Geostatistical Analysis of Synthetic Two-Dimensional Geospatial Point Surfaces Generated in GeoPandas

Written By: Juan C. Reyes

2022-02-10

GDAA2040 - Spatial Analysis and Geostatistics, Centre of Geographic Sciences, NSCC.

Instructor: Dr. Steven Edwards

Centre of Geographic Sciences
COGS | nscc

Table of Contents

1 Geostatistical Analysis in ArcMap and ArcGIS Pro

1.1 Geostatistics

Geostatistics is an useful methodology for applying statistical algorithms to interpolate spatial information. This is a very useful workflow for many disciplines such as meteorology, soil science, geology, oceanography, forestry, and many more[1]. The principle is to utilize spatial data to predict a target variable at other spatial locations which were not measured in the field. The analysis can be further extended to the prediction of spatiotemporal phenomena. The functionality of geostatistics is built with some underlying assumptions, the first, and arguably most important is Tobler's first law of geography introduced in 1969. It states that "everything is related to everything else, but nearby things are more related to distant things."[2]. This concept gives way to the notion of a spatially regionalized variable. To make best use of geostatistical tools, the phenomena and data we study must lie in between 'deterministic' and 'random'. For instance, the distribution of ocean salinity, air/water temperature, and soil concentrations are examples of variables which are regionally similar, but globally appear chaotic or stochastic. It is neither practical or possible to obtain a complete study of a geographic phenomenon. However, for geostatistical algorithms to yield confident reports, it is vital for geostatistical sampling to create a fair representation of the phenomenon of study[1]. The accuracy and method of data acquisition must be thorough and repeatable. Intensive sampling is useful at creating an accurate representation of the phenomena, but can be costly and yield diminishing returns. On the other hand, sparse data gathering may be cheap but lead to ineffective data. To ensure a high level of confidence in our results, there should be a optimal sampling approach to measure the regionalized variable[1].

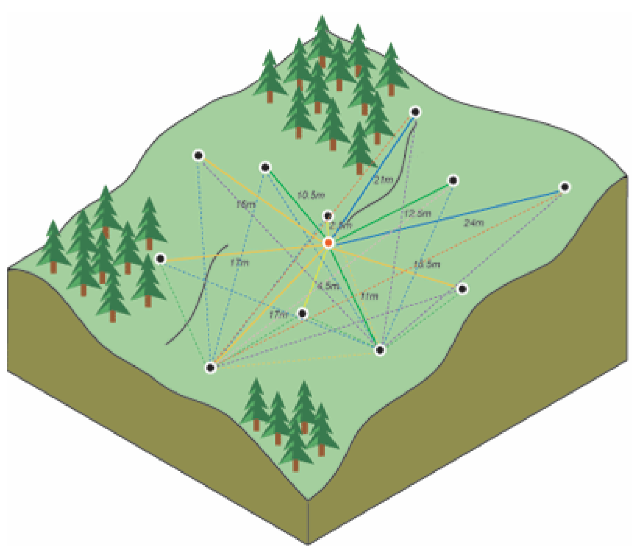


Figure 1: Pairing of a point (red) with other measured locations nearby. This is repeated for every other measured location. The process of fitting a spatial model is known as variography.[3]

1.2 Kriging Analysis

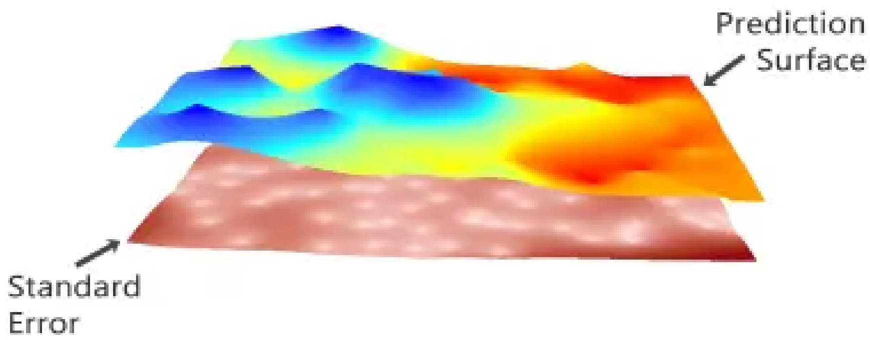


Figure 2: Prediction and Standard Error surfaces of a regionalized variable using geostatistical Kriging. [3]

Our study will take use of various geostatistical tools in ArcMap to justify a Kriging analysis of our underlying regionalized variable. Kriging is a Gaussian process regression algorithm which develops a *prediction surface* of the regionalized variable based on interpolation of a Gaussian process (stochastic process)[3]. That is, the Kriging prediction surface offers a prediction of the value of a function at a sample point s_0 by computing a weighted average of the known values of the function in the neighborhood of the point[3]. Kriging interpolation is formed accordingly to:

$$\hat{Z}(s_0) = \sum_{i=1}^{N} \lambda_i Z(s_i). \tag{1}$$

Where $Z(s_i)$ is the measured value of the regional variable at the ith location, λ_i is an unknown weight for the regionalized variable at the ith location, s_0 is the location to predict, and N is the number of measured values. It is important to note that like inverse distance weighted methods, the weight λ_i depends on the distance to the prediction location, but in addition is dependent on the spatial arrangements of the measured points. This is indicative that a spatial autocorrelation must be quantified and fitted to the model. [3] To fit a spatial autocorrelation function in Kriging refers to building a variogram function to model the statistical dependence of values amongst each other in our spatial model. The equation for a semivariogram is as follows:

$$Semivariogram = \gamma(d_h) = \frac{1}{2} avg(Z(s_i) - Z(s_j)). \tag{2}$$

Which is a least squares error function averaged over all pairs of locations s that are a distance h apart. This process is represented at a single in the image shown at the start of this section.

To compute a prediction surface, this procedure must be computed at every single coordinate value of our area of interest. Kriging requires the analyst to fit a continuous model to these points that make up the semivariogram. This will become essential for providing information regarding the spatial autocorrelation of the data[3]. With a semivariogram prepared, the analyst is ready to produce a prediction and error surfaces for their predictive analysis[4].

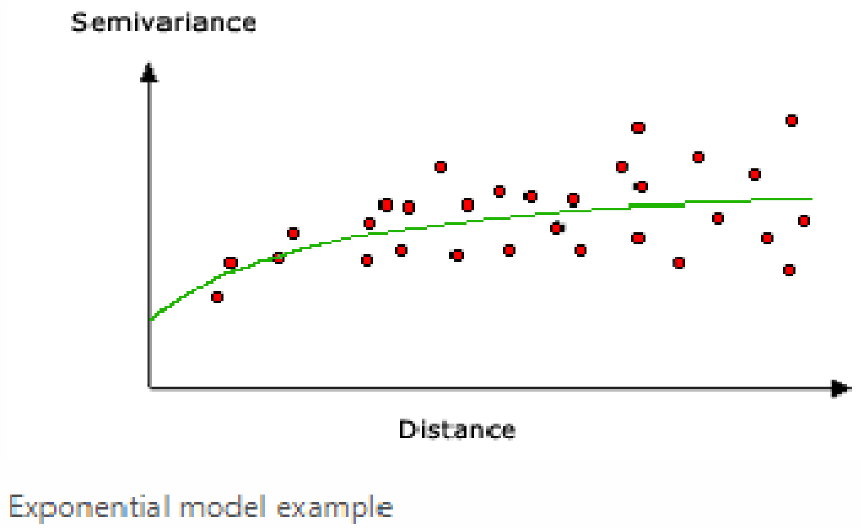


Figure 3: Fitting a continuous function to the semivariogram obtained from the data.[3]

1.3 Kriging with ArcGIS Pro

If the reader has not yet been deterred from the immense computational labour of computing a single kriging surface, they will be glad to learn that many modern software solutions have been optimally devised to compute geostatistical Kriging at solely the expense of careful data preparation and parameter fitting. Kriging results are made increasingly useful by simultaneously providing a standard error surface which is indicative of spatial locations with high and low prediction accuracy in the results. Most of the underlying mathematics and laborious and careful computation has been carefully taken care of by ESRI. The reader is left with the more complex task of how all the underlying parameters of the model play a role into fitting the right semivariogram and minimizing the prediction error. This paper serves as an introduction to the topic and will discuss some parameter fitting to produce a 'good' kriging surface.

This document varies from traditional papers of kriging analysis on real-world data in that the author has chosen to apply deterministic, two-dimensional functions to create synthetic ground truth data. This idea originated from a lack of understanding and experience of how Kriging analysis is produced. The idea is to test the predictive results of ArcGIS Pro's Geostatistical Wizard tools using well behaved functions to getting a better understanding of this subject.

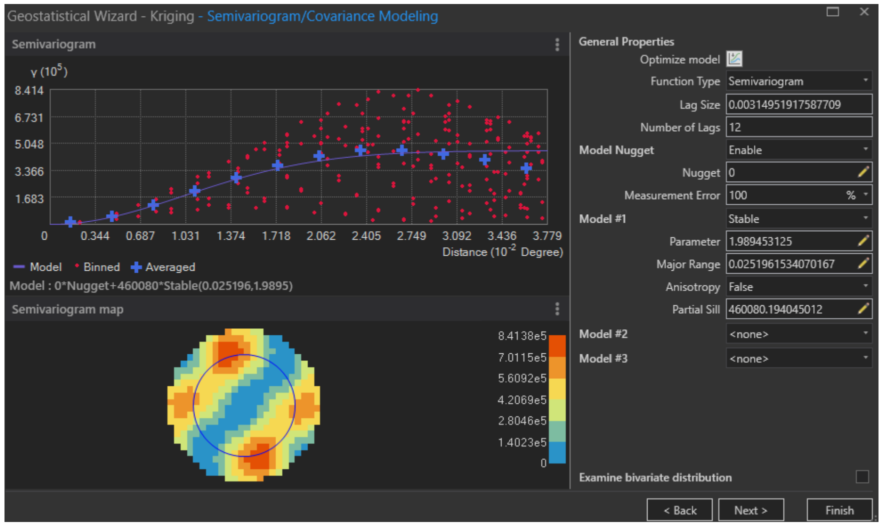


Figure 4: Fitting a continuous function to the semivariogram obtained from the data in ArcGIS Pro's Geostatistical Wizard.

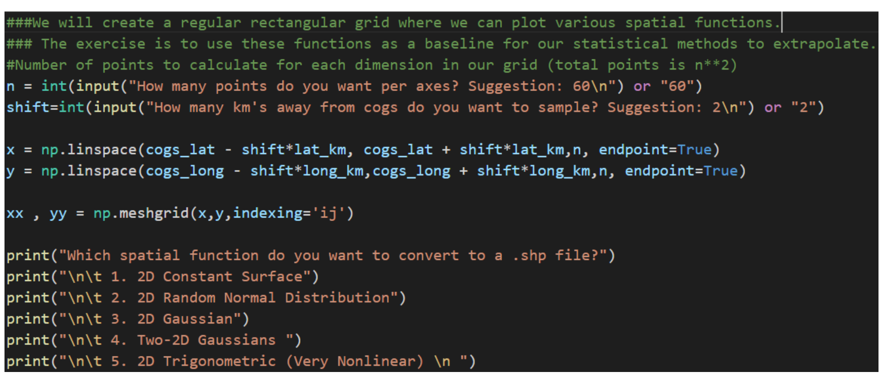
2 Data and Map Projection

2.1 Synthetic Data with Numpy and GeoPandas

The data generating script presented in this study is freely available on GitHub *here*.

You are welcome to experiment with the code directly on Google Colab *here*.

We will proceed by generating a nxn grid on Python representing latitudes/longitudes pairs with the numpy.meshgrid() function.



```
###We will create a regular rectangular grid where we can plot various spatial functions.
### The exercise is to use these functions as a baseline for our statistical methods to extrapolate.
#Number of points to calculate for each dimension in our grid (total points is n**2)
n = int(input("How many points do you want per axes? Suggestion: 60\n") or "60")
shift=int(input("How many km's away from cogs do you want to sample? Suggestion: 2\n") or "2")

x = np.linspace(cogs_lat - shift*lat_km, cogs_lat + shift*lat_km,n, endpoint=True)
y = np.linspace(cogs_long - shift*long_km,cogs_long + shift*long_km,n, endpoint=True)

xx , yy = np.meshgrid(x,y,indexing='ij')

print("Which spatial function do you want to convert to a .shp file?")
print("\n\t 1. 2D Constant Surface")
print("\n\t 2. 2D Random Normal Distribution")
print("\n\t 3. 2D Gaussian")
print("\n\t 4. Two-2D Gaussians ")
print("\n\t 5. 2D Trigonometric (Very Nonlinear) \n ")
```

Figure 5: Point Layer Created in GeoPandas Centered at COGS.

This function returns coordinate matrices from coordinate vectors which can then be fed into a function

$$z = f(x, y). \tag{3}$$

This variable will serve as our target variable and simultaneously define a continuous surface over our lattice of geographical coordinates.

```
# https://en.wikipedia.org/wiki/Gaussian_function#Two-dimensional_Gaussian_function
# This 2D gaussian function is centered about COGS.
# There are independent values for means and std devs in both the x and y axes.
# I had to play with the means (mx,my), standard deviations(sx,sy) to get them to look 'nice' on the graph.
# Feel free to tweak these parameters to adjust the shape of the distribution
# Added the shift parameter to further shift the bell shape along x, or y. This is useful for adding gaussians together.
def gaus2d(x=0, y=0, mx=cogs_long, my=cogs_lat, sx=0.009, sy=0.009,shift_x=0.01,shift_y=0.00):
    # Different amplitude parameters, wikipedia uses A=1, however the chosen amplitude is usually selected
      # for gaussian normalization (integral = 1)
    #A=1
    A = 1. / (2 * np.pi*sx * sy)
    #A1 = 1/ (np.sqrt(2*np.pi * sx**2 *sy**2))
    return A * np.exp(-((x - mx - shift_x)**2. / (2. * sx**2.) + (y - my - shift_y)**2. / (2. * sy**2.)))
```

Figure 6: Python Code for Generating a 2D Gaussian Distribution.

Once a z value has been assigned to every location on our grid, we proceed by creating a GeoPandas geodataframe with our latitude, longitude, and z values as a point object. This is then assigned the EPSG Geodetic Parameter Code '4326' to project our coordinates onto the World Geodetic System Datum WGS84 and exported into a .shp file which can be loaded directly into ArcGIS Pro.

```
#Create a geopandas geodataframe to begin transition to .shp file. Assign Point Data.
gdf = geopandas.GeoDataFrame(newdf, geometry=geopandas.points_from_xy(newdf.longitude, newdf.lat))
#Define a Geodetic Parameter. EPSG 4326 refers to WGS84 geodetic datum.
gdf.crs = datum
filename = input("Enter a name for the shapefile: \t\n")
gdf.to_file(driver = 'ESRI Shapefile', filename= filename + ".shp")
```

Figure 7: Point layer created in GeoPandas centered at COGS.

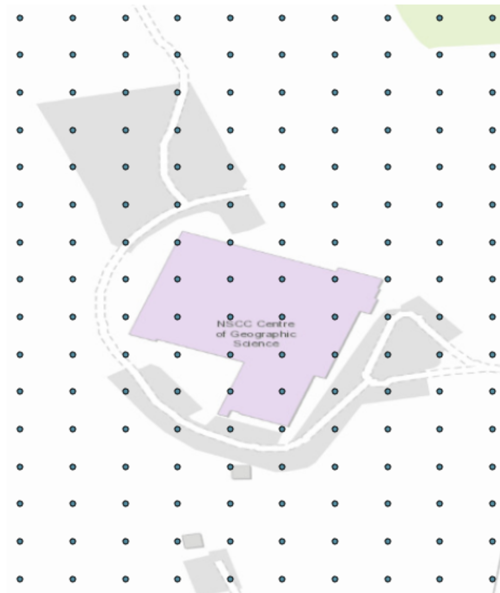


Figure 8: Point Layer Created in GeoPandas Centered and zoomed in on the Centre Of Geographic Sciences. This scale cannot represent the entire spatial distribution of the underlying data.

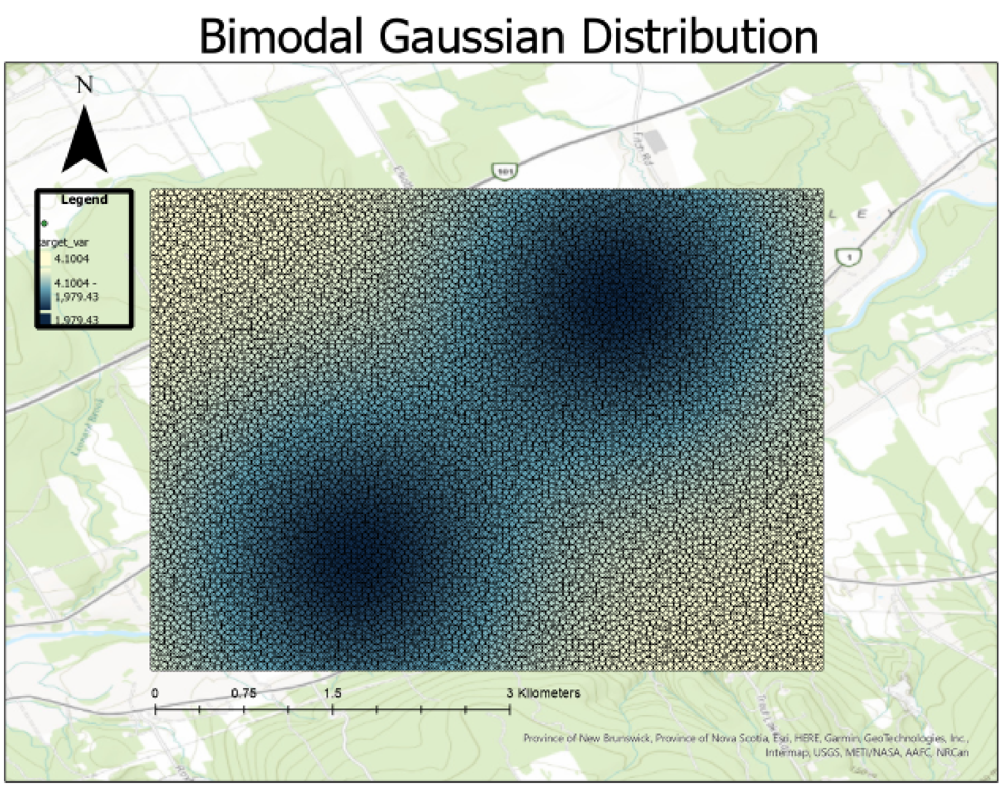


Figure 9: Point Layer Created in GeoPandas Centered at the Centre Of Geographic Sciences. Zooming out gives a global picture of the spatial function under consideration.

2.2 Spatial Functions

The following list of functions represent the functions which have been programmed into the python script pointdatagenerator.py. The script requires the python packages NumPy, Pandas, GeoPandas, and their associated dependencies. The script prompts the user for the total number of points n per spatial dimension, forming the nxn grid, as well as the total number of kilometers the bounding box shall stretch away from COGS (East,West,North,South). The program then prompts the user which spatial function they would like to fit to that bounding grid. The program is written such that "suggested" or default values are input into the functions if the user does not provide an entry.

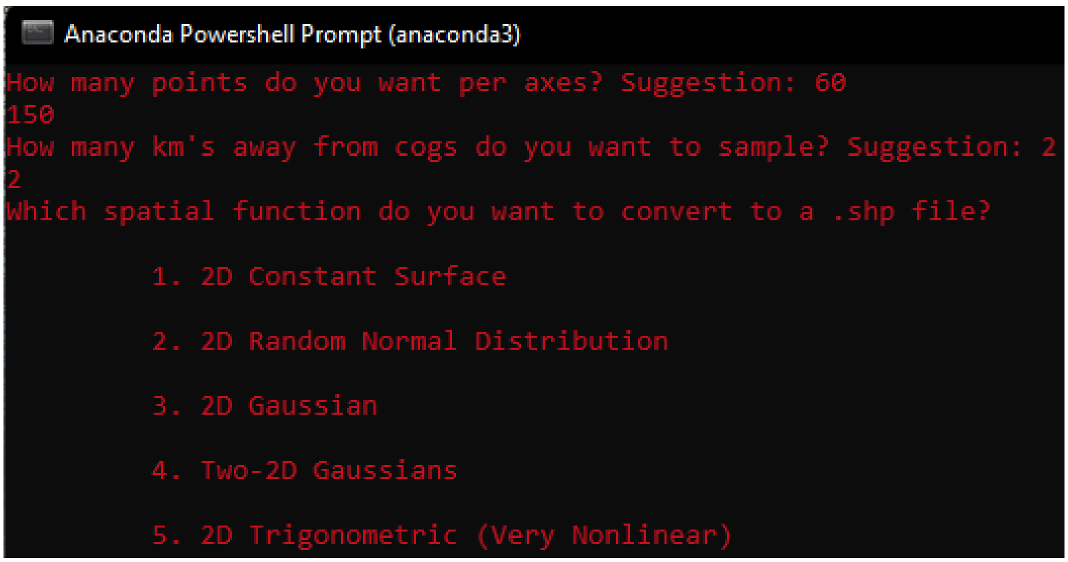


Figure 10: Terminal output of pointgenerator.py. The output is a .shp file which contains a spatial function fitted to a NumPy meshgrid.

2.2.1 Constant Surface Function

A constant function with 1 assigned to every latitude and longitude pair. This is a purely homogeneous and isotropic distribution.

$$z = f(x, y) = 1. \tag{4}$$

```python
#Constant function of '1s' at every spatial location. Homogeneous.
def ones_2d(size=n):
    return np.ones((n,n))
```

Figure 11: An nxn numpy array of ones. This represents a completely homogeneous distribution.

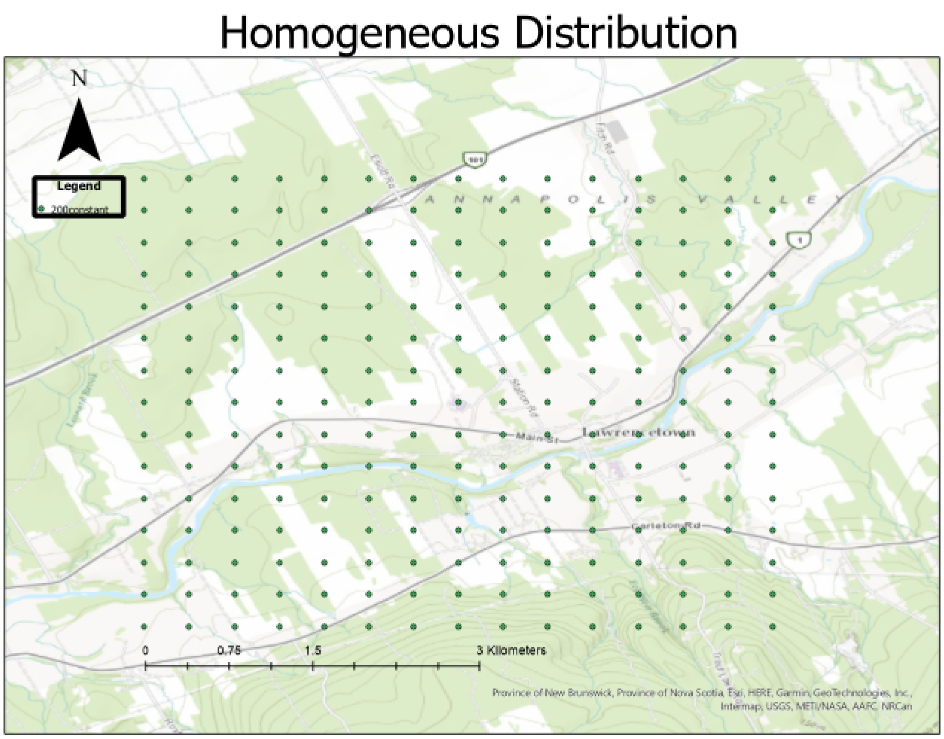


Figure 12: An 15x15 grid centered about COGS where each geographic location is assigned the value one.

