A VOI Web Application for Distinct Geothermal Domains: Statistical Evaluation of Different Data Types within the Great Basin

Whitney Trainor-Guitton[1], Sierra Rosado[1,2]

[1]National Renewable Energy Laboratory

[2]Amherst College

Keywords

INGENIOUS, Great Basin, value of information, Bayesian analysis

ABSTRACT

The Great Basin region contains different domains that have different structural and hydrothermal flow patterns. Depending on the characteristics of these patterns, certain data types may be more successful at detecting hidden geothermal resources. In this paper, we quantitatively evaluate if certain data types are more successful in certain domains. Given different aquifer, strain and structural conditions, we explore which data types statistically reveal positively labeled geothermal sites. We utilize value of information (VOI) metrics to help quantify the reliability of data types to discriminate against "positive" and "negative" labeled geothermal sites. We also evaluate how kernel density estimation can help generalize the statistics that inform VOI, which is necessary given the limited data in geothermal exploration. Except for the Carbonate Aquifer, the highest ranking of the $V_{imperfect}$ is the Local Structural Setting. Next, the slip and dilation tendency is first for Carbonate Aquifer and second for Central Nevada Seismic Belt and Western Great Basin. For the Carbonate Aquifer, heat flow is has the lowest $V_{imperfect}$ value compared to the other three domains, which is consistent with the understanding of how heat flow measurements are masked by regional groundwater flow.

1. Introduction

Play fairway analysis has been very successful in identifying hidden prospects via 2D regional analysis (Faulds et al. 2015; 2017; Craig et al. 2021) but has also helped bring consensus on the existence of domains within the Great Basin. Clustering analysis has further acknowledged and developed these ideas (Smith et al. 2021), which identified four domains: Western Great Basin, Walker Lane, Central Nevada Seismic Belt and the Carbonate Aquifer in the east (Figure 1).

As part of the INGENIOUS[1] data collection and modeling objectives, there is an updated focus by the regional play fairways workflow team to improve the play fairway analysis output by defining and examining geological domains. While further analysis will be done on refining these domains in other parts of the team's efforts, the hypothesis presented here is that certain data types may be more successful at discovering positive geothermal labeled sites in certain domains over others (Figure 1). Given the different data parameters for aquifer, strain and structural conditions, we explore which types of data types statistically distinguish more reliably between positive and negatively labeled geothermal sites. Examples of some of these data are shown in Figure 1: strain values from 2022 (grayscale diamonds), slip and dilation along faults (line features with black to yellow colormap) and the $\log_{10}$ well temperatures (circular viridis colormap).

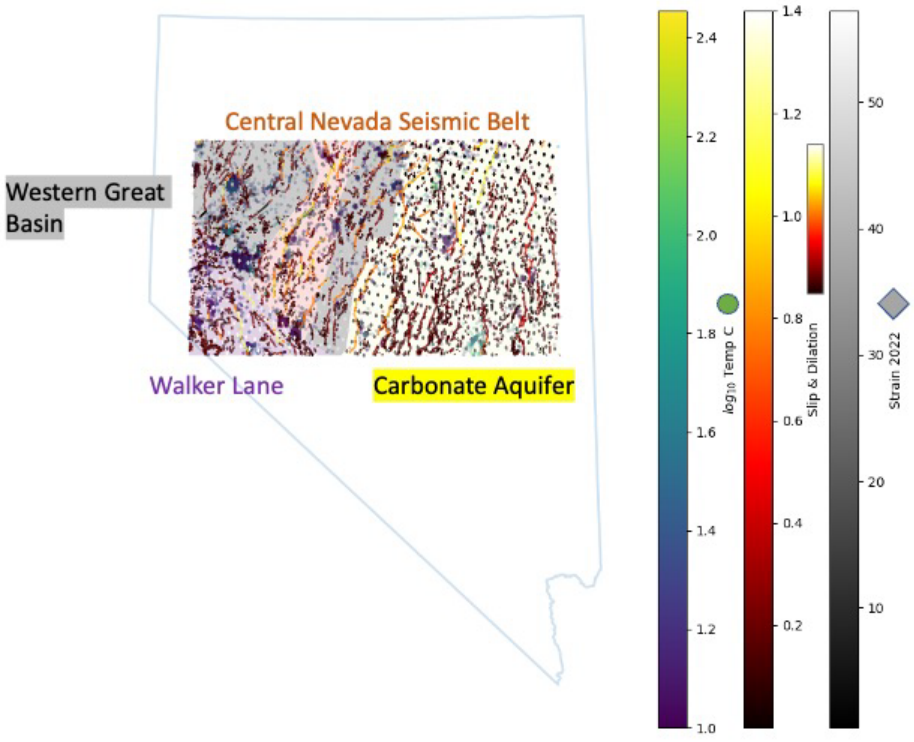

Figure 1: Four domains as background color with overlayed $\log_{10}$ well temps (circles), slip & dilation tendency (along fault lines) and strain (diamonds)

Using available datasets and labels from the Nevada machine learning project[2], we have developed codes that will be updated when all the INGENIOUS datasets and labels are available covering the Great Basin area, rather than just the initial machine learning area of central Nevada (shown in rectangle in Figure 1).

Specifically, we use value of information (VOI) metrics to quantify which data types are more reliable for each domain in discriminating between positive and negative geothermal labels (Trainor-Guitton 2014). VOI utilizes the labeled data by calculating likelihood functions for each

[1] https://gbcge.org/current-projects/ingenious/
[2] https://gbcge.org/current-projects/machine-learning/

combination of label and attribute, and then the likelihood can be transformed into a posterior as described in (Trainor-Guitton et al. 2014). VOI requires a decision to be identified where an uncertain parameter affects the outcome of the decision, e.g. temperature in a drilled well influences economic geothermal outcomes. Therefore, the value outcomes (often but not necessarily in monetary units) map the geothermal states (here positive/negative labels) to their rewards. Ultimately, the posterior is combined with these value outcomes such that each data attributes can be assigned a discrete value that accounts for how "cleanly" it distinguishes between positive and negative labels.

