Pragmatic Bayesian Kriging for Non-Stationary and Moderately Non-Gaussian Data

Konstantin Krivoruchko and Alexander Gribov

1 Introduction

Development of reliable real time automatic statistical interpolation model is a common issue in the GIS community. In the Geostatistical Analyst, we regularly update and improve the semivariogram model fitting algorithm and the most recent version of the software performs much better than the first one released in 2001 and described in [1]. However, there is little hope that the default semivariogram model will be close to optimal because in practice the observations may include erroneous values, measurement errors can be large and vary across space, the data are often contaminated by trend, and the data are rarely Gaussian.

In practice, interactive data exploration and modeling allows the researcher to find a reliable geostatistical model when the number of samples is relatively small, usually less than a thousand. However, a large number of GIS users prefer non-interactive interpolation for both small and very large datasets.

In the case of nearly perfect data (several hundred samples of nearly stationary Gaussian data without errors), a known solution for the automatic data interpolation problem is fully Bayesian kriging with non-informative or objective prior distributions for the parameters, although there are still some problems, including large computational time and understanding of the meaning of the prior distributions.

Since there are many Geostatistical Analyst users who asked for automatic interpolation, our goal became the development of a series of semi-automatic empirical Bayesian kriging models that work reasonably fast; do not require specification of the prior distributions for the model parameters; produce reliable outputs with default parameters; allow moderate local and large global data non-stationarity; allow for

K. Krivoruchko (✉) · A. Gribov
Environmental Systems Research Institute, 380 New York St, 92373 Redlands, CA, USA
e-mail: kkrivoruchko@esri.com

A. Gribov
e-mail: agribov@esri.com

E. Pardo-Igúzquiza et al. (eds.), *Mathematics of Planet Earth*,
Lecture Notes in Earth System Sciences, DOI: 10.1007/978-3-642-32408-6_15,

varying measurement error; locally transform data to Gaussian distribution, if needed; use explanatory variables; work with counts and proportions collected in points or polygons; simulate (transformed) Gaussian, binomial, and negative binomial fields; and potentially can be used on supercomputers or clusters.

A non-technical description of Geostatistical Analyst's empirical Bayesian kriging models can be found in [2] and an article with mathematical details is under review. In this paper we briefly describe our basic models and illustrate their usage using 1.35 billion samples collected using LiDAR technology.

2 Empirical Bayesian Intrinsic Random Function Kriging and Kriging with Local Data Transformation

There are several model candidates for semi-automatic interpolation, including objective Bayesian kriging [3], low-rank kriging [4], approximations by Gaussian Markov random fields [5], and Bayesian bootstrap [6]. Although all these models are valuable, we found that they cannot be safely used in general purpose commercial software for reasons specific to each model.

We decided to work on the Bayesian bootstrap model enhancements and generalization. The authors of [6] use informative priors simulated from the data by simple kriging with estimated covariance model. It was assumed that the data are normally distributed and the mean value is constant and independent from the covariance model parameters. The main differences between our implementation and the model discussed in [6] are the following:

- Our calculations are done locally on spatially-contiguous subsets, possibly overlapping, and the simulations from each subset are mixed together to capture the local spatial structure. The authors of [6] do calculations globally (on the entire dataset).
- All model parameters are estimated simultaneously.
- After each simulation, we do a Bayesian model update while the authors of [6] keeps simulating from the same model without updating. The difference between prior and posterior semivariograms is usually large as shown in Fig. 1.
- We support intrinsic random function kriging (IRFK) of order zero and one and also simple kriging with or without first order trend removal. The authors of [6] considers simple kriging without detrending. We use the following generalized covariance models: linear, power, thin plate spline (with IRFK) and the K-Bessel covariance model (with simple kriging.)
- The authors of [6] supports Box-Cox transformation. We support a much more flexible data transformation described in [7].
- Our software is interactive (the user can see a large number of graphs at the specified locations, including those that are shown in Figs. 1 and 2) and it works very fast due to large effort spent on algorithms optimization.

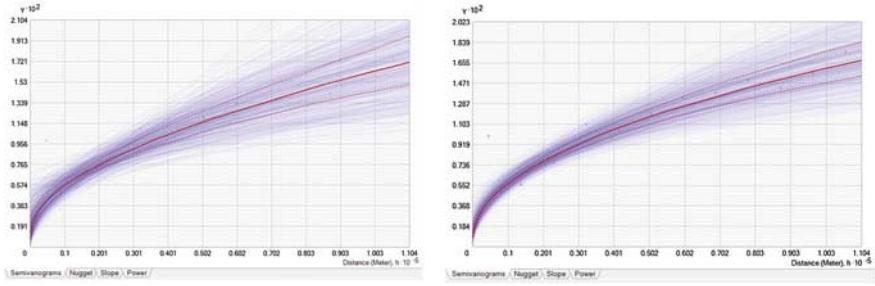

Fig. 1 An example of prior (*left*) and *posterior* (*right*) power semivariogram models. *Red lines* show quartiles of the distributions. *Prior* models are weighted equally. The weight of each *posterior* model is represented by line darkness

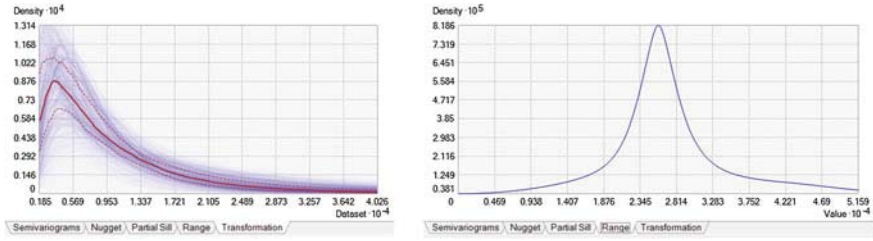

Fig. 2 Distributions of the data transformations (*left*) and the range parameter (*right*)

Prediction with variable model parameters is very useful when the number of samples is insufficient for accurate model fitting. It can be even more useful in the case of large datasets when the model is changing across space and the changes can be non-smooth, as in the case study in [8].

