Geostatistical Models for the Spatial Distribution of Uranium in the Continental United States

Sara Stoudt[1]

[1]Smith College, Unit 8697 1 Chapin Way Northampton, MA 01063,
Telephone: 724-464-3179
Email: sstoudt@smith.edu

1. Introduction

The United States Geological Survey has been working to sample geochemical properties across the United States (USGS). Despite its efforts, a complete picture of the amount of uranium found across the United States is not readily available. Because uranium is of interest to many government agencies, as uranium can both be harmful to the environment and be used to produce energy, an accurate interpolated surface of uranium would be useful to many parties (USDOE).

In this poster, we compare the performance of several non-parametric geostatistical models for uranium deposits including the k nearest neighbors method, local regression models, generalized additive models, and Gaussian Process models (kriging). In each case, we optimize model parameters using 15-fold cross validation on a training set, and choose the final, most accurate model by comparison of predictions with a test set. Evidence for successfully avoiding overfitting through this cross validation process is seen in the applicability of our optimal parameters for the prediction of substances other than uranium. We find that although each method produces an interpolation that is visually distinct, the performance of each on the test set, as measured by the root-mean-squared error, is only negligibly different from the others'.

2. Challenges

Modeling uranium deposits faces several challenges. First, the samples are not uniformly distributed across the United States, which introduces uncertainty to any model of sparsely sampled areas. Second, standard kriging is not appropriate for this data, since the distribution of uranium is neither symmetric nor normal, and furthermore cannot be easily transformed into a quasi-normal distribution. Third, the large sample size of over 40,000 uranium measurements makes traditional kriging almost impossible on a personal computer.

2.1 Dealing With the Large Data

Local regression and generalized additive models do not pose any computational problems given the size of our data. The parameter sweep and cross validation can be made faster through trivial parallelization. Standard kriging is too computationally intensive to work on this data set. We get around this by using Lattice Krig which implements a multi-resolution Gaussian Process model that takes advantage of sparse matrices to speed up the computations (Nychka).

2.2 Results

We find kriging to be the most effective method for predicting uranium (see Figure 1). The final root mean squared error on the test set is just under 6 parts per million (ppm). This is underwhelming, as most of the uranium samples are less than 5 ppm, but we will see that a large portion of this error comes from a few highly influential points. This method overestimates the value of uranium more often than it underestimates it, but we can balance the residuals by using a logarithmic transformation of the uranium prior to lattice kriging at the expense of a slightly larger RMSE. We also see large residuals in areas where we would expect more uranium from a geological point of view (see Figure 2). However, the residuals are not spatially correlated. The full results from other methods can be found in Table 1. The parameters that were determined to be optimal for uranium extend to the accurate prediction of other substances including aluminum, chromium, gallium, lithium, and magnesium.

