PROCEEDINGS, 46th Workshop on Geothermal Reservoir Engineering
Stanford University, Stanford, California, February 15-17, 2021
SGP-TR-218

Characterizing Signatures of Geothermal Exploration Data with Machine Learning Techniques: An Application to the Nevada Play Fairway Analysis

Connor M. Smith[1], James E. Faulds[1], Stephen Brown[2,3], Mark Coolbaugh[1], Cary R. Lindsey[1], Sven Treitel[4], Bridget Ayling[1], Michael Fehler[2], Chen Gu[2], and Eli Mlawsky[1]

[1]Great Great Basin Center for Geothermal Energy, Nevada Bureau of Mines and Geology, University of Nevada, Reno, NV 89557,

[2]Earth Resource Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139,

[3]Aprovechar Lab L3C, Montpelier, VT 05602, [4]Hi-Q Geophysical, Inc., Ponca City, OK, 74601

Correspondence: connormsmith@nevada.unr.edu, jfaulds@unr.edu

Keywords: Geothermal, Great Basin, Nevada, play fairway analysis, PFA, machine learning, permeability, exploration, neural networks, feature selection, principal component analysis, training sites, clustering.

ABSTRACT

We are introducing machine learning methods to the play fairway analysis to generate geothermal potential maps to support the evaluation of geothermal resource potential and the exploration for undiscovered blind geothermal systems in the Nevada Great Basin region. Our project aims to identify new ways to combine the play fairway data and empirically organize relationships between feature weights and labels in an improved workflow. As a means of doing this, we introduce machine learning methods to evaluate the influence of certain geological and geophysical features/feature sets in predicting geothermal favorability. This report highlights promising approaches based on supervised and unsupervised learning methods. First, we demonstrate a filter method applied to supervised classification modeling. The supervised filter method is based on permutation analysis to evaluate every possible feature combination/drop out scenario and rank feature influence based on the performance variance of supervised classification models. Additionally, we present an unsupervised factor analysis based on principal component analysis coupled with a semi-supervised k-means clustering algorithm. This analysis allows us to identify the optimal number of groups/clusters for training sites and structural settings to identify feature patterns including correlation, variance, and latent and dominant feature relationships. The results from these methods offer a promising avenue for identifying favorable sources of predictive information to identify the locations of blind geothermal systems and furthering our understanding of complex geothermal feature and label relationships in the Great Basin region and beyond.

1. INTRODUCTION

The Great Basin region is a world-class geothermal province with ~720 MWe of current gross generation from ~24 power plants. Studies indicate far greater potential for both conventional hydrothermal and EGS systems in the region (Williams et al., 2009).

Most geothermal systems in the Great Basin are controlled by Quaternary normal faults and generally reside near the margins of actively subsiding basins. Geothermal fluids commonly upwell along basin-bounding faults, flow into permeable subsurface sediments in the basin, and thus do not always daylight directly along the fault. Thermal springs may emanate many kilometers away from the deeper source, or thermal groundwater may remain blind with no surface manifestations (Richards and Blackwell, 2002). Blind systems are thought to comprise the majority of geothermal resources in the region (Coolbaugh et al., 2007). Thus, techniques are needed both to identify the structural settings that allow geothermal systems to form (e.g., Curewitz and Karson, 1997; Faulds et al., 2006; Faulds and Hinz, 2015) and to determine which areas may harbor subsurface hydrothermal fluid flow.

Geothermal play fairway analysis (PFA) is a concept adapted from the petroleum industry (e.g., Doust, 2010). As applied to geothermal exploration, PFA involves the integration of geological, geophysical, and geochemical parameters indicative of geothermal activity as a means to identify the most likely locations for significant geothermal fluid flow (i.e., play fairways). The Nevada PFA project (Faulds et al., 2017) focused on defining geothermal play fairways and generating detailed geothermal potential maps for ~1/3 of Nevada (Figures. 1 and 2; Faulds et al., 2017, 2018) to facilitate the discovery of blind geothermal systems. Thus far, the project has led to the successful discovery of two new blind geothermal systems (Faulds et al., 2018, 2019; Craig, 2018).

The original Nevada PFA incorporated ~10 geological, geophysical, and geochemical parameters indicative of geothermal activity. The linking of parameters was performed by multiplying each by a unique "weight", then combining weighted parameters into a linear summation (Figure 2). The weights used in that analysis were derived using a combination of statistics, including Bayesian-based weights-of-evidence and logistic regression (e.g., red numbers in Figure 2) through the analysis of 34 benchmark sites of known, relatively high temperature geothermal systems (≥130°C) in the study area and expert judgment (black numbers) due to known limitations of some datasets and small number of training sites. Our current efforts look to address some of the challenges and limitations of the original Nevada PFA study through the inclusion of new and improved datasets and principles and techniques of machine learning (Faulds et al., 2020, Brown et al., 2020).