2.2.2 Random Normal Distribution

This is function returns an nxn matrix with a random normally distributed value $\mathcal{N}$ with mean μ and standard deviation σ assigned to every latitude and longitude pair.

$$z = f(x, y) = \mathcal{N}(\mu, \sigma). \tag{5}$$

```
# Random normally distributed values, can set the mean and standard deviation. Can be scaled with the parameter A
def normal_dist_2d(mean=0,sd=1,size=n,A=1):
    return A*np.random.normal(mean,sd, size =(n,n))
```

Figure 13: An nxn numpy array of randomly normally generated values with a central mean and a standard deviation.

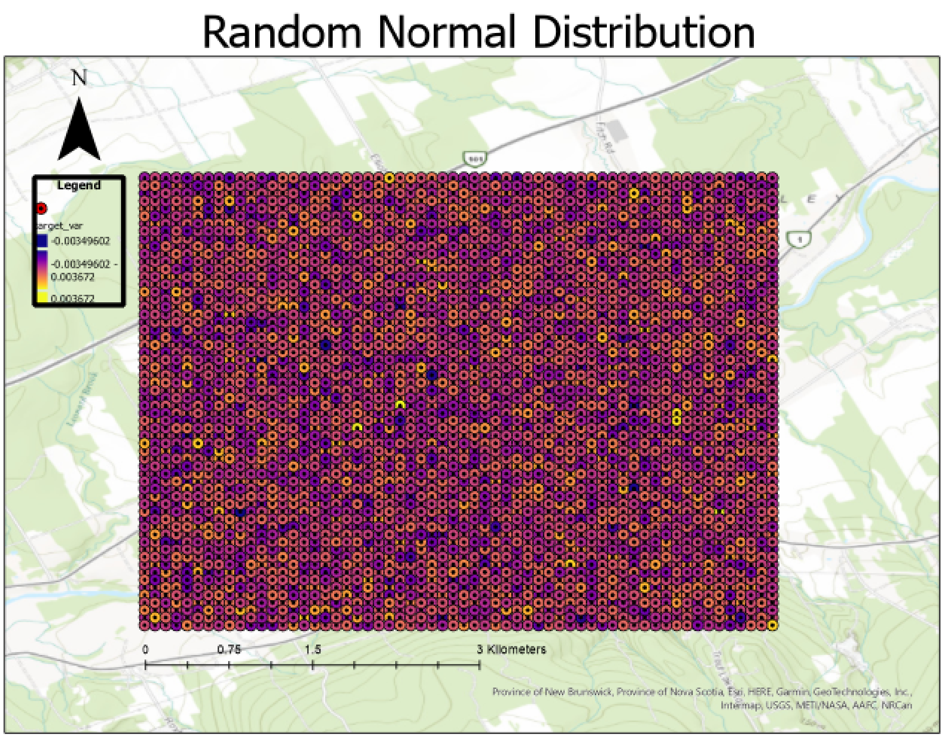


Figure 14: An nxn grid of random normally sampled points at each geographic location centered about COGS. The points have been enlarged for aesthetic purposes only.

2.2.3 2-D Gaussian Distribution

This is function returns an nxn matrix with a probability density function for random normally distributed value in two spatial dimensions with a horizontal mean (shift) μ_x and a vertical mean μ_y. Similarly, the horizontal standard deviation σ_x and vertical standard deviations σ_y are assigned to every latitude and longitude pair. This produces the standard "bell curve shape" in two dimensions.

$$z = f(x, y) = \frac{1}{2\pi\sigma_x\sigma_y} exp\left[-(\frac{(x-\mu_x)^2}{2\sigma_x^2} + \frac{(y-\mu_y)^2}{2\sigma_y^2})\right]. \tag{6}$$

```python
# https://en.wikipedia.org/wiki/Gaussian_function#Two-dimensional_Gaussian_function
# This 2D gaussian function is centered about COGS.
# There are independent values for means and std devs in both the x and y axes.
# I had to play with the means (mx,my), standard deviations(sx,sy) to get them to look 'nice' on the graph.
# Feel free to tweak these parameters to adjust the shape of the distribution
# Added the shift parameter to further shift the bell shape along x, or y. This is useful for adding gaussians together.
def gaus2d(x=0, y=0, mx=cogs_long, my=cogs_lat, sx=0.009, sy=0.009,shift_x=0.01,shift_y=0.00):
    # Different amplitude parameters, wikipedia uses A=1, however the chosen amplitude is usually selected
      # for gaussian normalization (integral = 1)
    #A=1
    A = 1. / (2 * np.pi*sx * sy)
    #A1 = 1/ (np.sqrt(2*np.pi * sx**2 *sy**2))
    return A * np.exp(-((x - mx - shift_x)**2. / (2. * sx**2.) + (y - my - shift_y)**2. / (2. * sy**2.)))
```

Figure 15: An nxn numpy array of randomly normally generated values with a central mean and a standard deviation.

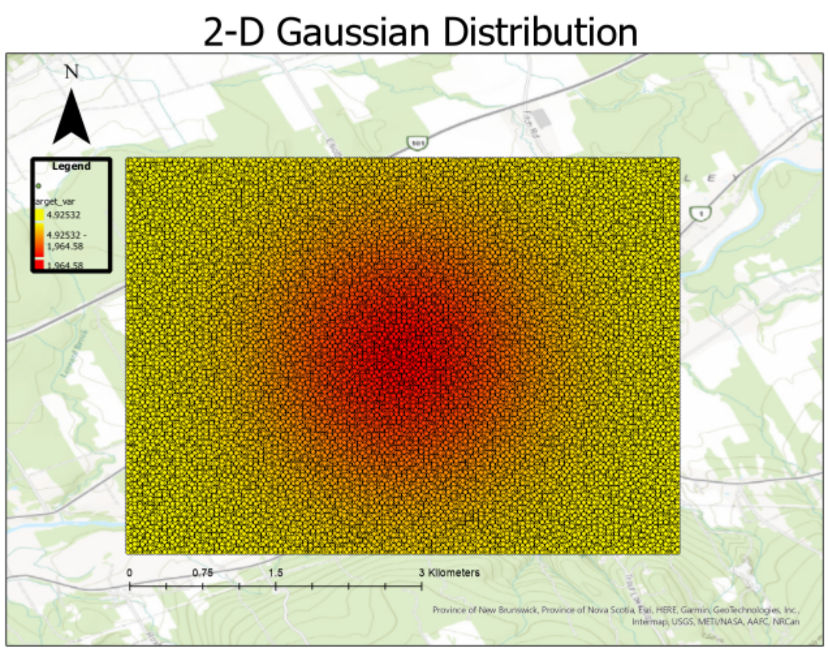


Figure 16: A 2-d Gaussian distribution with 40,000 points centered about The Centre of Geographic Sciences.

2.2.4 Bimodal Gaussian Distribution

A bimodal Gaussian distribution is observed near the merging of two separate Gaussian probability density functions.

We simply call the function above twice to generate two distributions and add them together.

$$z_1 = f(x, y) = \frac{1}{2\pi\sigma_{x_1}\sigma_{y_1}} exp\left[-(\frac{(x-\mu_{x_1})^2}{2\sigma_{x_1}^2} + \frac{(y-\mu_{y_1})^2}{2\sigma_{y_1}^2})\right].$$

$$z_2 = g(x, y) = \frac{1}{2\pi\sigma_{x_2}\sigma_{y_2}} exp\left[-(\frac{(x-\mu_{x_2})^2}{2\sigma_{x_2}^2} + \frac{(y-\mu_{y_2})^2}{2\sigma_{y_2}^2})\right].$$

$$z = f(x, y) + g(x, y) = z_1 + z_2. \tag{7}$$

Where σ_{x1} and μ_{x1} are the standard deviation and mean for the first Gaussian, and σ_{x2} and μ_{x2} are the standard deviation and mean for the second Gaussian.

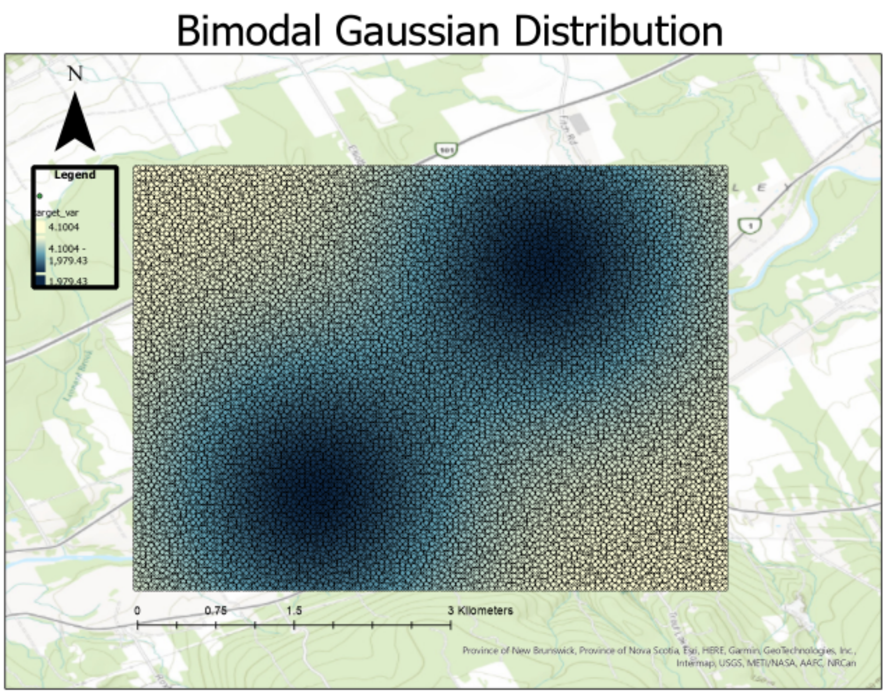


Figure 17: A bimodal Gaussian distribution with 40,000 points aligned off axis.

2.2.5 Trigonometric Distribution

The last programmed function is meant to be highly nonlinear displaying periodic phenomena about COGS.

```
# Trig function that I thought was cool... highly nonlinear so don't expect any methods to work..
def cool_trig_function(A=1,f=0.481,w=2.05,inner_pow = 4, outer_pow=2,cogs_long=cogs_long, cogs_lat=cogs_lat):
    return A*(np.sin( f*(yy-w*cogs_long)**inner_pow + f*(xx-w*cogs_lat)**inner_pow ) )**outer_pow/ np.sqrt(xx**2 + yy**2)
```

Figure 18: An nxn numpy array of a complex trigonometric function with many periodic properties.

$$z = g(x,y) = A\frac{sin\left[f(y-\omega\cdot longitude)^4 + f(x-\omega\cdot latitude)^4\right]^2}{\sqrt{x^2+y^2}} \tag{8}$$

Where in this context, the function has defaults of f=0.481 and ω=2.05.

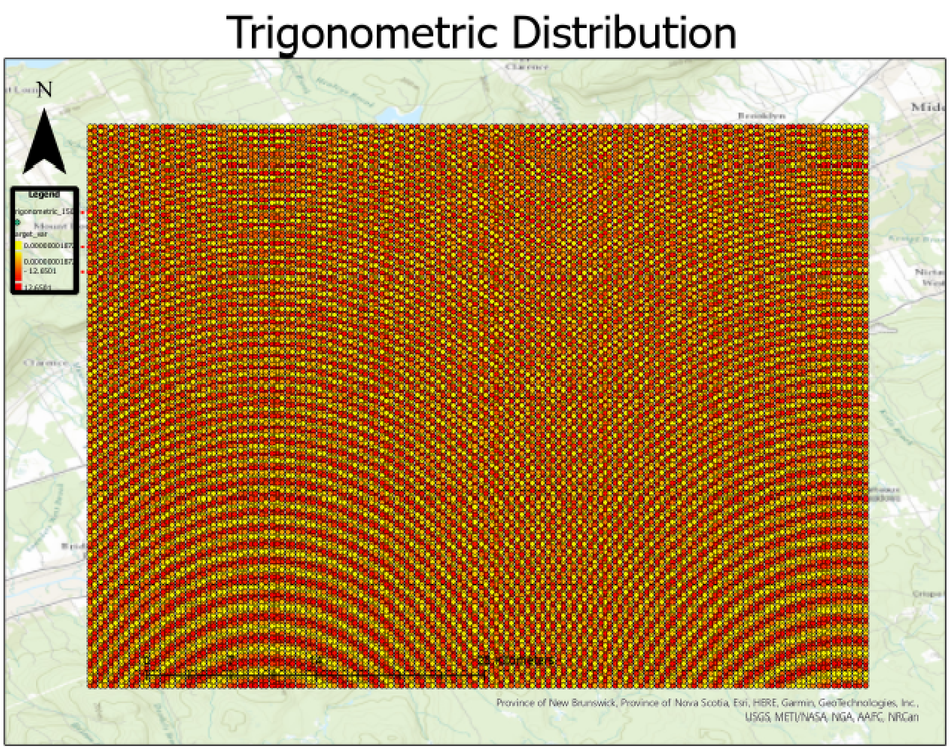


Figure 19: A trigonometric surface which reveals complex periodic phenomena at large scales.

With our functions readily defined in a script, the results are converted into a .shp file and are ready for exploratory data analysis to visualize our data and determine trends in our distributions.

3 Exploratory Data Analysis

By generating ground truth data, we are now able to study the predictive abilities of ArcGIS Pro's geostatistical wizard. Before undertaking to geostatistical Kriging it is best to conduct a preliminary data exploration of the data at hand. This is very useful for obtaining key statistical results like the mean, standard deviation, and variance which further aid our interpretation of Kriging results. For ordinary Kriging analysis, there are three main criteria for obtaining best results for the predictions. Ordinary Kriging interpolation is built under the assumption of normal distribution, stationarity, and lack of trends [GIS GEOGRAPHY 4]. Please note that these are *optimal* conditions for ordinary Kriging, and for many applied scientists, these optimal conditions are rarely encountered and recorded. Studies have been made and have additionally revealed that a normally distributed data set is not absolutely necessary [6]. If the data is not normally distributed there are some statistical methods such as transformations which may produce a normal distribution for the analysis. Later in this document we will intentionally relax the stationarity condition for Ordinary Kriging to perform Empirical Bayesian Kriging and compare the results of both statistical predictions when these conditions are relaxed. With our synthetic data at hand we will precisely be able to tell the limitations of the geostatistical tools given that we know the underlying function for each data set

3.1 Statistical Data Distibutions with Histograms

In this section we will take a look at a variety of sample distributions that we may encounter working with data. Please note we did not conduct an extensive EDA for every single function but rather are interested in highlighting interesting properties about the underlying distributions.

3.1.1 Homogeneous Distribution

The homogeneous distribution will show a single histogram column given that the entire of distribution of points is fixed at the value (+1 in this paper).

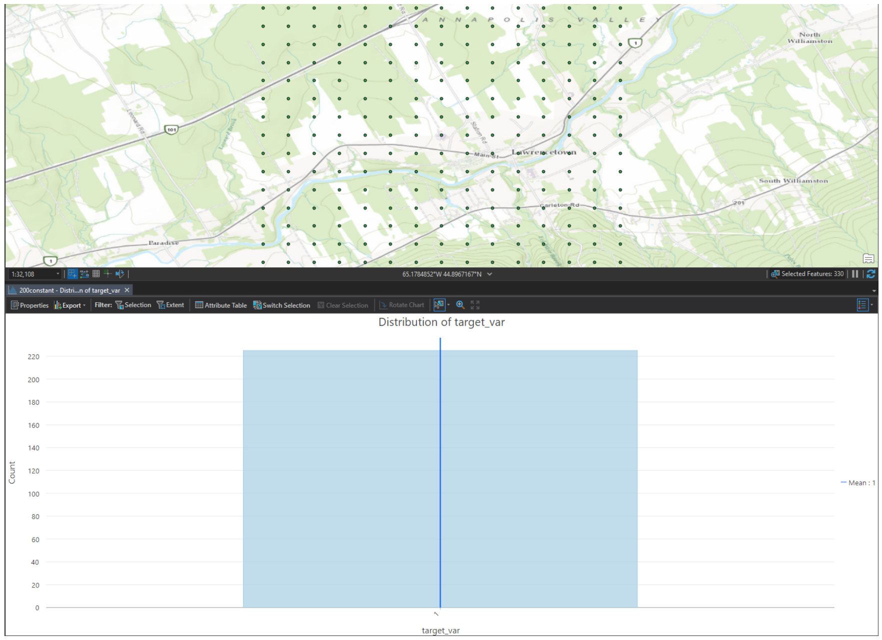


Figure 20: A single column for the entire data set is indicative of homogeneity.

3.1.2 Random Normal Distribution

The random normal distribution is a measure of randomness. It is indicative of many real-world processes like heights of individuals in a population, dice rolling, grades in examinations, and many more. For our synthetic data, each point has been assigned a continuous random variable which represents a variable whose probability is described by a normal distribution (bell shape) with a mean μ and standard deviation σ. We can confirm this with our synthetic data given the histogram generated in ArcGIS Pro shown below.

3.1.3 Gaussian Distribution

Gaussian distributions show a skewness in their distribution. This is associated to the sharp drop off in value counts from the height of the central peak of the bell curve towards the tail ends of the function as can be seen by the population of points highlighted at the far end of the histogram tail.

3.1.4 Bimodal Gaussian Distribution

Similarly, a bimodal Gaussian distribution will display the same skeweness as above, except due to averaging and mixing the distribution of intermediate values is increases displaying new peaks in the histogram.

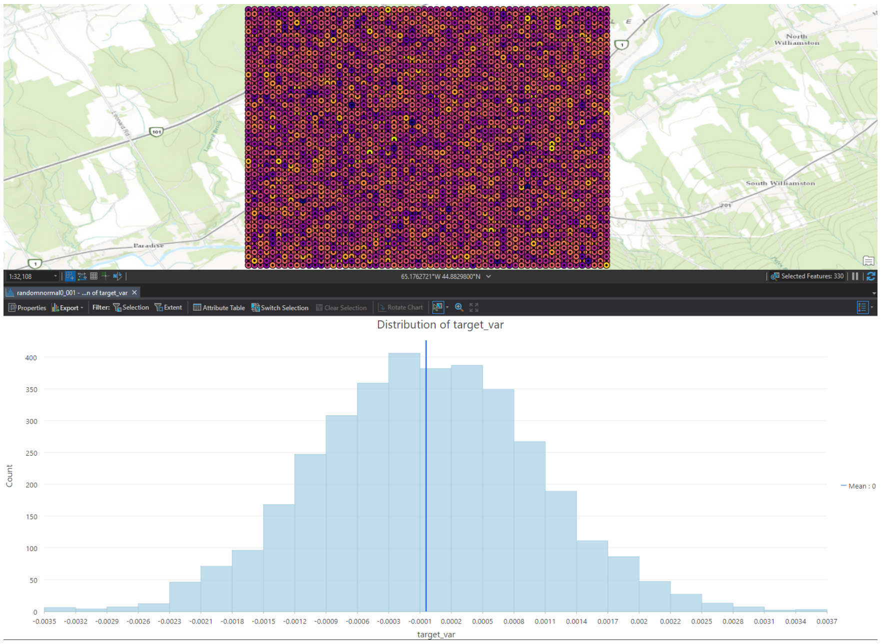


Figure 21: A bell shaped histogram is indicative of sampling a random normal variable.

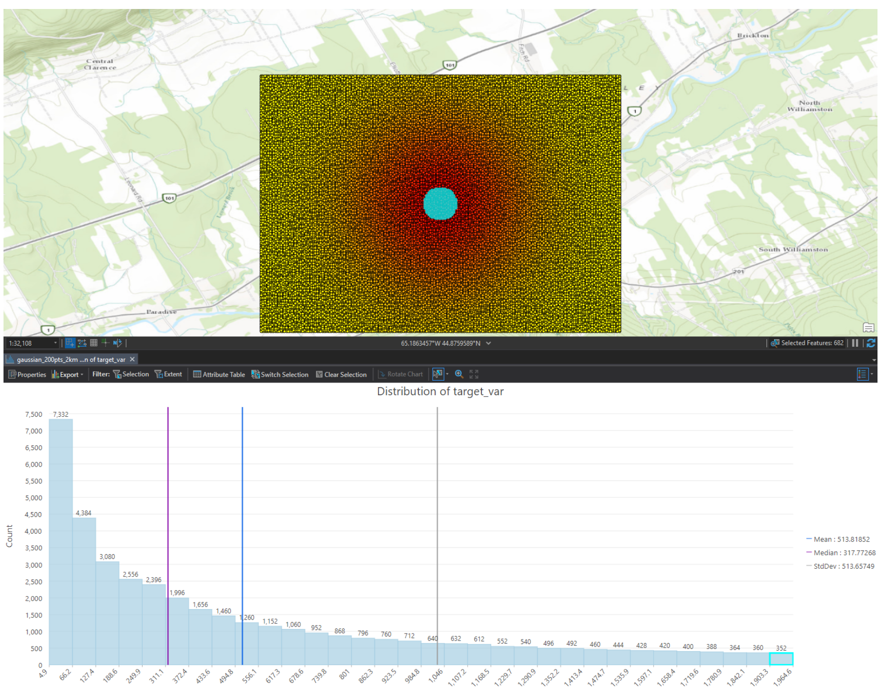


Figure 22: A right tail skewed distribution is generated by a Gaussian peak. Highlighted points at the peak of the graph correspond to the smallest count in the histogram due to their localized spatial distribution.

