Using Data from Fossil Fuels to Phase Out Fossil Fuels

Owen McClure
10 May, 2021

Abstract

With the world looking for renewable energy resources to displace the various other methods of power generation such as oil and natural gas, there is a strong case to be made for geothermal power generation. Geothermal energy is a virtually renewable and clean source of energy. Also the skill-set required to build and maintain geothermal plants is very similar to the skill-set of oil and gas well workers which this type of energy generation is poised to replace. One of the major reasons geothermal energy is not more common in the United States is because there simply is not a whole lot of data to determine just how effective it could be. With that said, the push for clean energy has generated more research into the possibility of geothermal power. The Southern Methodist University has aggregated a massive data set of oil and gas wells across the country including temperature measurements of the bottom of each well. With this data I intend to expand my skills in the topic of statistical spatial analysis and determine how predictable the temperature at the bottom of a well can be given it's location. In order to be used as a tool for the construction of geothermal power production facilities.

Introduction

A general understanding of the way in which geothermal power is generated is necessary to be able to understand and work with the existing data about the heat in the Earth's crust. Geothermal power generation relies on harvesting the heat which rises up through the crust from deep within the planet. This form of energy is considered virtually unlimited, and does not produce any harmful by-products.

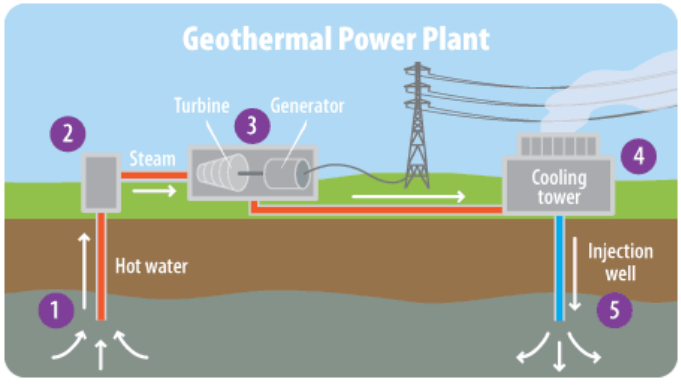

Figure 1: How a geothermal plant works. [5]

There are many different specific methods to be able to collect the energy from the crust, but Figure 1 shows a common setup for a geothermal power plant. It is important to note here that Figure 1 shows that water is turned to steam to drive a turbine. However, using water requires that the ground is more than 100°C. This constraint has severely limited the possible locations which are suitable for power plants in the past. Recently new methods have been developed which are not constrained by this 100°C limit. This can open up vastly more area which is suitable for geothermal energy. However, it can be generally assumed that higher ground temperatures will lead to a more profitable power plant.

Therefore, the largest roadblock to setting up and building a geothermal power generation infrastructure stems from the uncertainty about the temperatures beneath our feet. Since the 1970s many different methods have been used to model this uncertainty. Most of these studies focus on hydro thermal basins like the ones found in Yellowstone National park. This is because data about these areas is relatively easy to collect and the profitability of power generation is considered high [10]. This paper will instead focus on modeling heat flow in areas which have a large amount of oil and natural gas well drill sites. Although these areas may not be considered the most profitable locations for geothermal power, they will undoubtedly play a key role in reducing the world's dependency on fossil fuels.

Exploratory Data Analysis

The main data source in this paper is all the temperature measurements taken at the bottom of well drill sites in the US from the Southern Methodist University [1]. This data has been curated and compiled into a massive list of more than 40,000 wells across the United States and in the Gulf of Mexico. The first step to understanding this data was to visualize it. Importing the data into R and generating a histogram plot of the temperatures yields the result shown in Figure 2.

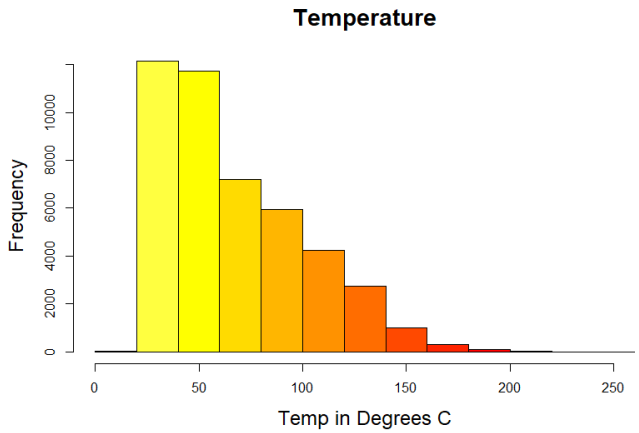

Figure 2: Histogram of observed temperatures.