This paper will describe the VOI equations in detail and how $V_{imperfect}$ compares for the different domains. These are preliminary results, as the domains and data coverage will be updated with INGENIOUS progress. We demonstrate how these domain-specific VOI calculations are implemented into a Streamlit app[3], which allows for Python code to be shared via custom web apps. By building and deploying a VOI app, INGENIOUS members can build intuition on how VOI works and eventually allow members of the general geothermal public to do the same.

2. Methodology

We motivate this domain-focused work by presenting the distributions of strain and bottom hole temperatures for the different domains. These are examples of the distributions that can be put into the VOI equations. The next subsection presents and describes the VOI equations, demonstrating the current implementation of each in the VOI Streamlit app. The last part of the methodology describes how smoothing the histogram can be performed, which helps to generalize the statistical distributions and greatly affects the VOI metrics.

2.1 Example Attributes for the Different Domains

The 2022 strain data for the four domains are shown in Figure 2. The highest strains (33 to 58 units) are within Walker Lane, while the lowest strains are within the Carbonate Aquifer (0 to 3 units), based on the bar chart of strain versus counts. The Central Nevada Seismic Belt appears to have a bimodal distribution, and the Western Great Basin seems to have a peak at strains slightly higher than the Central Nevada Seismic Belt. The right y-axis of Figure 2 demonstrates the "likelihood" or probability density function (pdf) for the strain for each domain and is represented by the dashed lines. They are normalized versions of the counts and provide a "smoothing" factor compared to the bar counts. This smoothing will be described in detail later.

[3] https://docs.streamlit.io/

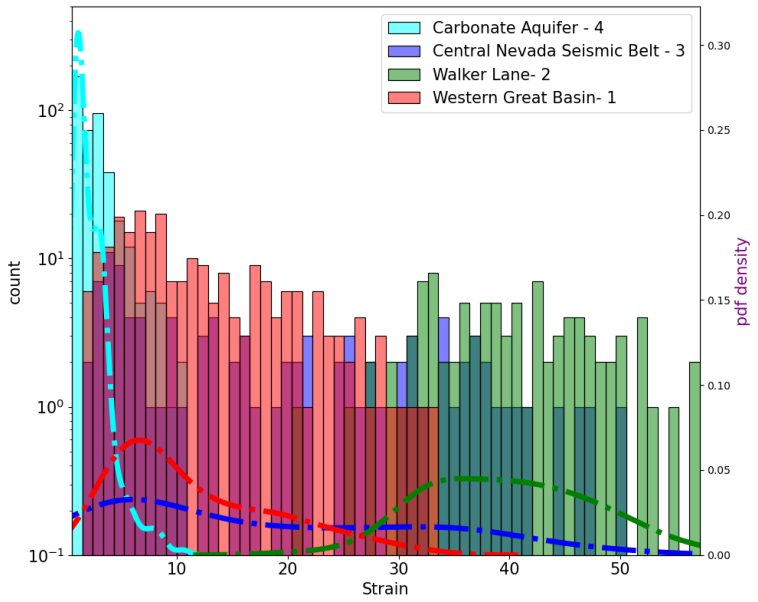

Figure 2: Histogram of strain values within each domain. Left y-axis are raw count values corresponding to bars; right y-axis are normalized probabilities corresponding to the dash-dot lines.

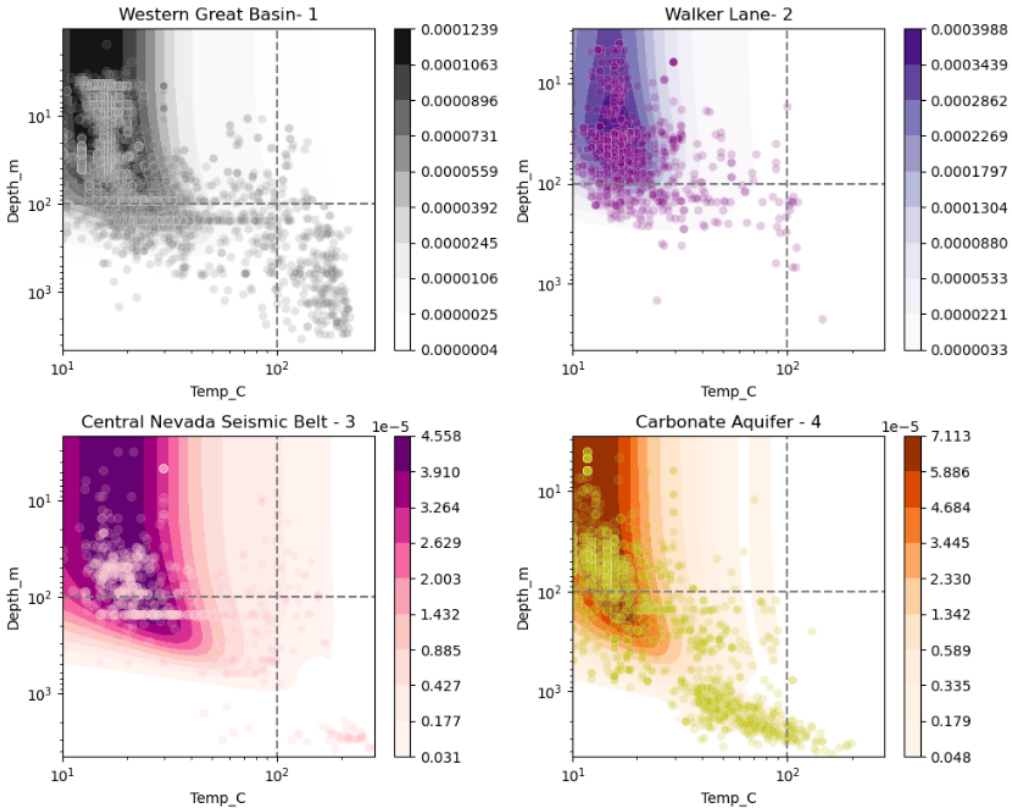

Figure 3: Distribution of well temperatures (BHT: bottom hole temperature) for each domain. Empirical data shown with circle markers, and kernel density estimation (kde) is the shaded domain (smooth version of density of data points).