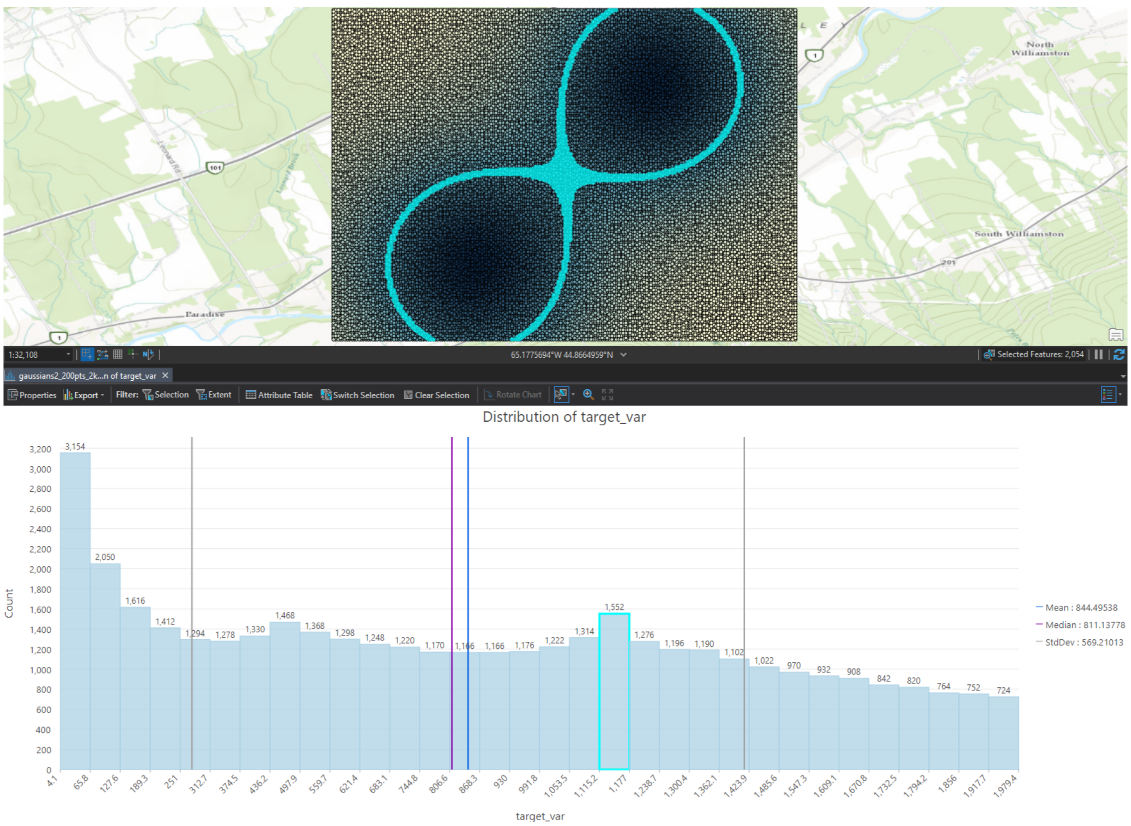


Figure 23: Gaussian addition generates averaged results at the center of the distribution due to the interaction of both Gaussians. This is reflected as the small highlighted peak on the histogram.

3.1.5 Trigonometric Distribution

The trigonometric distribution displays symmetry in its histogram with two sharp peaks at each end of the graph. Interestingly enough, the distribution of points which lie in between these peaks appear to be random. However, the points which correspond to each peak correspond to what appears to be dominant periodic trends in the trigonometric function. The author speculates that this is likely due to the additive nature of wave superposition.

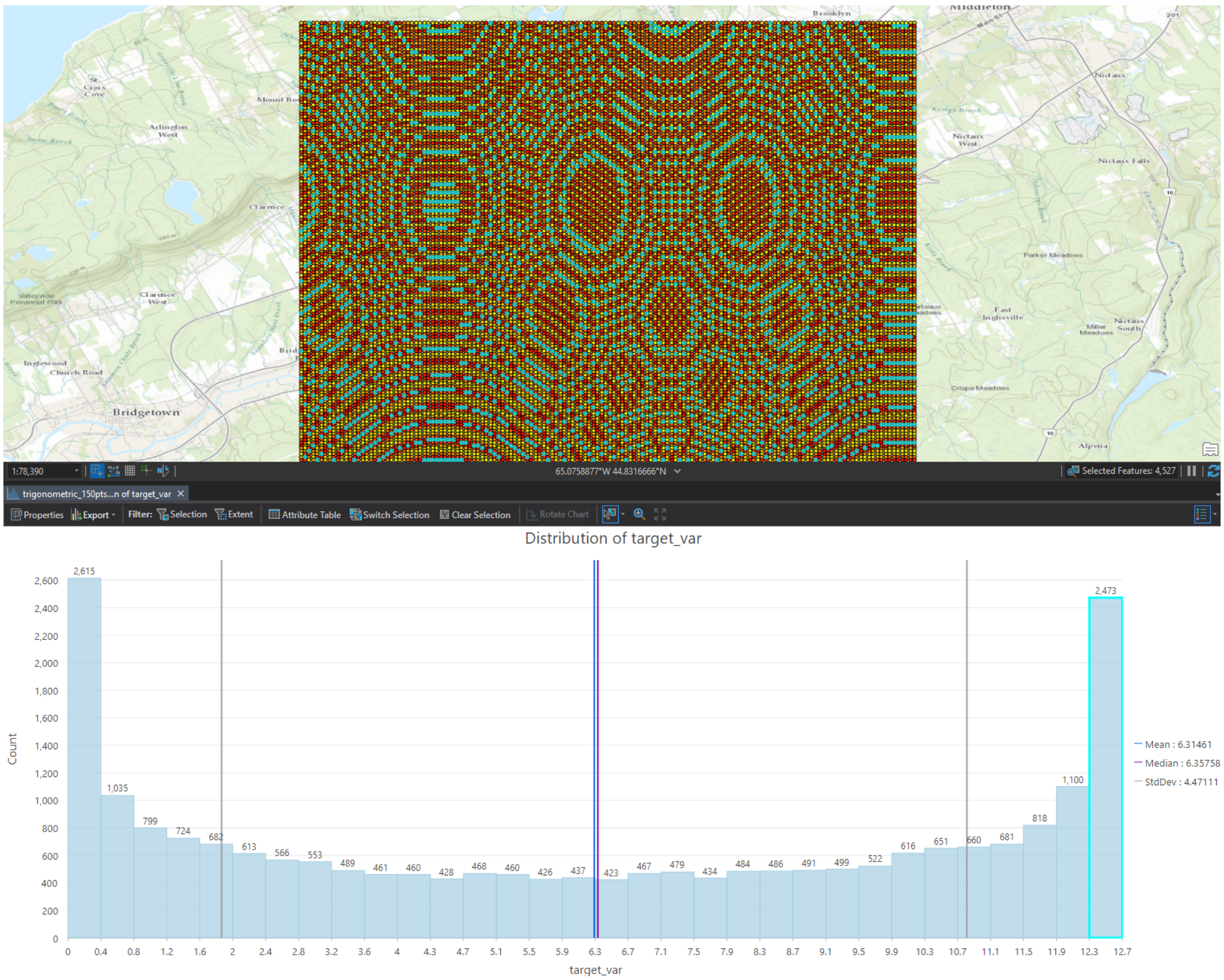


Figure 24: The trigonometric distribution shows symmetry in the histogram. The peaks at the tail ends (far right is selected) display dominant harmonic modes in the underlying data.

3.2 Normal QQ Plots

QQ plots are an useful tools for determining the normality of the underlying data. The QQ plot gives a measure of normal distribution given how much of the data falls on the reference line.

3.2.1 Random Normal Distribution

For our generated random normal distribution we expect to see the majority of the data lie on the reference line. This is indeed what is observed.

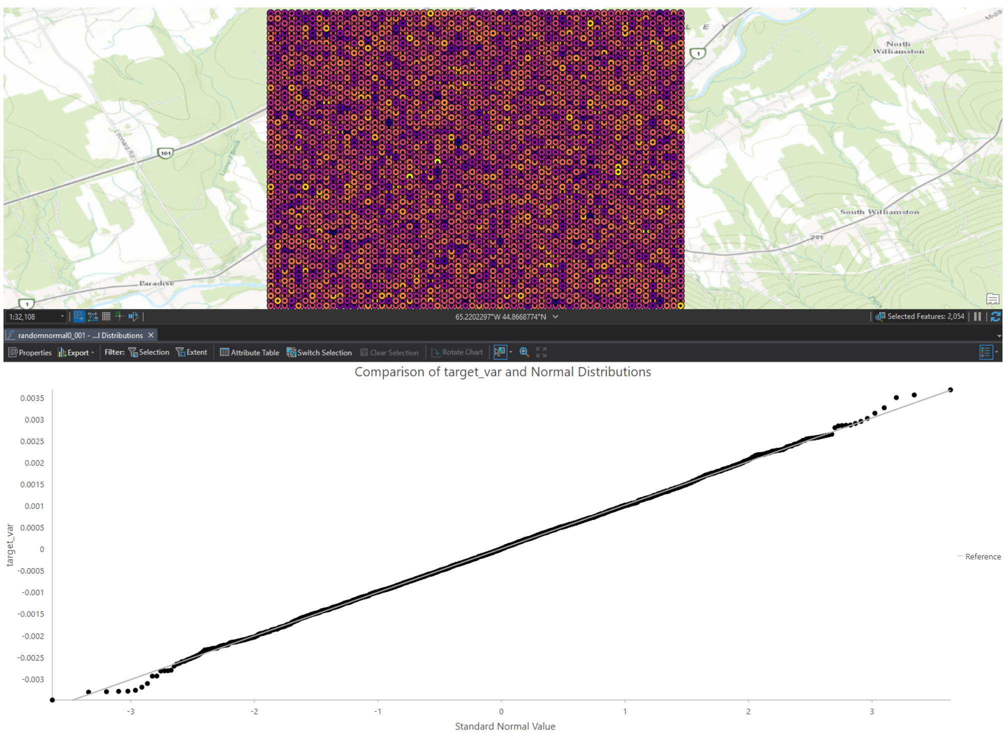


Figure 25: A straight line is observed in the QQ plot of the underlying data confirming a normal distribution.

3.2.2 Bimodal Gaussian Distribution

The QQ plot for the bimodal Gaussian distribution forms a symmetric skew about the reference line. This might indicate that simple kriging might not be the best for our interpolation technique.

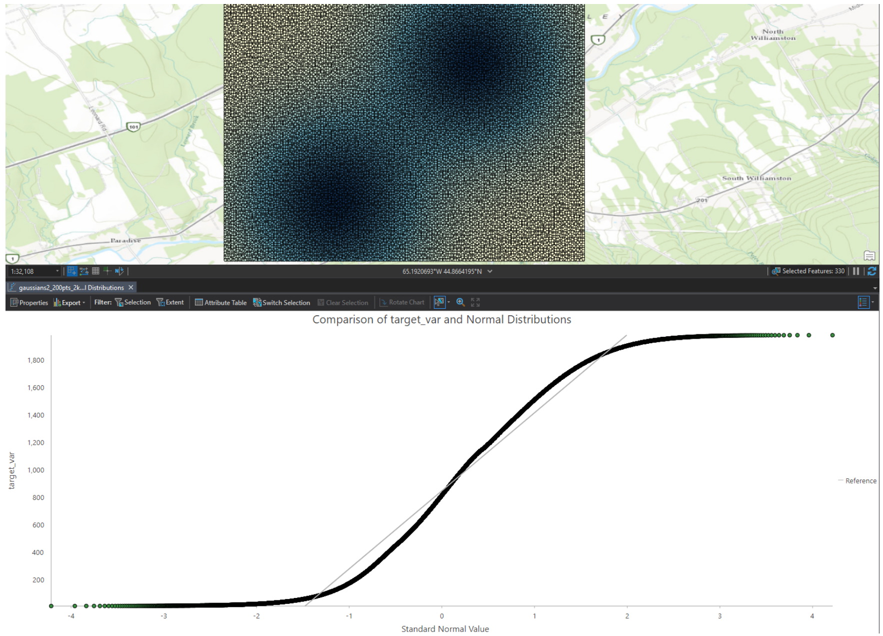


Figure 26: A symmetric line is generated through the reference line. There is indication away from normalcy in our generated data.

3.3 Trend Analysis in ArcMap

3.3.1 Random Normal Distribution

Trend analysis is important to perform in order to determine change over the entire study area. It generates a three-dimensional perspective of the data which is ideal for our modeling of 2-D surfaces. The locations of all the generated point values are plotted on the (x,y) plans with the target (synthetic) variable given as a height at every point as the z dimension. What is important to note from this tool are the projected trend lines on the walls of the graph (xz and yz planes). What is key to remember when using the trend analysis tool, is that if there is no underlying trend (flat projected lines on the wall) then no trend exists. As seen in the trend analysis graph of the normal distribution data:

3.3.2 Bimodal Gaussian Distribution

Trend analysis for the bimodal Gaussian distribution reveals the true nature of the 2D surface consisting of two Gaussian peaks interacting with each other and averaging the region in between. The "front" of the data is pictured above, while the "back" of the graph is shown to clearly display the negative quadratic trends which have been projected on the walls of the graph. These quadratic lines suggest that a second order polynomial can be fit to the data. [7] From our function it is clear that the direction of the trend is skewed such that its strongest influence lies from the center of the Gaussian peaks, towards the border regions in the outer edges.

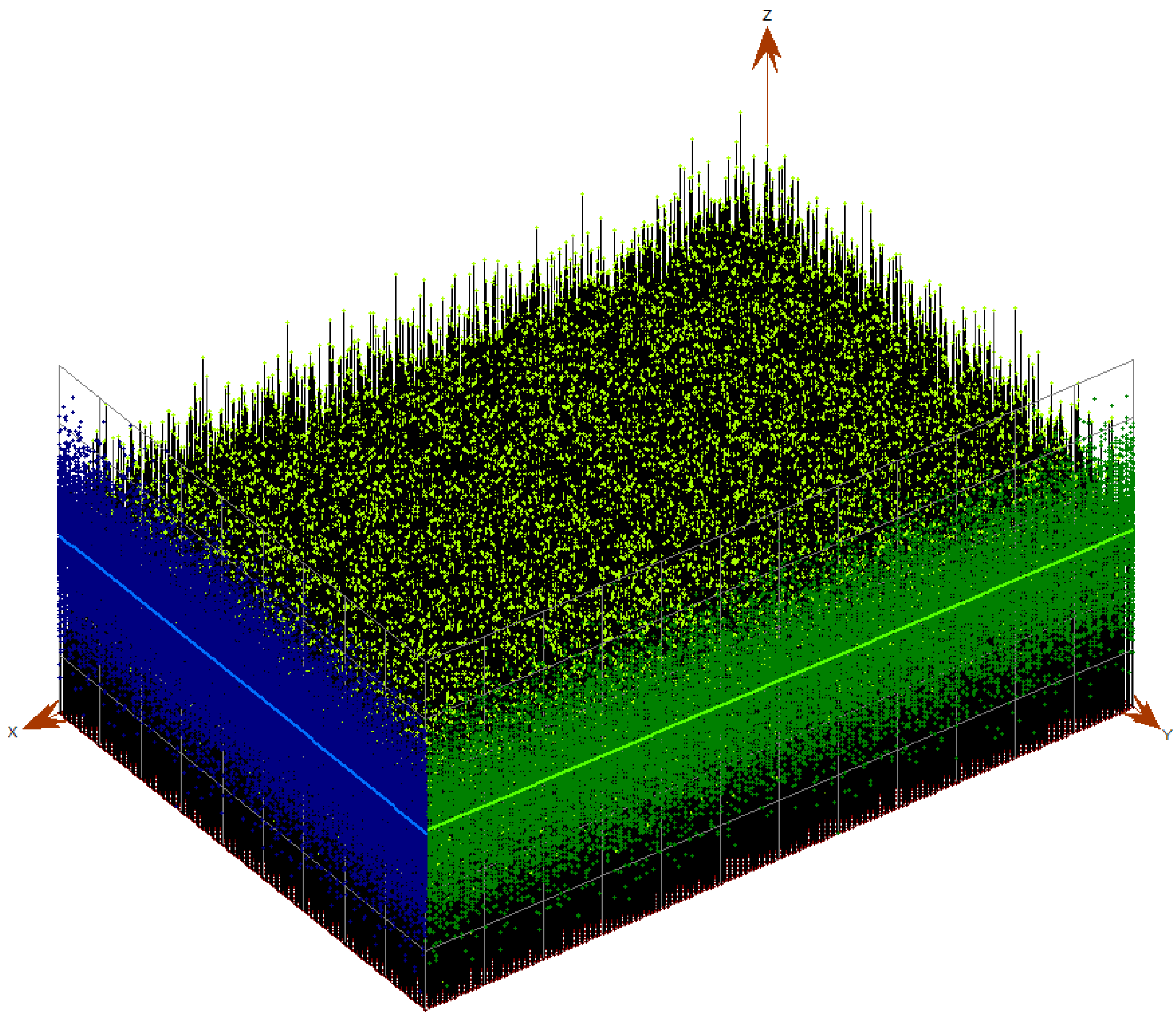


Figure 27: A trend analysis in ArcMap of 40,000 sampled points from random normal data. Straight lines are visible on the projected axes indicating the lack of any trends in the data.

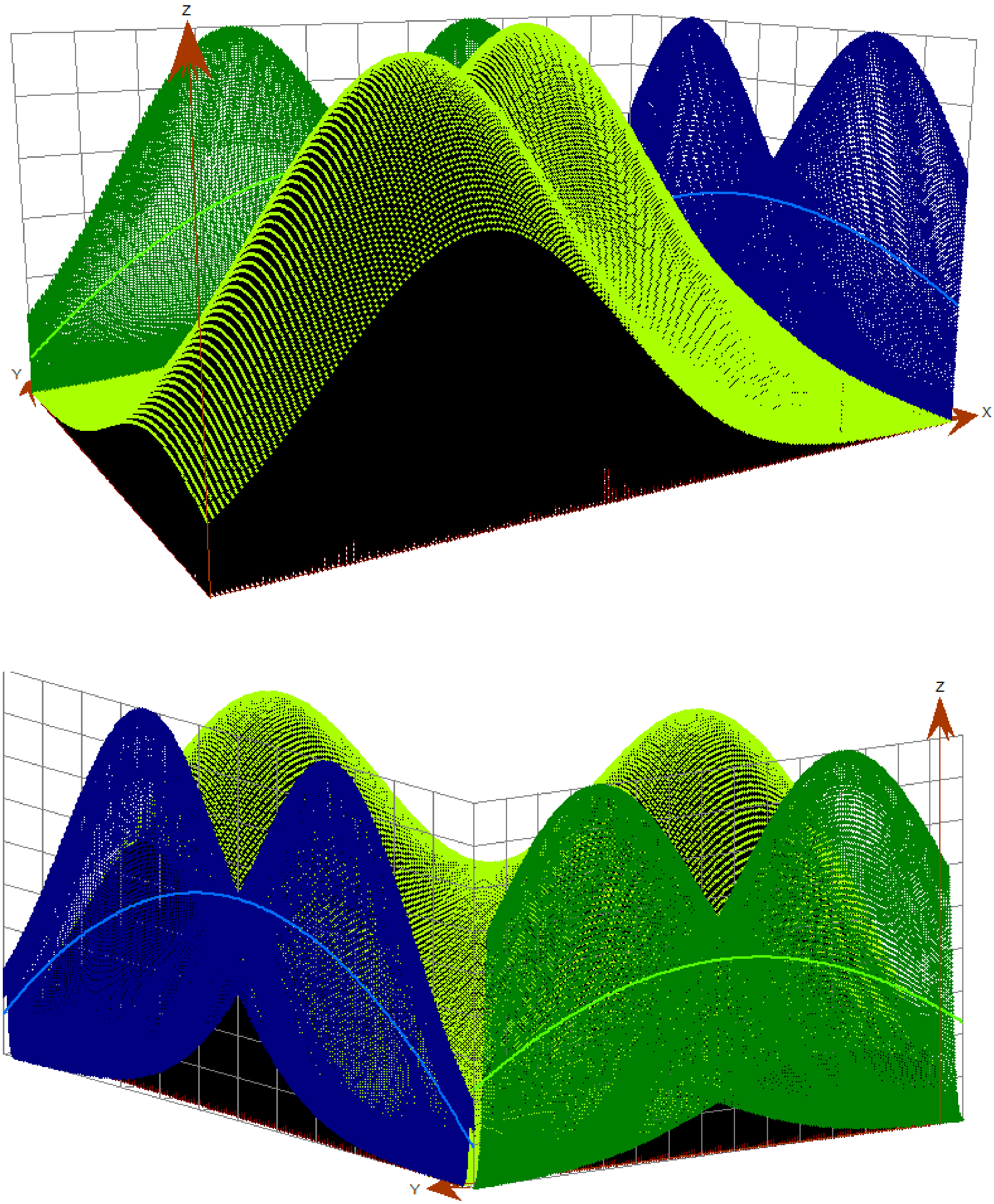


Figure 28: 40,000 points make up the Bimodal Gaussian peaks shown in the ArcMap trend analysis tool. Further revealing second order trends through projections onto the axes of the graph.

3.4 Voronoi Maps in ArcMap

Voronoi maps are useful geostatistical tools to examine the local variation and clustering of the underlying data. This is done by creating polygons in our map which ensure that every location within the polygon are closer to the sample point than any other sample point. The reason for this type of study is that ordinary kriging does not respond well to sudden changes in values. It is assumed that nearby values will have similar differences and increase the further apart the measured points become. This is one of the main assumptions behind ordinary Kriging, known as stationarity. We will test for stationarity, or local variation with an entropy Voronoi map.

An entropy map will group all polygons into $n = 5$ classes based on the quantiles of the data. The value of the Shannon entropy H assigned to a polygon is the entropy calculated from the polygon and its neighbours [8] and is defined by,

$$H = -\sum_{i=1}^{n=5} (p_i \cdot log(p_i)) \tag{9}$$

where p_i refers to the proportion of polygons assigned to each class value.

3.4.1 Homogeneous Distribution and Sparsity

We begin by testing the Voronoi tool with our homogeneous distribution. We should expect to find no local (or global) variation in the data.

We can confirm this by noting from the graph that every single point has been labelled the same colour (class) and due to the regular spacing of the data, has defined a well spaced array of polygons around each synthetic point.

We can relax the spatial homogeneity constraint of our synthetic data by collecting a random subset of these points. By generating sparse coordinate data from the homogeneous distribution, we can create a new Voronoi map that demonstrates the way that the polygons are defined. We import a randomly selected subset of 75% of these coordinate points to generate the second homogeneous valued, sparse Voronoi entropy map.

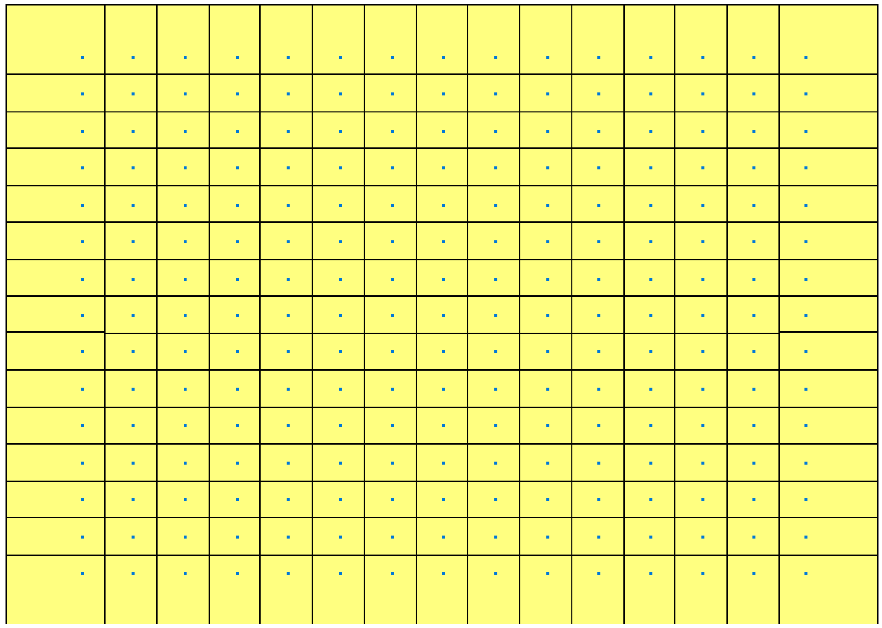


Figure 29: An entropy Voronoi map displays a uniformly distributed lattice of coordinates uniformly tesselated with a homogeneous value (+1).