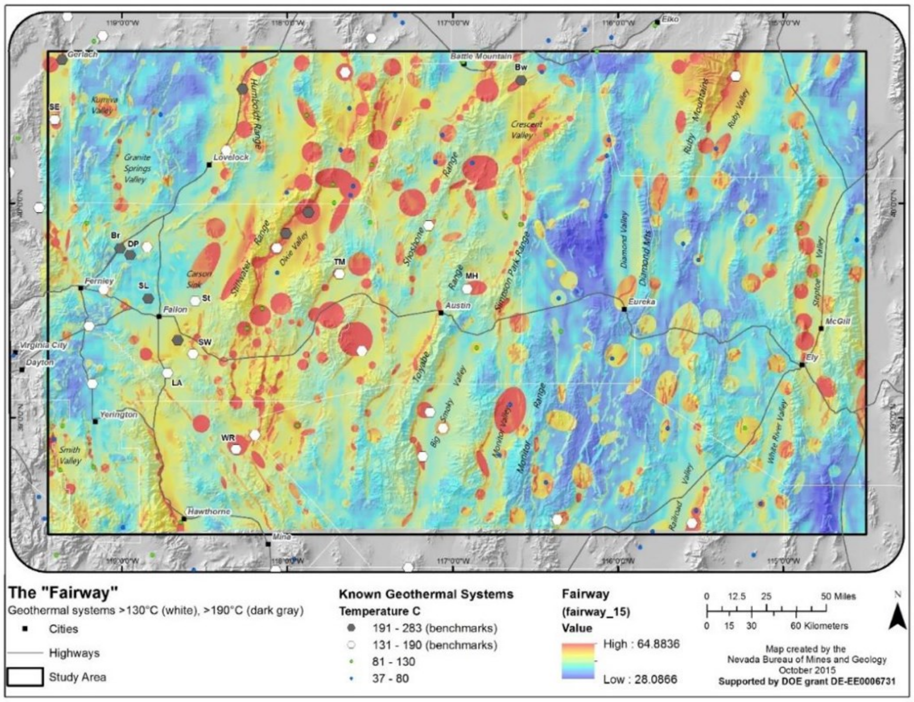

Figure 1: The initial play fairway model of the study area in west-central to eastern Nevada. Plotted within the study area are fairway prediction values ranging from a low ~28 to a high of ~ 65. Known geothermal systems comprising 34 relatively high-temperature (>130°C) benchmarks are shown in dark gray and white (modified from Faulds et al., 2018).

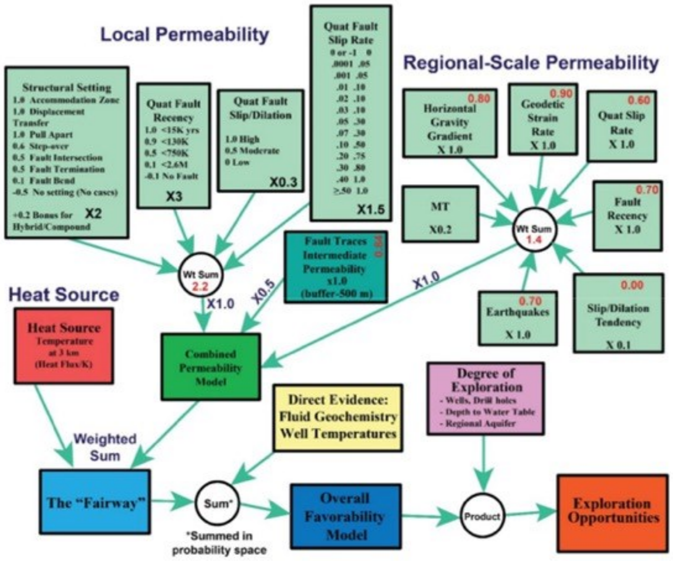

Figure 2: Nevada PFA workflow. Note the mixture of numerical and categorical/ordinal features, each tied to geographic positions on a map with varying scales of resolution (modified from Faulds et al., 2017).

1.1 Machine learning Efforts

The goal of our machine learning (ML) efforts is to identify the simplest models so as to understand what defines the key relationships between geological and geophysical data inputs and label outputs of predicted favorability values. We have identified several key advantages in replacing the combined statistically based (utilizing multiple datasets) and expert-driven methods for determining model parameters (weights) with ML optimization methods. First, the outcome can be cast as a probability, defensible through validation tests. Second, careful implementation can reduce or eliminate biases in the choices of the most appropriate feature set and in the choices of the network architecture controlling how the features are combined. Finally, the algorithms can be easily automated, generalized, refined, and extended to accommodate new data sources.

Ongoing work (Brown et al., 2020) has demonstrated a workflow utilizing a supervised learning approach, in particular with the use of artificial neural networks (ANN) and data augmentation as being capable of recreating and improving the original results of the play fairway study. Supervised learning involves an algorithm that is optimized to associate pairs of measurable features and labels by providing it with many examples. Our efforts include translating the PFA datasets (original and enhanced) and adding some new datasets into a form suitable for machine learning algorithms and exploring a variety of ANN architectures. Major challenges encountered during our efforts include addressing:

- A small number and potential imbalance of training examples (initially only 34 positive benchmarks were available).
- Variable data types (a mixture of categorical and numerical)
- Complex feature and label relationships

In the course of our work, we explored remedies for each of these issues. Below, we highlight the introduction of new training data and ML methods to evaluate complex feature input and label output relationships to improve our future workflow.

2. METHODS

2.1 Training Sites

One issue that we have encountered in our work is how to approach training data. For optimal benefits, ML methods commonly employ a much larger number of training data than the inventory of benchmark sites used in the initial PFA analysis. Additionally, many supervised learning methods require a balanced inventory of positive and negative training samples. Small numbers of samples and imbalanced data can lead to over-fitting and a corresponding reduction in the capability for generalization (being able to accurately predict the labels of datasets not used in the training process).

Remedies explored for this problem include data augmentation or simulation (Brown et al., 2020) and maximizing a set of training data from our regional data inventory. Our efforts have led to a training site inventory (Figure 3) that includes 83 positive sites from known geothermal systems (≥39°C) and 62 negative sites from deep and cool wells (mostly from oil and gas exploration). In addition, we evaluate the locations of favorable structural settings (e.g., Faulds and Hinz, 2015) as training data for our unsupervised learning (discussed below in sec 2.2).

Using a broader temperature range (≥37°C), additional positive geothermal sites became available. In the case of negative sites, criteria were established to select them from a relatively large number (>250) of relatively deep (>1 km) oil and gas wells in the region that do not show temperature anomalies. In our criteria, the distribution and depth of the carbonate aquifer (Brooks et al., 2014) in eastern Nevada was reviewed for its possible impact of disguising geothermal anomalies. It was decided to use wells in a regional database that were at least 2 km deep in areas underlain by the carbonate aquifer, and 1 km deep outside the carbonate aquifer. Those wells meeting these conditions were then evaluated on the basis of temperature. The temperature assigned to a well was compared to the predicted temperature at the bottom of the well based on the regional heat flow and temperature gradient map used for the play fairway analysis (Faulds et al., 2017). If the regional predicted temperature was greater than or equal to the assigned temperature (i.e., no temperature anomaly), the well was considered as a potential negative training site. Next, a de-clustering algorithm was developed to reduce the number of possible negative training sites in areas with a high-density of drilled holes. This de-clustering involved the following conditions:

- A negative training site always corresponds to a given well location and its attributes.
- No two wells selected as negative training sites could be closer than 5 km to each other, or closer than 5 km to a positive training site.