Additionally, well temperatures and their depths were visualized for the four domains (Figure 3). The shaded regions in Figure 3 represent kernel density estimations or kde plots which represent and smooth out the density of the distribution and are described in detail below. The scatter plots (circle markers) alone make it difficult to understand the relative density since many could plot on top of each other. For example, the Carbonate Aquifer looks like it has many deep hot temperatures, but they are a small number of the total measurements. Thus, the density values you see on the colorbars are specific to each subdomain dataset (depends on number of data and how they are spread out). Next, we describe the kde functionality for smoothing histograms to help generalize the statistics, and ultimately, we show how it affects the VOI metrics.

2.2 KDE Histogram Estimation

Kernel density estimation (KDE) is a non-parametric method for estimating the probability density function of a given random variable, or specifically here for the observations of the subsurface features. The traditional name of KDE is the *Parzen-Rosenblatt Window* method, named after those who formalized the technique. Given a sample of independent, identically distributed (i.i.d) observations ($x_1,x_2,\ldots,x_n$) of a random variable from an unknown source distribution, the kernel density estimate, is given by:

$$\Pr(x) = \frac{1}{nh}\sum_{j=1}^{n} K\left(\frac{x - x_j}{h}\right) \quad (1)$$

where *K(.)* is the kernel function and *h* is the smoothing parameter, also called the bandwidth. Various kernels are possible, but we use the default Gaussian.

When thinking of histograms as estimates of the density function, it is reasonable to suggest that bin size controls the smoothness of the function since bin size controls how the data are modeled, which then affects how the data are interpreted. Histograms using an ideal bandwidth will have a density estimate that can most accurately approximate the true distribution of the data, and importantly for geothermal, smooth out gaps in the data due to sparseness or incomplete data.

2.2.1 Ideal Bandwidth Calculation with Naïve Bayes

We began exploring how the bin size (bandwidth *h* in Equation 1 above) for the attribute affects the posterior probability calculation, the marginal, and thus the value with imperfect information $V_{imperfect}$. We determine the ideal bandwidth by performing a grid search with a Naïve Bayes classifier. Naïve Bayes is a generative predictor, in other words it calculates the posterior, then assigns a class (positive or negative) according to which one has a high posterior probability (VanderPlas 2016; Powers, Trainor-Guitton, and Hoversten 2022). The grid search performs the Naïve Bayes classification for 20 different bandwidths then compares the predicted class with the true class. The bandwidth that results in the highest accuracy in Naïve Bayes is deemed the ideal bandwidth.

$$\text{Accuracy} = \frac{True\ Positives + True\ Negatives}{All\ Samples} \quad (2)$$

The accuracy for 20 different bandwidths of heat flow is plotted in Figure 4; the bandwidth that results in the highest predictive accuracy is equal to 7. This grid search is performed for each attribute when calculating the VOI within the app.

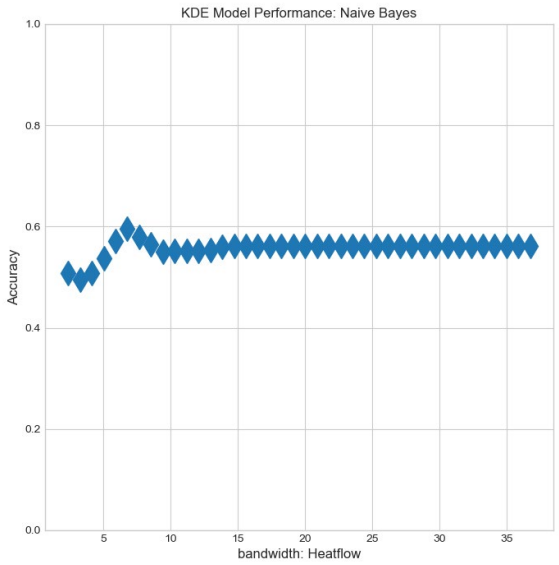

Figure 4: Naive Bayes accuracy versus heat flow bandwidth: bandwidth = 7 results in best accuracy for predicting negative or positive labeled geothermal sites

2.3 Value of Imperfect Information: Quantifying Past Performance of Different Data Types to Identify Positive/Negative Geothermal Labels

Currently, the VOI app begins with a demonstration problem, where the user can toggle the "odds in the geothermal lottery" by changing the *a priori* probability: $\Pr(\Theta = \theta_i)$. This is shown on the right side of Figure 5. A plot demonstrating the "prior value" as a function of the dry hole cost (cost of when the decision action is to drill, and no geothermal production resulted). The prior value is calculated with the following expression:

$$V_{prior} = \max_{a} \left[\sum_{i=1}^{2} \Pr(\Theta = \theta_i)\, v_a(\Theta = \theta_i) \right] \quad a = drill, don't\ drill \tag{3}$$

$\Pr(\Theta = \theta_i)$ is the probability of being either a positive or negative geothermal site (indexed by *i*), and $v_a(\Theta = \theta_i)$ represents the profits or costs (values) when action *a* is taken and the geothermal state turns out to be θ_i. Example values for this are shown in **Table 1**, whereas the x-axis in Figure 5 shows a range of values from -$1,000,000 to 0 are shown for $v_{a=drill}\left(\Theta = \theta_{negative}\right)$. In Figure 5, the prior probability toggle is set at $\Pr\left(\Theta = \theta_{positive}\right) = 15\%$.