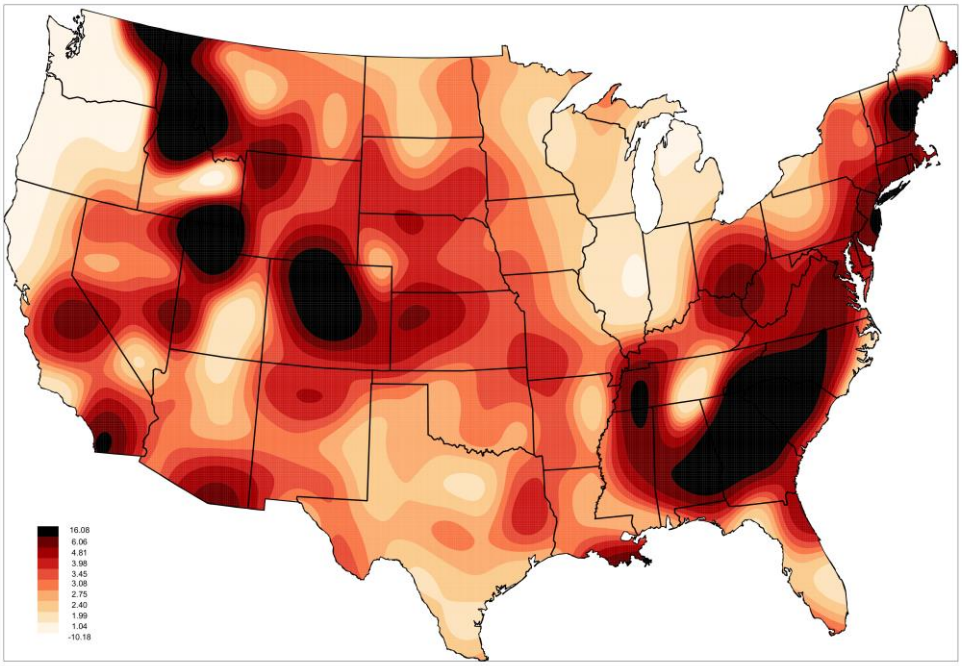

Figure 1. Interpolation using Optimal Lattice Krig Method

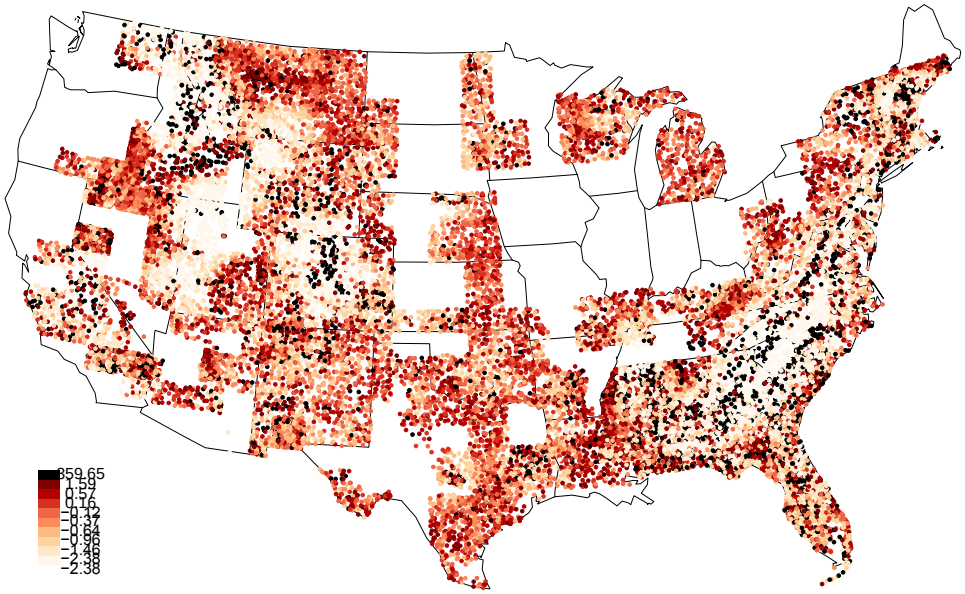

Figure 2. Residuals using Optimal Lattice Krig Method

Method	Test RMSE (ppm)	Max (ppm)	Qualitative Assessment
Block Krig	5.930	31.42	granularity issue
LK3	5.967	16.08	most detailed yet smooth interpolation
Smooth BK	6.021	31.42	visually more reasonable than original BK
Local Regression	6.033	29.13	can be marginally better than LK by using $\alpha < 0.2$ but give up some generalizability
Log LK3	6.035	> 1000	more balanced residuals but some unrealistically large values
Log LR	6.118	237	more balanced but some unrealistically large values
KNN	6.182	18.50	in sparse regions relies heavily on samples that may not be representative
GAM TE1	6.268	34.47	smallest standard errors
Z-Score Krig	6.240	16.18	normality adjustment does not help us
Local Copula	6.280	NA	computationally intensive for full interpolation
N-Score Krig	6.930	365.70	large predictions do not correspond with a true extreme values

Table 1: Results on Test Set

3. Strategies for Overcoming the Normality Assumption

The method that performs the best assumes that the data follows a Gaussian distribution. Transformations can help remove patterns in the residuals, but we want to find adaptations of the standard methods that do not require normality.

3.1 Transformations

Through a logarithmic transformation we can increase our RMSE to a bit over 6 ppm yet remove the pattern of overestimating more than underestimating. A Box Cox transformation gives similar results with a slightly larger RMSE. Transformations can help remove patterns in the residuals, but we fail to lower the overall RMSE.

3.2 Indicator Kriging

Indicator kriging does not have any distributional assumptions as it predicts the probabilities that the response variable is less than a certain value (Cressie). We use Lattice Krig to implement this by predicting probabilities for the response lying in ten quantiles of the uranium distribution. We then combine these predicted values to determine an expected value of uranium for each location. Note that to determine the indicator quantiles and to calculate the expected value, we rely on the true quantiles of uranium in our training sample. The expected value interpolation looks reasonable and follows the same types of patterns of high areas found by other methods. However, the RMSE is very poor (67.4 ppm), and the residuals reach very large values throughout. These large overestimates are most likely due to the many standard adjustments that need to be made to make the estimates fit within the rules of probability.

3.3 Disjunctive Kriging

We can force the uranium data to follow a normal curve exactly by lattice kriging the z-scores. We can then un-transform to assess performance. This leads to a RMSE of 6.24 ppm on the test set. We can try a similar method based on ranked data. N-Score kriging ranks the sample data and assigns each sample a value based on the expectation of the order statistic of the same rank in a standard normal random variable. We then lattice krig on the assigned values and again un-transform to assess performance. This gives us a RMSE of 6.93 ppm. We can also do the assignments based on the expectation of the order statistic of the same rank in a beta random variable or an extreme value random variable.