It was found that the distribution of such sites was not too complicated, such that it was possible to determine the optimal selection of sites in a reproducible manner from careful visual inspection. Finally a detailed quality review of positive and negative sites was completed, including a check of spatial location, temperatures, and depths.

Given the extent of our study area (96,000km^2), the population of training sites represents a relatively high spatial density compared to most geothermal provinces around the world. However, ML problems rely on tens to perhaps hundreds of thousands of labeled examples from which to train, develop, and test network algorithms and architecture. Our initial deep learning efforts relied primarily on data augmentation (e.g. generative adversarial networks (Goodfellow et al., 2016)) to produce large simulated datasets. Now that we can expand the training datasets, we are able to obtain a better understanding of physical controls for geothermal favorability and can better support ML modeling approaches (e.g., data augmentation, regularization, transfer learning). To facilitate this work and provide context in feature/label analysis, we also outlined principal domains that reside within the PFA study area (Figure. 3). General characteristics of these domains are listed below:

Western Domains:

- The Walker Lane: a northwest trending belt of largely transtensional dextral motion (Stewart, 1988; Faulds and Henry, 2008) that accommodates ~20% of the right-lateral motion between the Pacific and North American plates (Dixon et al., 1995, 2000; Hammond et al., 2009; Kreemer et al., 2012). This domain hosts the highest strain rates in the study area, a higher density of earthquakes, and noticeably lower fault slip and dilation tendency along its northeastern margin relative to the rest of the study area.

- Central Nevada Seismic Belt: a north-northeast trending belt of high crustal strain rates and strong earthquakes (e.g., Caskey et al., 2004). Along with earthquake and strain signals, this domain hosts many faults with relatively recent offsets, as documented by fault recency values.

- Western Great Basin: represented by two regions respectively west and east of the central Nevada seismic belt, hosting relatively moderate to high strain rates of crustal extension with some dextral transtensional motion, high heatflow, and high slip and dilation tendency values.

Eastern Domain:

- Carbonate Aquifer: a regionally extensive, relatively cool and deep aquifer system that occupies most of the eastern Great Basin. Crustal strain in this area is generally a few tenths of millimeters per year (Hammond et al., 2009), significantly less compared to the western Great Basin. There are notable spots of low geodetic strain rates and recent faulting in the north half, and high earthquake density and low heatflow in the south. Additionally, most of this region is at a higher average elevation than the western domains.

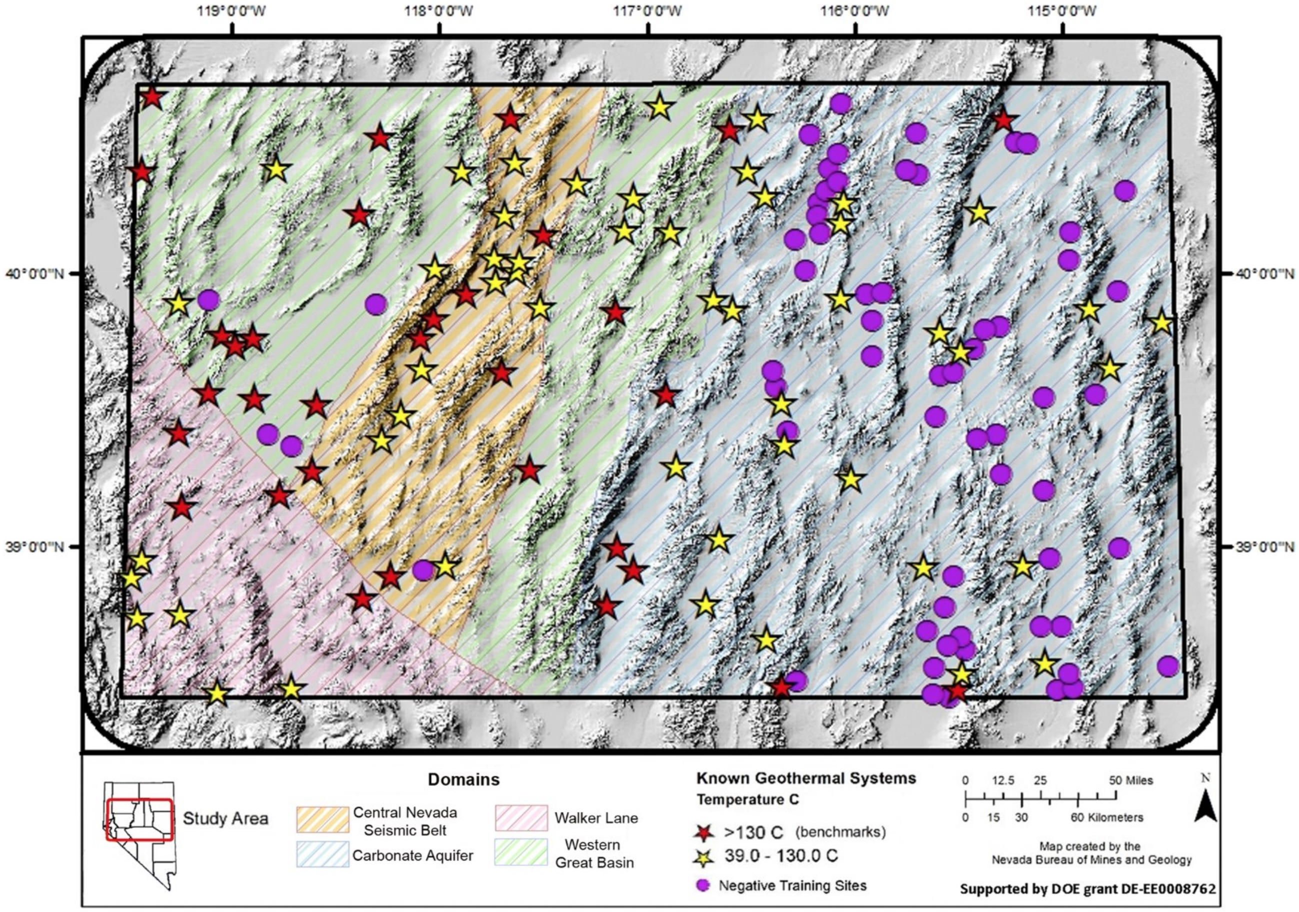

Figure 3. Distribution of positive (red and yellow stars) and negative (purple circles) training sites along with the extent of the structural domains (central Nevada seismic belt, carbonate aquifer, Walker Lane, western Great Basin)