Table 1: Value array $v_a(\Theta = \theta_i)$ for binary geothermal (positive/negative, columns) and binary decision (don't drill/drill, rows)

	$\theta_{i=NEGATIVE}$	$\theta_{i=POSITIVE}$
a= don't drill	$0	$0
a= drill	X-axis in Figure 5 Default: -$1,000,000	+$1,000,000

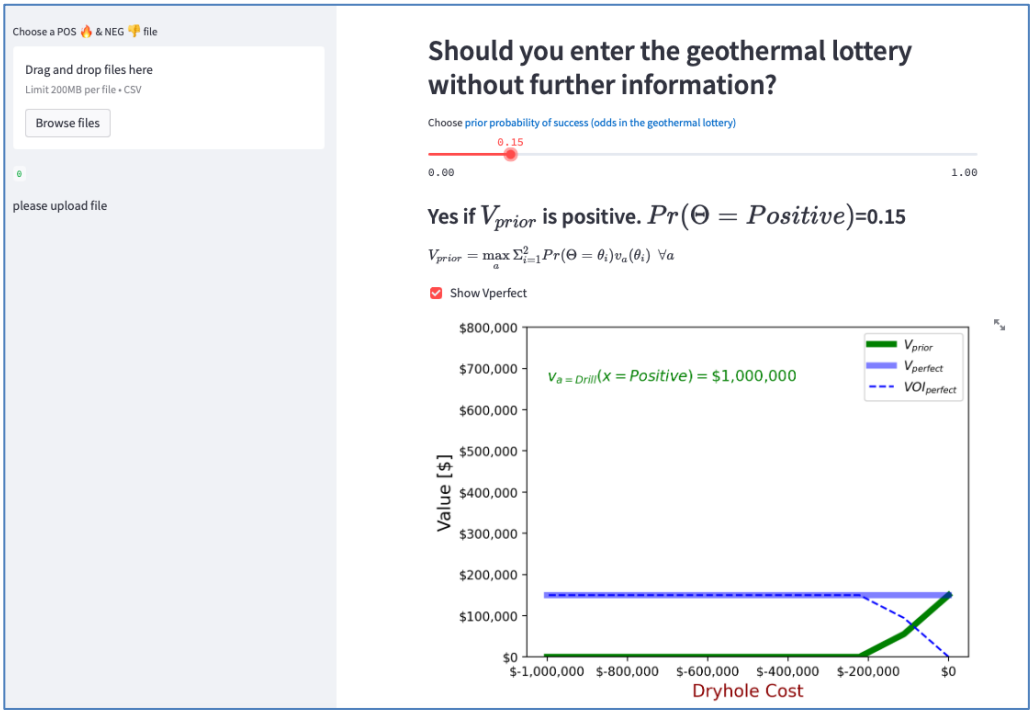

Figure 5: Starting page of the app. On the right side is the demonstration VOI, where V_{prior} as a function of dry hole cost is plotted (green) and the user can change the prior probability (red slider above plot). Also, the user can choose to show the upper bound on information by checking to plot $V_{perfect}$ (blue) and value of perfect information (blue dash). On the left is where you upload site specific data.

When V_{prior} is <0, we do not want to enter the geothermal lottery without further information. Thus in Figure 5, V_{prior} indicates that no drilling should be done unless a dry hole cost is less than -$200,000. If the prior probability toggle is increased ($\Pr(\Theta = \theta_{positive}) > 15\%$), then the V_{prior} curve also increases.

Given these values and the prior probability defined, the upper bound on information is calculated via "value with perfect information":

$$V_{perfect} = \sum_{i=1}^{2} \Pr(\Theta = \theta_i) \max_a [\, v_a(\Theta = \theta_i)\,] \quad a = drill, don't\ drill \tag{4}$$

Compared with V_{prior}, $V_{perfect}$ says that we can choose the best action (*max a*) for each possible geothermal scenario (positive or negative), as it has moved inside the equation compared to V_{prior}. $V_{perfect}$ is the weighted average of those value outcomes where the weights are the prior probabilities. $V_{perfect}$ tells us the maximum value ***any*** information source may have given the risk (economics and probability) of the decision we (the geothermal decision maker and VOI app user) are faced with. Figure 5 plots $V_{perfect}$ in a blue solid line. Since this plot is for $\Pr(\Theta = \theta_{positive}) = 15\%$, $V_{perfect}$ = 0.15*$1,000,000 + 0.85*$0 = $150,000. This is the most value any type of information can have, given the value of drilling in a positive site as defined in **Table 1**.

The value ***of*** perfect information is the increase of $V_{perfect}$ over $V_{prior.}$ As the decision has less consequences (e.g., lower dry hole cost), $VOI_{perfect}$ has a decreasing value: the blue dashed line in Figure 5.

The left side bar of Figure 5 asks the user to upload a *Negative* and *Positive* file, where these are csv files (comma separated value files) that contain the features, attributes, or data types that are considered within a neighborhood of surrounding *Negative* and *Positive* geothermal labels. Currently, the app loads domain-specific csv files. After showing a preview of the files on the left panel, the user chooses which attribute (data type) they wish to evaluate. Figure 6 shows what is in the app after the csv files are chosen; in this example *slip & dilation* from the Carbonate Aquifer domain are shown. Currently, the app displays a bar chart of the raw data: positive in green and negative in red. In this case, there is no data around 0.4 in the observed data, although we would expect with more observations this would occur. A smoothed, continuous, and normalized histogram of the chosen attribute is also plotted, with the label on the right y-axis. These are considered the likelihoods: $\Pr(X = x_j | \Theta = \theta_i)$, from the labeled data. The smoothing helps generalize the potential data distribution since we have incomplete data, such as empty data bins.