The observed temperature data shown in Figure 2 has a mean value of 67.3°C and is heavily skewed to the left. Again the fact that most of the temperatures are not more then 100°C does not rule them out as candidates for geothermal power. Not only are new methods available which do not rely on boiling water, but these measurements can be used to predict higher temperatures that may occur at lower depths. The effect that the depth of the well has on the temperature is shown in Figure 3.

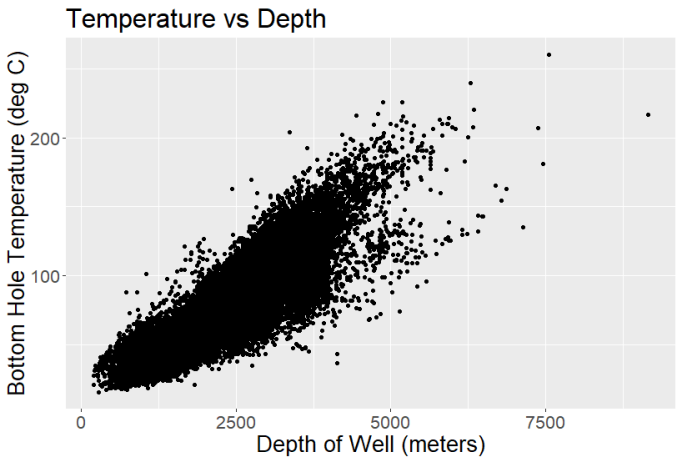

Figure 3: As depth increases so does the temperature.

In fact the measured temperature at the bottom of these wells is merely a metric used to calculate the amount of heat flowing towards the surface. The measured temperature data can not be considered clean either. Most of the

time these measurements are taken very soon after the hole was drilled. So this measured temperature does not always accurately reflect the equilibrium temperature (that is the temperature that would be observed if the well was never drilled)[14]. Entire papers have been written to solve this very problem. The widely regarded best method used to determine the equilibrium temperature measurement is to use a Horner plot [11]. Unfortunately this plot requires multiple measurements of the temperature over time, and the data from SMU only includes measurements at a single point in time. Another method to correct these temperatures was developed by William Harrison in his 1982 paper [6]. This 'Harrison Correction' is simply the second order polynomial function of depth in meters shown in Equation 1. This equation was found by fitting known equilibrium temperatures to observed temperatures from well drill sites in Oklahoma.

$$\Delta T = -16.51 + 0.018z - 2.34E10^{-6}z^2 \tag{1}$$

Here the ΔT is the change from the measured temperature to the equilibrium temperature, and the value of z is the depth of the well in meters. Since this data set contains the depths of each temperature observation, calculating the equilibrium temperatures is as straight forward as writing a simple script in Python which implements Equation 1.

Heat Flow Model

This temperature data is only the first step towards modeling the heat flow. In order for a true idea of what is going on beneath the crust be be established, we must know the surface heat flow. That is the movement of heat from the interior of earth to the surface. However, in order to calculate this it is necessary to know the average annual surface temperature of each of these locations. This new measurement, along with the depth and the equilibrium temperature can be used to calculate the heat flow using Equation 2.

$$\frac{dT}{dz} = \frac{T_{EQ} - T_S}{z} \tag{2}$$

Where T_{EQ} is the previously calculated equilibrium temperature, T_S is the average annual surface temperature and z is the depth of the hole. In order to implement this equation, the mean annual surface temperature was retrieved for each state from the National Centers for Environmental Information [12]. Then the gradient $\frac{dT}{dz}$ was calculated and added as a column to the data set.

$$Q_S = k(\frac{dT}{dz}) \tag{3}$$

This new data can then be combined with the thermal conductivity of the rock which is also provided by SMU. Combining these two measurements as shown in Equation 3 where k is the thermal conductivity (provided in the SMU

data set). Gives the true values for the heat flow towards the surface: Q_S. Which will be analyzed to model a continuous prediction of the heat flow value across the US.