2.2 Data Compilation

As can be seen in the original workflow diagram (Figure 2), we have sets of data divided by their perceived information content indicating local, intermediate, and regional scale permeability, as well as heat. Individual parameters are known in map form throughout the study area and are referred to as "features". In our machine learning analysis, these features are normalized to allow comparison between strongly contrasting magnitudes using the z-score transformation:

$$z = \frac{x - \bar{x}}{s_x} \tag{1}$$

where x is the feature to be transformed, $\bar{x}$ is the sample mean, s_x is the sample standard deviation, and z is the transformed feature. Z-score normalization transforms any sample distribution to a corresponding feature with a mean of zero and a standard deviation of one, while retaining rank information. We selected ten numerical features pertaining to regional permeability for this study, where each feature map represents continuous real numbers known at every grid block (250 m grid block size with >1.6 million grid blocks) in the study area. Color contoured maps of these features are shown in Figure 4 and include the data that were incorporated in the initial PFA (geodetic strain rate, Quaternary slip rate, fault recency, Quaternary fault slip and dilation tendency, and earthquake density), as well as data that were improved (augmented) (Quaternary fault density, horizontal gravity gradient), and new types of data that have been newly integrated during our machine learning study (horizontal magnetic gradient, heatflow (which replaced heat source at 3 km), and a 30 m digital elevation model.

For this study, we have converted the Quaternary fault traces map representing intermediate permeability from a categorical feature (1's and 0's) to a continuous numerical feature by calculating fault density using a gaussian filter (Brown et al., 2020). The original fault layer is populated quite densely and uniformly in the study area, and thus is easily transformed for this study.

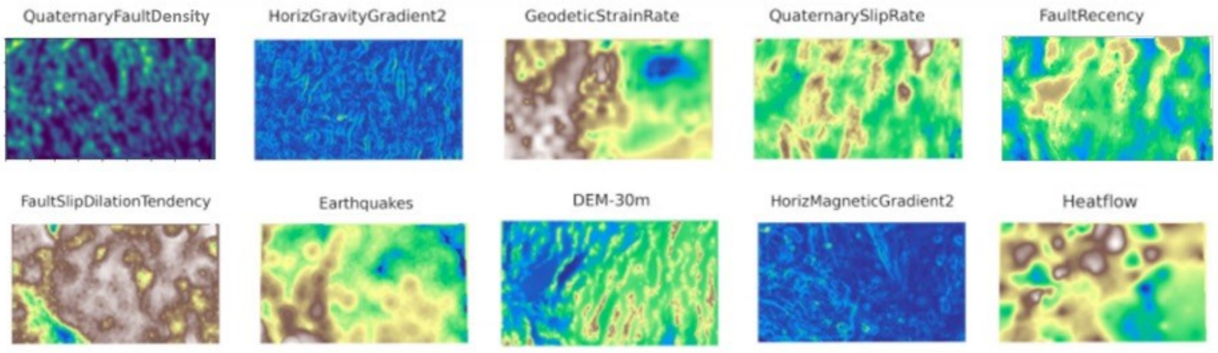

Figure 4: Select numerical features explored in this study. Warmer colors are associated with higher numerical values.

We also explored integrating additional categorical features (fluid geothermometry data, paleo-geothermal deposits [sinter, tufa, travertine], and local permeability features) by evaluating continuous numerical transformations (e.g., conversion to density/distance); however these were identified as too sparsely distributed to introduce alongside intermediate and regional features for this study. Of all the categorical data examined, local permeability factors, in particular structural setting types, were the most relevant. Local permeability was used as the highest weighted feature set in terms of predicting geothermal favorability in the original play fairway analysis. The challenge that these features present is that numerical values for all local permeability features, such as the structural settings, are known only in elliptical regions that are heterogeneously distributed throughout the study area (Figure 5). We attempted to integrate numerical values of this feature set directly. However, when evaluated at training sites, each local permeability feature does not contribute enough variance (close to zero) to produce meaningful results. With ~85% of positive sites being within the margins of structural setting ellipses and ~90% of negative sites falling outside these margins, local permeability features have nearly the same values in all negative and positive samples, respectively.

Given the knowledge the local permeability feature set offers, as a work around we demonstrate below an example of treating the locations of these settings as training samples in our unsupervised analysis. This is done by grouping the grid blocks within each structural setting area (ellipse) and taking the mean value for each numerical feature, thus converting each ellipse into a single representative block.

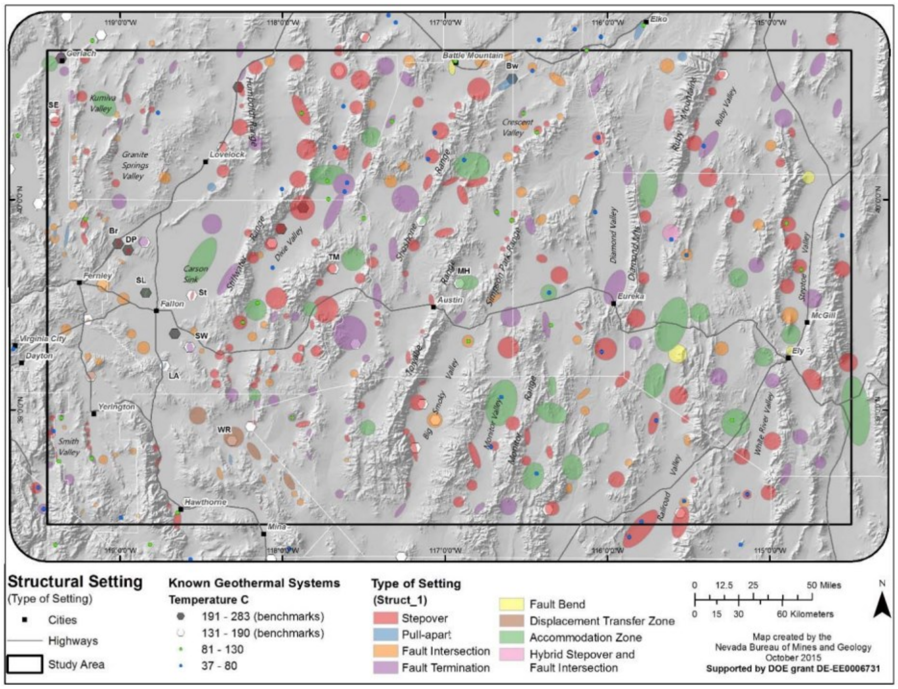

Figure 5: Favorable structural settings identified in the study area and color coded according to type of setting. About 375 settings were recognized, which incorporate nearly 12% of the area. This includes 174 step-overs, 76 fault terminations, 76 fault intersections, 30 accommodation zones, 9 pull-aparts, 6 displacement transfer zones, and 4 fault bends. (Faulds et al., 2015)