3.4 Generalized Gaussian Processes- Copulas

Another option is to use a copula as a replacement for the Gaussian Random Field in the kriging (Kazianka and Pilz). The benefit of using this method is that we know that our marginal distributions are non-Gaussian. Instead of fitting a parametric copula to our data, which is extremely computationally intensive, we can create a data driven empirical copula. This is still computationally intensive, but it is reasonable to use on a subset of the data at the state level or locally. We choose to look at Colorado. We build an empirical copula using the following steps:

1. Find empirical cumulative distribution of uranium F(y).

2. Pick a distance **h** and find locations y_i and y_j in the training set that are separated by **h**. The value of **h** and the width of the band for approximation must be chosen.
3. Create a set of pairs representing locations that are separated by **h** using the empirical cumulative distribution values for the uranium amounts in each location: $(F(y_i), F(y_j))$
4. This set will contain coordinates that lie within the unit square. When plotted, these will form our bivariate density, or copula, of interest. Now we can use this empirical copula to predict uranium values for locations in the test set.
5. For a test point **s** we find the 10 nearest neighbors in the training set. The use of 10 neighbors is a choice that could be further optimized. For each neighbor **n** we draw a random value from the copula conditioned on the empirical cumulative density value for **n**, F(n). We can choose to increase the number of random values drawn and aggregate them in some way.
6. We must choose how to aggregate the values from each neighbor. A first step is to use the mean across the neighbors.

There are many choices in this method including the choice of **h**, the bandwidth around **h**, the number of neighbors to use in prediction, and the aggregation method, that we optimized using a training and test set. Our best set of parameters yielded a RMSE of 6.28 ppm.

4. Future Work: Extreme Values

In Figure 3 the bottom ten blocks show the influence on our results of extreme values in. while the map shows the location of these influential points. The top block represents the combined effect of the rest of the points. All of our methods fail to predict extremely large values of uranium. For context, the point with the largest influence occurs in a location that used to be a commercial uranium mine. Future work will include the exploration of extreme value methods to tackle these few, but influential samples.

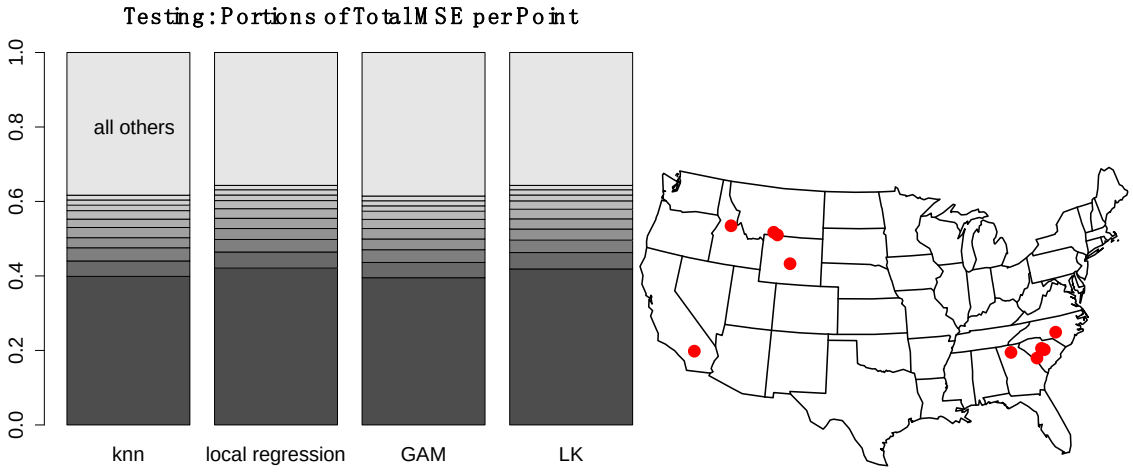

Figure 3: Top 10 Most Influential Points

5. Acknowledgements

Thank you to Ben Baumer, Nick Horton, and Antonio Possolo for advice and guidance on this project. Thank you to NSF Travel Support for funding my participation in this conference.

6. References

Cressie, Noel A. C, 1993, Statistics for Spatial Data. *Wiley Series in Probability and Mathematical Statistics*

Kazianka, Hannes, and Pilz, Jurgen, 2010, Geostatistical modeling using non-gaussian copulas. *Accuracy Symposium*.

Nychka, Douglas, Bandyopadhyay, Soutir, Hammerling, Dorit, Lindgren, Finn, and Said, Stephen, 2013, A multi-resolution Gaussian process model for the analysis of large spatial data sets. http://nldr.library.ucar.edu/repository/assets/technotes/TECH-NOTE-000-000-000-875.pdf

Nychka, Douglas, 2014, Package "Lattice Krig". http://cran.r-project.org/web/packages/LatticeKrig/LatticeKrig.pdf

United States Department of Energy (USDOE), 2014, DOE Submits Its Defense-Related Uranium Mines Report to Congress. http://energy.gov/lm/articles/doe-submits-its-defense-related-uranium-mines-report-congress

United States Geological Survey (USGS), 2004, The National Geochemical Survey - Database and Documentation. http://mrdata.usgs.gov/geochem/doc/home.htm

Wood, Simon N, 2006, Generalized Additive Models: An Introduction with R. *Chapman and Hall/ CRC*