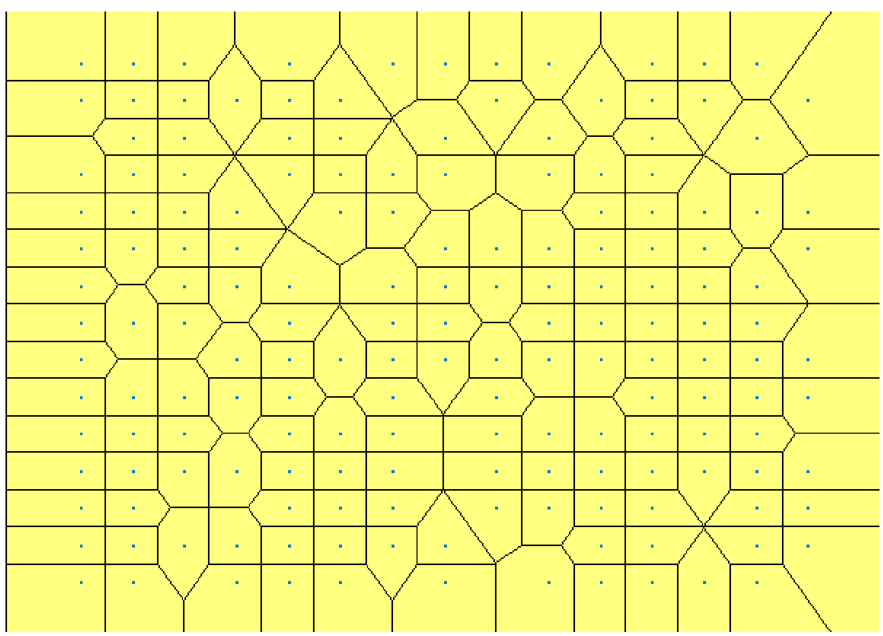


Figure 30: An entropy Voronoi map of a random subset of 0.75 the homogeneous and isotropic data from the last image. A non-regular tessellation is produced with a homogeneous distribution of values (+1). This is useful for understanding the effects of data sparsity in relation to the prediction of results in local point neighborhoods.

3.4.2 Random Normal Distribution

In contrast to the homogeneous distribution, we take a look at the entropy Voronoi map with random normally distributed data. As we saw in the trend analysis, there was no evidence of any trends in any spatial direction and so due to the nature of random selection of a normally distributed value, this data is stationary.

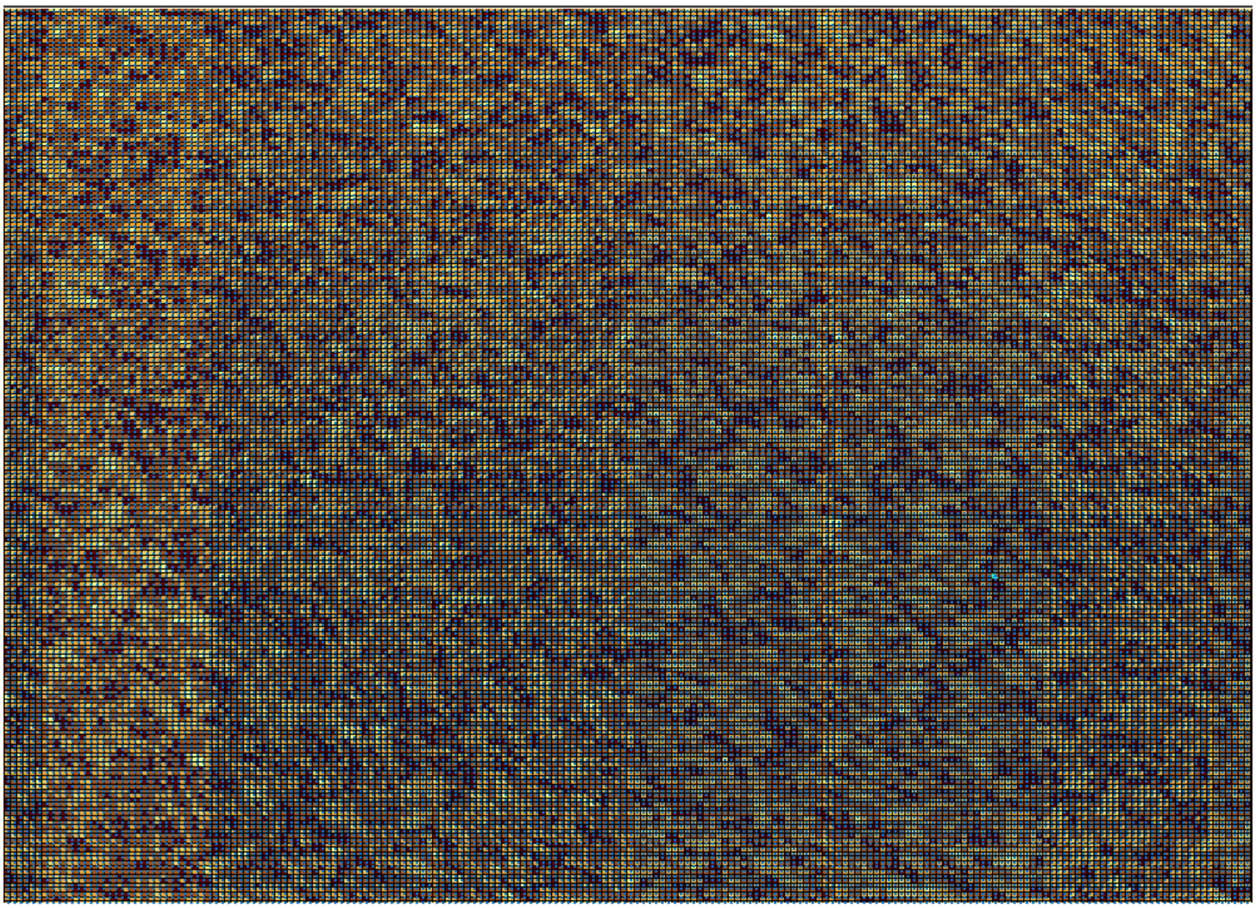


Figure 31: An entropy Voronoi map displaying the local variation of the data set. There is no evidence of trends, and constitutes a near perfect stationary data set.

3.4.3 Bimodal Gaussian Distribution - Entropy Voronoi Map

By creating a entropy Voronoi map of the bimodal Gaussian distribution, we obtain a clearer understanding of how the entropy model works, and thus how it labels data in order to create a measure of variation. Entropy is a term stemming from the physical sciences which is used for describing systems with a measurable state of disorder, or randomness. Using the entropy method, all polygons are categorized into up to five classes (smart quantiles). The value that is assigned to a polygon is purely dependent on the proportion of distinct neighbours in other classes. In general, maximum entropy value corresponds to a system (data set) where every neighbour polygon is of a different class. In contrast, a minimum entropy value corresponds to a configuration where nearby polygon values are all of the same class. From the entropy map generated by the bimodal Gaussian distribution we see that the majority of our data consists of low entropy rankings. Meaning that much of the spatial distribution of the data is locally similar to itself. However, the algorithm picks up on regions of rapid decay for the Gaussian distributions. These closed lines are topological *level sets* and hint high variability within those regions of space.

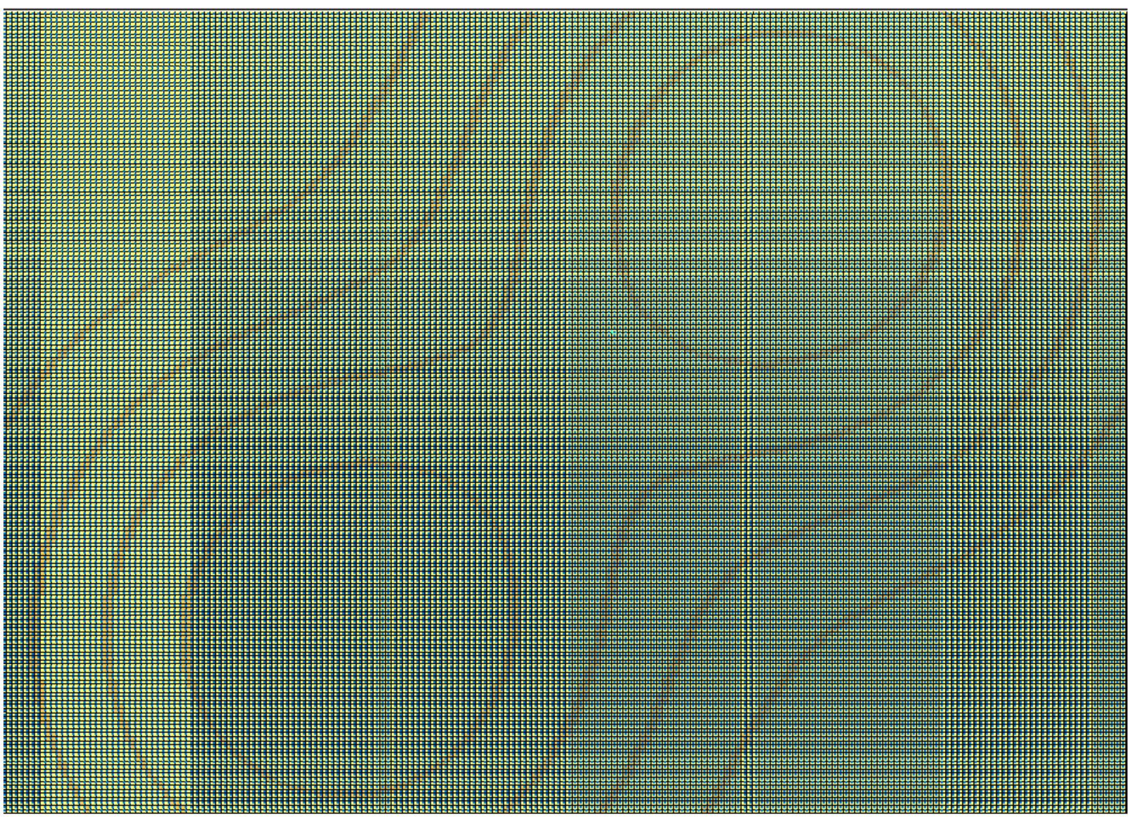


Figure 32: 2D entropy Voronoi Map of the bimodal Gaussian distribution. Variance is quite uniform throughout the data except for band regions where the values of the function decrease abruptly.

3.4.4 Bimodal Gaussian Distribution - Standard Deviation Voronoi Map

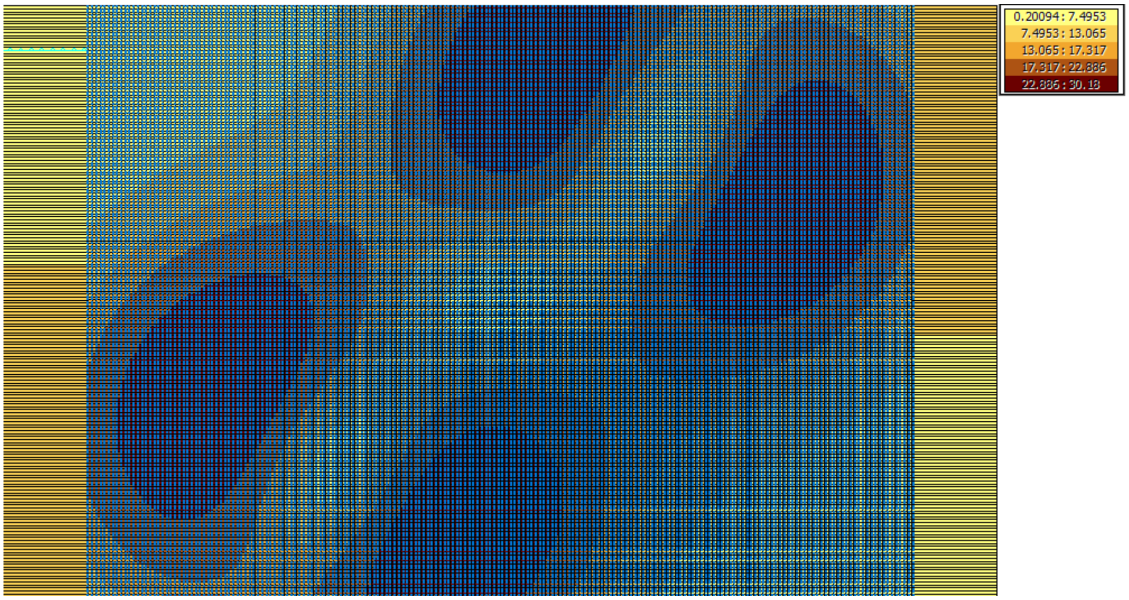


Figure 33: 2D entropy Voronoi Map of the bimodal Gaussian distribution. This map represents the standard deviation of the cell and its neighbours. Dark colours indicate high variability (standard deviatio), while low colours correspond to low variation.

3.5 Semivariogram Cloud

The final step before starting our Kriging analysis, is the fitting on a semivariogram using the Semivariogram / Covariance Cloud Tool on ArcMap. It is useful for determining the existence of spatial autocorrelation of the sample data to determine if there is anisostropy in our data. In particular, we can test Tobler's First Law of Geography by studying the trends between points that are closer to each other (leftmost in the graph) and points that are farther apart. In particular we expect a similar values for points that are closer and differing values for points that are father apart. If the points produce a horizontal straight line in the semivariogram there may be no spatial correlation at all.

3.5.1 Random Normal Distribution

The semivariogram for a random normally distributed data set shows no preferred correlation at shorter distances. As we can see the variance of data is approximately uniformly random at every single distance.

3.5.2 Bimodal Gaussian Distribution

The semivariogram for a bimodal Gaussian distribution appears like a bell curve due to the symmetry of the underlying distribution. In general, we can see evidence of spatial

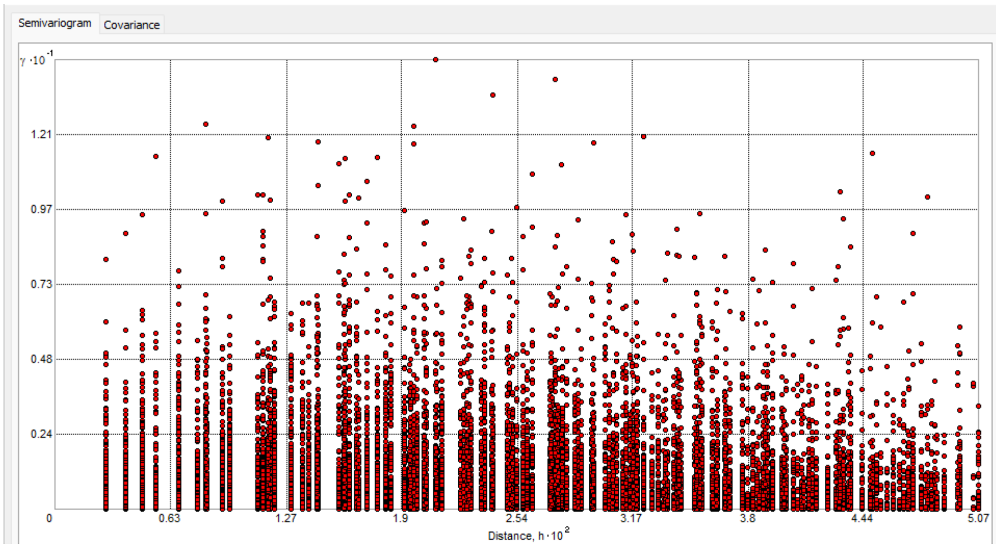


Figure 34: A semivariogram cloud for a random normally distributed data set. We can see high variability in the data independent of spatial proximity.

autocorrelation given that far left points appear close (vertically) and increasingly develop more variance the further apart the points become (further along the x axis). For this graph a lag size of 10 has been selected which controls the number of bins for the data.

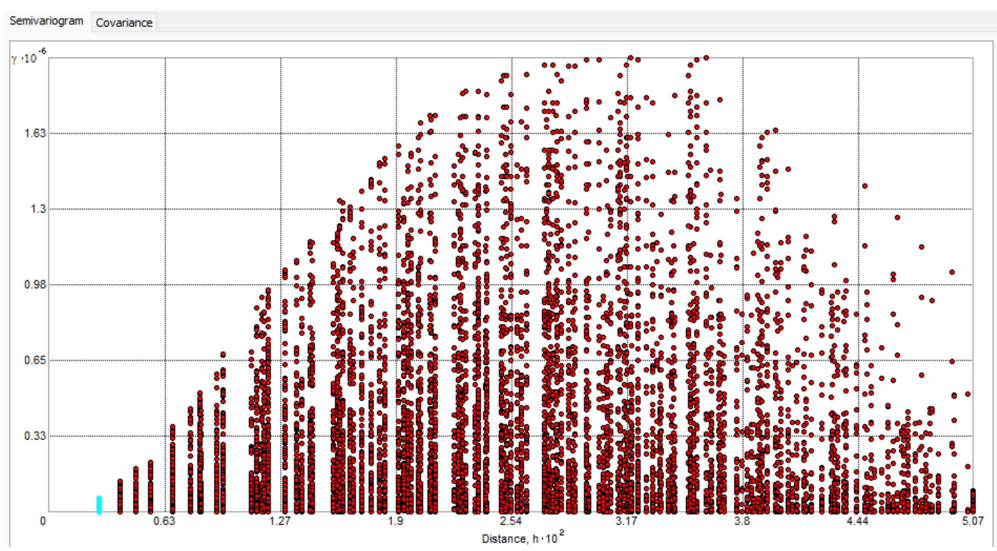


Figure 35: A semivariogram cloud of a bimodal Gaussian distribution shows spatial autocorrelation at short distances.

4 Kriging Analysis

In this section we will highlight important characteristics of Kriging interpolation. Out of the generated functions we have decided to demonstrate Kriging on random normal data as well as bimodal Gaussian data. Three kriging prediction maps have been generated for each graph. Each graph demonstrates a decreasing amount of sampled points considered by the Kriging algorithm to produce a prediction surface for each point in our bounding grid.. As we will see, Kriging results are most optimal given an evenly distributed sampling of the data.

4.0.1 Kriging Results - Random Normal Distribution

The following graphs demonstrate ordinary Kriging on random normal data. As we will see, data sparsity is central to the computation of a final prediction surface.

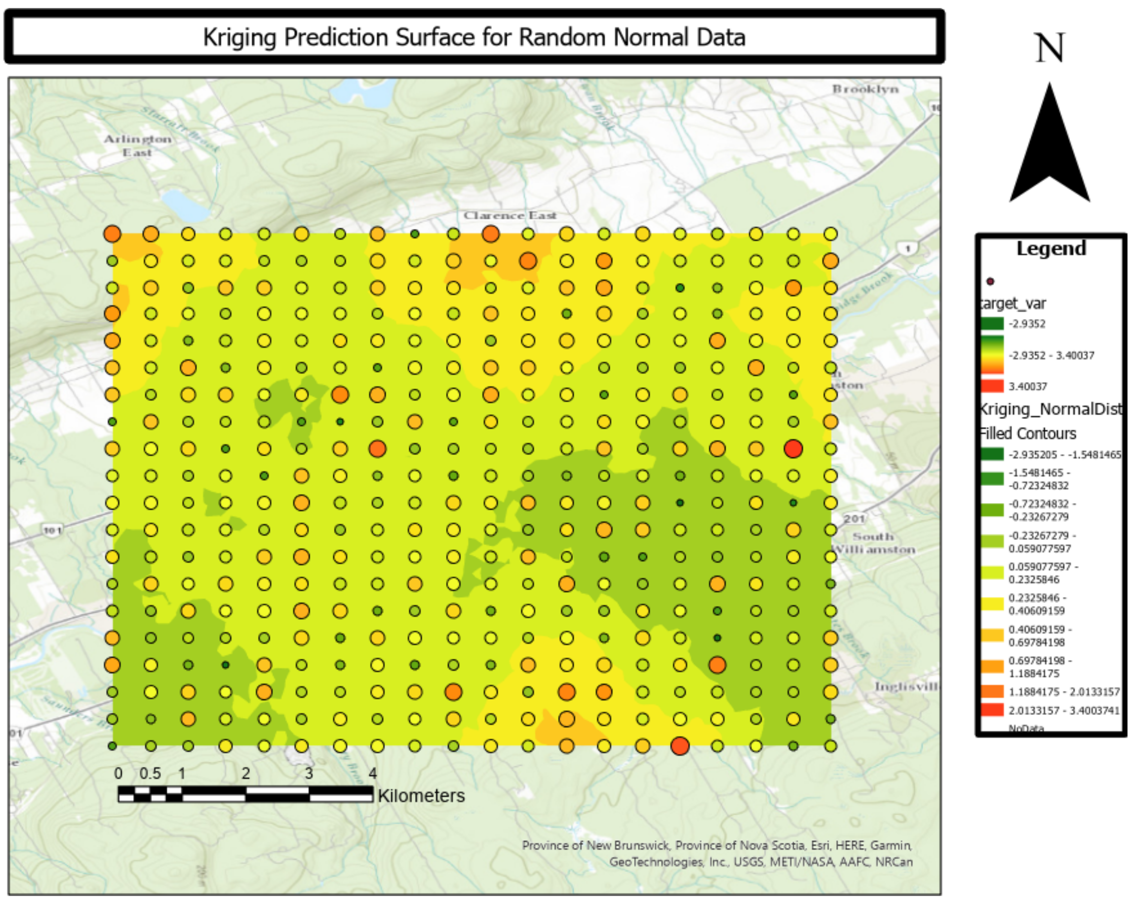


Figure 36: Ordinary Kriging on a complete data set of random normally distributed data.