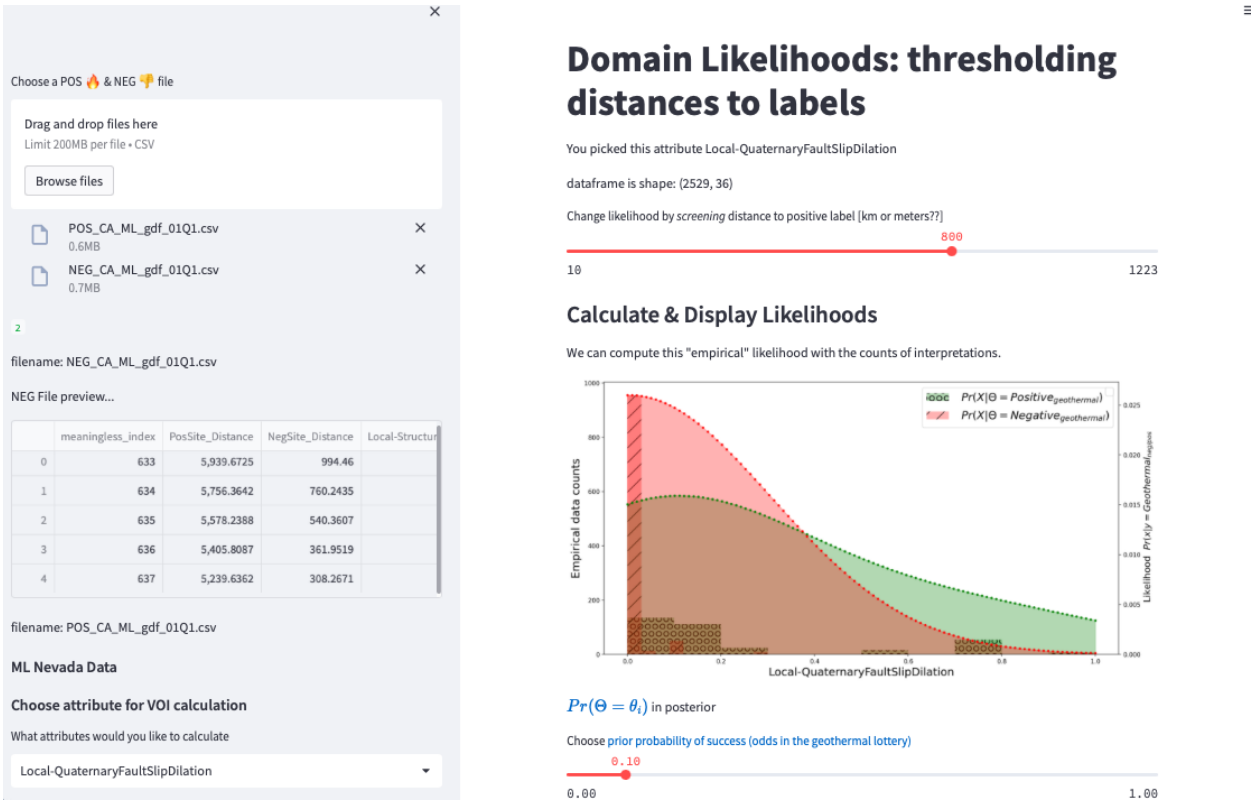

Figure 6: VOI Streamlit app showing the likelihood plots of the chosen attribute. In this example, likelihoods for slip and dilation from positive (green) and negative (red) are shown from the Carbon Aquifer. The red slider bar allows the user to include more or fewer data according to distance from the labels (negative and positive sites)

The next step in the VOI process is to calculate the posterior probability, which combines the likelihood and prior probability to a probability of either a positive or negative geothermal resource to occur given a certain observation was made in the field: $\Pr(\Theta = \theta_i | X = x_j)$. Bayes law mandates that this is a scaled version of the likelihood (e.g., the data observations) with the prior:

$$\Pr(\Theta = \theta_i | X = x_j) = \frac{\Pr(\Theta = \theta_i)\,\Pr(X = x_j |\, \Theta = \theta_i)}{\Pr\,(X = x_j)} = \frac{Prior\ x\ Likelihood}{Marginal} \tag{5}$$

The posterior can be thought of the reliability of a data type to identify a positive or negative geothermal resource. In other words, it chronologically reverses the likelihood, taking into account the prior probability assigned. Importantly here, it is used in the value with imperfect information ($V_{imperfect}$):

$$V_{imperfect}(\boldsymbol{u}) = \sum_{j=1}^{N} \Pr(X = x_j) \left[\max_{a} \left[\sum_{i=1}^{2} \Pr(\Theta = \theta_i | X = x_j)\, v_a(\Theta = \theta_i) \right] \right] \tag{6}$$
$$a = drill, don't\ drill$$

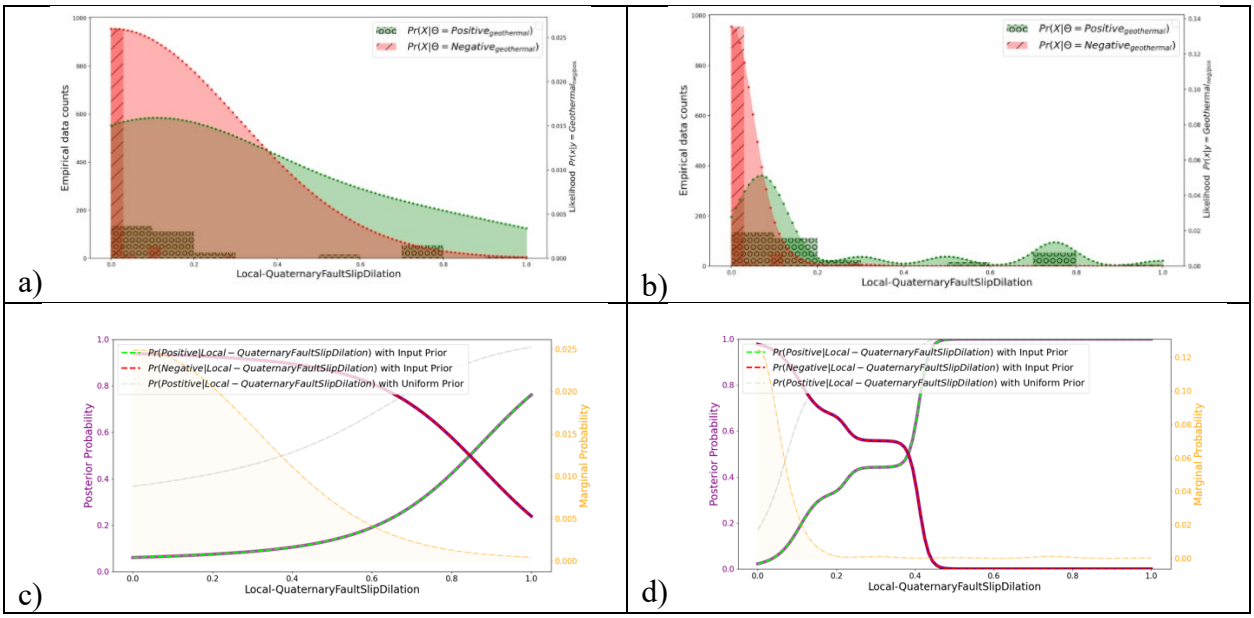

Figure 7: Slip & Dilation from Carbonate Aquifer a) and b). a) Smoothed Likelihood with 0.3 b) Smoothed Likelihood with Ideal Bandwidth 0.05 (Secction 2.2) c) Posterior (green & red) and marginal (orange) from default smoothed likelihood a) d) Posterior (green & red) and marginal (orange) from ideally smooth likelihood (b)

Figure 7 shows two examples of both the likelihood and posterior that are plotted in the app; the left side uses a default smoothing parameter while the right shows an idealized smoothing. Other default values used are a prior probability $\Pr(\Theta = \theta_{positive}) = 10\%$ and a distance to a geothermal label of 800 km.