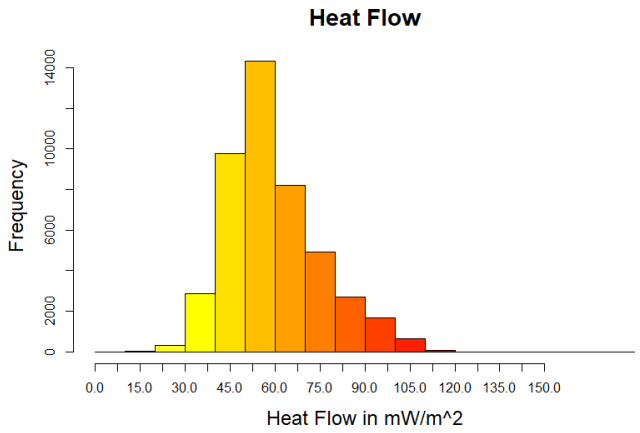

Figure 4: A Histogram of the calculated Heat Flow values.

Figure 4 shows the results of the previous calculations. Unlike the histogram of observed temperatures this distribution is centered more closely around its mean. Also as a reference the average heat flow value for Iceland (an area with huge geothermal resources) is around $100mW/m^2$ [2].

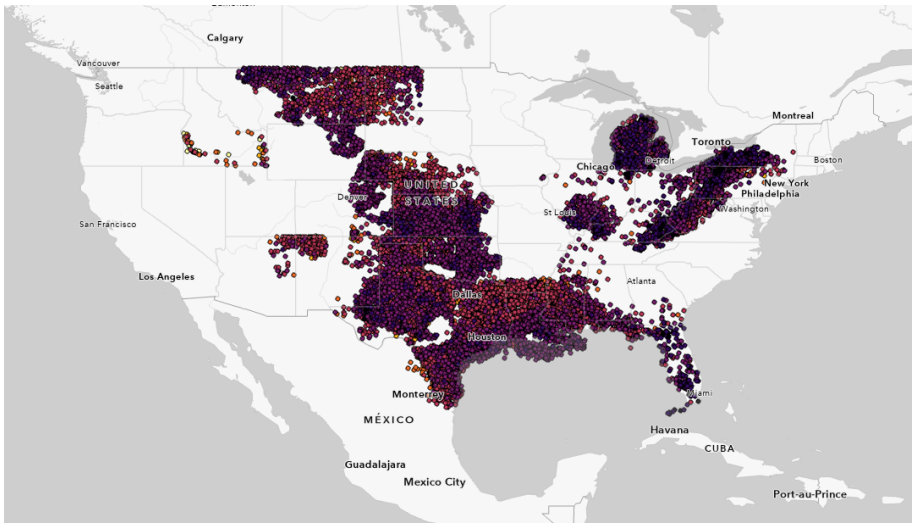

Figure 5: The Heat Flow values across the US, color coded by standard deviations from the mean.

Next, using ArcGIS Pro this data set was imported and overlaid onto a map to see is there are any general trends which relate to the geographical area which the data comes from. This led to an important discovery about the feasibility of generating a statistical model which covers the entire United States. As shown in Figure 5, not every state has well data than can be used. This means that only areas which are near to large grouping of wells can be used to accurately model heat flow. Uncertainty about the predicted heat flow value should increase in areas with fewer data points.

Analysis

Analyzing this data in order to predict the value of the heat flow at any given point presents some challenges. The main issue is that a method like kernel density estimation which would be to model an unknown distribution on a geographical surface will not be effective here. This is because the density of the observations found in the data set does not have any bearing on the heat flow. Because of this characteristic, a different approach is necessary to model the distribution.

Natural Neighbor Model

The Natural Neighbor Model is a good fit to model this data, because it is built for data which has varying densities [13]. The method of natural neighbors was developed by Robin Sibson in the 1980s and continues to be a useful way to interpolate discrete sets of points in space. This method uses a Voronoi tessellation which is a way to divide the the area of study (the entire US) into sets (or cells) where each set contains the points closest to a corresponding point in the data. For this data set it is important to use geographical distance, to account for the curvature of the earth. Using this metric of distance the tessellation shown in Figure 6 can be generated using most GIS software packages. This tessellation makes it possible to model the heat flow present in the earth's crust for arbitrary points which are not present in the data set. The Natural Neighbor Model is very simple to implement as shown in Equation 4.

$$\hat{Q}(\phi, \lambda) = \sum_{i=1}^{n} f(\phi_i, \lambda_i) \frac{A(\phi_i, \lambda_i)}{A(\phi, \lambda)} \tag{4}$$