When the dataset is large, the software creates subsets with specified number of samples M (M can vary from 30 to 500) recursively by splitting the data locations on two or three subsets. The division on two subsets is a straight line which divides the data into two equal parts in such a way that the sum of the squares of the deviation of each point from the corresponding center of mass is the lowest. In the case of three subsets, three beams from the center of the area divide the plane. For each subset with the number of samples not equals to M, the samples are added or removed to get exactly M samples, if possible. When the prediction searching neighborhood includes observations from several data subsets, the datum contribution to the prediction is divided between the subsets.

Figure 3 shows an example of predictions and prediction standard errors produced using more than a billion samples collected using LiDAR technology in part of Ohio, USA [9]. We used a model with the least number of parameters, IRFK, with linear semivariogram (this model describes Brownian motion, which can be relevant to elevation formation). It is interesting that the variation of the prediction standard

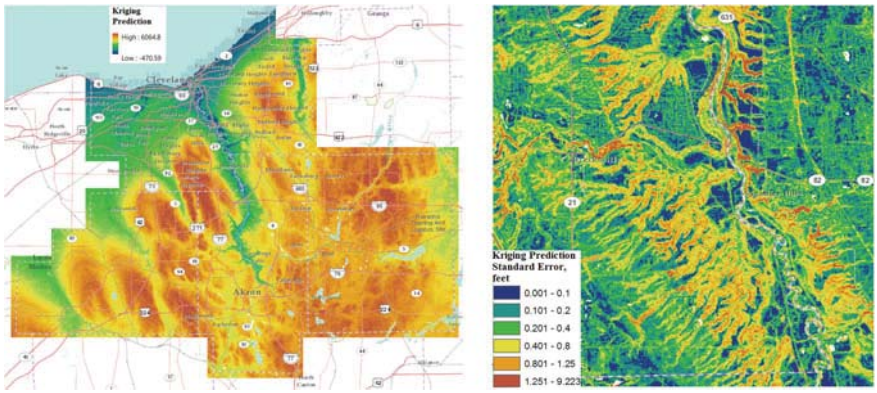

Fig. 3 Predictions using 1.35 billion samples collected using LiDAR technology in part of Ohio, USA (*left*) and enlarged prediction standard errors in the central part of the data domain (*right*). The data are from [9]. The units are feet. IRFK with linear semivariogram was used

errors is much larger than the variation of the predictions (however, the result of elevation modeling, a DEM, is almost always available as a prediction surface only.)

Extensive testing using a large variety of data showed that our model is a reliable automatic interpolator, which produces accurate predictions with non-stationary and non-Gaussian data even when the data vary non-smoothly across space.

References

1. Gribov, A., Krivoruchko, K., & Ver Hoef, J. M. (2006). Modeling the semivariogram: New approach, methods comparison, and simulation study. In T. C. Coburn, J. M. Yarus, & R. L. Chambers (Eds.), *Stochastic modeling and geostatistics: Principles, methods, and Case Studies*, *Volume II: AAPG Computer Applications in Geology 5* (pp. 45–57).
2. Krivoruchko, K. (2012). Empirical Bayesian kriging. ArcUser, Fall 2012. http://www.esri.com/news/arcuser/1012/empirical-byesian-kriging.html
3. Kazianka, H., & Pilz, J. (2012). Objective Bayesian analysis of spatial data with uncertain nugget and range parameters. *Canadian Journal of Statistics*, *40*(2), 304–327.
4. Katzfuss M. (2013). Bayesian nonstationary spatial modeling for very large datasets. (wileyonlinelibrary.com) doi:10.1002/env.2200.
5. Lindgren, F., Rue, H., & Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach. *Journal of the Royal Statistical Society, Series B*, *73*(4), 423–498.
6. Pilz, J., & Spöck, G. (2008). Why do we need and how should we implement Bayesian Kriging methods. *Stochastic Environmental Research and Risk Assessment*, *22*(5), 621–632.
7. Gribov, A., & Krivoruchko, K. (2012). New flexible non-parametric data transformation for trans-Gaussian kriging. *Geostatistics Oslo 2012, quantitative geology and geostatistics* (Vol. 17, Part 1, pp.. 51–65). Netherlands: Springer.
8. Krivoruchko, K. (2012). Modeling contamination using empirical Bayesian kriging. ArcUser, Fall 2012, http://www.esri.com/news/arcuser/1012/modeling-contamination-using-empirical-bayesian-kriging.html
9. Ohio Geographically Referenced Information Program (OGRIP) (2011). Ohio statewide imagery program II. http://gis3.oit.ohio.gov/geodata