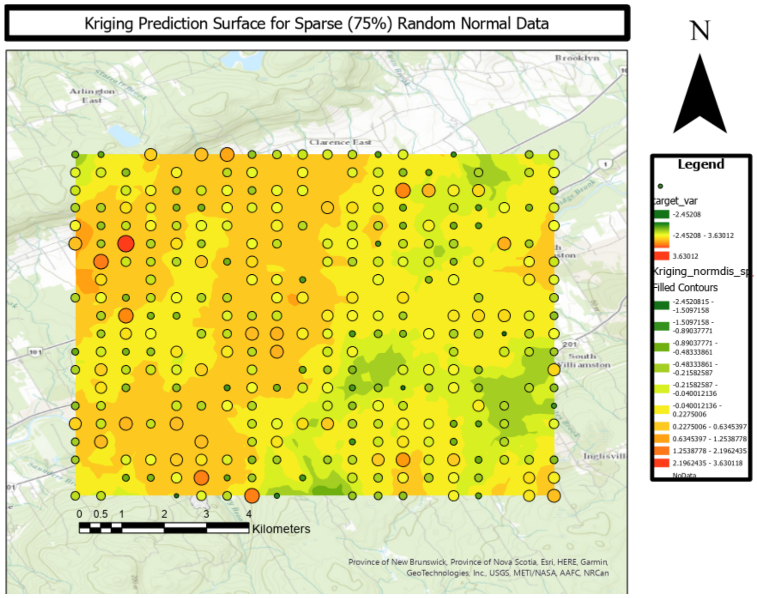


Figure 37: Ordinary Kriging sampling (3/4) of a set of random normally distributed data. We begin to see the effects of data sparsity by noting how regions which are missing values are coloured (weighted) by the colour or value of nearby points.

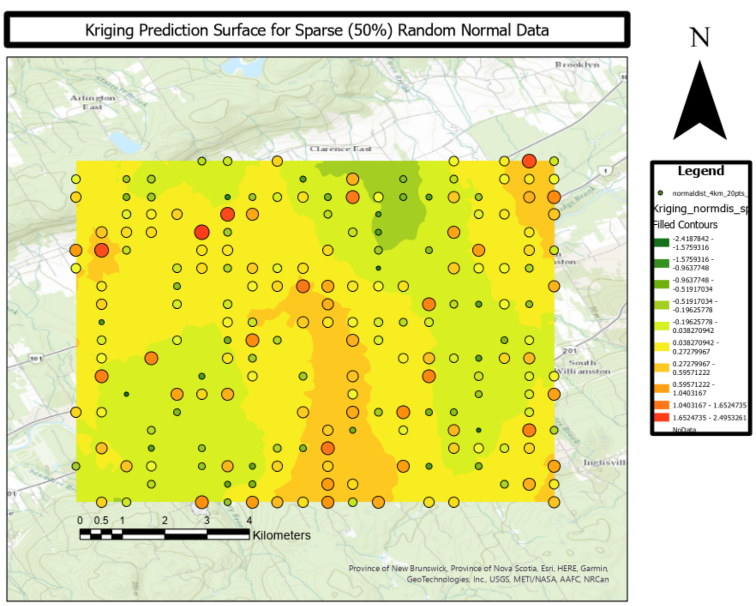


Figure 38: Ordinary Kriging sampling (1/2) of a set of random normally distributed data. Much of the underlying data has been removed and thus the predictive surface must be accompanied by a standard error map to give us an indication of how certain the algorithm is at points of generating the correct value.

4.0.2 Kriging Results - Bimodal Gaussian Distribution

Similarily, we display Kriging prediction surfaces for bimodal Gaussian data with decreasing levels of random sampling. These results are simpler to interpret given that we fabricated the underlying data distribution. It is impressive to see the Kriging algorithm is able to approximate the bimodal shape of the distribution. Note that the bimodality orientation, as well as the shape of the peaks and the general decay has been extracted quite well. Again, it is important to note the sampling of the data, given that generally we expect that the less data we have the, worse our prediction surface will be.

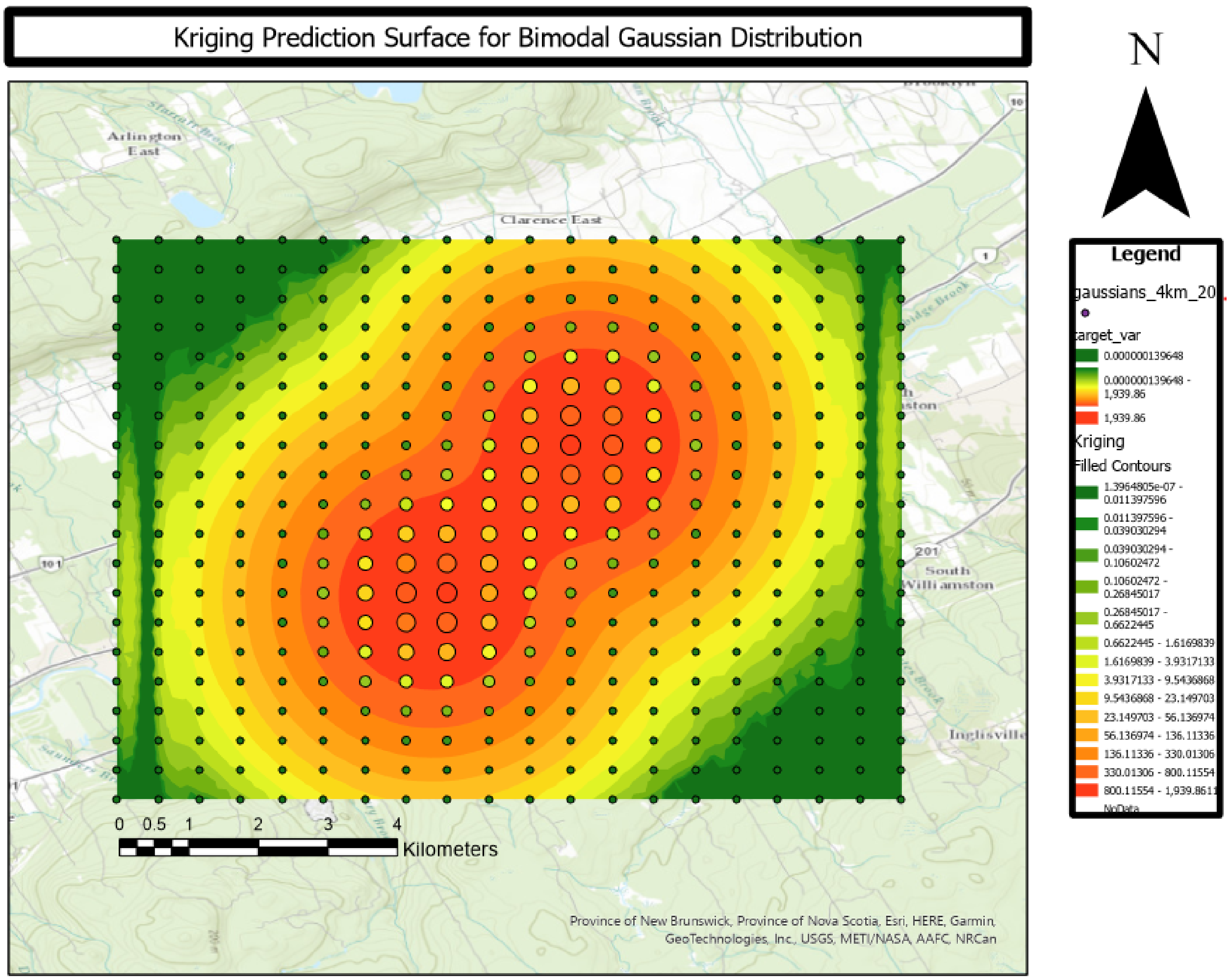


Figure 39: A bimodal distribution is seen through the Kriging prediction surface for the dual Gaussian data.

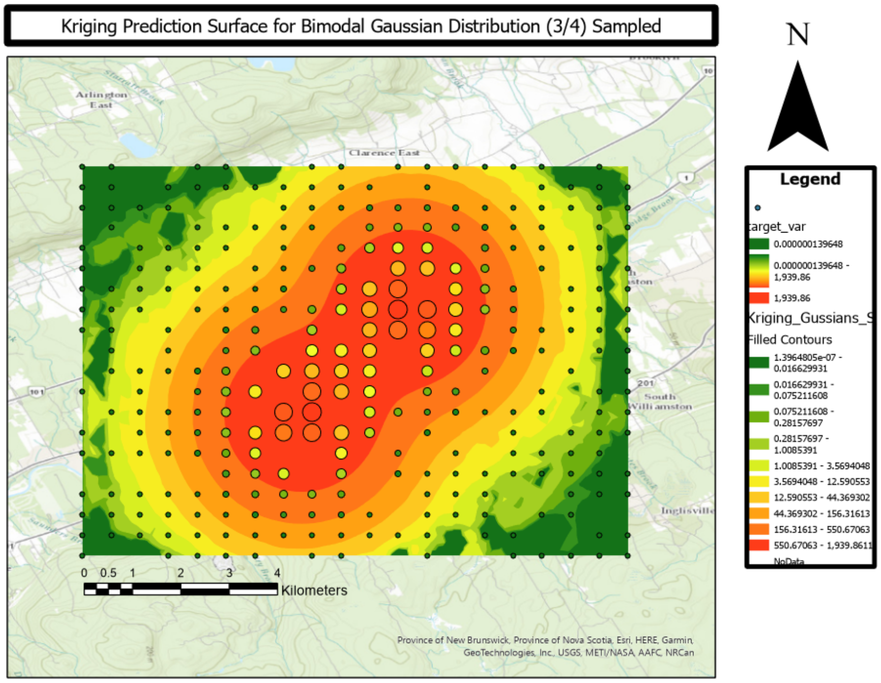


Figure 40: The data in this prediction map retained 3/4 of the original data to demonstrate the effects of sparsity in our data. In general, the inner distribution close to the Gaussian peaks is smooth and well defined, however there is plenty of noise generated in the edges of the predicted surface.

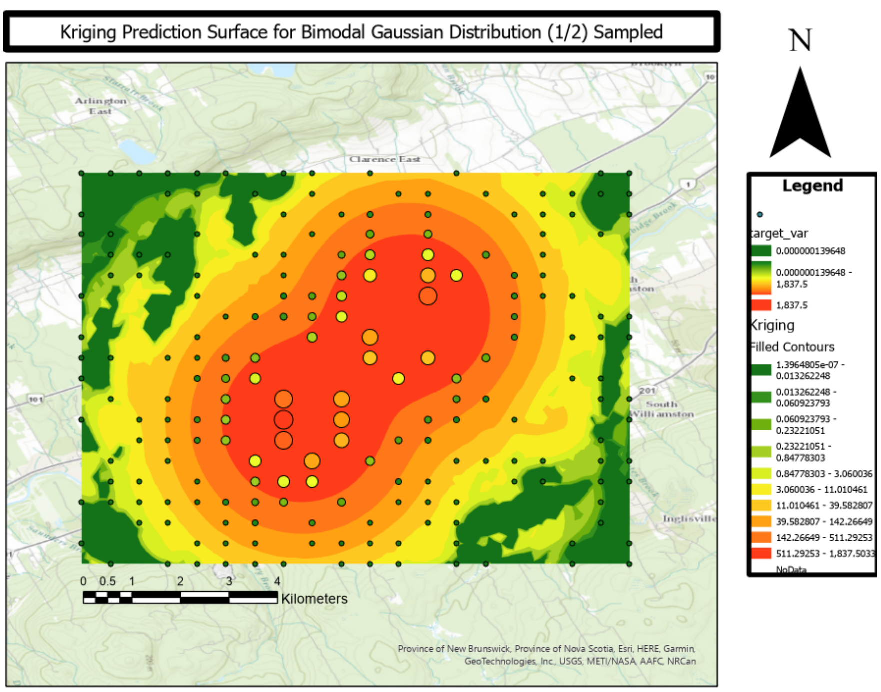


Figure 41: A prediction map with half of the original data is used to further demonstrate the effects of sparsity in our data. Noise has begun to take part inside the inner layers close to the Gaussian peaks.

4.1 Kriging Results - Error maps

The following section studies the error accuracy generated by each Kriging interpolation. Each error map generates a 2-D error surface indicating low error (white) to high error (dark red). We can see the effects of data sparsity influencing the error surfaces. In general, the more regular our data collection is, the more accurate the Kriging prediction will be.

4.1.1 Error Map - Random Normal Distribution with 1/2 Sampling

The following is the error map for the random normal distribution with a 50% sampling of the underlying data. The error seems to be centralized to the center of the image. This is likely to be attributed to having more neighbours to sample from compared to the edges of the surface.

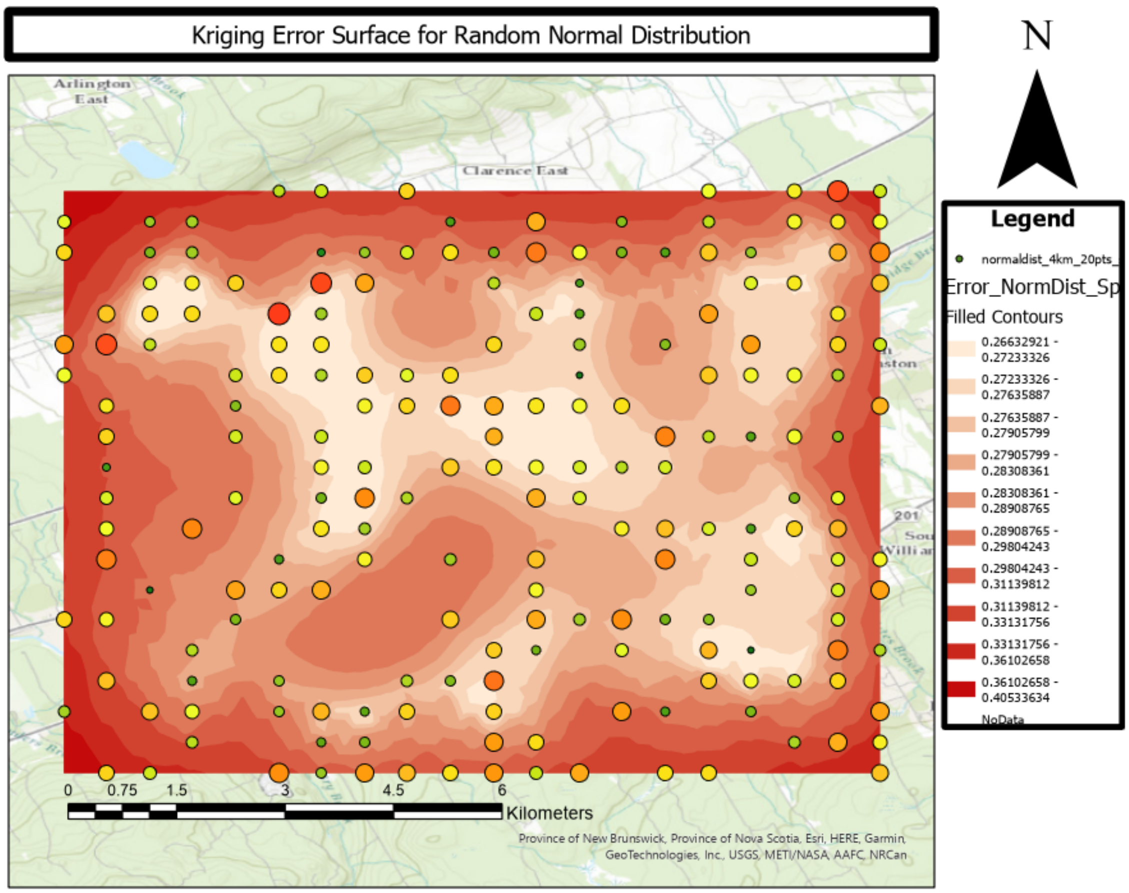


Figure 42: A Kriging error surface for (1/2) sampled random normal data. Error is maximal at the edges of the surface, while error is minimized where the most data samples are provided.

4.1.2 Error Map - Bimodal Gaussian Distribution

The next few images represent Kriging error surfaces for the binomial Gaussian distribution with decreasing sampling. The first image demonstrates an error surface with the entire data set provided, the following images represent the same underlying data but with 75% and 50% data sampling. It is interesting to note the uniform error distribution given the complete data set and evenly sampled points. All models seem to produce the highest errors near the edges of the sampling area.

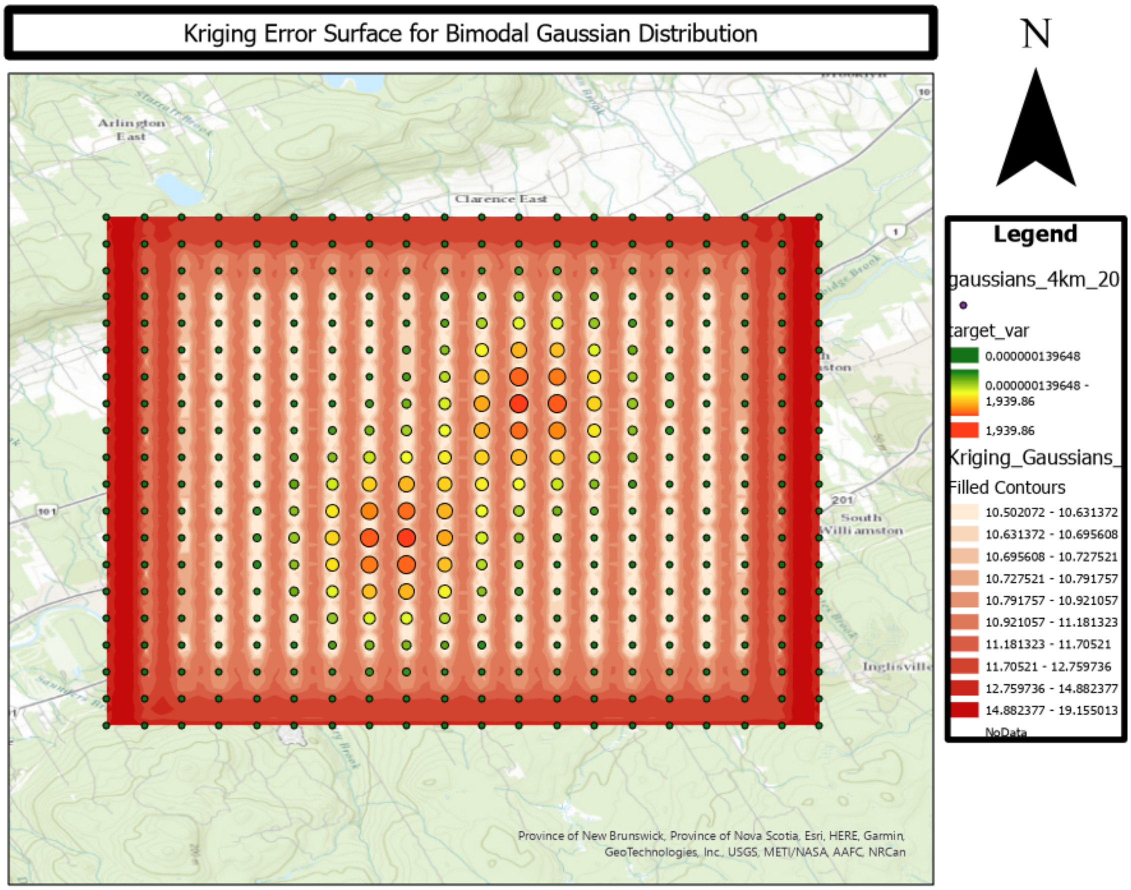


Figure 43: A Kriging error surface for a complete data set of bimodial Gaussian data. There is an uniform error distribution near the center of the image, with maximal error at the edges of the bounding box.

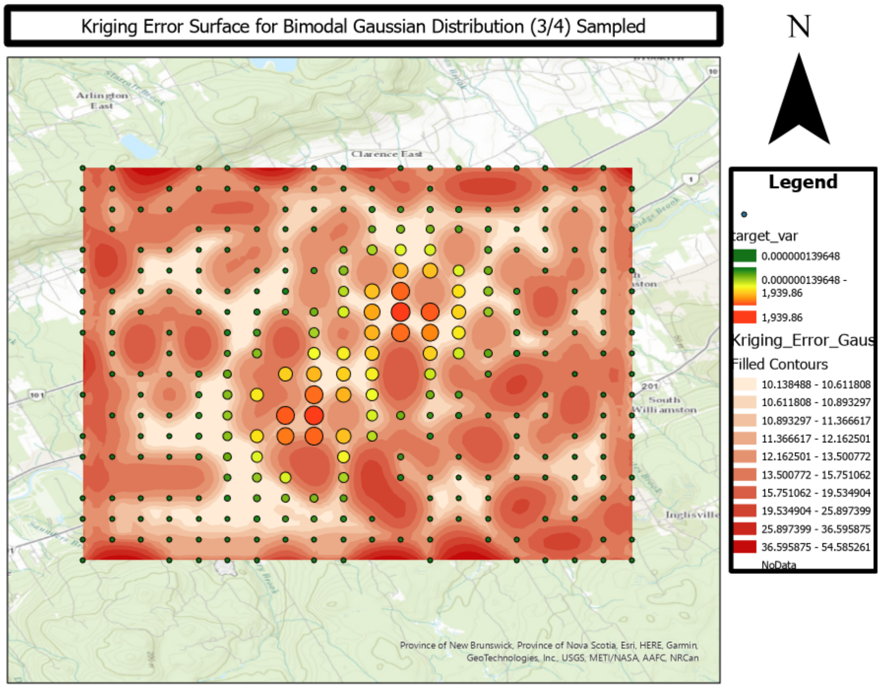


Figure 44: A Kriging error surface for a (3/4) sampled data set of bimodal Gaussian data. Error appears to be minimal in regions where there are high concentrations of labeled points, and maximal where there is missing data.

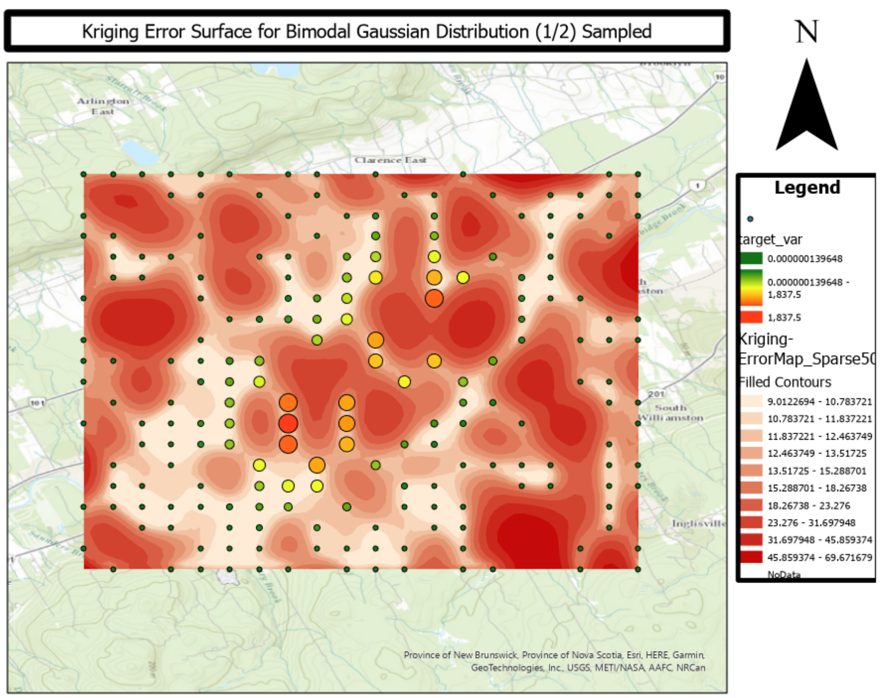


Figure 45: A Kriging error surface for a (1/2) sampled data set of bimodal Gaussian data. Error appears to be minimal in regions where there are high concentrations of labeled points, and maximal where there is missing data.