2.3 Analysis

In this study, we use a combined supervised and unsupervised approach to arrive at observations of certain features and feature sets at training data and structures. Modeling is performed using python and the Sklearn machine learning library (Pedregosa et al., 2011). After selecting numerical continuous features and preparing the data for machine learning compatibility, we first introduce the permutation supervised model filter method to identify feature dependence in the classification of positive and negative training and test data. This analysis gives an idea of which features might best pertain to favorability analysis in training data. In our unsupervised modeling we introduce principal component analysis (PCA) to identify which features offer the most independent information, and cluster the reduced dataset using the k-means semi-supervised algorithm to identify spatial patterns in our data (PCAk). We describe each of these methods in the following sections.

2.3.1 Feature Selection Analysis: Permutation Importance Filtering

Supervised feature selection methods are commonly used in high dimensionality data problems for isolating features that may or may not be relevant in the overall predictive performance of a model. A traditional approach considered effective if the number of available examples (training sites) is relatively small (as is our case) is the use of filter methods (e.g., Radivojac et al., 2004). We present the use of a filter method based on permutation dropout, whereby we evaluate every possible feature set scenario as each feature is dropped out one at a time and a model is fit to the remaining features. Permutation importance is defined to be the difference between the baseline metric and metric from permutating the feature column. The implementation of this method is independent of our model and metric choice, so we compared returned accuracy as our metric from several different supervised classification models. We found nearly identical results in each model (artificial neural network, support vector machine, random forest ensemble). For this experiment, the data of positive and negative training sites was split into separate training and test sets. The training set was used to optimize each model, and the separate test set (25% of the total training data) was held out from model training to enable evaluation of which features may contribute the most to the generalization power of the model. During this process we increased the number of permutations until we had consistent results. The results from permutation analysis of an artificial neural network (100 hidden layers, rectified linear unit activation function, and the stochastic gradient-based optimizer) which returned fairly high results during training (90.74%) and testing (91.89%) with the full feature set are shown in Figure 6.

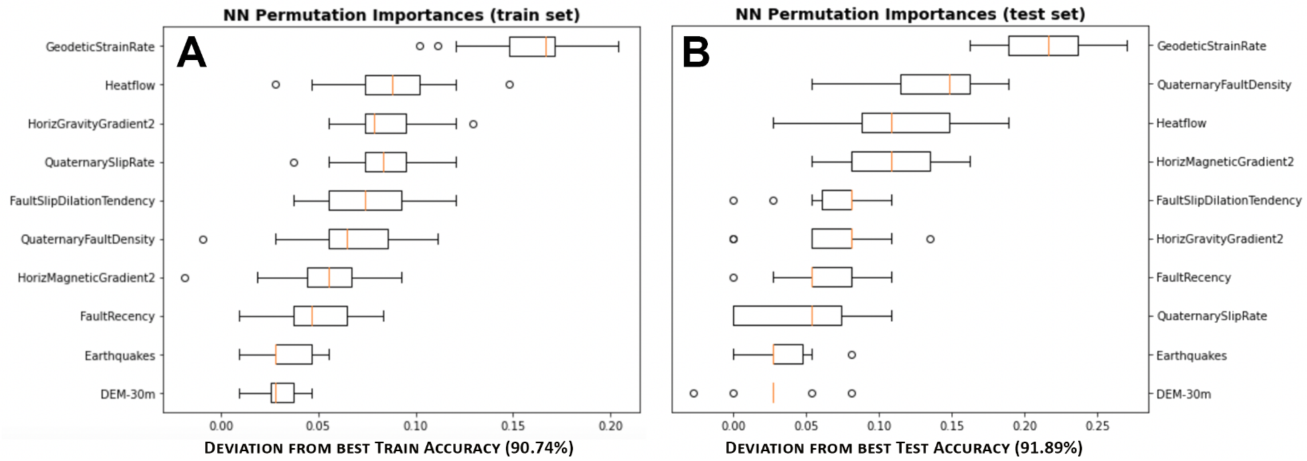

Figure 6: Boxplots of permutation importance scores for training (A) and test data (B) based on a multilayer artificial neural network model. Scores represent feature influence on model performance if dropped out during training/testing. Orange line represents mean values, the box bounds represent the mean value +/- standard error, and whisker bounds represent the mean value +/- standard deviation. Dots represent outlier sample values.

Both the training and test experiments indicate that geodetic strain and heatflow are principal features in dictating model performance. Additionally, Quaternary fault density appears to be a key feature in the test set, which may indicate that this feature is better suited in supporting the generalization power of our model in testing versus directly fitting our data in the training process. Conversely, we can also identify the features that are less relevant to the classification of positive and negative sites using ML, including the DEM-30m and earthquake feature maps. Because permutation importance does not reflect the intrinsic predictive value of a feature by itself, but how important the feature is for a particular model and task, it is difficult to interpret why these features may receive their relative rankings. Given the nature of our problem, where each of these features is uniquely structured and representative of different geologic and geophysical characteristics, it is clear that we cannot make direct observations of how they relate or what significance they have in distinguishing positive and negative sites without providing content into their spatial relationships.

