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

Abstract:

Although the United States Geological Survey works to sample geochemical properties across the country, a complete understanding of the distribution of uranium remains elusive. Such an understanding would be useful to many government agencies since uranium can both be harmful to the environment and used to produce nuclear energy. We compare the performance of several non-parametric models for uranium deposits including the K Nearest Neighbors method, Local Regression models, Generalized Additive Models, and Gaussian Process models (kriging). We optimize model parameters using cross validation on a training set, and choose the final, most accurate model by comparison of predictions with a test set. We recommend using a Lattice Krig model with an optional logarithmic transformation for uranium interpolation. 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.

1 Introduction

Using statistics to answer geological questions has occurred since the introduction of ``geostatistics" in the 1960s. Danie G. Krige, one of the developers of geostatistics' main tool (kriging), was motivated by the search for gold [5]. It is unsurprising that the use of kriging was originally focused on the evaluation of reserves in mining deposits. Predicting the amount of uranium started as a business venture with the goal of minimizing exploration expenditure for mining operations [6]. Government interest soon followed, and there became a heavier emphasis on integrating different data sets and analysis methods [21]. Until the 1980s the primary goal of geostatistics was to explain spatial patterns and predict the value at unsampled locations. In the mid 1980s there was a shift towards using geostatistical methods to model uncertainty of the estimates at unsampled locations [9]. This was in part because an increase in computational resources made simulations using these methods more feasible. As data becomes more abundant, strategies for working with and better visualizing spatial ``big data" have also become a priority [8].

Kane et. al. performed an extensive optimization of the parameters for inverse distance weighting interpolation of geochemical properties. However, this method is only appropriate for regular sampling patterns and breaks down when samples are not uniform across the region of interest [15]. Wu et. al. compared different kriging methods for skewed data, but they replaced extreme values of the data with the median value instead of incorporating them into their analysis [24]. The interpolation of uranium has been done to meet various motivations, but often on a small scale. For example, Garza et. al. focused on uranium prediction in a portion of New Mexico [11].

Our contribution to the literature is an extensive comparison of not only kriging methods, but other methods that are less widely used in geostatistics yet are still reasonable tools for the problem of modeling uranium. We do this on a large scale, using samples from the continental U.S. that are not uniformly distributed across the region and without removing any ``troublesome" large values.

2 Background

Uranium is most well-known for its use in nuclear weaponry and energy, but it also naturally occurs in low concentrations in soils and sediments throughout the United States. Since the 1960s the U.S. Geological Survey (USGS), located in the U.S. Department of the Interior, and other organizations, both public and private, have performed geochemical sampling around individual mines, in small areas, and even on the state scale. The USGS continues to collect geochemical data and compile geochemical data from many different sources.

Despite the efforts of the USGS, 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.

We use data that the USGS made publicly available for this research (see Figure 1). Before working with the data, we went through standard cleaning procedures, including the removal of all entries that had a missing value in a coordinate or uranium value. Values of uranium that were obviously wrong had already been removed before the data was made publicly available. There are some large values relative to neighboring values, but they are within the realm of possibility, so we did not clean for potential outliers [20].

There were 5,450 duplicates in coordinates that were sampled twice; this is about ten percent of the final data set. When creating the data set, the median of uranium values at duplicate coordinates replaced the duplicate values. We checked to make sure that the duplicate measurements

were not wildly different from one another. The majority of duplicates had very similar uranium values. The data contains latitude and longitude coordinates, but we chose to transform these coordinates using the Lambert Conformal Conic projection. This minimizes the distortions that arise due to projection from the 3-D globe to the 2-D plane. Note that all maps use this projection unless otherwise specified.

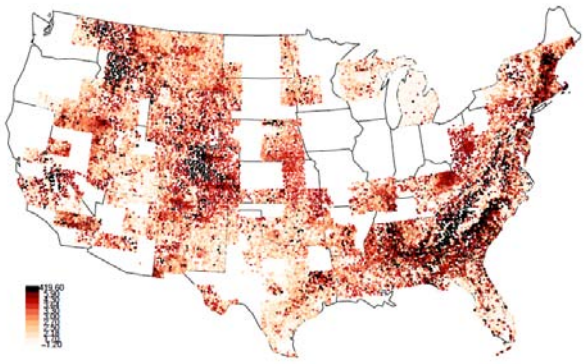


Figure 1: Uranium Distribution in our Sample

3 Goals

Our goal is to be able to predict the amount of uranium at any point within the continental United States with minimal error and minimal uncertainty i.e. with both high accuracy and precision. Modeling uranium deposits faces several challenges. First, the known values of uranium are not sampled uniformly across the United States, which introduces uncertainty to any model of sparsely sampled areas. Second, the standard method of spatial interpolation, 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.

We explore different methods for spatial interpolation, optimizing the parameters over a training set, and finding the best of the best on a test set. In cases where the methods are visually distinct yet have similar results for our criteria of ``best" we are able to use geological intuition as well as qualitative assessments of the interpolations to make recommendations.

We want a model for uranium as a function of spatial coordinates X:

$$y \sim f(X) + \epsilon$$

that minimizes the error:

$$min_f \; min_p \|y - \hat{y}\|^2,$$

where $\hat{y} = f_p(X)$. Note that we first find the parameters p of each model f that produce the minimum sum of squared differences between the true values of uranium and the predicted values of uranium and then find the model f that has the smallest sum of squared differences overall.

3.1 Contributions

We make three contributions to the existing knowledge of uranium in the lower 48 states.

1. We develop a scheme for determining the optimal set of parameters for predicting uranium values that does not compromise the generalizability of our results. This includes the use of a training and test set, c fold cross validation, and a parameter sweep, yet avoids overfitting. The avoidance of overfitting is demonstrated by the fact that the optimal parameters chosen per model are similar if not exactly the same as those chosen for the prediction of other substances including aluminum, chromium, gallium, lithium, and magnesium.
2. We compare and contrast the K Nearest Neighbors method, Local Regression, Generalized Additive Models, and Gaussian Process (kriging) models as well as modifications to them that mitigate untenable assumptions.
3. We make our results interactive through the Google Earth interface.

4 Methods

We compare the performance of several non-parametric geostatistical models for uranium deposits. Two classes of models, K Nearest Neighbors and Local Regression, make only weak assumptions about the distribution of uranium. Two other classes of models, Generalized Additive Models and kriging-based models, make flexible assumptions about this distribution. In the latter case we employ disjunctive kriging, adjusted indicator kriging, and the use of an empirical copula to obtain estimates of uranium that rely minimally on these assumptions. In each case, we tune model parameters using 15-fold cross validation, a procedure which is only feasible through the use of parallel computing techniques.