In the case of smoothing of 0.3, $V_{imperfect}$ = \$0; this is displayed below the posterior plot in the app. Both posterior (the red and green dashed lines with purple backing) and marginal (orange, Pr $(X = x_j)$) are highly affected by the bin smoothing parameter. In Figure 7c, the posterior probability for a positive geothermal site (green dash) very gradually increases with increasing slip and dilation index. However, when the posterior is calculated with a likelihood smoothed by a bandwidth of 0.05 (Figure 7b), the posterior indicates that positive geothermal sites are highly likely with slip and dilations greater than 0.45. $V_{imperfect}$ increases to \$25,973.

2.2.1 Example calculations

To build intuition for the reader on how the smoothing out the probability estimates (e.g., posterior and marginal) influences the final $V_{imperfect}$, we present a few example calculations, assuming six data bins for simplicity. These six data bins are represented by rows in Table 2 and Table 3. Specifically, we focus on the influence of the posterior and the marginal probabilities on the final $V_{imperfect}$.

Table 2: Almost ideal likelihood in turquoise, scaled by prior (90 negative/10 positive) in blue, resulting marginal in orange and posterior in purple. Average drilling value and max outcome in green. Resulting $V_{imperfect}$ below.

Data bins	Pr $(X \mid \theta_{neg})$	Pr $(X \mid \theta_{pos})$	90%*Pr $(X \mid \theta_{neg})$	10%*Pr $(X \mid \theta_{pos})$	Pr($X=x_j$)	Negative Posterior	Positive Posterior	a = drill [$]	max a
$x_1 = 60$	0%	0%	0%	0%	0%	0%	0%	$0	$0
$x_2 = 70$	0%	0%	0%	0%	0%	0%	0%	$0	$0
$x_3 = 80$	100%	0%	90%	0%	90%	100%	0%	-$1,000,000	$0
$x_4 = 90$	0%	100%	0%	10%	10%	0%	100%	$1,000,000	$1,000,000
$x_5 = 100$	0%	0%	0%	0%	0%	0%	0%	$0	$0
$x_6 = 110$	0%	0%	0%	0%	0%	0%	0%	$0	$0
								$V_{imperfect}$	$100,000

Table 2 represents an almost ideal likelihood case: data bin x_3 is 100% associated with negative sites and x_4 is 100% associated with positive sites. Per Equation 5, this is scaled by the prior (here we assume 10% probability of a positive geothermal site). Finally we can compute the posterior where both x_3 and x_4 give "perfect" indications of either negative or positive. The next calculation is the average outcome for each action column, where the posterior is used as the weights in the average; since the weights in this case are 0 or 1, the values are exactly the values seen in Table 1. After looking at both action columns, we choose the action with the highest average outcome. In this ideal case, $V_{imperfect}$ is equal to perfect information: $100,000.

However, the case is not ideal as we are missing data in several of the data bins, hence our motivation to smooth the likelihood.

Table 3: Imperfect posterior in blue, average value outcomes in green, & 3 different possible marginals in purple, with resulting $V_{imperfect}$'s below them

Data bins	Pr $(X \mid \theta_{neg})$	Pr $(X \mid \theta_{pos})$	90%*Pr $(X \mid \theta_{neg})$	10%*Pr $(X \mid \theta_{pos})$	Pr($X=x_j$)	Negative Posterior	Positive Posterior	a = drill [$]	max a
$x_1 = 60$	0% / 9%	0%	0% / 8%	0%	0%/ 8.2%	0% / 100%	0%	$0 /-$1,000,000	$0
$x_2 = 70$	50% /45%	0%	45%/ 41%	0%	45%/ 40.9%	100%	0%	-$1,000,000	$0
$x_3 = 80$	50% /45%	50% /45%	45% /41%	5% / 4.5%	50% / 45.5.%	90%	10%	-$800,000	$0
$x_4 = 90$	0%	50% /45%	0%	5% / 4.5%	5% / 4.5%	0%	100%	$1,000,000	$1,000,000
$x_5 = 100$	0%	0% / 9%	0%	0% / 1%	0% / 0.9%	0%	0% / 100%	$0 / $100,000	$0 / $100,000
$x_6 = 110$	0%	0%	0%	0%	0%	0%	0%	$0	$0
								$V_{imperfect}$	$50,000 / $54,545

Table 3 contains an "imperfect" likelihood, where one out of the six bins (x_3) gives a 50/50 split of a positive or negative geothermal site existing but a clear message for x_2 (negative) and x_5 (positive). Once scaled by the prior probability, we see that only in the case of observing x_4, would you drill, thus $V_{imperfect}$ is reduced to $50,000 (0.05 * 1,000,000). However, if we think about smoothing this likelihood out, just so that 9% of the labeled positive data has occupied data bin x_5

and 9% of labeled negative data occupies x1, this increases our $V_{imperfect}$ to \$54,545 because they are exclusively associated with one of the labels. These values are shown in red in Table 3 where they change from the previous imperfect case.

3. Results

We use the $V_{imperfect}$ metric to evaluate which data types are more successful for the different domains and to test sensitivity to the bandwidths. $V_{imperfect}$ is calculated for the default bandwidth (0.3) and for the idealized one according to Naïve Bayes. This was done for the four domains: Carbonate Aquifer, Western Great Basin, Seismic Belt, and Walker Lane. The results are shown in Table 4 through Table 7. The current version of the VOI app performs this grid search to find the idealized bandwidth according to the Naïve Bayes, then uses this bandwidth for the VOI calculations.