2.3.2 Principal Component Analysis and Clustering

As an alternative to feature selection, we introduce unsupervised learning using principal component analysis (PCA) alongside semi-supervised k-means clustering. This work draws from similar studies (e.g., Lindsey et al., 2018; Pepin 2019; Vesselinov et al., 2020), which utilized matrix factorization algorithms paired with clustering algorithms to characterize and classify signatures of permeability and heat at geothermal systems. The unsupervised learning aspect of PCA and other matrix factorization algorithms (e.g., non-negative matrix factorization, singular value decomposition) is that they are able to generate (learn) a reduced representation of a data matrix in the form of a weighted linear combination of a mixing matrix and feature matrix (Figure. 7).

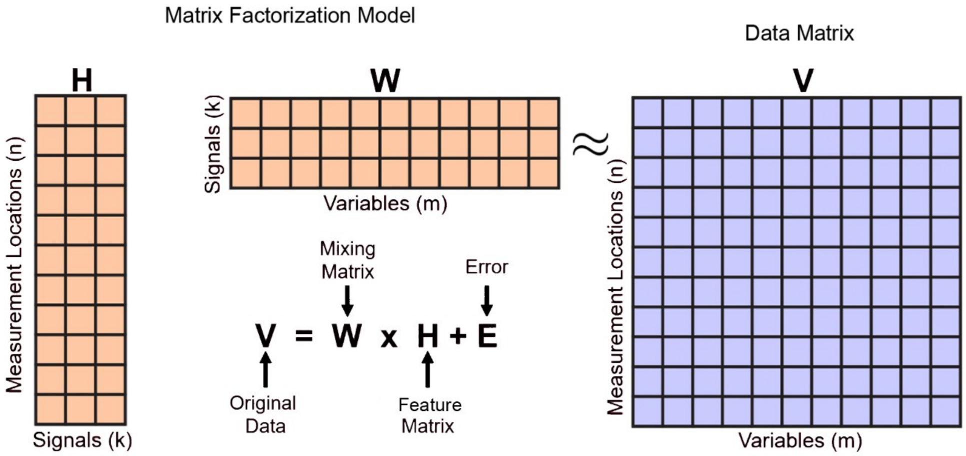

Figure 7: Matrix factorization schematic and equation. The data matrix (X) is decomposed into two related matrices that represent its sources of variation: a mixing matrix (W) and feature matrix (H), where the rows of the mixing matrix quantify the sources of variation among measurements (n), and the columns of the feature matrix (m) quantify the sources of variation among the measurements.

The goal of our PCAk analysis is to (1) decompose a matrix X of size (n,m) into a feature matrix H of size (n,k) and mixing matrix W of size (k,m), and (2) find the optimal number signals (k) to cluster our measurements. With PCA, signals (k) are known as principal components (PCs), which are ordered by their proportion of dataset variance explained (i.e., PC1 explains the largest proportion of variance, PC2 the second largest, and so on). The solutions for PCs are constrained to be orthogonal to one another. There are as many PCs as there are variables considered, and they take the following form:

$$PC1 = \beta_1 V_1 + \beta_2 V_2 + \cdots + \beta_m V_m \, , \tag{2}$$

where PC1 is the first principal component value for a given data point (i.e. scores), V_m are the original variables considered, $ß_m$ are the PC loadings or weights for the first PC, and m is the total number of variables considered in the analysis. The loadings vary between -1 and 1 and therefore scale the influence of the original variables based on their contribution to each PC (Pepin, 2019). PCA is often useful in exploratory data analysis for screening large numbers of variables, some of which may be correlated (i.e., some variables may not be independent) to identify a subset of factors that best represents the behavior of the group (Lindsey et al., 2018).

2.4 PCAk Results

In this section, we present the results of PCAk on select features at training sites and at structural settings described in section 2.1 and 2.2. Three major cluster groups (k=3) were identified for each analysis. Figure 8 shows results from our training inventory, and Figure 9 shows results from mean numerical values at structural settings. Cluster results are visualized by color. In each figure we project spatial locations of each cluster member onto a map of the DEM-30m layer (where darker colors indicate higher elevation) that also includes the labeled domains and their boundaries described in sec 2.1.

PCA results are displayed using a biplot of the first two principal components (with % variance explained in axes), whereby data are represented as points (the closer together two points are, the closer their common denominators), while features are shown as vectors. A vector is defined from the center of the plot to the vector vertex (endpoint), and the length of the vector is proportional to the fraction of the total variance explained by that feature, where larger vectors have a higher influence on data position. The arrowhead on each vector corresponds to high values of that particular variable. The cosine of the angle between any two vectors is approximately equal to the correlation coefficient between the two variables; therefore, two vectors that are separated by a small angle represent variables that are likely to be positively correlated, two vectors that are orthogonal to each other represent features likely to be independent, and vectors that form angles greater than 90° are negatively correlated (vectors at 180° have a high negative correlation) (Otero et al., 2005).

The k-means analysis alongside PCA modeling (PCAk) is performed based on the first three principal component scores, which account for over 60% of the overall variance in both the training site and structural setting datasets. Determining k in the k-means clustering involves experimenting with a range for the number of clusters and evaluating the compactness of the resulting solution; an elbow in a plot of within-group sum of squares (WSS) as a function of the number of clusters is commonly used to denote the appropriate number of clusters present in the data (Everitt et al., 2011).

Figure 8 shows the biplot and spatial clustering of positive and negative training sites based on PCAk analysis. In the biplot (Figure 8A), the distribution of most feature vectors and positive sites is in the positive PC1 direction, while most negative sites and the DEM-30m vector are distributed in the negative PC1 direction. Correlating these training site cluster members with the structural domains shows that sites in cluster-1 (cyan/blue) are predominantly located in the western domains, especially the central Nevada seismic belt (CNSB). Most training sites in this cluster are known geothermal systems and are linked to high values of geodetic strain, earthquakes, heatflow, fault recency, and low dem-30m values. Cluster-2 training sites (green) are predominantly located in the carbonate aquifer domain. Most sites in this cluster are negative linked to high dem-30m values (higher elevation), and negatively correlated to every other permeability features besides fault slip and dilation tendency. Cluster-3 training sites (violet) are also located mostly in the carbonate aquifer domain. Cluster-3 hosts a balanced mix of positive and negative sites and is mainly associated to high horizontal gravity gradient and Quaternary fault slip rate values.