4.1.3 Probabiity Map - Bimodal Gaussian Distribution

We have included probability maps for the three bimodal Gaussian data sets that each represents the likelihood that each pixel value will exceed a predefined value. For each graph, the mean of the data has been selected.

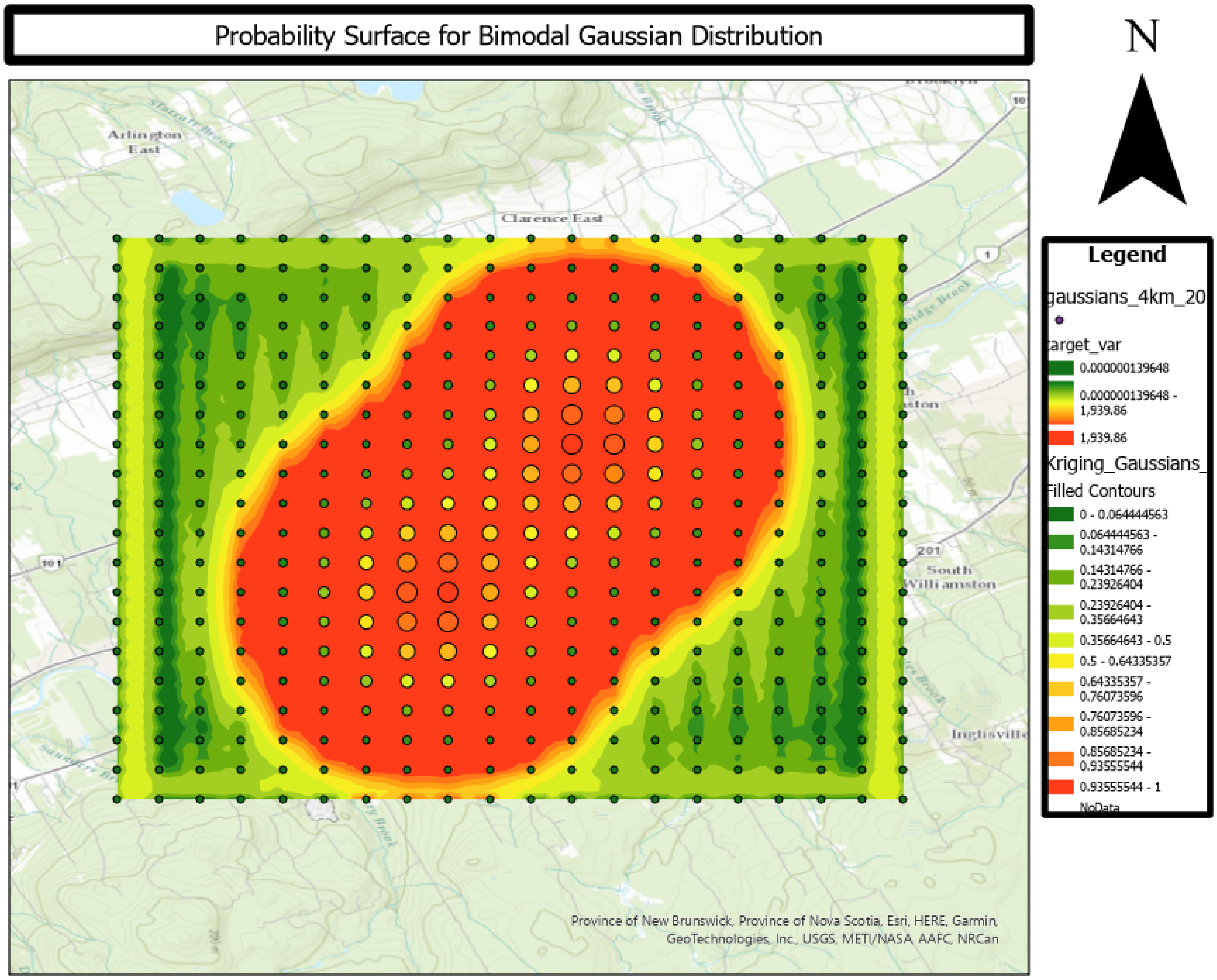


Figure 46: A Kriging probability map displaying the likelihood that a point on the map will have a value higher than the mean. This map represents the entire data set. We can see that due to the symmetry of the underlying function, the probability distribution map is also symmetric.

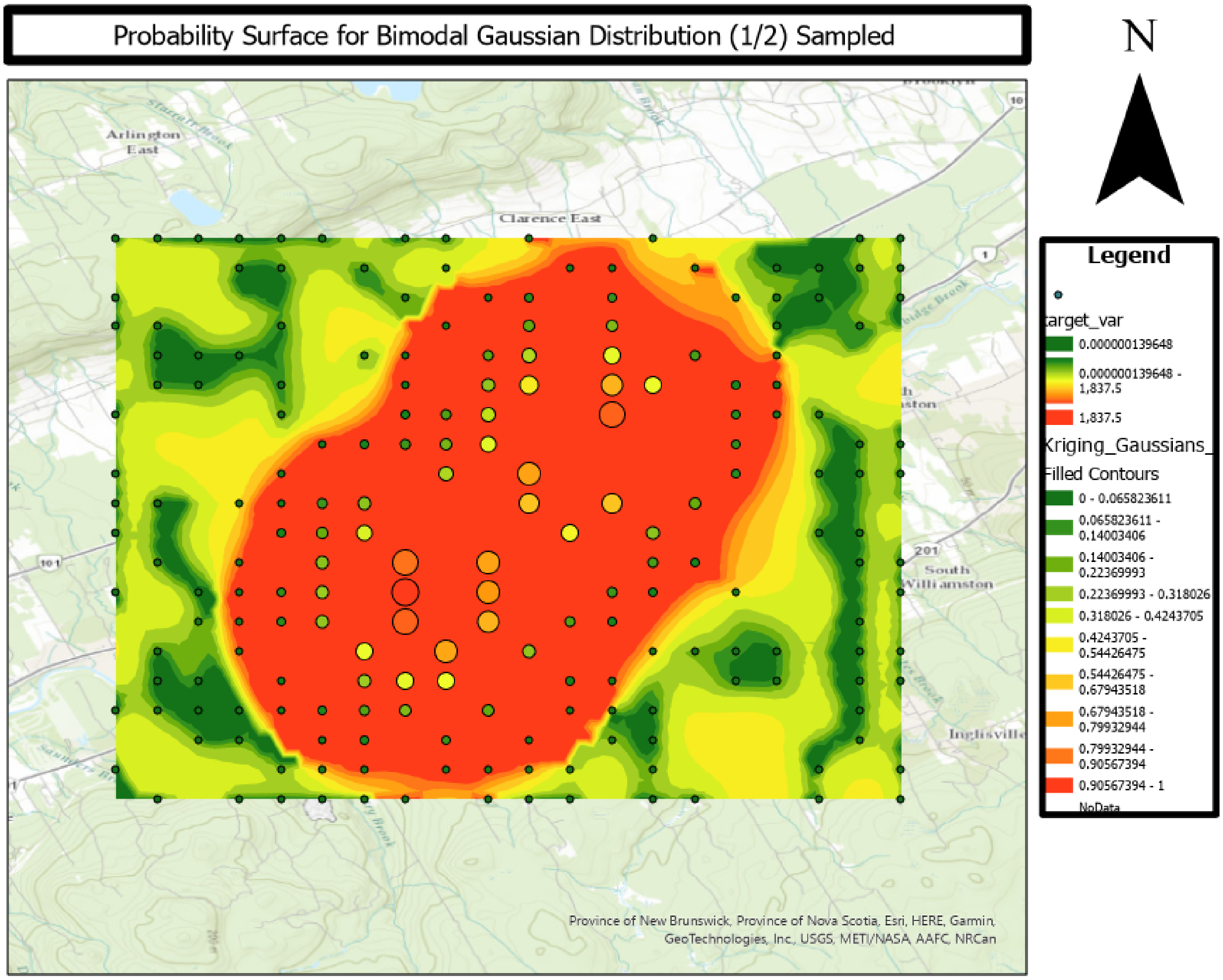


Figure 47: A Kriging probability map displaying the likelihood that a point on the map will have a value higher than the mean. This probability map was created with a random sampling of (1/2) of the underlying data. Due to the irregularity of the sampled points, the symmetry of this map has been broken. The most accurate results are from regions which have close measured samples.

5 Methodology

5.1 Parameter Fitting

We will take a look at the optimal parameters for Kriging the bimodal Gaussian distribution. Ordinary Kriging is available on ArcGIS Pro through the Geostatistical Wizard. and its use will be described in detail going forward.

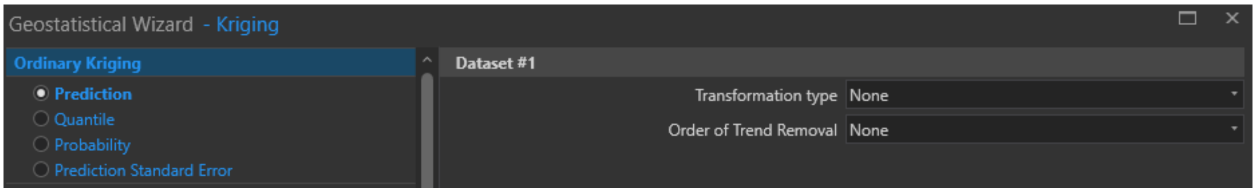


Figure 48: A GUI for ArcGIS Pro Geostatistical Wizard. The Kriging maps from the previous section were generated from selecting these options.

Once the user has selected a type of transformation to apply on the data set and the order of trend removal, the user is taken to the following screen which allows the varying of various statistical parameters to fit the model's semivariogram discussed in section 1.2 This report has been conducted with ordinary Kriging which assumes to transformation type or trend removal.

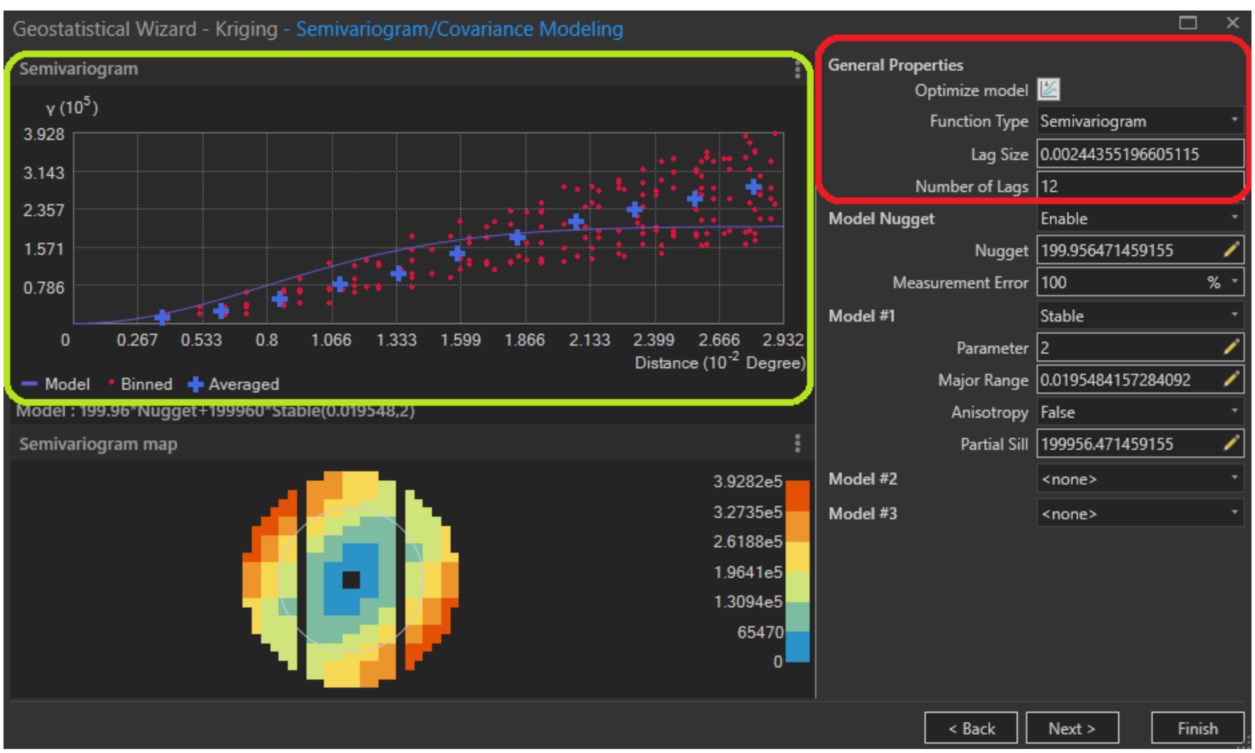


Figure 49: The Semivariogram/Covariance Modeling Menu. Parameters affecting the semivariogram fitting are shown here along with a visual representation of the semivariogram.

In the red box we are able to see the types of functions we can apply to our statistical model, for ordinary Kriging we will be fitting a semivariogram seen in the green box. The semivariogram has a number of lags parameter which allows the user to decide the distance size for which pairs of locations will be analyzed. This helps in reducing the total number of possible combinations to fit the model to [3] From this diagram the nugget, range, and partial sill can be tweaked.

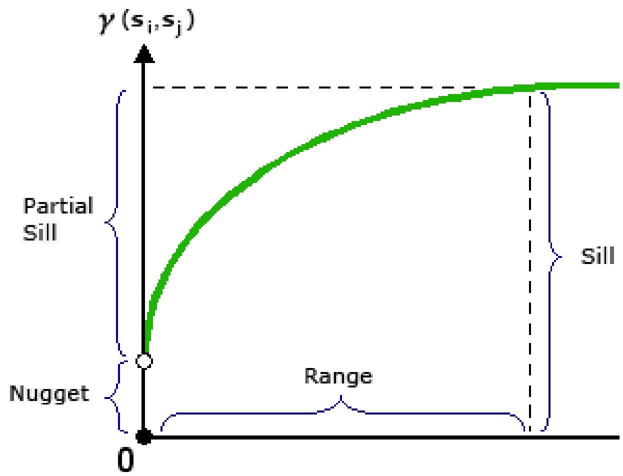


Figure 50: Fitting a theoretical model to a semivariogram[3].

In the next menu, the user is prompted to determine the searching neighborhood for the algorithm. Recall that ordinary Kriging is an algorithm which fits weights to points according to neighbours of each test point. If the neighborhood type selected is “Standard” then the user will have the option to determine the maximum and minimum number of neighbours to sample for the weight calculation. Note that increasing the maximum number of neighbors will generally lead to more accurate results but simultaneously lead to longer computation times, as well as the risk of overfitting the data. For this example, the author has chosen to select a smooth neighborhood type due to the continuous nature of the underlying data. A smoothing parameter can be fitted to select a “degree” of smoothing in the results.

The final screen is the cross validation menu which gives the statistical results of the model. Cross validation is used to give a measure of how well the model is at predicting unknown values. This is done by intentionally omitting a point, using the Kriging model to predict a value, and then compares the value that was predicted to the true value [9]. On the right hand side we are able to see summary statistics such as the number of points modeled, the mean value, RMS, mean standardized, RMS standardized, and average standard error. On the left hand side, a series of graphs are generated in addition to fitted lines to the data. We can access error graphs as well as the QQ plot and distribution graphs generated earlier in this document. To understand the true results of the statistical results requires a fair bit of statistical experience. However, generally the mean error should be close to 0, RMS standardized error close to 1, and a root mean square error and average standard error as small as possible[9].

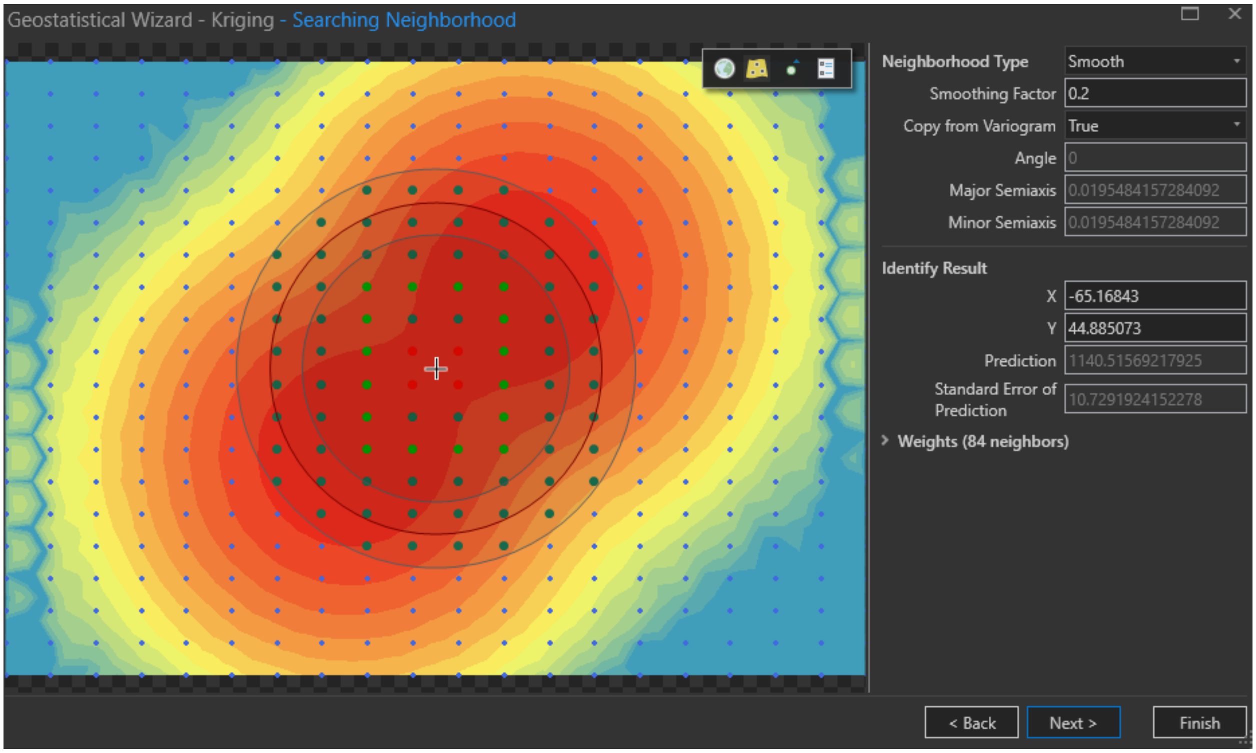


Figure 51: The searching neighbohood menu allows the user to change the number of neighbors parameter to influence Kriging results.

5.1.1 Error - Bimodal Gaussian

The error for our bimodial Gaussian using the Kriging technique had suboptimal results, which was expected given the tailed distribution of this data violating the normal distribution assumption for ordinary Kriging. The model resulted in a satisfactory mean error of -0.32 but a quite unsatisfactory RMS standardized error of 0.05. The RMS and average standard errors should always be minimized but we obtained values of 1.41 and 24.05 respectively. Although the model did not perform optimally,it still gave a decent image of the underlying data distribution. Further Kriging techniques should be considered to model this function more accurately.

5.1.2 Error - Random Normal Data

Ordinary Kriging resulted in a good fit to our random normal data (Figure 14) given that this data set was constructed optimally for this tool. Let's take a look at the error results.

From our error statistics we are able to see a mean error value of -0.01 and a RMS standardization error very close to unity (1.04). In addition from the generated graphs, we can see that the regression line for the error regression lies quite well in accordance to the trend in the data. Similarly, the generated data fits the QQ plot quite well among the reference line.

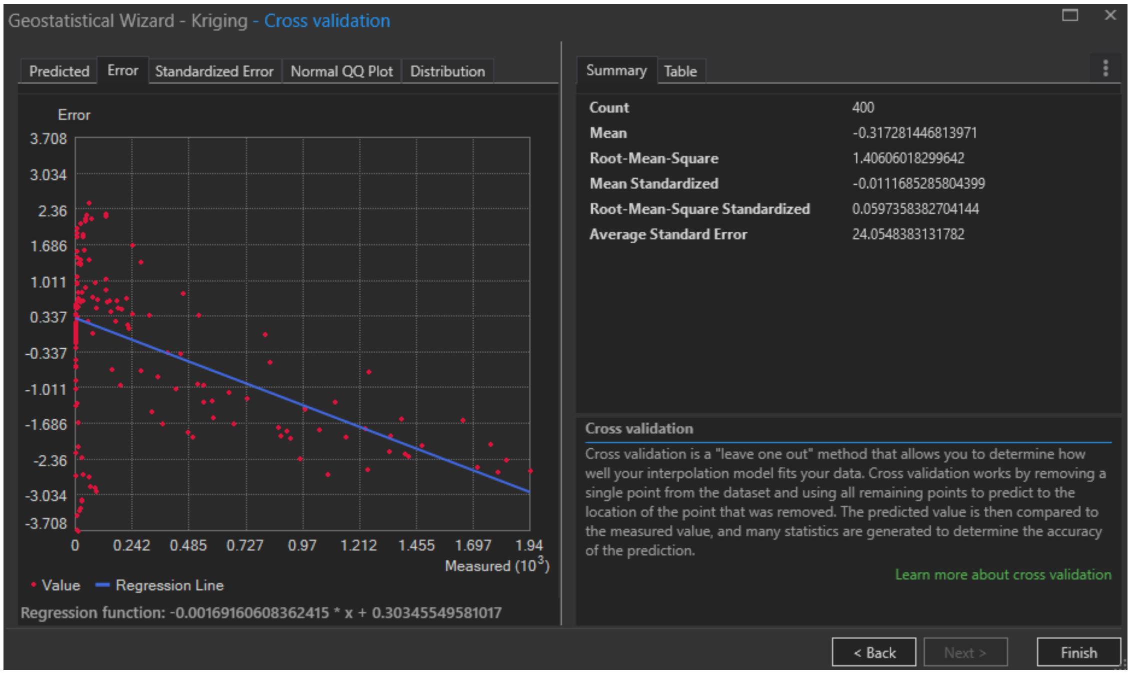


Figure 52: Cross validation step in ArcGIS Pro's Geostatistical Wizard for analyzing error in the Kriging results for the bimodal Gaussian distribution sample.