4.1 Cross-Validation Scheme

We use cross-validation to avoid overfitting [10]. Overfitting occurs when a model confuses random noise in the data with actual trends in the data. This makes the model less generalizable, and predictions are often poor for data not used to build the model.

- Randomly split data into training (75% with sample size n) and test (25% with sample size m) sets.
- Find tuning parameters for each interpolation method through c-fold cross-validation.
  - Randomly split training data into c equal sections of size $q = n/c$.
  - Train model on $c - 1$ sections and test on the remaining section (getting predictions $\{\hat{y}_1, \hat{y}_2, ..., \hat{y}_q\}$).
  - Repeat c times so each section is a test section ($\hat{y}_{i,j}$ represents the ith prediction on the jth fold).
  - Get c distributions of residuals ($\{y_1 - \hat{y}_{1,j}, y_2 - \hat{y}_{2,j}, ..., y_q - \hat{y}_{q,j}\}$) for each set of parameters.

4.2 How To Determine Optimality

For each of the four methods:

- Determine reasonable ranges of values for parameters.
- Create all possible combinations of these parameters (number of combinations is P).
- Test each set of parameters on the training set using c-fold validation.
  - Have c sets of residuals from c-folds.
  - Calculate the root mean squared error for each fold (r_{c,p_i} where c is the fold and p is the vector of parameter values) [22].
  - Take the median of those c values.
- Choose the combination p_i of parameters that yields the smallest median of root mean squared errors (RMSE) from the c-folds.

$$min\{\text{median}(r_{1,p_1}, ..., r_{c,p_1}), ..., \text{median}(r_{1,p_P}, ..., r_{c,p_P})\}$$

Justification: Why Define Smallest Median RMSE of Folds as Best?

Another possible definition of ``optimal" is the minimum root mean squared error for the entire training set:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$$

.

However, this pools the information from the c-folds. We want to avoid a situation where a method that performs really well on a few folds, and really poorly on a few other folds, wins over a method that does fairly well on all folds. We do not want the good performance on some folds to outweigh poor performance on other folds. This could still happen while using the median if over half of the folds do well yet the remaining folds do extremely poorly, but we at least ensure that a few extreme values of RMSE do not impact the choice of optimal parameters.

How to Choose c

Often, when one wants an uncertainty quantification, the bootstrap, either parametric or non-parametric, is used [7]. In this application we cannot use the parametric bootstrap as our methods are all non-parametric models. We also cannot use the non-parametric bootstrap where we sample with replacement from our data to create new samples. In this context it does not make sense to have duplicate measurements as having values in duplicate locations does not add any additional information to train the model on. Therefore, the only way to get an uncertainty quantification is to use samples of a smaller size than the original sample. This is our motivation for using the c-fold validation as each fold is smaller than the original sample. However, we must decide on a value for c.

If c is set to be equal to the size n of the training data set, this is known as a ``leave-one-out" procedure. The estimates yielded by the ``leave-one-out" cross validation is approximately unbiased, but it can have high variance because the n folds are so similar. This idea leads to the bias-variance tradeoff [10].

Smaller folds will introduce more bias as they are less representative of the full data set. However, by increasing the number of folds, they become more similar, leading to higher variance. Large folds contain many of the same points; the model is therefore trained on similar values. Predictions between the folds are more highly correlated than for small folds that are trained on data that overlaps less with other folds. Higher variance comes from the fact that the mean of many highly correlated values has higher variance than the mean of less correlated values. The increase in number of folds also leads to increased computational complexity and becomes less feasible on large data sets [14].

A common choice for c in practice is 10. We chose $c = 15$ so that our training data would be split evenly among the folds ($n/c \in \mathbb{N}$). Bengio et. al. show with sample sizes in the hundreds that the variance levels off as the number of folds approaches 20. With our data set of thousands we should not run into the problem of highly correlated training sets [2].

4.3 K Nearest Neighbors

K Nearest Neighbors is a straightforward learning method that does not rely on any model assumptions. To predict the value of uranium at a certain location, the K Nearest Neighbors method uses the average of the k sample points in the training set that are the closest in Euclidean distance to the point of interest [10].

Parameter to Choose

- number of neighbors k to use

Model	Parameter	Values Considered	Optimal Value
KNN	k	5-30 by 5	5

Table 1: KNN Parameter Summary

4.4 Local Regression

Local Regression allows us to maintain the use of neighbors while making smoother and more sophisticated predictions. Each candidate point x_i is predicted by fitting a low-degree polynomial to a subset of the known data. The polynomial is fit using weighted least squares, giving more weight to points near (in Euclidean distance) to x_i and less weight to points further away. The subsets used in the fitting are determined by a nearest neighbors algorithm [16].

Parameters to Choose

- α: percentage of neighbors to use
- degree: degree of the local polynomial
- weight function: assigns weights based on distance
- scale: controls the relative amounts of smoothing in each explanatory variable

Model	Parameter	Values Considered	Optimal Value
Local Regression	α	0.2-0.6 by 0.05	0.2
Local Regression	degree for longitude	1, 2	2
Local Regression	degree for latitude	1, 2	2
Local Regression	weight function	tricube, rectangular, triweight, triangular, exponential, bisquare, Gaussian	tricube
Local Regression	scale for longitude	1, 0.124 (standard deviation)	1
Local Regression	scale for latitude	1, 0.061 (standard deviation)	1

Table 2: Local Regression Parameter Summary

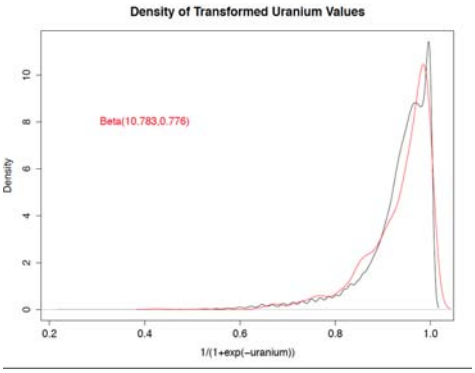
4.5 Generalized Additive Models

A generalized additive model (GAM) relaxes the assumption of normality in the response variable, allows for the response variable to follow distributions other than the normal distribution, and is specified as a sum of smooth functions of predictor variables. This makes the model more flexible than a generalized linear model which would only include one smooth function [23].

