Namely, their Tesla GPU processors allow for the Fortran progams to have tremendous speedup.

When R, Fortran, and MPI were combined, the best results occurred on the Stampede and Bridges supercomupters (Table 1).

Implementing the kriging process was quite lengthy. Note, R is not designed for large data sets, so in order to work around that restriction, files were written back and forth, but that was not particularly successful. So, the pbdDMAT package, which is the R version of a high performance LAPACK was used. pbdDMAT allows large data to be brought into R directly from Fortran programs, using R for matrix multiplication, inversion, and matrix-vector multiplication. The analysis then utilized the high performance of Stampede and Bridges.

The smallest vector performs well for the Mac. Presumably in this case, communication overhead is slowing down the supercomputer. But in every other instance, Stampede is out-performing the Mac by far,

Table 1 Processing time (in seconds).

Vector	Mac	Stampede	Bridges
1,000	3.9	9	6
2,000	11.7	11	5
5,000	80	15	7
10,000	975	35	17
20,000	5594	184	95
25,000	12702	349	147

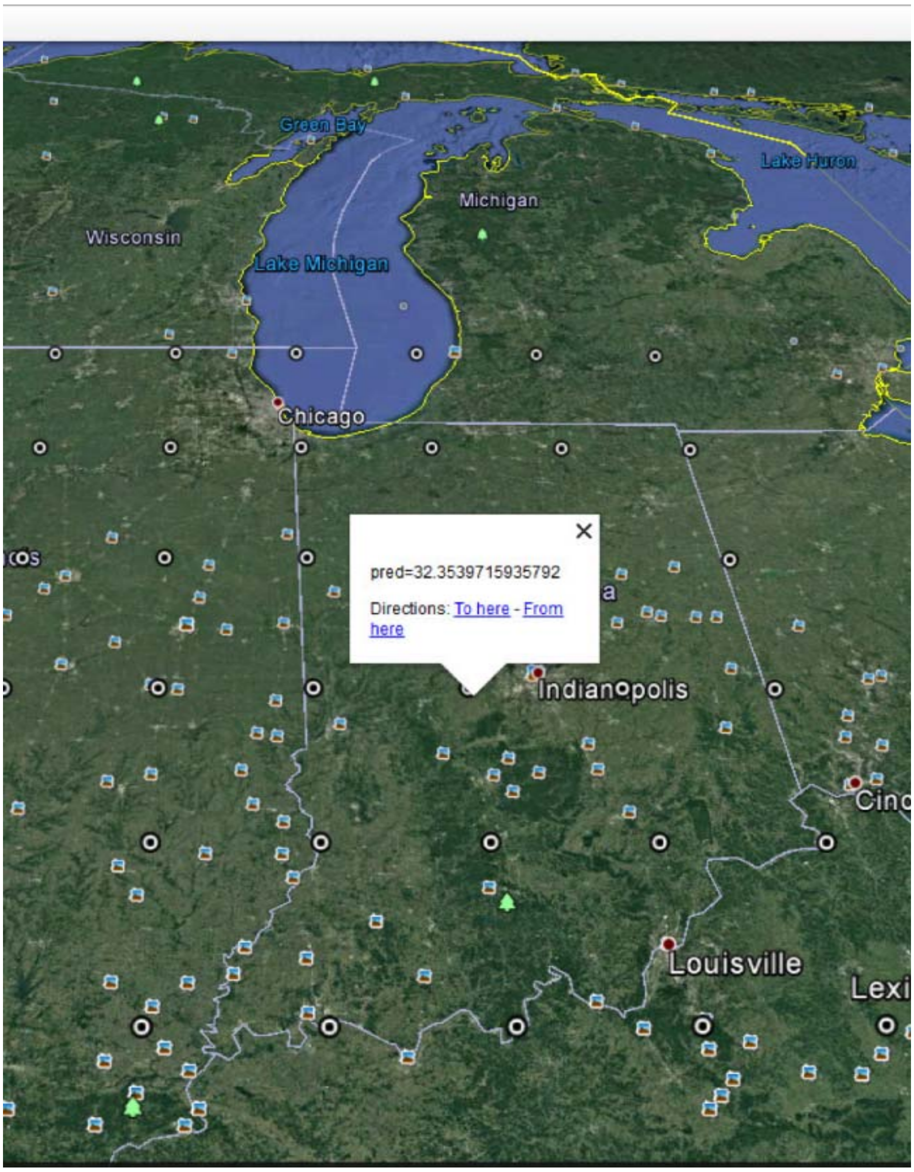

Fig. 2 Forecast output.

with Bridges surpassing all.