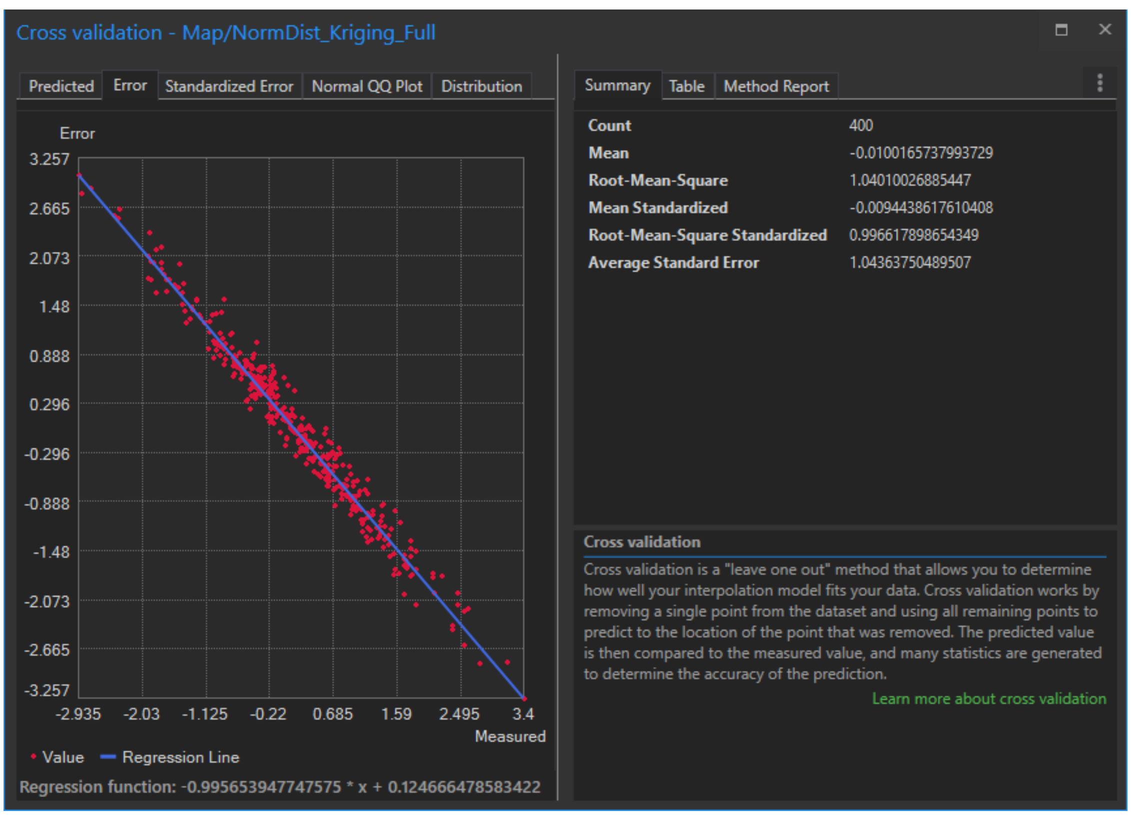


Figure 53: Error regression line and summary statistics for random normal data.

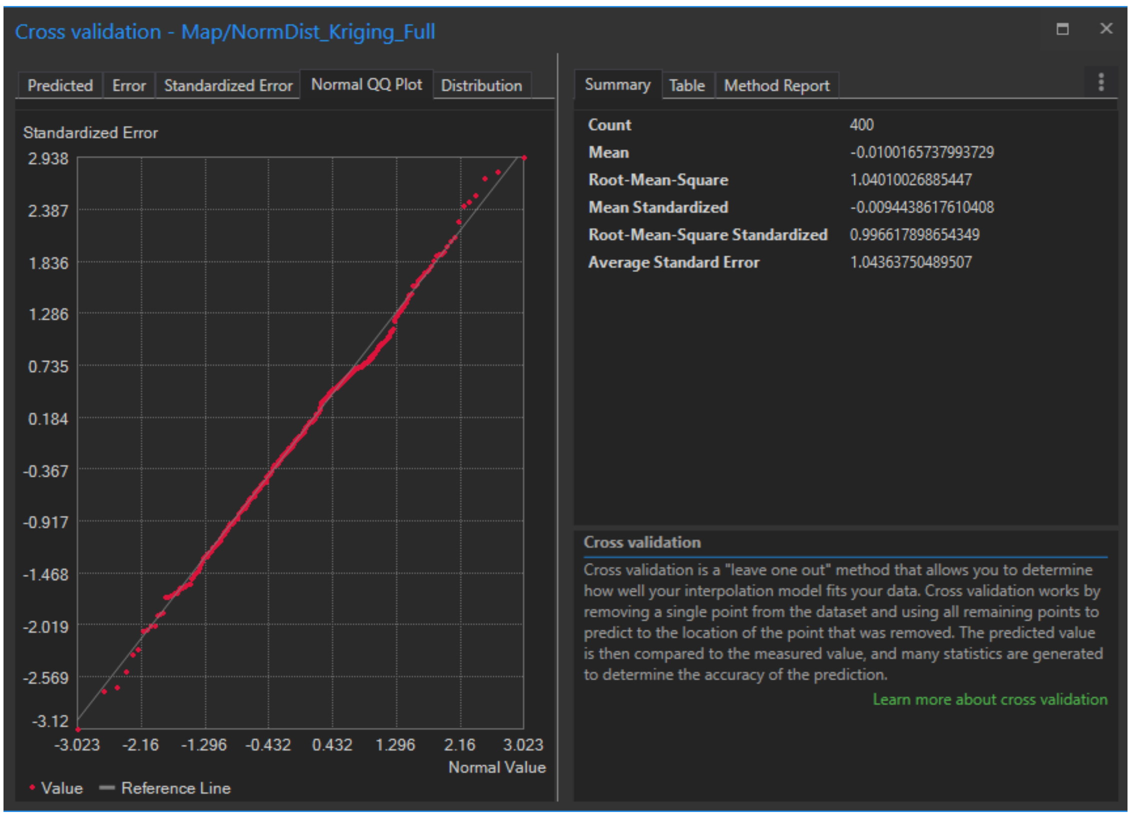
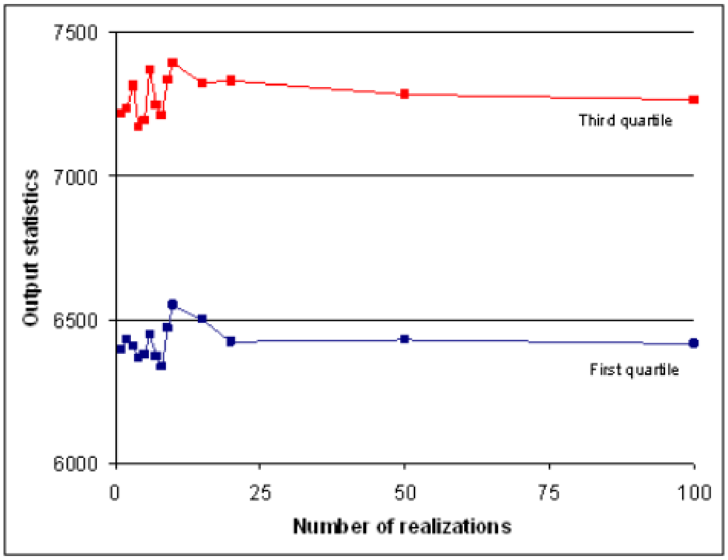


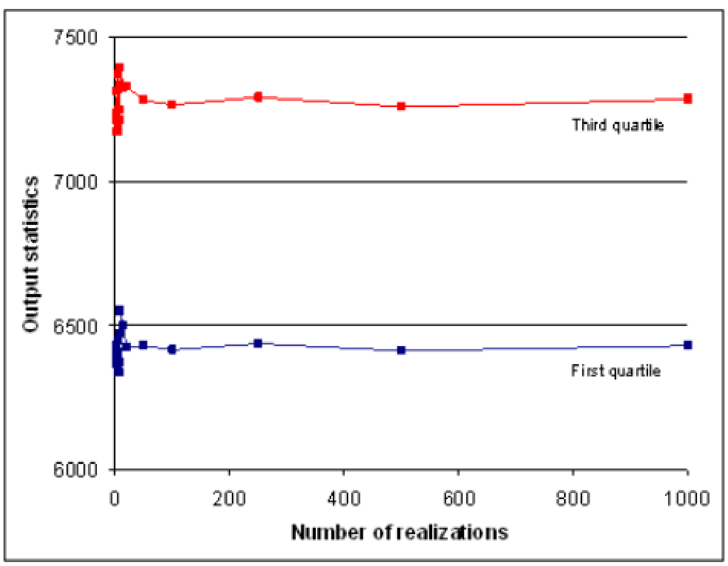
Figure 54: Cross validation step in ArcGIS Pro's Geostatistical Wizard for analyzing error in the Kriging results for random normal data.

6 Gaussian Geostatistical Simulation

A geostatistical simulation (GGS) is a refinement of Kriging where it does not provide the typical smoothed out average of the data. This is desirable given that most real-life processes are not well-behaved and thus have smooth distributions. A GGS produces a better representation of the local variability in the data because it will re add the lost variability from Kriging back into the generated surface [10]. GGS is dependent on a simple Kriging model with normally distributed data. This is because simple Kriging returns a generated estimate and variance that define the underlying distribution at every spatial location [10]. GGE takes this information to produce a "realization" of the random function (the unknown, sampled surface) [10]. In practice GGS will produce different realizations which can be stacked as output layers. Collecting a large number of these representations offer a way to study the uncertainty for the unsampled locations in space[10]. GGS emphasizes the uncertainty for a decision and risk, as opposed to producing a single prediction surface with the best unbiased prediction at each un sampled location (ordinary Kriging)[10]. In short, by iterating and thus, generating the random function (sampled surface), and stacking the output results, this allows the user to gather statistics from each resulting layer. By generating many realizations (often 100-1000), the user can compare the range of statistics gathered from the entire set of realizations. What the user should expect to see are statistical values to show some initial degree of variability, but as the number of iterations take tend towards large numbers, converge towards a fixed value[10].



Effect of the number of simulations on output parameter values; graph of the first 100 simulations



Effect of the number of simulations on output parameter values; graph of 1,000 simulations

Figure 55: Stabilization of statistical values over the sampling of many realizations. [10]

6.1 Settings and Parameters

We performed GGS on random normal data that has been sampled to retain 1/2 of the total data. The ordinary Kriging surface looks as follows:

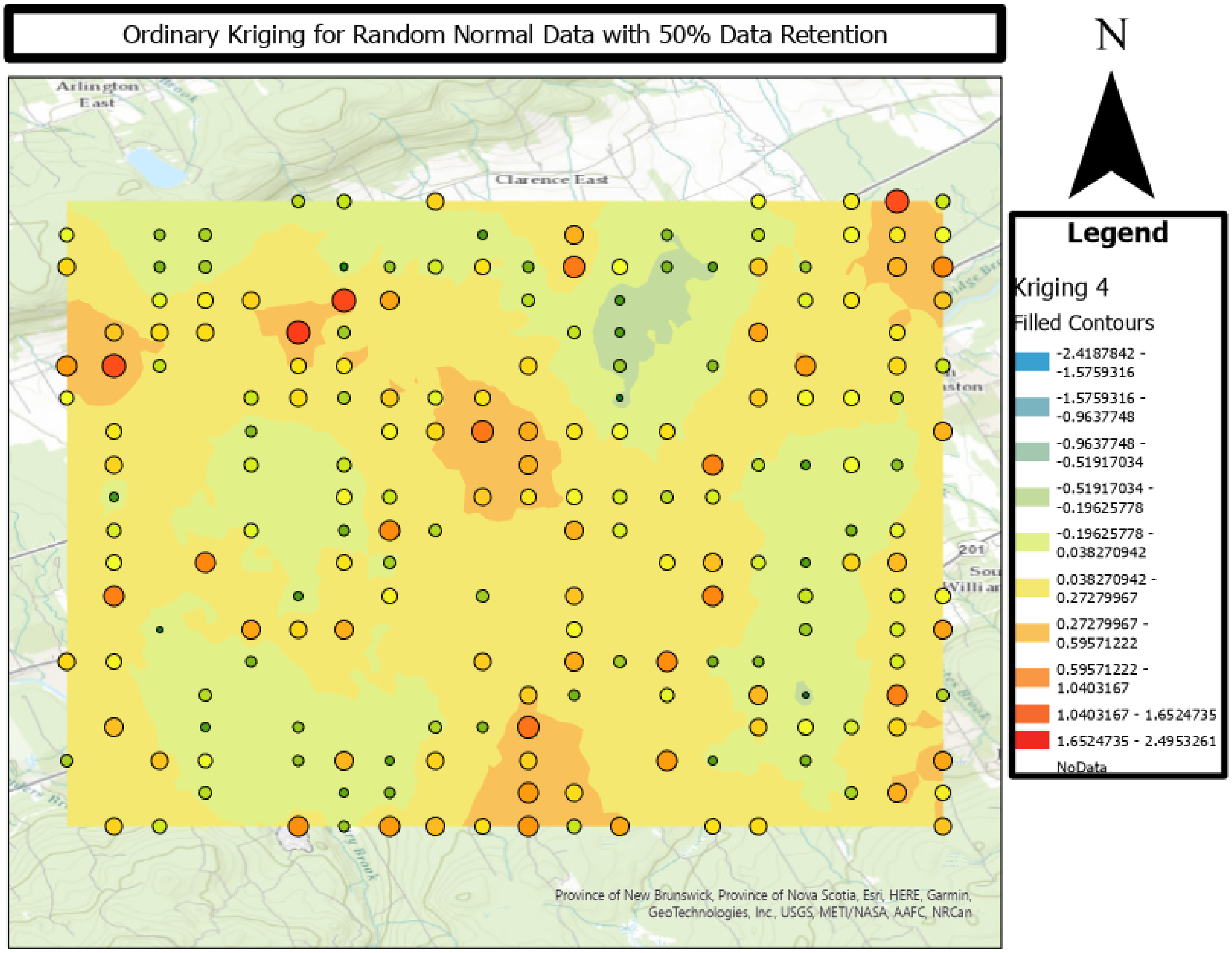


Figure 56: Ordinary Kriging prediction surface for random normal data sampled to retain (1/2) of the original data.

Note that the Gaussian geostatistical Simulations tool in ArcGIS Pro takes as input an Ordinary Kriging prediction surface to operate on.

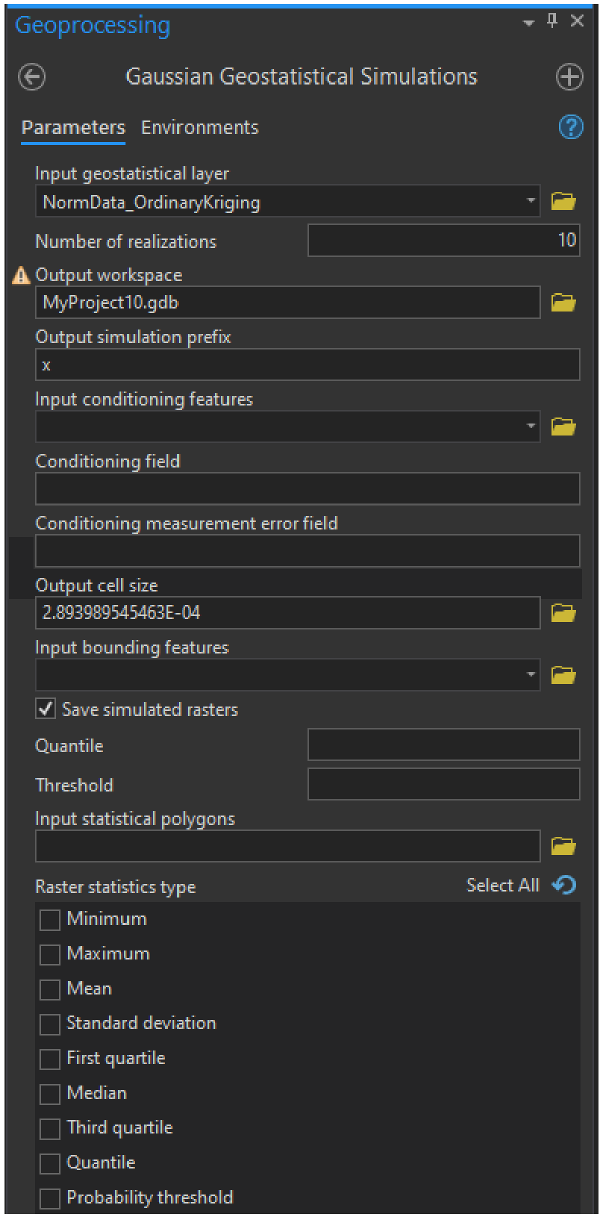


Figure 57: Gaussian Geostatistical Simulations tool in ArcGIS Pro.

The number of realizations parameter allows the user to select how many different interpretations to generate. A few sample generations are shown below.

6.2 Series

6.2.1 Mean Realization

The output yields the minimum and maximum values generated at each cell including the max, min, mean, standard deviation, and quantiles[10]. For the realization of the mean, the mean of all the cells in all the realizations that fall within the polygon are calculated. We display the realization, as well as the distribution of generated data. We confirm that the simulation output lies in a reasonable range similar to that of the original data (mean) as well as a normal distribution (bell shape) representative of the original random normal distribution.

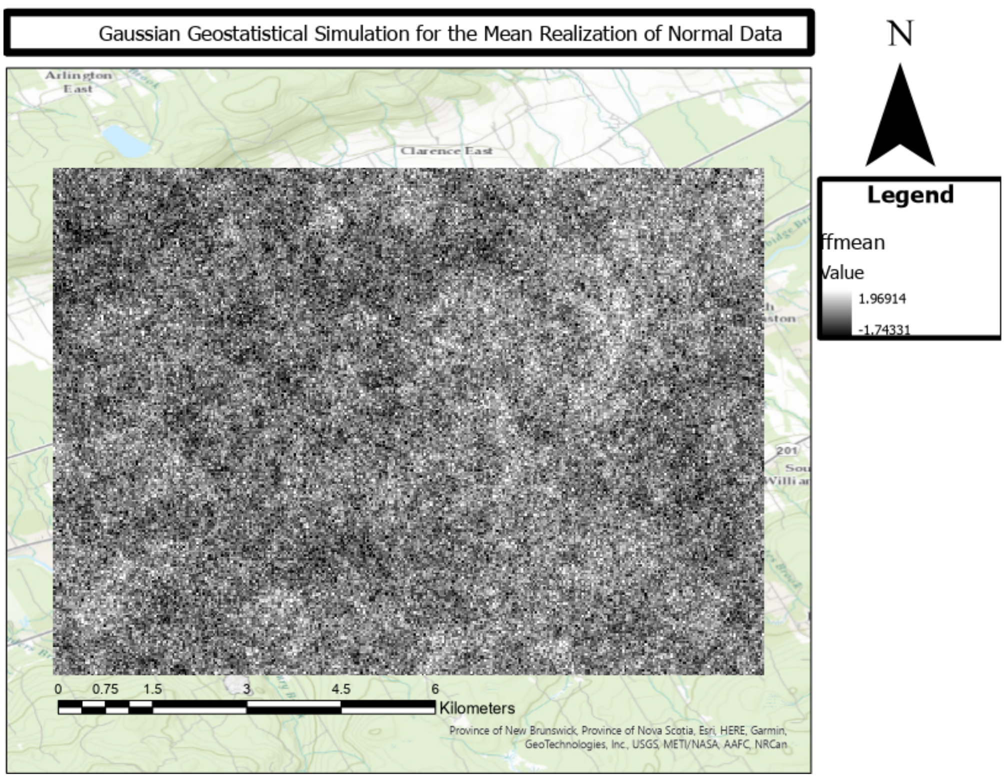


Figure 58: Realization of the mean given a random normal data ordinary Kriging prediction surface.

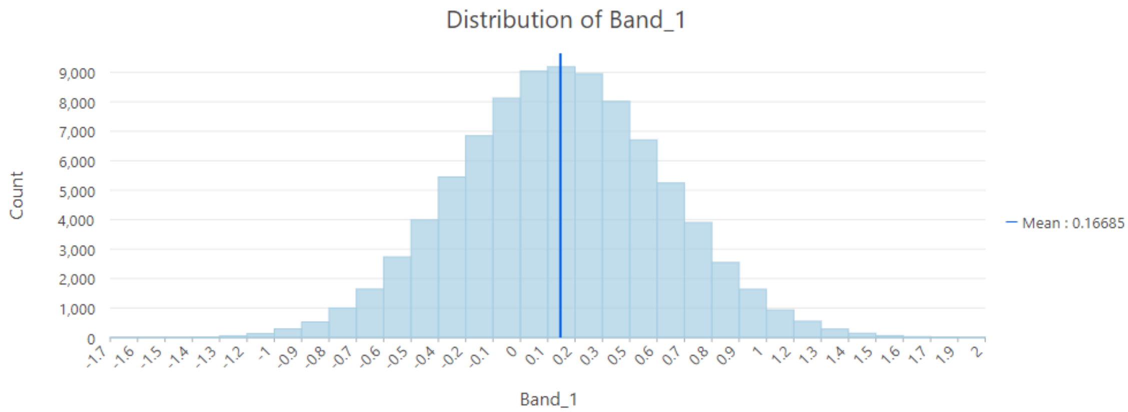


Figure 59: A bell curve shape in the distribution of the simulated data.