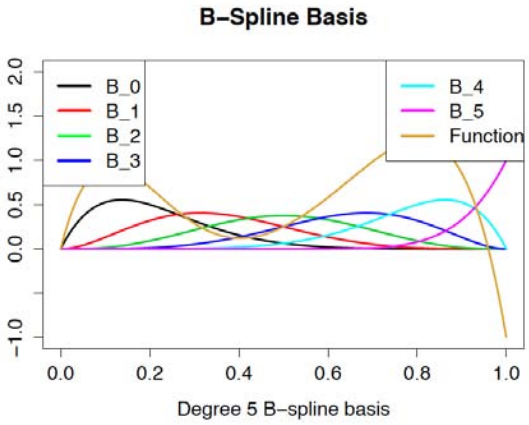
Parameters to Choose

- family: distribution that the response variable follows
- link function: varies linearly with the predicted values instead of requiring the response variable itself to vary linearly, maps the mean of the response to the linear predictors in the model
- type of smooth: spline based, tensor product based
- basis for smooth: type of splines doing the actual smoothing
- penalization: smoothness parameter λ that penalizes ``wiggliness" in the least squares fitting criteria

For the exploration of GAMs we map the uranium values to $[0, 1]$ so that we can use the Beta distribution and the logit link function. When we perform this transformation, the distribution of uranium follows a Beta distribution well (as seen in the left of Figure 2). We find this to be true in many geological contexts as the Beta is a flexible distribution. We use the coordinates as covariates and can smooth them in different ways (separately or together) based on different types of smoothers. Each smooth function is built up from splines. Splines allow us to build up a complicated curve by optimally weighting a series of basis curves. An example is shown on the right in Figure 2.



(a) Best Fit Beta Distribution of Transformed Uranium



(b) B Spline Basis, Degree 5

Figure 2: GAM Components

4.6 Gaussian Processes: Kriging

Gaussian Processes allow us to think of the coordinates as more than just covariates and to take advantage of their spatial nature. The Gaussian Process model is defined by a covariance matrix which takes into account relationships between neighboring sample points.

The value of uranium at a location s can be modeled as a random field $Z(s)$ with a trend component, $m(s)$, and a residual component, $R(s) = Z(s) - m(s)$. Kriging estimates the residual at s as a weighted sum of residuals at surrounding data points. Kriging weights are derived from a covariance function. Similar to local regression when we had an α parameter that determined what proportion of the data was to be considered neighbors, in kriging s_α represent points in the known data that lie within a neighborhood used to estimate the value at an unknown point s. We can think of this neighborhood as containing points within a distance of s where spatial correlation is still noticeable [3]. Traditional kriging is fundamentally a matrix inversion problem; to find the optimal weights, the covariance matrix is inverted. With a large data set, this can get computationally expensive as inverting an $n \times n$ matrix is $O(n^3)$ where n is the number of data points in our sample [12]. There are a few different strategies for dealing with this computational hurdle.

4.6.1 Block Kriging

Block kriging estimates the average value in given blocks rather than at a set of given points. Regular, rectangular blocks and irregular polygons such as U.S. counties can be used. The process of finding covariances between each s and every point in each block has an initial increased computation, but overall, the computation time is reduced as one must only solve one system of linear equations rather than one for each block [13].

Parameter to Choose

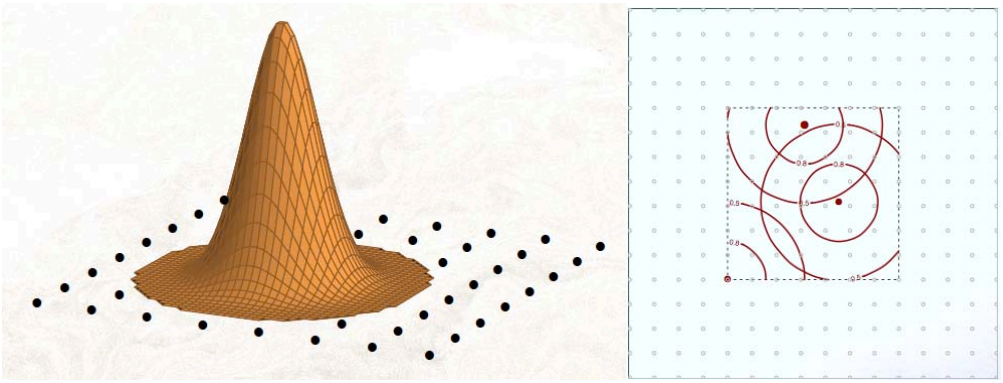
- covariance structure

Model	Parameter	Values	Optimal Value
GAM TE1	basis for (longitude, latitude)	cubic regression spline, p-spline, thin plate regression spline	thin plate regression spline
GAM TE1	is (longitude, latitude) basis penalized	yes, no	no
GAM S2	basis for longitude	cubic regression spline, p-spline, thin plate regression spline	p-spline
GAM S2	is longitude basis penalized	yes, no	no
GAM S2	basis for latitude	cubic regression spline, p-spline, thin plate regression spline	cubic regression spline
GAM S2	is latitude basis penalized	yes, no	no
GAM TE2	basis for longitude	cubic regression spline, p-spline, thin plate regression spline	thin plate regression spline
GAM TE2	is longitude basis penalized	yes, no	no
GAM TE2	basis for latitude	cubic regression spline, p-spline, thin plate regression spline	thin plate regression spline
GAM TE2	is latitude basis penalized	yes, no	no
GAM TE3	basis for longitude	cubic regression spline, p-spline, thin plate regression spline	cubic regression spline
GAM TE3	is longitude basis penalized	yes, no	no
GAM TE3	basis for latitude	cubic regression spline, p-spline, thin plate regression spline	thin plate regression spline
GAM TE3	is latitude basis penalized	yes, no	no
GAM TE3	basis for (longitude, latitude)	cubic regression spline, p-spline, thin plate regression spline	thin plate regression spline
GAM TE3	is (longitude, latitude) basis penalized	yes, no	no

Table 3: GAM Parameter Summary

4.6.2 Lattice Krig

This method is a combination of multi-resolution analysis and fixed-rank kriging, which finds the kriging weights by working with a fixed number m of basis functions where m is much less than the number of data points n. The method takes advantage of sparse matrices to decrease the runtime of the calculations to find the kriging weights [17]. Like the GAMs, we optimally weight a relatively small number of basis functions to build up the spatial random field, creating the covariance matrix to maintain sparsity (see Figure 3).



(a) Example Basis Function (b) Building Up Random Field

Figure 3: Lattice Krig Components
*Figure Credit: Doug Nychka NCAR

Parameters to Choose:
Note: There are four different ways to fully specify a Lattice Krig model using the LatticeKrig R package [18].