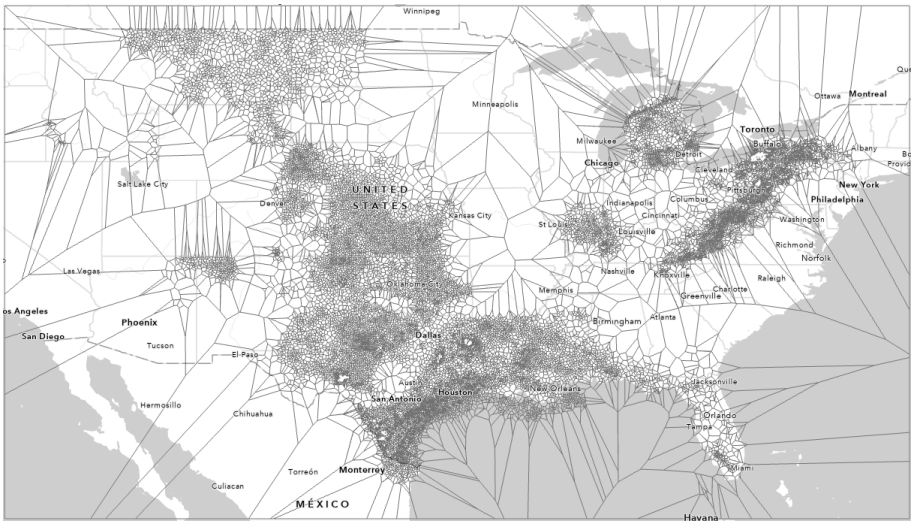

Figure 6: Voronoi diagram of the US, generated with ArcGIS Pro.

The function $\hat{Q}(\phi, \lambda)$ estimates the value of the heat flow at a given point in the United states, here ϕ and λ are used in to represent a longitude and latitude coordinate, respectively. The $f(\phi_i, \lambda_i)$ term is the known data about heat flow, and the remaining term $(\frac{A(\phi_i,\lambda_i)}{A(\phi,\lambda)})$ represents the weight given to the known data in order to calculate the new value. This is why the Voronoi tessellation is necessary. It can be helpful to think of the heat flow estimates at a given point as a weighted average of the points around it. These weights are calculated as a fraction of the area that a new cell would take from existing cells if the newly calculated point were to get its own cell. This concept is demonstrated in Figure 7. Using this method to generate the weights guarantees that for any given point the weights will sum to 1. This way bias that would occur from some points having more neighbors than others can be avoided.

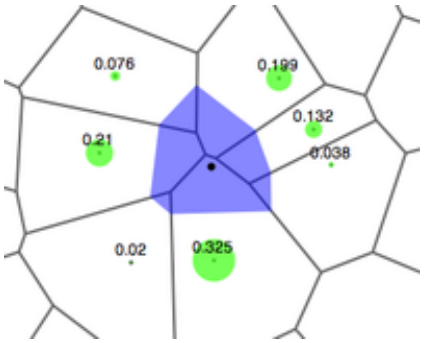

Figure 7: Weights calculation [15].

Putting all this together, a heat map of the entire region represented by the data can be generated to allow for estimates of heat flow at any arbitrary point in the US. This map is shown in Figure 8.

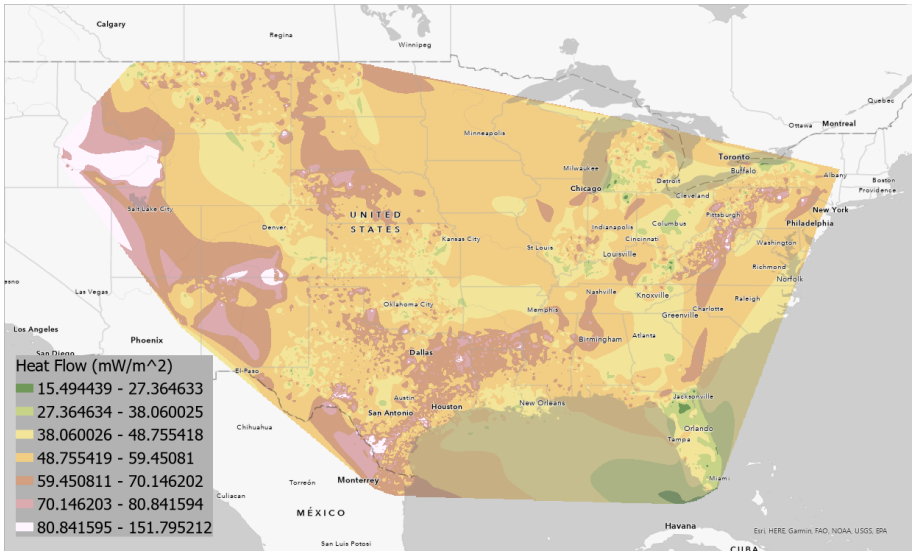

Figure 8: A map of predicted heat flow across the US.