Table 4: Carbonate Aquifer

Attributes	**Min**	**Max**	**Ideal Bandwidth (Accuracy)**	**$V_{imperfect}$ with 0.3 bandwidth [$]**	**$V_{imperfect}$ with ideal bandwidth [$]** (ranking)
Quaternary Fault distance	0	40	3.1 (71.1%) [3]	0	0 (5)
Quaternary Slip / dilation	0	1	0.05 (90.8%) [1]	2,536	25,973 (1)
Local Structural Setting	0	1.2	0.06 (90.84%) [1]	115	15,599 (2)
Horizontal Gravity Gradient	0	0.0116	0.0025 (59.0%) [4]	0	0 (4)
Heat flow	62.26	110.33	7.74 (58%) [5]	21,230	7,135 (3)

Table 5: Central Nevada Seismic Belt

Attributes	**Min**	**Max**	**Ideal bandwidth (accuracy)**	**$V_{imperfect}$ with 0.3 bandwidth [$]**	**$V_{imperfect}$ with ideal bandwidth [$]** (ranking)
Quaternary Fault distance	0	27	1.69 (62.01%)	15,677	11,856 (4)
Quaternary Slip / dilation	0	0.8	0.04 (78.5%)	2,077	66,127 (2)
Local Structural Setting	0	1.2	0.06 (81.66%)	21,817	79,842 (1)
Horizontal Gravity Gradient	0.0002	0.0122	0.002 (82.63%)	0	8,669 (5)
Heat flow	78.47	114.27	6.04 (82.09%)	84,702	59,305 (3)

Table 6: Walker Lane

Attributes (Walker)	**Min**	**Max**	**Ideal bandwidth (accuracy)**	**$V_{imperfect}$ with 0.3 bandwidth [$]**	**$V_{imperfect}$ with ideal bandwidth [$]** (ranking)
Quaternary Fault distance	0	31	1.5 (59.4%)	11,933	10,838 (4)
Quaternary Slip / dilation	0	1	0.05 (72.7%)	6,660	42,055 (3)
Local Structural Setting	0	1.2	0.06 (72.7%)	30,015	83,678 (1)
Horizontal Gravity Gradient	0.0001	0.0087	0.0014 0.01 (72.67%)	0	6,310 (5)
Heat flow	69.2	101.4	1.6 (82.0%)	92,376	68,786 (2)

Table 7: Western Great Basin

Attributes (WGB)	**Min**	**Max**	**Ideal bandwidth (accuracy)**	**$V_{imperfect}$ with 0.3 bandwidth [$]**	**$V_{imperfect}$ with ideal bandwidth [$]** (ranking)
Quaternary Fault distance	0	1	1.7 (66.7%)	4,570	2,090 (4)
Quaternary Slip / dilation	0	0.8	0.04 (82.3%)	0	38,345 (2)
Local Structural Setting	0	1.2	0.06 (85.3%)	17,254	61,400 (1)
Horizontal Gravity Gradient	0.0001	0.0123	0.0005 (58.9%)	0	1,670 (5)
Heat flow	70.33	111.99	2.07 (63%)	46,093	18,033 (3)

We see that depending on the scale of the attribute values, the change in bandwidth can greatly affect $V_{imperfect}$. For example, the horizontal gravity gradient has a max value on the order of 10^{-2}, therefore the base case bin value of 0.3 would put all into the same bin.

In general, the higher the accuracies calculated with the ideal bandwidth, the higher the $V_{imperfect}$, which makes sense given that Naïve Bayes uses the posterior to make predictions and $V_{imperfect}$ uses posterior to map and weigh the economic outcomes of Table 1. Heat flow seems to be the only one *not* to have benefited from the ideal bandwidth calculations: reducing $V_{imperfect}$ in all four domains. Its behavior is opposite from the example shown in Section 2.3.1; the extremely small bandwidth of 0.3 however, is not helpful in generalizing the behavior of heat flow with respect to positive and negative sites. Figure 8 displays the likelihoods and posteriors for heat flow from the Carbonate Aquifer for both bandwidth options. The larger bandwidth helps generalize the marginal shape, but a too small of bandwidth can give a false sense of precision for an attribute to perfectly predict a positive or negative geothermal site. The smaller bandwidth results in a higher $V_{imperfect}$ because the marginal models higher frequency for higher heat flow measurements (compare the right y-axis for plots Figure 8c and Figure 8d).