- $\alpha[j]$: weights at each grid level, controls the spatial dependence and must be greater than or equal to 4
- ν: constrains α to make the weights proportional to $exp(-2 \cdot j \cdot \nu)$ where ν controls decay of the α weights, theory suggests that ν is analogous to the smoothness parameter from the Matérn family [17]
- λ: the ratio of the nugget variance to the parameter controlling the marginal variance of the process
- range: can constrain the covariance structure to have a certain range where spatial correlation does not exist past it

Model	Parameter	Values Considered	Optimal Value
Block Krig	Covariance Structure	Matern, Exponential, Spherical, Gaussian	Matern
LK1	α	4.01, 4.1, 4.5, 5	5
LK1	ν	0.05, 0.5, 1, 3, 5	0.05
LK1	λ	0.0001, 0.001, 0.01, 1	1
LK2	ν	0.05, 0.5, 1, 3, 5	0.05
LK2	λ	0.0001, 0.001, 0.01, 1	0.0001
LK2	range	0.1, 0.2, 0.4	0.1
LK3	α	4.01, 4.1, 4.5, 5	5
LK3	ν	0.05, 0.5, 1, 3, 5	0.05
LK4	ν	0.05, 0.5, 1, 3, 5	0.05
LK4	range	0.1, 0.2, 0.4	0.1

Table 4: Krig Parameter Summary

4.7 Commonalities in Optimal Parameters

The K Nearest Neighbors and Local Regression methods have a parameter that captures a sense of neighborhood. For both methods we find the optimal parameters to be those that correspond to the smallest neighborhood. This gives us the least smooth interpolation locally.

For the Generalized Additive Models, in each specification we find the optimal bases to be un-penalized, allowing for local ``wiggliness." Although this method does not explicitly have a parameter representing the size of a neighborhood, the fact that we do not penalize curvature locally, gives us the same result: we value the ability to model local variation.

For the kriging methods, we treat the coordinates as truly spatial instead of as covariates. The large α value and small range value that we get to be optimal translates to a rough spatial field and the least smooth covariance structure used to make predictions. Again, we see that our optimal parameters allow for local variation while preserving a smooth interpolation overall.

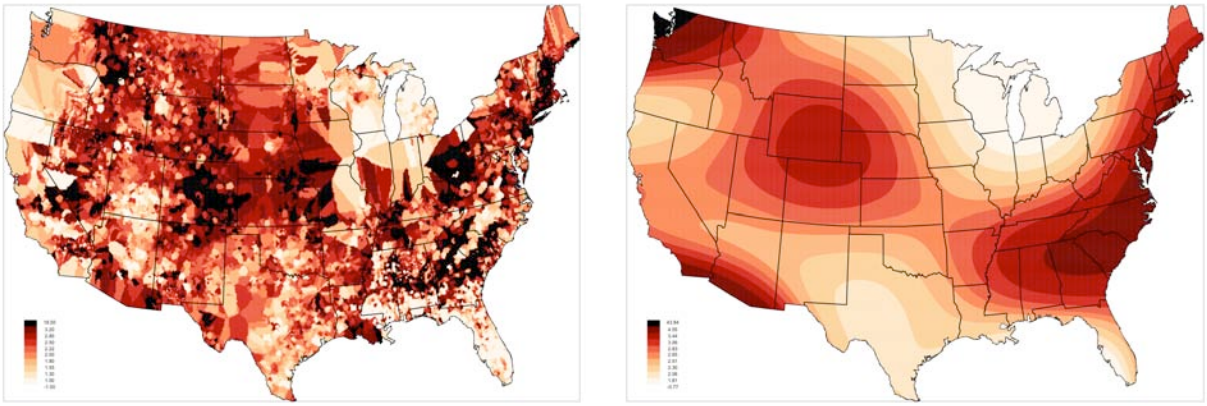
It makes sense that we want our interpolation to be flexible locally as the true distribution of uranium is likely not completely smooth. However, we want to be careful when fitting to the details locally. If we allow the model to be too flexible locally, it could be fitting to noise in our particular data set instead of to the true variability in uranium. This is why we are especially careful about overfitting. Along with the cross validation process, we only allow the number of neighbors in the KNN method to go as low as five and the percentage of neighbors in the Local Regression to go as low as twenty percent. We can get marginal benefits in the RMSE by going lower than these values, but we build in this extra guard against overfitting.

5 Results

After using a training set and 15-fold cross validation to optimize the parameters in each modeling method we predict using the best set of parameters for each method against one another on a test set to find the ``best of the best".

We show a subset of the interpolations to illustrate the visual differences.

In Figure 4 we can see on the left that the interpolation given by the K Nearest Neighbor method is not smooth at all while on the right, the interpolation given by a Generalized Additive Model is too smooth.



(a) KNN: Smallest Median Root Mean Square Error Interpolation

(b) GAM TE3: Smallest Median Root Mean Square Error Interpolation

Figure 4: Comparing Interpolations

In Figure 5 we can see that the interpolation given by the Block Krig does not predict at a fine enough granularity to be able to predict at *any* point within the continental U.S.

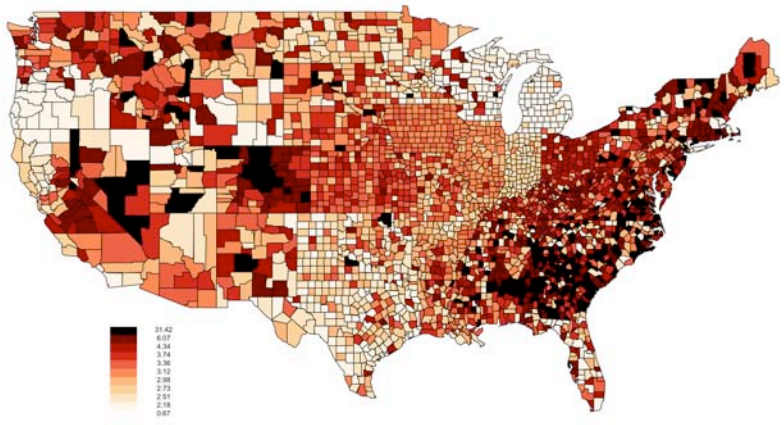


Figure 5: Block Krig: Smallest Root Mean Square Error Interpolation
*not projected

Figure 6 shows the interpolation for the method that we deem ``best," the Lattice Krig. This method yields a smooth interpolation that has the most detail.