Figure 9 shows the biplot and spatial clustering of structural settings based on PCAk analysis. A total of 68 out of the 83 positive sites fall within structural ellipses. Many of the same feature, cluster, and domain relationships seen in Figure 8 also appear in Figure 9; however, there is a more balanced distribution of variance captured by each feature in the PC1 and PC2 direction, and certain feature vectors have changed in orientation and length. Structures representing cluster-1 (cyan/blue) are predominantly located in the western domains, especially the Walker Lane and southwestern part of the central Nevada seismic belt (CNSB). Cluster-1 is linked to high values of earthquakes, geodetic strain rate, fault density, and low values of slip and dilation tendency. Cluster-2 structures (green) are predominantly located in the carbonate aquifer domain, with some groupings in the southern part of the western Great Basin (WGB) domain. Structures in cluster-2 are associated to low values of most permeability features, especially fault recency and Quaternary fault slip rate. Cluster-3 structures (violet) are located mostly in the western domains, with some structures in the northern part of the carbonate aquifer region and along its western margin. Most structures in cluster-3 are correlated to high values of Quaternary fault slip rate, fault recency, and dem-30m.

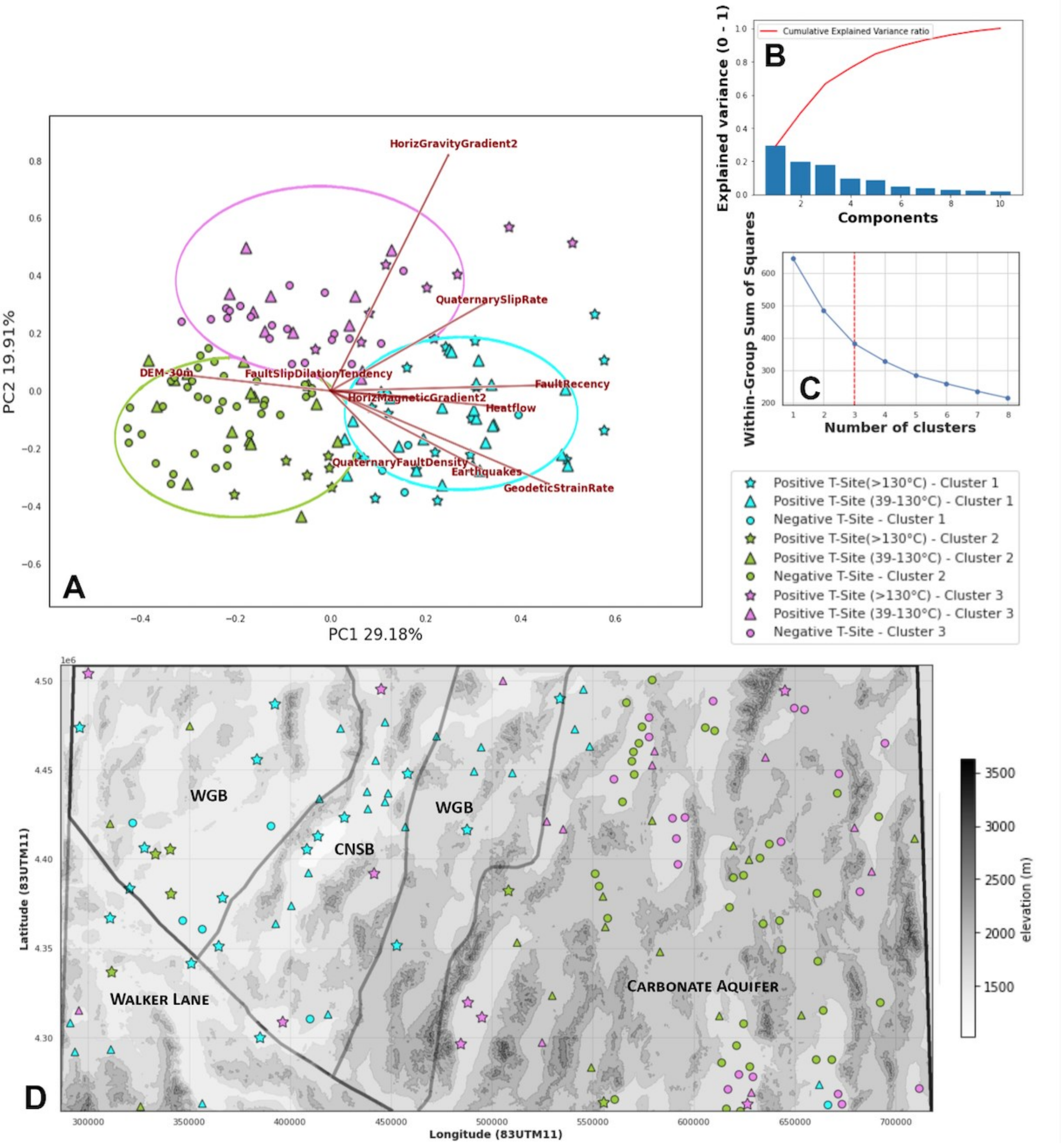

Figure 8: (A) PC1 vs. PC2 biplot; (B) PCA explained variance bar plot; (C) WSS cluster plot; (D) and study area map with domains that present training data PCA and k-means clustering results. Stars are higher temperature systems (>130°C), triangles lower temperature systems (≥37°C), and circles are negative training sites. Abbreviations: T-site = training site, CNSB = central Nevada seismic belt, WGB = western Great Basin.