Looking at Figure 8 it is clear that a map like this could be useful for individuals looking to find that best spots for geothermal power generation. There are massive hot spots in the Western United States. This is near Yellowstone National park and has been known to have vast geothermal resources. In fact the entire Western US is on top of the fault line between the Pacific and North American plates which are currently moving away from each other. Therefore this area is a prime spot for geothermal power going into the future [3]. This map also shows that there is a stripe of heat near that Appalachian mountain chain. This area is known as a geothermal basin, there is much less research into the geothermal potential of this area. However, according to this Natural Neighbors model there is definitely potential here [13].

The usefulness of a model like this when it comes to the large upfront costs of constructing geothermal power plants cannot be understated. However, this model has its drawbacks. Because of the simplicity of the natural neighbors calculation, there is no clear way to model the variability or uncertainty of the predictions made by the model. On top of this, the model generated for this data set did not account for varying depths. Modeling the depths is not entirely necessary on this data set since nearly all of the data is for depths between 1 and 3 kilometers. But the extra dimension could provide more insight into the areas of interest previously mentioned.

Kriging

Kriging is another method to model data which has varying densities. Since kriging works with the residuals of the data rather than the data itself it has more advantages. The residuals of the data is the difference between a point's true value and the data's mean. In this paper, the function $r(\phi, \lambda)$ as shown in equation 5 will be used to represent the residual for a point in the US.

$$r(\phi, \lambda) = Q(\phi, \lambda) - E\{f_{\phi,\lambda}\} \tag{5}$$

By predicting residuals, finding the variability of the predicted measurement is built into this model. So it will be useful to see which areas have greater uncertainty when generating a map of predicted values. This method can also be considered better than natural neighbors because it aims to reduce bias in the predictions even further. For example, consider a point that needs to have its value predicted. If this point has three data points to the north but only one to the south, it would make sense that the three to the north could be redundant and should not have the same weight as the one point to the south. Kriging can address this redundancy problem by attempting to minimize the estimation variance. Estimation variance can be defined as the squared difference between the predicted value and the unknown true value for the residuals. This is shown in Equation 6.

$$\begin{aligned} &E\left([\hat{r}(\phi, \lambda) - r(\phi, \lambda)]^2\right) = ... \\ &= E\left([\hat{r}(\phi, \lambda)]^2\right) - 2E\left(\hat{r}(\phi, \lambda) r(\phi, \lambda)\right) + E\left([r(\phi, \lambda)]^2\right) \end{aligned} \tag{6}$$

Looking at the product in Equation 6 there are some substitutions which can be made. We can replace the predicted residual function ($\hat{r}(\phi, \lambda)$) with something similar to the predictor in Equation 4, but the weights term will be made to be more generic (represented as the variable w) [9]. This substitution is shown in Equations 7 and 8. To reduce clutter the location coordinates are not shown as (ϕ, λ) and are simply represented as L for the location.

$$\hat{r}(L) = \sum_{i=1}^{n} r(L_i) w_i \tag{7}$$

$$\begin{aligned} &E\left([\hat{r}(L) - r(L)]^2\right) = E\left([\hat{r}(L)]^2\right) - 2E\left(\hat{r}(L) r(L)\right) + E\left([r(L)]^2\right) = \\ &= \sum_{i=1}^{n} \sum_{j=1}^{n} w_i w_j E\{r(L_i) r(L_j)\} - 2 \sum_{i=1}^{n} w_i E\{r(L) r(L_i)\} + Var(L) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{n} w_i w_j C(L_i, L_j) - 2 \sum_{i=1}^{n} w_i C(L, L_i) + Var(f_L) \end{aligned} \tag{8}$$

The final result of equation 8 includes the function $C(L_I, L_j)$ this is used to represent the co-variance of the data at each of these locations. Since the co-variance is present for all combinations of the data points, this term is what addresses the redundancy problem stated earlier. Also the $C(L, L_i)$ is the co-variance between the data at location L_i and an unknown location. Finally the last term $Var(F_L)$ is simply the variance of the data set. These substitutions are only made possible because the problem was re-framed to work with the residuals of the data in equation 5. All that the kriging model aims to do is to find the weights w_j, which minimize equation 8 [9]. Therefore the best way to do this is to take the partial derivative of equation 8 with respect to the weights. By doing this and setting the result to be equal to zero, the weights can be solved for by using the system of equations shown in 9.