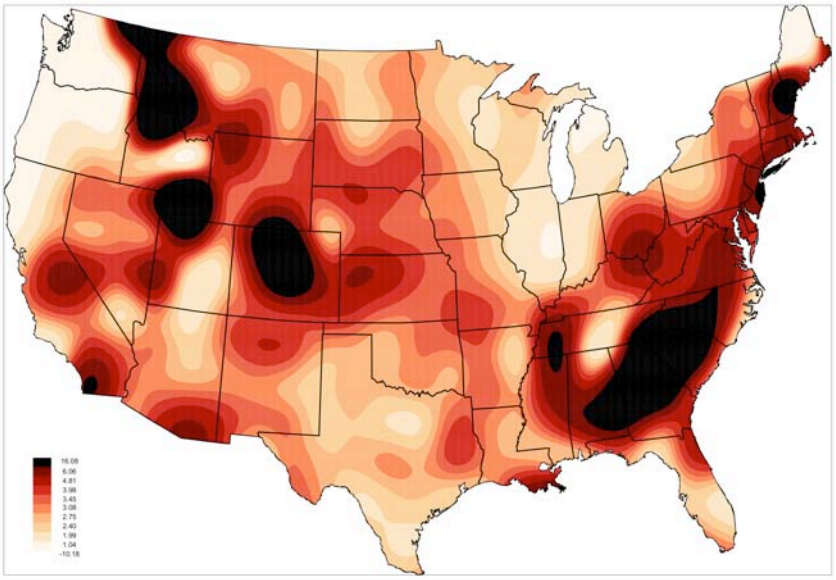
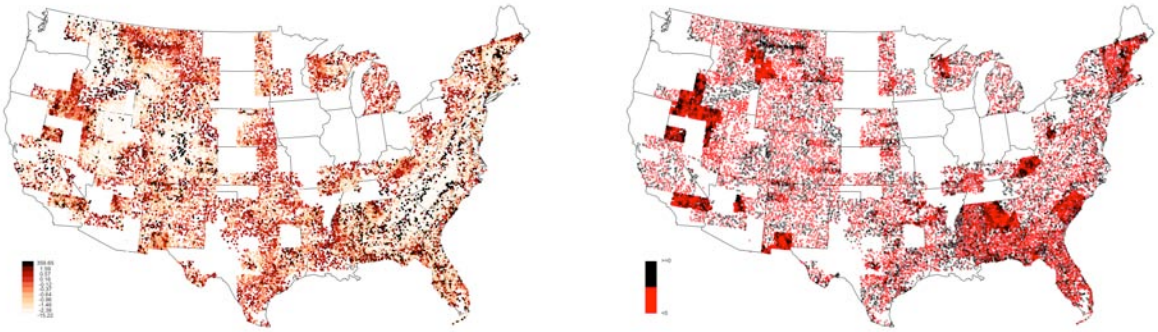


Figure 6: LK3: Smallest Median Root Mean Square Error Interpolation

The residuals for this method can be seen in Figure 7; here we can see that the method overestimates much more than it underestimates. This pattern in the residuals can be removed by first taking the logarithmic transformation of the uranium values and then performing the Lattice Krig method. This comes at a cost of a slightly higher RMSE value, but is a reasonable fix if we are worried about unbalanced residuals.



(a) LK3: Residuals $(y - \hat{y})$ on Training

(b) LK3: Over or Under Estimate $(y - \hat{y})$ on Training

Figure 7: LK3 Training Residuals

In Table 5 we summarize our results. We find that the Lattice Krig methods produce the inter-

polations with the most detail while remaining smooth. To get more balanced residuals and the ability to predict larger values of uranium, using the logarithmic transformation prior to performing the kriging method is recommended. This transformation comes at the price of a slightly higher RMSE as by increasing the values of uranium that are possible to predict we expose ourselves to greater risks if we wrongly overestimate.

We can see that our RMSE values are similar yet our interpolations from each method look visually distinct. We put our results in context by creating Google Earth layers with each of our interpolations. Figure 8 shows an example. Using Google Earth, we can zoom in on places of interest and see the geological settings there. We can also put two interpolations on top of one another and use transparency options to better compare and contrast them. By using Google Earth to create a more interactive exploration of our results, we make our work accessible to interested parties who may not be statisticians, including geologists and policymakers.

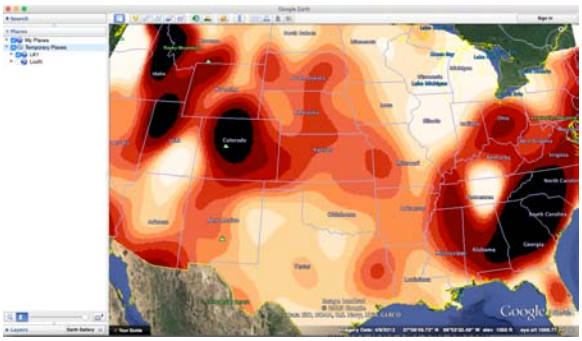


Figure 8: Sample Google Earth Layer

6 Discussion

The overall winner on the RMSE criteria is the un-smoothed Block Krig, but due to the lack of a fine granularity for predictions, we recommend using the Lattice Krig for uranium interpolations. However all of these methods yield a RMSE between 5.93 and 6.94 ppm on the test set. The RMSE is in the units of the response (ppm of uranium). Since most of the uranium values are less than 5 ppm, these results are a bit unsatisfying. Although the residuals do not seem to be spatially correlated, we have large residuals in areas of the United States that typically have more uranium, i.e., we are not predicting high enough values of uranium to match more extreme values. Our models are insufficiently flexible to predict these rare and large values of uranium.

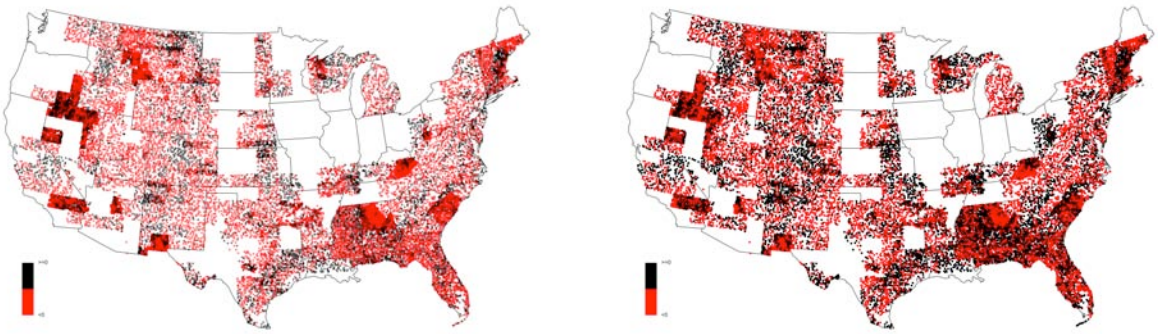
6.1 Assumptions

In trying to improve the RMSE, we addressed untenable normality assumptions through various transformations and more general methods that do not require normality of the residuals. These methods did not drastically improve our RMSE on the test set, but they provided further insight into where we could further improve our methods. We were most worried about lack of normality, but in reality, our main problems do not come from this.