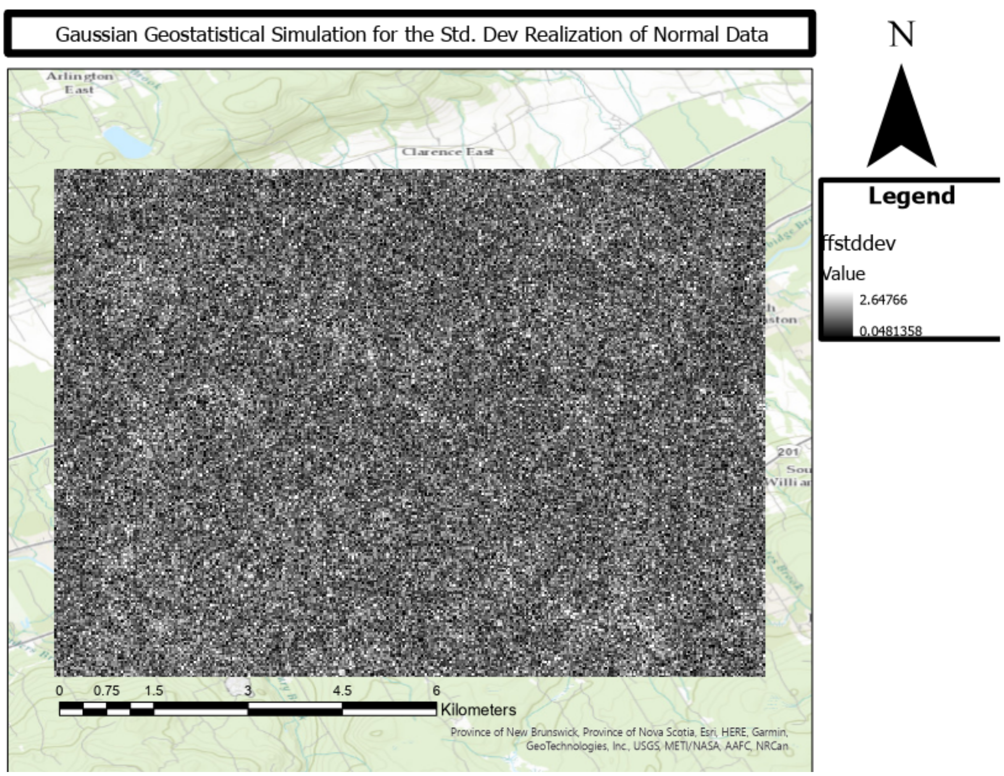
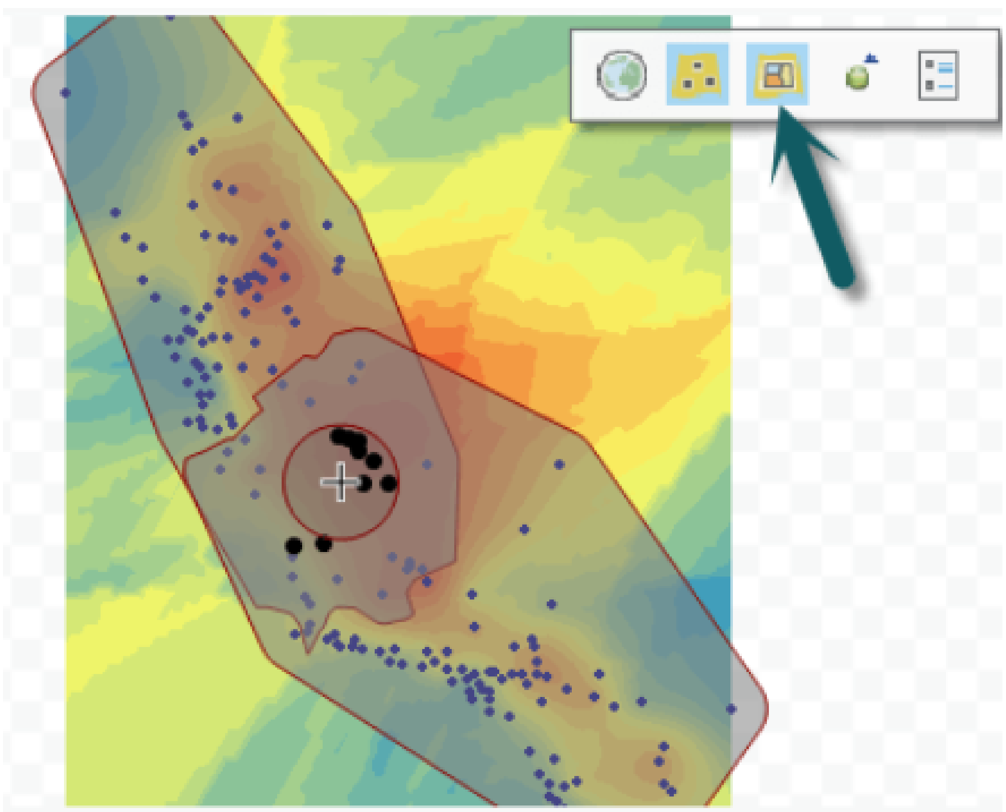


Figure 60: Realization of standard deviation given a random normal data ordinary Kriging prediction surface.

7 Empirical Bayesian Kriging Regression Prediction

Empirical Bayesian Kriging (EBK) is a tool that automates most of the difficult aspects of building a Kriging model [11]. Many of the exercises which we covered in this document requires the user to adjust parameters to create a valid semivariogram to create various Kriging models. EBK automatically does these taks for the user in order to get good results automatically. The difference between EBK and Ordinary Kriging is that EBK accounts for the error by estimating the underlying semivariogram [11]. The idea is to create a large number of semivariograms and plot them together, forming an empirical distribution of semivariograms. These semivariograms are an estimate of the true semivariogram. For each prediction location, the prediction is generated by using a new empirical semivariogram associated to the semivariograms of the points in the neighborhood. The goal is to use the collection of all semivariograms of regions with more neighbours to have more influence on the predicted value[11]. The following image from ESRI demonstrates this procedure for one region of interest. We will go forward and apply EBK to some of the data sets in this document.



Predictions are generated from neighboring subsets.

Figure 61: A prediction location at the center of the crosshairs define a search neighborhood which determine overlapping polygons that contain the subset of points used to generate the prediction[11].

7.1 EBK - Periodic Data

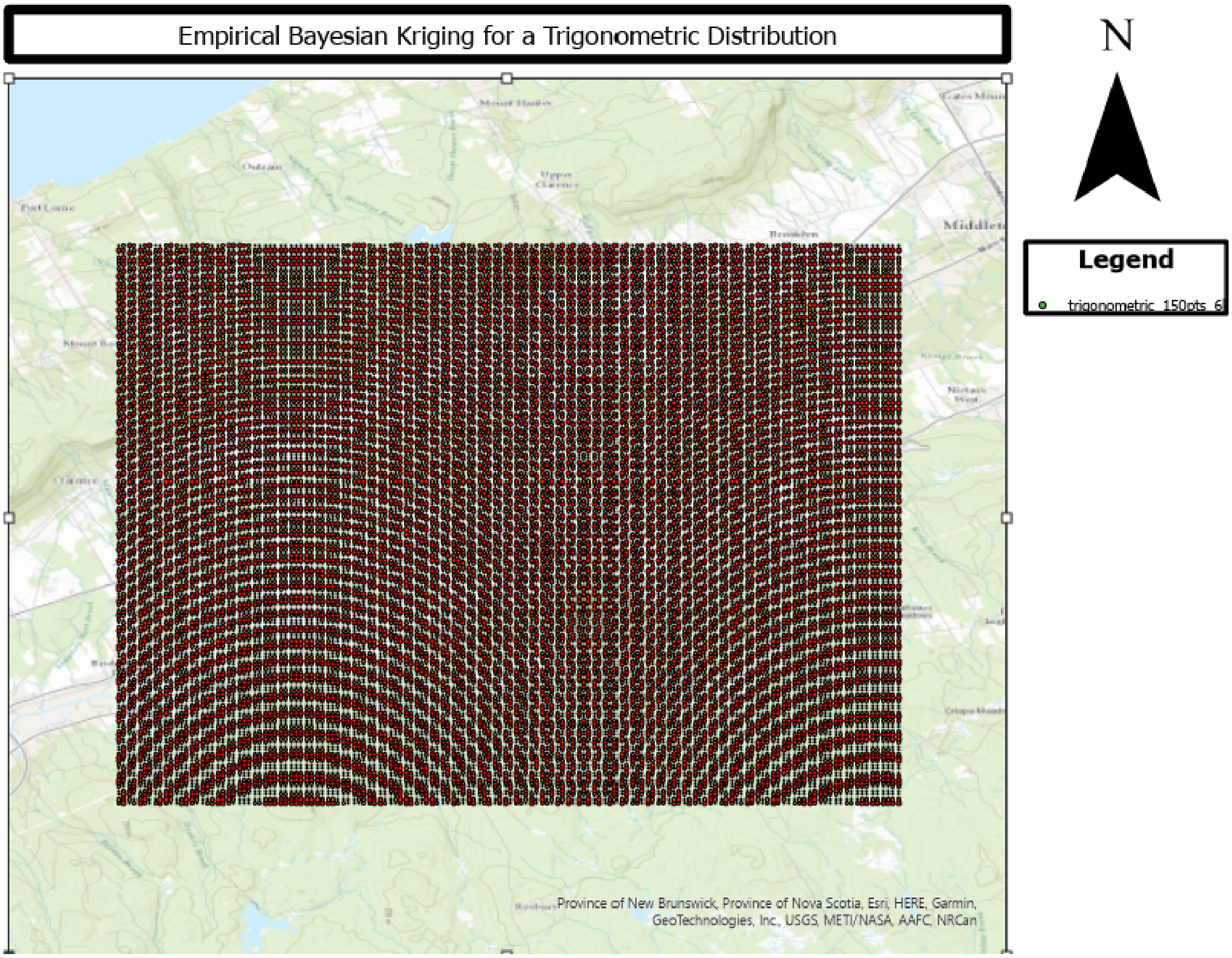


Figure 62: Trigonometric underlying data. Complex periodic patterns appearing like wave sources can be seen near the bottom of the bounding box.

7.2 Output Surfaces

As we can see, EBK was able to pick up on the sources of the wave-like patterns. Regions of wave interference demonstrate challenges for the algorithm to properly predict values. EBK surfaces for the random normal distribution, as well as bimodal distributions have been included for completeness.

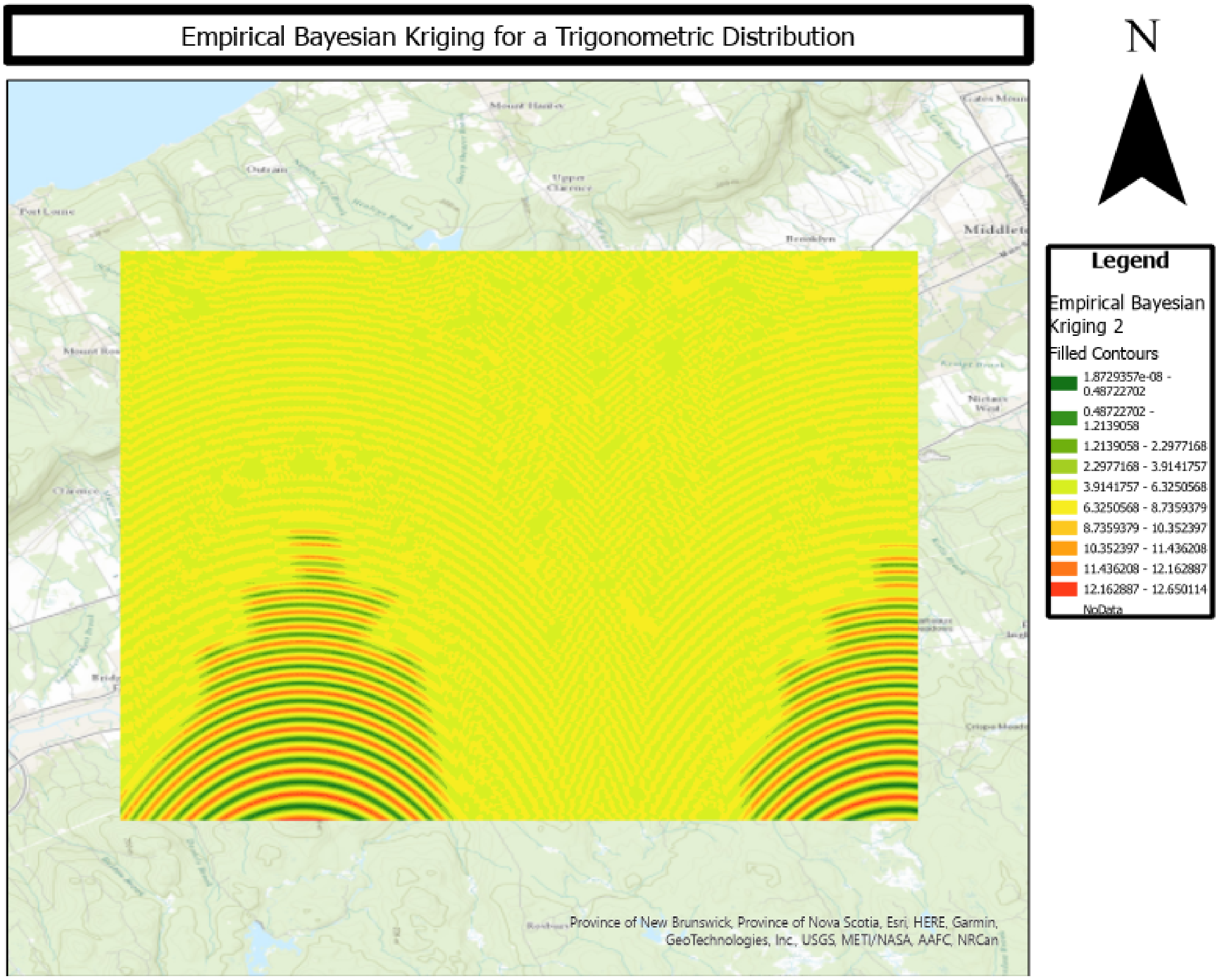


Figure 63: EBK prediction surface for the periodic data. Some periodic behaviour has been picked up by the algorithm. It appears to struggle with regions of wave interference.

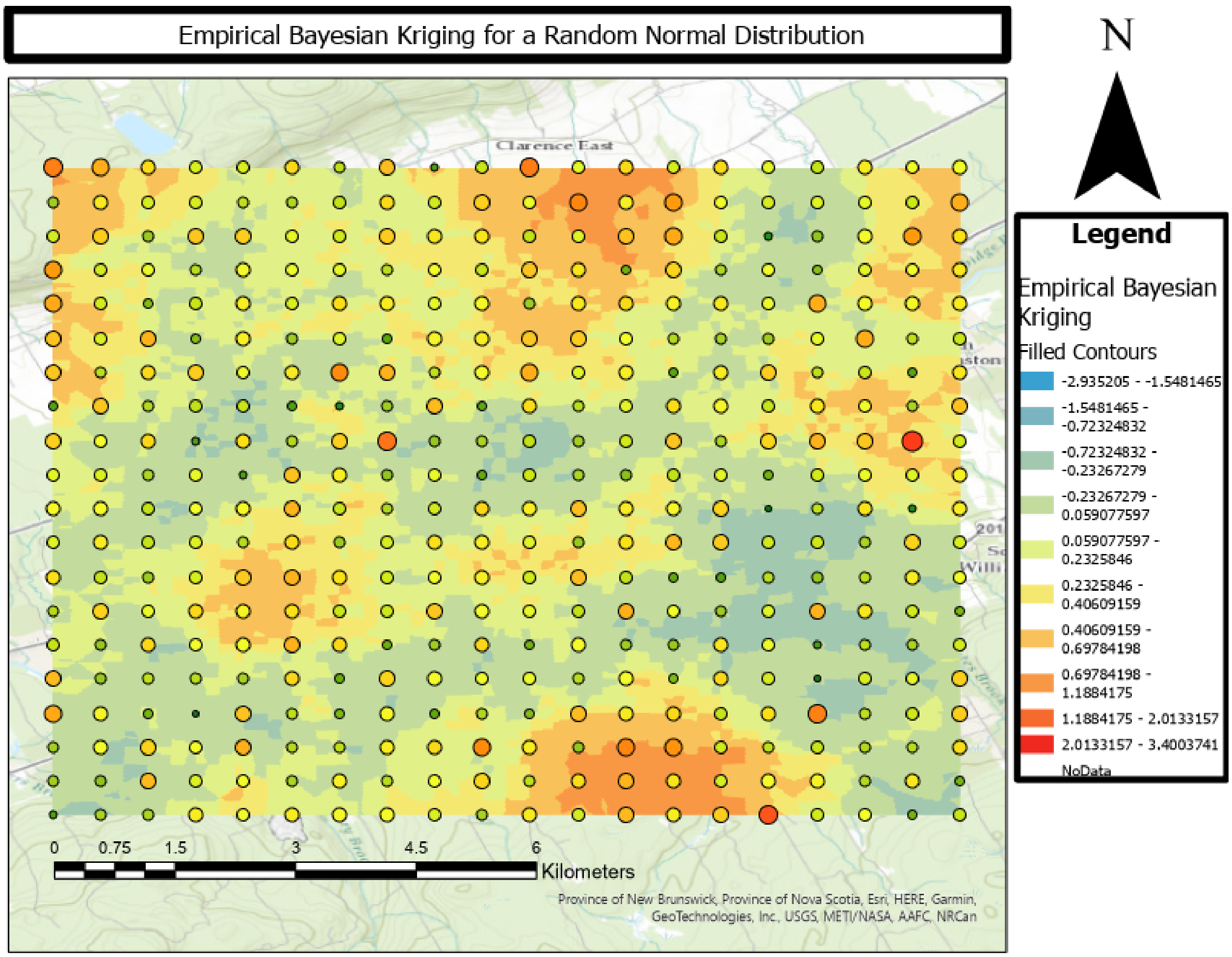


Figure 64: EBK on the random normal data set with no sparsity. EBK was able to yield this prediction surface with very little input from the user.

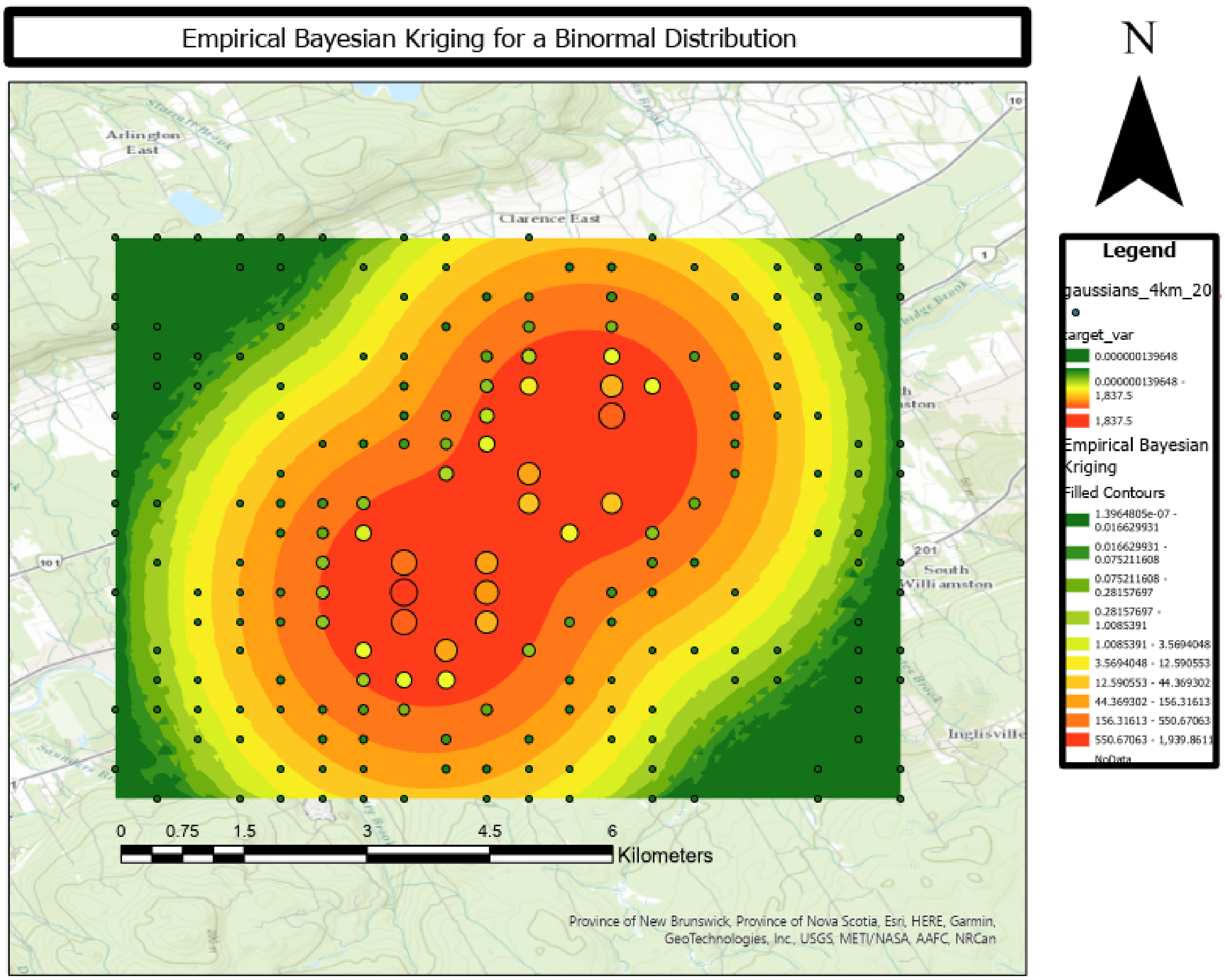


Figure 65: EBK on the Binormal Distribution. The inner region with two Gaussian peaks show very little noise. The noise grows as the function values decay near the edges of the bounding box.

8 Conclusion

We have analyzed various Kriging tools and techniques in ArcGIS Pro as well as ArcMap on a wide range of underlying data. Synthetic data generation in GeoPandas is useful in allowing the analyst to test the true limits of the computational algorithms provided by ESRI (or other open source algorithms!) and serve as a theoretical baseline for much more complex and stochastic natural processes. We have presented a non-exhaustive list of key concepts and parameters when fitting prediction surfaces using Kriging. Error interpretation plays a big role in learning which Kriging model is accurate and thus useful to an executive team which needs to run data-oriented operations. We observed that Ordinary Kriging, as well as more advanced techniques such as Gaussian Geostatistical Simulation, depend on many of the same underlying assumptions about the collected data in order to yield correct results. However, there are methods where certain assumptions about the data can be relaxed, such as in Empirical Bayesian Kriging which has been optimized and automated by ESRI in order to aid the analyst in creating Kriging prediction surfaces in a more streamline fashion. Nevertheless, plenty of work and study must still be applied in order to truly understand how to use geostatistical algorithms and their results.

9 References

[1] Fraczek, W (2007). Making Effective Use of Geostatistics. USDA Forest Service. https://www.esri.com/news/arcuser/0807/geostatistics.html

[2] GIS Geography (2022). What is Tobler's First Law of Geography?

https://gisgeography.com/tobler-first-law-of-geography/#:~:text=Everything%20is%

20related%20to%20everything,Tobler%20in%201969.

[3] How Kriging Works. ESRI ArcMap Documentations. https://desktop.arcgis.com/en/arcmap/10.3/tools/3d-analyst-toolbox/how-kriging-works.htm

[4] GIS Geography (2022). Kriging Interpolation – The Prediction Is Strong in this One. https://gisgeography.com/kriging-interpolation-prediction/

[5] Wikipedia Kriging: https://en.wikipedia.org/wiki/Kriging

[6] Bagheri, M (2017). Is it necessarily a normally distributed data for kriging? A case study: soil salinity map of Ghahab area, central Iran.

https://jdesert.ut.ac.ir/article_69125_2da835a19663f71e76d5aa11f9ff2a37.pdf

[6] ESRI. Trend Analysis https://desktop.arcgis.com/en/arcmap/latest/extensions/geostatistical-analyst/trend-analysis.htm#:~:text=The%

20Trend%20Analysis%20tool%20provides,stick%20in%20the%20z%2Ddimension.&text=

Polynomials%20are%20then%20fit%20through%20the%

20scatterplots%20on%20the%20projected%20planes.

[7] ESRI. Examining Global Trends Through Trend Analysis.

https://desktop.arcgis.com/en/arcmap/latest/extensions/geostatistical-analyst/

examining-global-trends-through-trend-analysis.htm

[8] ESRI. Examining local variation. https://desktop.arcgis.com/en/arcmap/latest/extensions/geostatistical-analyst/examining-local-variation.htm

[9] ESRI. ArGIS 9 - ArcGIS Geostatistical Analyst Tutorial http://webhelp.esri.com/arcgisdesktop/9.2/pdf/Geostatistical_Analyst_Tutorial.pdf

[10] How Gaussian Geostatistical Simmulations Works https://pro.arcgis.com/en/pro-app/latest/help/analysis/geostatistical-analyst/

how-gaussian-geostatistical-simulations-works.htm

[11] ESRI. What is Empirical Bayesian Kriging? https://pro.arcgis.com/en/pro-app/latest/help/analysis/geostatistical-analyst/what-is-empirical-bayesian-kriging-.htm

10 List of Figures

List of Figures