$$\sum_{j=1}^{n} w_j C(L_i, L_j)) = C(L, L_i),\ i = 1, ..., n \tag{9}$$

This is known as simple kriging, and it has one drawback that does not make it ideal for this data set. This is because the mean of heat flow for the entire US is used, but the US is large and the mean is not representative of any arbitrary data point. Instead, a better approach is to calculate the mean locally for a set of points near to the point that needs to be estimated. This approach is known as ordinary kriging. The only other difference between simple and ordinary kriging besides the locally calculated mean, is the additional constraint that all of the weights must sum to one.

In order to implement kriging properly a variogram must be used to find the parameters to use for the model [4]. The ArcGIS software will do this automatically when it is asked to generate it's own model. However, for the purpose of demonstrating what a variogram is, one was generated separately in R. The graph shown in Figure 9 is a semi-variogram where the semivariance is modeled with a spherical function. The spherical function models how the the semivariance increases as the distance between points grows. A spherical function was chosen for this data because it's linear slope appeared to fit the data better than exponential or Gaussian functions. Initially the semivariance (and variance) between points that are close together is expected to be small, but not zero. This non-zero value is known as the "nugget", which comes from when variograms were used to model the best places to dig for gold [9]. As the distance increases it will eventually reach a value where the semivariance does not change anymore, this value is known as the "sill". In this model the value for the nugget is 96.03, and the sill value is 110.4. The range is another parameter which describes the distance before the semivariance stops increasing. The value in this model is 468km, or 290 miles. There is not a huge difference between the a variogram and semivariogram. The semivariance is simply a measurement of variance below the mean, and is often better suited for kriging applications [2].

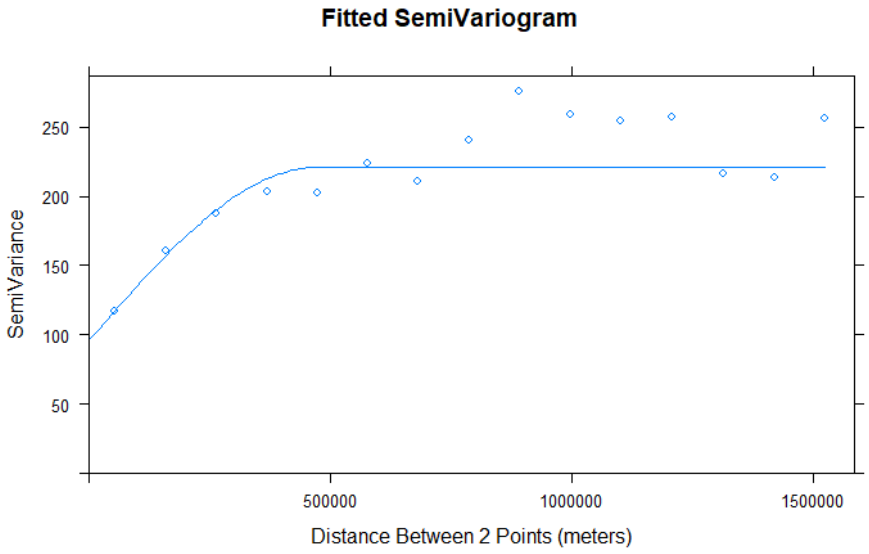

Figure 9: The SemiVariogram fitted in R.

Now that the semivariogram has been fit and and the process of kriging is well understood, it is possible to generate a kriging model. This model can be generated in ArcGIS Pro and then potential values can be interpolated for arbitrary data points. The map generated is shown in Figure 10.