Interestingly the non-symmetric and non-normal distribution of uranium is not the main culprit as adjustments to our models to account for the non-normality do not give us markedly better results. Our RMSE is inflated due to test samples that have extreme (large) values of uranium as we are not predicting large enough values.

6.1.1 Local Regression Assumptions

Local Regression assumes that the response distribution is symmetric. We can see evidence that this assumption is not met as the residuals are not balanced. In Figure 9 on the left we can see that we overestimate much more frequently than we underestimate using Local Regression. We can easily perform a logarithmic transformation on the uranium values before performing Local Regression; on the right of Figure 9 we can see that the residuals become more balanced. This same method balances residuals for the Lattice Krig models as well.



(a) Local Regression: Over or Under Estimation on Training

(b) Local Regression Log Transform: Over or Under Estimation on Training

Figure 9: Comparing Residuals after Transformations

6.1.2 Generalized Additive Model Assumptions

Generalized Additive Models assume that there are no interaction effects between the covariates. We are only using longitude and latitude as covariates, yet it is very likely that there are interaction effects between the two. There is no easy fix for this. A variation of GAMs with spatial awareness may help model these interaction effects. Although there are many frameworks for incorporating spatial information into GAMs, they are not particularly effective in predicting uranium values. We see the same problem arise that we did in the plain GAMs: the methods are not able to predict extremely large values of uranium.

This property comes from our original mapping. We first map the data to $[0, 1]$ using an inverse-logit transformation in order to make the data fit a Beta distribution. This takes extremely positive values of uranium and maps them to values close to 1. We predict on this scale and then transform back to ppm units using the logit function. To get a large prediction in ppm, we need a transformed prediction extremely close to 1.

$$1 = \log \frac{x}{1-x} \text{ has no solutions}$$

In fact, there is a singularity at 1, which explains the methods' problems with predicting large values of uranium. The values of uranium have too wide a range to be effectively modeled by

Generalized Additive Models with the Beta family and logit function.

6.1.3 Kriging Assumptions

Lattice Krig assumes that the distribution of the response variable is normal. We can use a logarithmic transform on the uranium values, but the distribution is still not normal. There are also other methods of kriging that do not rely on normality assumptions including adjusted indicator kriging (where we bin the data and predict the probability that at a given point the true uranium value lies within that bin) and disjunctive kriging (which relies on applying a nonlinear transform to the data to achieve normality). We can also replace the Gaussian Random Field in the kriging with an empirical copula that is data driven and does not make assumptions about the data distribution. We found these methods to be ultimately ineffective, but we outline them for reference.

Adjusted Indicator Kriging

The adjusted indicator kriging method is much more computationally intensive as it requires a kriging step for each bin. In our case we use ten quantiles as the bins. Instead of attempting to optimize the parameters for this method with 15-folds (150 kriging steps per parameter trial), we use the optimal parameters found for Lattice Krig.

First we calculate the probabilities that each location has a uranium value within each quantile. 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 full 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 adjustments that need to be made to make the estimates fit within the rules of probability (see Figure 10).

Indicator	order violation	correction	`set order`
independent	$\hat{p}_i < 0$	$\tilde{p}_i = 0$	`1-4`
independent	$\hat{p}_i > 1$	$\tilde{p}_i = 1$	`1-4`
categorical, open	$\sum_{i=1}^{n} \hat{p}_i > 1$	$\tilde{p}_i = \hat{p}_i / \sum_{i=1}^{n} \hat{p}_i$	2
categorical, closed	$\sum_{i=1}^{n} \hat{p}_i \neq 1$	$\tilde{p}_i = \hat{p}_i / \sum_{i=1}^{n} \hat{p}_i$	3
cumulative	$\hat{p}_i < \hat{p}_{i-1}$	$\tilde{p}_i - \tilde{p}_{i-1} = 0$	4

Figure 10: Order Relation Corrections [19]

Disjunctive Kriging

Z-Score Krig

This method forces the uranium data to follow a normal curve exactly by performing Lattice Krig on the z-scores:

$$z = \frac{y - m}{s}$$

where y is a vector containing the sample values of uranium, m is the mean of the uranium values in the training set, and s is the standard deviation of the uranium values in the training set.

We can then un-transform to assess performance.

N-Score Krig

This method 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 perform Lattice Krig on the assigned values and un-transform to assess performance. 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.

Empirical Copula

The kriging method assumes that the response value of the observations come from a Gaussian random field. A random field G is a Gaussian random field if $G(s_1), ..., G(s_n)$ is multivariate normal for any s_i [4].

An alternative is to use a copula which generalizes Gaussian random fields. Instead of assuming the form of a multivariate distribution, copulae provide a way to combine marginals —which are more easily approximated from data —into a joint distribution.

We can create an empirical copula to make uranium predictions [1]. This is computationally intensive, but it is reasonable to use on a subset of our data. We use the data from Colorado as an example.

1. Find the empirical cumulative distribution of uranium in the training set $F(y)$.
2. Pick a distance h and find locations in the training set that are separated by h. The value of h and the width of the band for approximation needs to be specified.

   $$\text{Dist}(s_1, s_2) \approx h$$

   where the value at s_1 is y_1 and the value at s_2 is y_2.
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:

   $$(F(y_1), F(y_2))$$
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 optimized. For each neighbor n_i we draw a random value from the copula conditioned on $F(v_i)$ where v_i is the value of uranium at n_i. We can choose to increase the number of random values drawn and aggregate them in some way.
6. We must also choose how to aggregate the values from each neighbor. A first step is to use the mean across the neighbors.

One possible issue with this method is that we over-smooth by aggregating across neighbors, causing a loss of detail and conservative estimates of uranium that do not reach extreme enough values.

There are many choices in this method that can be explored more fully.

- h: A small h yields a more detailed copula while a large h yields a smoother copula.
- bandwidth for h: A smaller bandwidth will decrease the number of pairs included for comparison to unknown locations.
- number of neighbors: This could draw from our work on the KNN method where we saw that the smallest number of neighbors in our reasonable parameter space was optimal to avoid over-smoothing.
- the number of values to draw from conditional distribution: We want to make sure that extreme prediction values are possible.
- the aggregation method: Since aggregating necessarily smooths, we need a way to avoid over-smoothing.