The number of cores, or central processing units, is not extremely large, and both supercomputers perform very well. Stampede has restrictions on the number of cores, while Bridges permits the maximum for all vectors. This new process is an excellent analysis tool. Multiple cores were used to produce 80-97% speedup.

5. A Real world Example

A data set of 2,681 observations was obtained from a study done by the US Geological Survey called the National Geochemical Survey. In this survey the levels of various soil sediments were measured. This real world example used the lead level variable. The new process was implemented and a map was produced (Fig. 2). The new process and map took 11 seconds on Stampede and 7 on Bridges—forecasting 50 points, which represent lead levels in parts per million (ppm) onto the output map.

Forecast values are located with a click on the round dots. Furthermore, the forecasts are in a gridded convex hull around the original data set.

There are alternatives to the ordinary kriging method; IDW (inverse distance weighted) interpolation, and linear regression. These methods use different assumptions than the kriging method. The formula for IDW, as found in Ref. [5], is

$$\hat{Z}(s_0) = \frac{\sum_{i=1}^{n} w(s_i)Z(s_i)}{\sum_{i=1}^{n} w(s_i)},$$

We have

$$w(s_i) = \| s_i - s_0 \|^{-p}$$

This assumes no nugget effect. Also, the R function for IDW does not calculate the error variance.

Finally, the longitude and latitude could be used as components in a multiple linear regression. This is typically not a good plan for spatial data because items which are closer together tend to have a stronger relationship than those farther apart.

The IDW and regression models performed on the Mac, Stampede, and Bridges found nearly identical results on all; roughly 2 seconds each. Yet, proceeding with these models should be done with great caution.

The designers of the gstat [6] package added LAPACK to the code in late 2015. However it still does not impact large data sets. Similarly, others have created maps to determine the spatial distributions of lead with the Kriging method (Rogozan, Micle, & Sur, 2016). Yet, no researcher has discussed the time burden for creating such maps.

6. Conclusion and Future Work

Our new method is an effective tool for speeding up kriging and forecasting. R, Fortran, MPI, and the pbdDMAT package were used efficiently in order to produce a useful analysis tool. Future work will consist of an online course and an R package.

Also, this analysis tool will be extended to the spatial-temporal arena. It too, is quite slow in processing speeds. This is a natural extension of the current research.

Acknowledgments

This work used the XSEDE (Extreme Science and Engineering Discovery Environment), which is supported by the National Science Foundation (USA) grant number ACI-1053575. We thank Kevin Chen, Kay Hunt, Nancy Wilkins-Diehr for their assistance which was made possible by the XSEDE ECSS (Extended Collaborative Support Service) program and the Campus Champions Fellows program.

References

[1] Matheron, G. 1963. “Principles of Geostatistics.” *Economic Geology* 58: 1246-66.

[2] R Foundation for Statistical Computing. *R: A Language and Environment for Statistical Computing*, Vienna, Austria: [ref. July 29, 2016]. URL https://www.R-project.org.

[3] Schmidt, D., Chen, W.-C., Ostrouchov, G., and Patel, P. 2012. *pbdDMAT: Distributed Matrix Algebra Computation*. R Package, URL http://cran.r-project.org/package=pbdDMAT.

[4] Rossiter, D. *Geostatistics and Open source Mapping*, online Course [ref. July 29, 2016]. URL http://www.css.cornell.edu/faculty/dgr2/teach/degeostats.html.

[5] Bivand, R., Pebesma. E., and Gomez-Rubio, V. 2013. *Applied Spatial Data Analysis with R*. 2nd Edition.

[6] Pebesma, E. 2004. “Multivariable Geostatistics in S: The Gstat package.” *Computers and Geosciences* 30: 683-91.

[7] Hiemstra, P., Pebesma, E., Twenhofel, C., Heuvelink, G. B. M. 2009. “Real-Time Automatic Interpolation of Ambient Gamma Dose Rates from the Dutch Radioactivity Monitoring Network.” *Computers and Geosciences*.

[8] Rogozan, G. C., Micle, V., and Sur, I. M. 2016. “Maps of Heavy Metals in Cluj County Soils Developed Using the Regression-Kriging Method.” *Environmental Engineering and Management Journal* 15 (5): 1035-9.