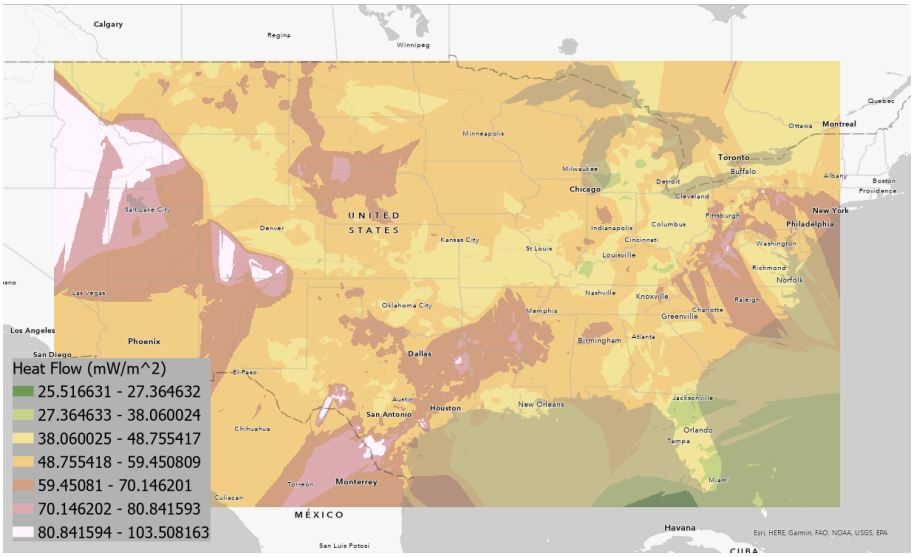

Figure 10: A map of predicted heat flow across the US. Using Kriging.

A comparison of this map to the one generated using the natural neighbors model shows that they are very similar. This similarity is reassuring, and shows that both models have been implemented correctly. All of the hot spots that were mentioned previously are also present in the kriging model. However, there are some differences between the two maps. The natural neighbors map appears to be less smooth. This phenomenon is most likely caused by the differences between both methods. For the natural neighbors model the weights are calculated using only the points which have adjacent voronoi cells. For areas like Eastern Texas, where there is a lot of wells all very close together, this causes the prediction of heat flow to vary quite a bit for small distances. The kriging model is known to smooth out the data because it attempts to reduce variability of any estimation. This is most likely the reason why the kriging model generates smoother results.

Another difference between the two maps is that the kriging map in figure 10 seems to have these strange jagged areas which look like sheets of ice. An example of what is meant is near the Appalachian Mountain range which juts out into North Carolina. The reason for this is unknown, although it most likely has to do with the lack of data in these areas. In fact, any area which does not have a sufficient amount of data points should not be considered to have a good prediction of surface heat flow. As previously mentioned, the kriging model calculates the variability of the predictions it makes. The areas which have a higher variability for predictions can be plotted onto a map and are shown in Figure 11

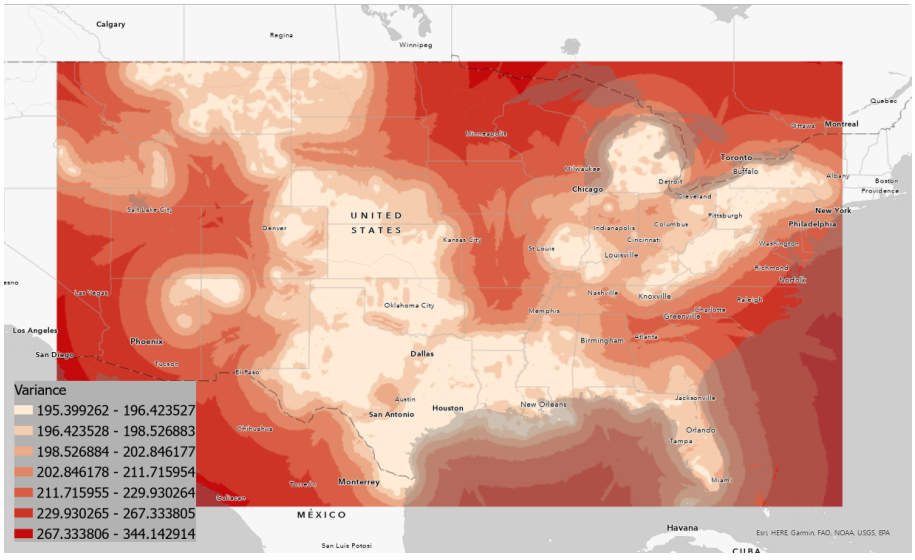

Figure 11: The Kriging model's calculations of expected variance.

The calculations of variance are not very surprising. The areas with the

most amount of data points shown in figure 5 also have the lowest variance. A map like this one could be very helpful for the construction of a geothermal power plant. Essentially it would be very unwise to begin construction in any of the areas colored red. In the future if more data was collected for these red areas the variability of predicted heat flow would be expected to decrease, and it would be feasible to consider those sites for power production as well.