We use the same method and criteria for optimizing these parameters over the Colorado data set (for computational reasons). Here $c = 3$ so there are many fewer folds as we have a smaller sample size.

Scaling Up: Using a Copula on the United States Data

For each coordinate in the test set, we make and simulate from an empirical copula using the above parameters and the training set data. Creating and simulating from this copula will be computationally feasible as it will only use a relatively small subset of the training data at each prediction point. Think of this as a local copula. In the case that the test point is far away from other training data and there are not enough training points within h's bandwidth to build an empirical copula we can expand the bandwidth and repeat until a successful empirical copula is built. A complete interpolation across the United States using this method is computationally intensive.

6.2 Checking for Consistency: Testing a Subset of the U.S.

We go through the same process to find the ``best" set of parameters for each model type for the data found in Colorado (chosen because we have a better idea of what the uranium distribution is here from the National Institute of Standards and Technology). We find that our methods for the continental U.S. extend to analysis at the individual state level. Many of the ``best" parameters for each method chosen by Colorado and the continental U.S. coincide. Our optimal number of neighbors for the KNN method is the same. The only differences in optimal parameters for local regression between the Colorado and the U.S. cases are in the degree for longitude and scale for latitude. This could be an artifact of the change in scale and projection. The optimal parameters for the generalized additive models remain unchanged. The only difference in optimal parameters for the Lattice Krig models is in λ. In the U.S. case a large value of λ was optimal, but for Colorado a very small value is best. This makes sense as we expect the marginal variance of the Gaussian Process to be smaller when covering a smaller area. Note that we changed the possible values for the range to { 1, 2, 3 } for Colorado as range is dependent on the size and units of the coordinates that cover the area of interest. These parameters match those of the United States (after the adjustment for the scale of the range). This is reassuring as we would like the methods chosen for a larger data set to be applicable to subsets of that data set.

Method	RMSE (ppm) on Test Set	Maximum Prediction Value (ppm)	Qualitative Assessment
Block Krig	5.930	31.42	does not predict at as fine of a level
LK3	5.967	16.08	produces most detailed yet smooth interpolation (as opposed to that of the KNN)
Smooth Block Krig	6.021	31.42	looks visually more reasonable than the original block krig
Local Regression	6.033	29.13	has potential to be marginally better than a Lattice Krig method by using $\alpha < 0.2$ if you are willing to give up some generalizability
LK3 Log	6.035	$>$ 1000	more balanced residuals at the expense of predicting some unrealistically large values
Local Regression Log	6.118	237	more balanced residuals at the expense of predicting some unrealistically large values
KNN	6.182	18.50	in sparse regions this approach relies too heavily on samples that are ``closest" but may not be representative
GAM TE1	6.268	34.47	smallest standard errors
Z-Score Krig	6.240	16.18	as lack of normality is not our main concern in this scenario we do not need to sacrifice our RMSE to use this method
Local Copula	6.280	NA	extremely computationally intensive to calculate a full interpolation across the U.S.
N-Score Krig	6.930	365.70	high predictions do not in general correspond with a true extreme values

Table 5: Final Results Summary

6.3 Extension: Looking at Other Substances

We went through an extensive cross validation process to avoid overfitting. We want to ensure that our work can be extended to other data sets and be applicable to other problems of this type. We re-optimize parameters for five different substances, aluminum, chromium, gallium, lithium and magnesium, whose distributions are shown in Figure 11.

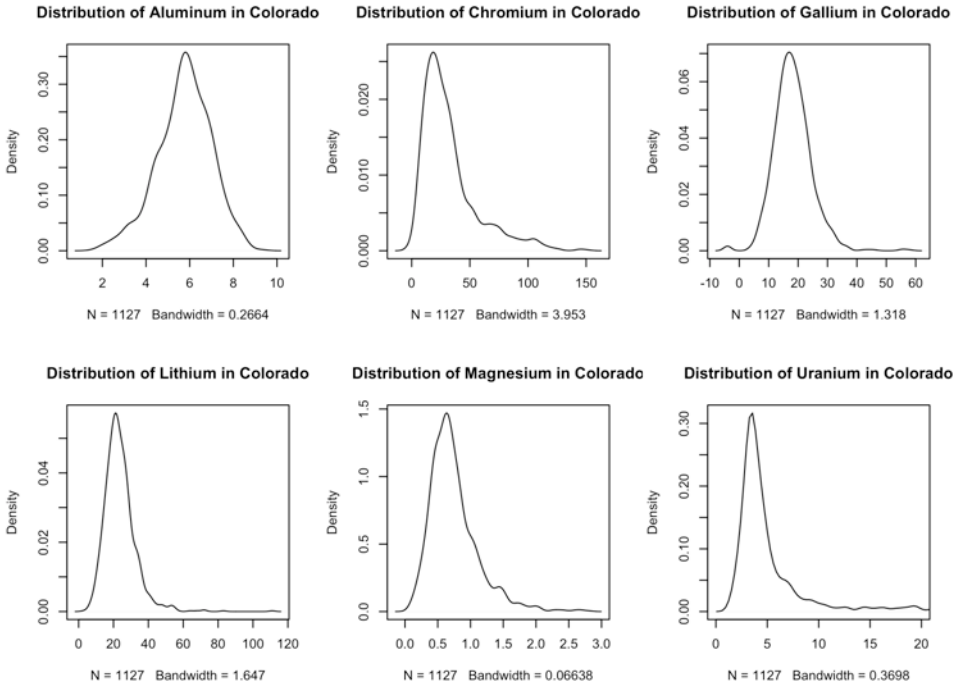
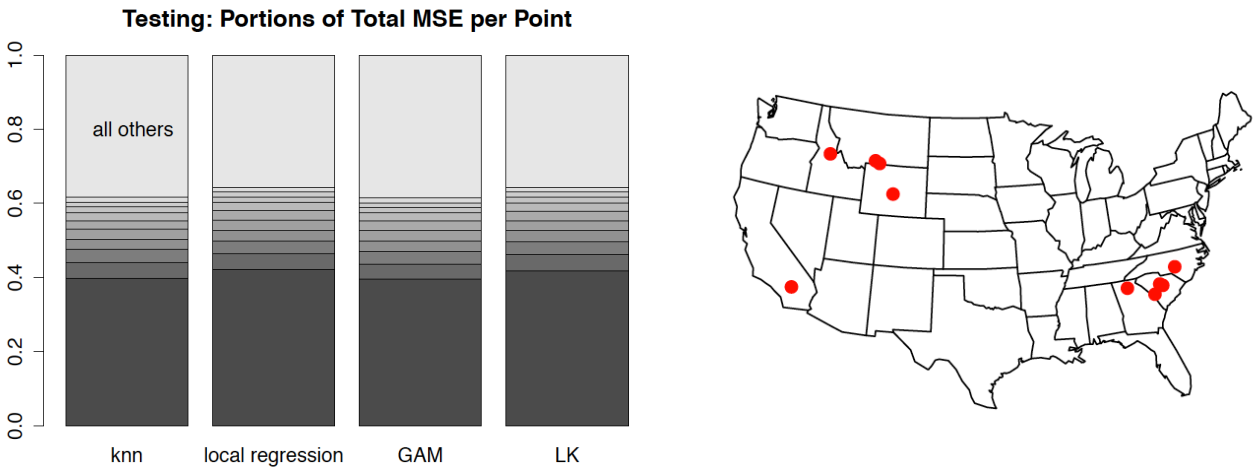