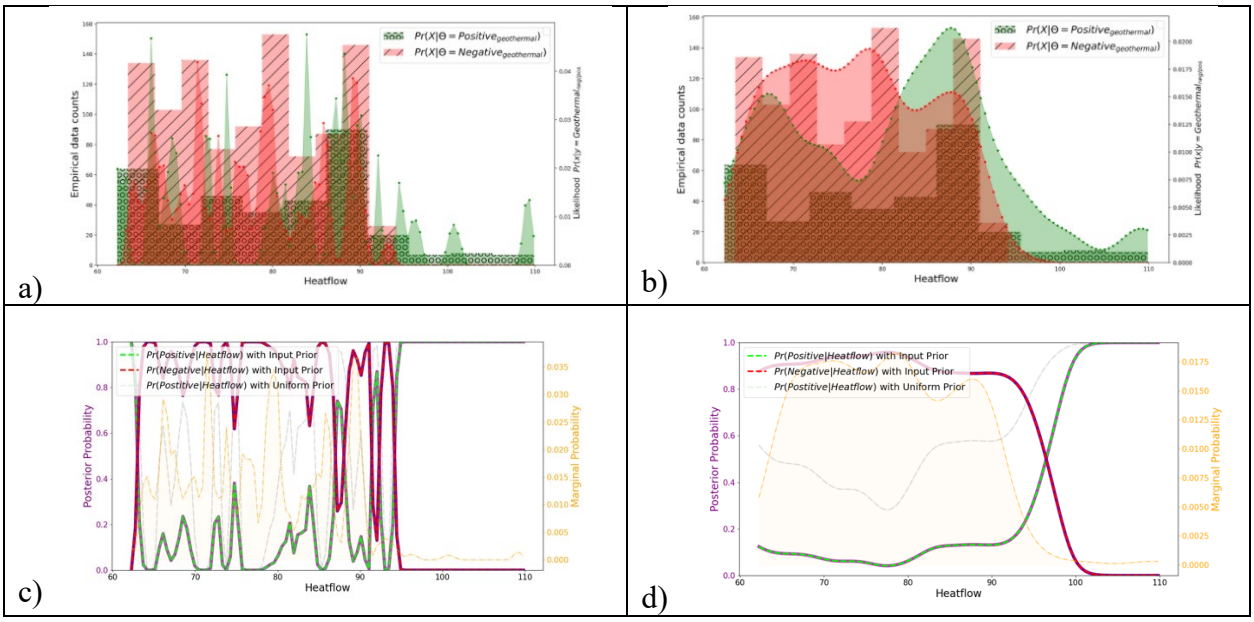

Figure 8: Likelihood (top row) and Posteriors (bottom row) for Carbonate Aquifer Heat flow. a) Likelihood with 0.3 smoothing b) Likelihood with 7.74 smoothing c) Posterior with 0.3 smoothing d) Posterior with 7.74 smoothing

Except for the Carbonate Aquifer, the highest ranking of the $V_{imperfect}$ is the Local Structural Setting. Next, the slip and dilation tendency index is first for Carbonate Aquifer and second for Central Nevada Seismic Belt and Western Great Basin. Heat flow is second for Walker Lane. For the Carbonate Aquifer, heat flow has the lowest $V_{imperfect}$ magnitude of the four. This physically is consistent with the conceptual understanding of how the Carbonate Aquifer masks the heat flow measurements.

Another consideration to keep in mind is the availability of labels within each domain. Currently, the distances needed for negative geothermal labels in the Walker Lane are greater than in the other domains.

4. Conclusions & Future Work

While additional analysis will be done to refine the domains, this work focuses on evaluating if certain data types are more successful in the domains as defined currently. We emphasize these are preliminary results, that will be improved and refined as the VOI app is further developed. We have demonstrated that kernel density estimate is important for geothermal as it can smooth out gaps in the data due to sparseness or incomplete data. The app has the functionality to choose an appropriate bandwidth for different attributes that vary across orders of magnitudes.

Future work includes expanding the data used to calculate the likelihoods for the INGENIOUS area, rather than just the initial machine learning area of central Nevada (shown in rectangle in Figure 1). We hope to also use more temperatures at depth, and not just bottom hole temperatures, which may reveal more convective versus conductive patterns in different domains. Lastly, we would like to build out functionality in the app to evaluate more than one attribute at a time.

Acknowledgement

This work was authored in part by the National Renewable Energy Laboratory, operated by Alliance for Sustainable Energy, LLC, for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308. This project is funded by the U.S. Department of Energy - Geothermal Technologies Office under award DE-EE0009254 to the University of Nevada, Reno for the INnovative Geothermal Exploration through Novel Investigations Of Undiscovered Systems (INGENIOUS). This work was supported in part by the U.S. Department of Energy, Office of Science, Office of Workforce Development for Teachers and Scientists (WDTS) under the Science Undergraduate Laboratory Internship (SULI) program. The views expressed in the article do not necessarily represent the views of the DOE or the U.S. Government. The U.S. Government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this work, or allow others to do so, for U.S. Government purposes.

REFERENCES

Craig, Jason W., James E Faulds, Nicholas H. Hinz, Tait E. Earney, William D. Schermerhorn, Drew L. Siler, Jonathan M. Glen, Jared Peacock, Mark F. Coolbaugh, and Stephen B. Deoreo. 2021. "Discovery and Analysis of a Blind Geothermal System in Southeastern Gabbs Valley, Western Nevada, USA." *Geothermics* 97 (December): 102177. https://doi.org/10.1016/j.geothermics.2021.102177.

Faulds, James E, Nicholas H Hinz, M. F. Coolbaugh, Sadowski, Andrew J, Shevenell, Lisa A, Mcconville, Emma, Jason W. Craig, Sladek, Chris, and D.L. Siler. 2017. "Progress Report on the Nevada Play Fairway Project: Integrated Geological, Geochemical, and Geophysical Analyses of Possible New Geothermal Systems in the Great Basin Region." In *Proceedings*. Stanford University.

Faulds, James E, Nicholas H Hinz, Mark F Coolbaugh, Lisa A Shevenell, Drew L Siler, M Craig, William C Hammond, et al. 2015. "Integrated Geologic and Geophysical Approach for Establishing Geothermal Play Fairways and Discovering Blind Geothermal Systems in the Great Basin Region , Western USA : A Progress Report." *Geothermal Research Council Transactions* 39: 691–700.

Powers, Hayden, Whitney Trainor-Guitton, and G. Michael Hoversten. 2022. "Naïve Bayesian Classification of Cumulative Oil Production from Stochastic Amplitude Variation with Angle Inversion Attributes: As Applied to SEAM and a West Africa Field." *Geophysical Prospecting* 70 (4): 801–14. https://doi.org/10.1111/1365-2478.13190.

Smith, Connor M, James E Faulds, Stephen Brown, Mark Coolbaugh, Cary R Lindsey, Sven Treitel, Michael Fehler, Chen Gu, and Eli Mlawsky. 2021. "Characterizing Signatures of Geothermal Exploration Data with Machine Learning Techniques : An Application to the Nevada Play Fairway Analysis." In .

Trainor-Guitton, Whitney J. 2014. "A Geophysical Perspective of Value of Information: Examples of Spatial Decisions for Groundwater Sustainability." *Environment Systems and Decisions* 34 (1): 124–33. https://doi.org/10.1007/s10669-013-9487-9.

Trainor-Guitton, Whitney J, G Michael Hoversten, Abelardo Ramirez, Jeffery Roberts, Egill Juliusson, Kerry Key, and Robert Mellors. 2014. “The Value of Spatial Information for Determining Well Placement : A Geothermal Example.” *Geophysics* 79 (5): W27–41.

VanderPlas, Jake. 2016. *Python Data Science Handbook: Essential Tools for Working with Data*. First edition. Python / Data. Beijing Boston Farnham Sebastopol Tokyo: O’Reilly.