Conclusion

The biggest hurdle towards a more sustainable and geothermal powered future is the uncertainty and large upfront costs associated with geothermal power plants. The initial costs of these plants can be very high, and when coupled with the chance that there may not be a profit, many companies do not take the risk. In this paper, hopefully the associated risk of starting construction in an unsuitable area has been minimized. With the proper modeling and geo-statistical techniques to analyze existing data, perhaps this risk could be brought so low that more geothermal plants are constructed than oil or gas wells. However, until that day comes every time a new oil or natural gas well is drilled the modeling techniques described in this paper will get better at reducing the risks associated with geothermal power plant construction. Oil and gas wells do not need to be the only sources of data used to predict heat flow either. Measurements of groundwater and even emerging remote sensing technologies may be able to model surface heat flow in a more accurate manner [3].

In the next decade geothermal energy is expected begin to grow [7]. With the recent invention of technologies which reduce the expenses for plant production and more analysis of geospatial data this energy market is set up for success. Many of the same skills used by technicians and engineers currently in the oil and natural gas markets are applicable to geothermal drilling as well. In the last decade Germany has seen a massive growth in their geothermal power market from studies similar to this one [8]. Hopefully this paper and the methods described in it can be applied to increase the growth of geothermal plants in the US.

References

[1] Gloria Aguirre. “Geothermal resource assessment: A case study of spatial variability and uncertainty analysis for the states of New York and Pennsylvania”. In: (2014).

[2] Daniele Cinti, Monia Procesi, and Pier P. Poncia. “Evaluation of the Theoretical Geothermal Potential of Inferred Geothermal Reservoirs within the Vicano-Cimino and the Sabatini Volcanic Districts (Central Italy) by the Application of the Volume Method”. English. In: *Energies* 11.1 (2018).

Copyright - Copyright MDPI AG 2018; Last updated - 2020-05-29; SubjectsTermNotLitGenreText - Italy, p. 142.

[3] Mark Coolbaugh et al. "A Geothermal GIS for Nevada: Defining Regional Controls and Favorable Exploration Terrains for Extensional Geothermal Systems". In: *Transactions - Geothermal Resources Council* (Jan. 2002).

[4] André Dauphiné. "7 - Models of Basic Structures: Points and Fields". In: *Geographical Models with Mathematica*. Ed. by André Dauphiné. Elsevier, 2017, pp. 163–197. ISBN: 978-1-78548-225-0. DOI: https://doi.org/10.1016/B978-1-78548-225-0.50010-5.

[5] *Geothermal Energy*. URL: https://archive.epa.gov/climatechange/kids/solutions/technologies/geothermal.html.

[6] William E Harrison et al. *Geothermal resource assessment in Oklahoma*. Tech. rep. 1982.

[7] *Initial Results from the 2020 U.S. Geothermal Power Production and District Heating Market Report*. URL: https://www.nrel.gov/docs/fy21osti/77774.pdf.

[8] Oliver Kastner et al. "The deep geothermal potential of the Berlin area". In: *Environmental earth sciences* 70.8 (2013), pp. 3567–3584.

[9] GeostatsGuy Lectures. *12b Geostatistics Course: Kriging*. Oct. 2018. URL: https://www.youtube.com/watch?v=-Bi63Y3u6TU.

[10] Cary Lindsey. "Exploratory and Spatial Statistics for Evaluating Heat and Mass Transfer in Geothermal Areas". English. Copyright - Database copyright ProQuest LLC; ProQuest does not claim copyright in the individual underlying works; Last updated - 2020-11-09. PhD thesis. 2018, p. 78. ISBN: 978-1-392-02418-8.

[11] M Macenić, T Kurevija, and I Medved. "Novel geothermal gradient map of the Croatian part of the Pannonian Basin System based on data interpretation from 154 deep exploration wells". In: *Renewable and Sustainable Energy Reviews* 132 (2020), p. 110069.

[12] *National Centers for Environmental Information*. URL: https://www.ncdc.noaa.gov/.

[13] Elaina Shope. "A detailed approach to low-grade geothermal resources in the Appalachian Basin of New York and Pennsylvania: heterogeneities within the geologic model and their effect on geothermal resource assessment". In: (2012).

[14] George R Stutz et al. "A well by well method for estimating surface heat flow for regional geothermal resource assessment". In: *Proceedings of thirty-seventh workshop on geothermal reservoir engineering, Stanford. SGP-TR-194*. 2012.

[15] Wikipedia. *Natural neighbor interpolation — Wikipedia, The Free Encyclopedia*. http://en.wikipedia.org/w/index.php?title=Natural\%20neighbor\%20interpolation&oldid=950949003. [Online; accessed 07-May-2021]. 2021.