Figure 11: Distribution of Other Substances in Colorado

We find that the optimal parameters for uranium are similar to those found to be optimal for the prediction of the other substances. Using the methods that we determined to be optimal, others can study the distributions of other substances across the continental U.S.

6.4 Future Work: Influential Points

Upon closer investigation of our final RMSE we determined that about 60% of our RMSE comes from ten points in our test set. In Figure 12 the bottom ten blocks show the proportion of the test set mean squared error that the most influential points contribute. The top block represents the combined influence of the rest of the test points. For example, the largest block corresponds to a location in Bear Valley, Idaho that used to be a commercial uranium mine in the late 50s. If we omit the Idaho sample, our RMSE on the test set improves by about 1.4 ppm. Note that even by omitting this point, we still do not get a RMSE that is less than the median amount of uranium in the U.S. (3 ppm). These influential points are outliers in the sense that they are more extreme than the rest of the data but not in the sense that they are somehow ``wrong". Since we are interested in knowing where uranium is, and are especially interested in areas where it is plentiful, removing them from our sample would be counterproductive. As each method fails to predict extremely large values of uranium, future work will include the exploration of extreme value methods to tackle these few, but influential samples.



(a) Influence Per Point By Method

(b) Influential Test Points

Figure 12: Influential Points Location and Impact on Testing Mean Squared Error

References

[1] Andras Bardossy. Copula-based geostatistical models for groundwater quality parameters. *Water Resources Research*, 42(11), 2006.

[2] Yoshua Bengio and Yves Grandvalet. No unbiased estimator of the variance of k-fold cross-validation. *The Journal of Machine Learning Research*, 5:1089--1105, 2004.

[3] Geoff Bohling. Kriging . `http://people.ku.edu/~gbohling/cpe940/Kriging.pdf`, October 2005. Kansas Geological Survey. Last Accessed: January 30, 2015.

[4] Moo K. Chung. Gaussian Random Fields. `http://www.stat.wisc.edu/~mchung/teaching/stat992/ima01.pdf`, December 2003. University of Wisconsin, Last Accessed: January 30, 2015.

[5] Noel Cressie. The origins of kriging. *Mathematical Geology*, 22(3):239--252, 1990.

[6] M.W. Drew. US Uranium Deposits: A Geostatistical Model. *Resources Policy*, 3(1):60 -- 70, 1977.

[7] Bradley Efron and Robert J. Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

[8] Michael R. Evans, Dev Oliver, Kwang Soo Yang, Xun Zhou, and Shashi Shekhar. Enabling Spatial Big Data via CyberGIS: Challenges and Opportunities. *CyberGIS: Fostering a New Wave of Geospatial Innovation and Discovery. Springer Book*, 2013.

[9] Pierre Goovaerts. *Geostatistics for natural resources evaluation*. Oxford University Press, 1997.

[10] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer New York, 2nd edition, 2009.

[11] Octavio R. Hinojosa de la Garza, Maria Elena Montero Cabrera, Luz H. Sanin, Manuel Reyes Cortes, and Enrique Martinez Meyer. Spatial analysis techniques applied to uranium prospecting in Chihuahua State, Mexico. *AIP Conference Proceedings*, 1607:116--122, 2014.

[12] Kh.D. Ikramov. Inversion of a Matrix. `http://www.encyclopediaofmath.org/index.php/Inversion_of_a_matrix`, February 2007. Last Accessed: January 30, 2015.

[13] Edward H. Isaaks and R. Mohan Srivastava. *An Introduction to Applied Geostatistics*. Oxford University Press, 1989.

[14] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning*. Springer New York, 2013.

[15] Victor E. Kane, Connie L. Begovich, Todd R. Butz, and Donald E. Myers. Interpretation of regional geochemistry using optimal interpolation parameters. *Computers & Geosciences*, 8(2):117 -- 135, 1982.

[16] Clive Loader. *Local Regression and Likelihood*. Springer-Verlag, 1999.

[17] Doug Nychka, Soutir Bandyopadhyay, Dorit Hammering, Finn Lindgren, and Stephen Sain. A multi-resolution Gaussian process model for the analysis of large spatial data sets. *National Center for Atmospheric Research*, 2013. `http://nldr.library.ucar.edu/repository/assets/technotes/TECH-NOTE-000-000-000-875.pdf`, Last Accessed: April 10, 2015.

[18] Doug Nychka, Dorit Hammerling, Stephan Sain, and Nathan Lenssen. Lattice Krig: Multiresolution Kriging based on Markov random fields. http://CRAN.R-project.org/package=LatticeKrig. R package version 3.4, 2014.

[19] Edzer J. Pebesma. gstat User's Manual. http://www.gstat.org/gstat.pdf, April 2014. Utrecht University, Last Accessed: April 10, 2015.

[20] Peter Schweitzer. History of the National Geochemical Survey: Background: National Geochemical Surveys. http://mrdata.usgs.gov/metadata/nurehssr.faq.html#how.2, December 2014. U.S. Geological Survey, Last Accessed: April 10, 2015.

[21] Tang Shenghuan, Xue Yuxuan, and Meng Jinqing. Application of the geostatistical analyses to uranium geology. *Geological Data Integration Techniques*, page 219, 1988.

[22] Cort J. Willmott. On the validation of models. *Physical Geography*, 2(2):184--194, 1981.

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

[24] Chunfa Wu, Jiaping Wu, Yongming Luo, Haibo Zhang, Ying Teng, and Stephen D. DeGloria. Spatial interpolation of severely skewed data with several peak values by the approach integrating kriging and triangular irregular network interpolation. *Environmental Earth Sciences*, 63(5):1093--1103, 2011.