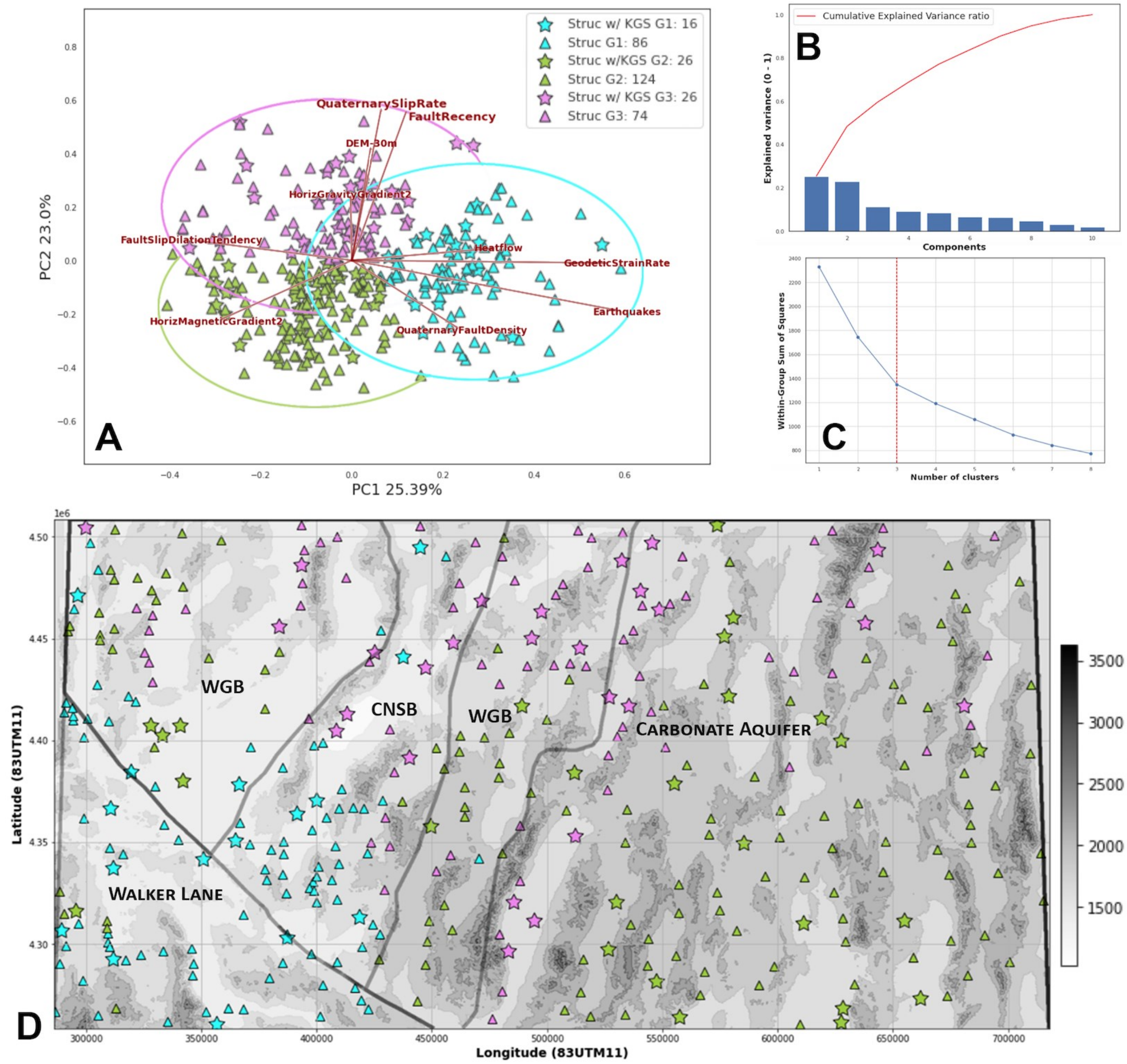

Figure 9: (A) PC1 vs. PC2 biplot; (B) PCA explained variance bar plot; (C) WSS cluster plot; (D) study area map with domains representing the structural setting PCA and k-means clustering results. Stars are structures that host a known geothermal system/positive site (≥37°C) and triangles structures without identified systems. Abbreviations: Struc = favorable structural setting, KGS = known geothermal system, CNSB = central Nevada seismic belt, WGB = western Great Basin.

Both the PCAk analysis with training sites and favorable structural settings reveal spatial feature patterns that appear to define subtypes of geothermal systems in each cluster set. Many of these patterns emerge due to a strong contrast between the eastern and western domains, including differences in geodetic strain rate, earthquakes, fault slip rate, and heatflow. Cluster group-1 (cyan/blue colors) hosts systems that are linked to low elevation zones with high values of geodetic strain, earthquakes, and heatflow and primarily occupies the Walker Lane and central Nevada seismic belt. The other cluster groups (2 and 3) are more broadly distributed and overlap considerably. Known geothermal systems in cluster group-2 (green) are generally located in the central and southern portions of the western Great Basin and carbonate aquifer and are closely linked to zones of low values of fault recency (i.e., older Quaternary faults), heat flow, and fault slip rate. Known systems in cluster group-3 (violet) are generally linked to high values of fault recency, Quaternary fault slip rate, and horizontal gravity gradient. With further analysis we can start to examine possible relationships between 1) different types of structural settings and the cluster groups, 2) structural settings that host known systems and those without identified systems, and 3) the most diagnostic features of known systems in the individual structural domains.

3. DISCUSSION AND CONCLUSIONS

The combined results from our supervised and unsupervised methods demonstrates how we can utilize ML to estimate the relative importance of each feature in a set based only on a relatively small sampling population of training samples and/or structures. First, we demonstrate the results of a simple classification problem. Our supervised ML filter method shows how a classification model may rank features at our training sites in terms of importance during the training and testing stages. Commonly, supervised methods may be joined to statistics, ranks, and knowledge into how to best design a model, transfer learning approach, etc., as illustrated in ongoing work in our project by Brown et al. (2020). The key takeaway from this work is that supervised feature selection methods provide an automated way to diagnose how features in our training samples may be represented in a learning problem, where certain features are better at directly fitting our data in the training process and others better at generalizing during the testing process to approximate a positive or negative prediction. Although we have demonstrated that a supervised network can return promising metrics (e.g., high classification accuracy) with our training samples, it is difficult to draw significance from the permutation importance scores of each feature (as these values are non-unique) without context from how each feature is spatially structured and related.

This is where our unsupervised dimensionality reduction methods like PCAk come into play. By ignoring target variables, we can better identify principal correlations between features and measurements. Unsupervised learning in our case focuses on distinguishing groups of positive and negative training sites (Figure. 8) and groups of favorable structural settings with and without known systems (Figure. 9) in order to evaluate which features offer the most independent information and which are highly corelated and thus redundant. For example, we can relate observations from PCAk with training sites back to our supervised permutation filtering (sec 2.3.1) which demonstrated that geodetic strain rate, heatflow, and Quaternary fault density are key features in classifying positive and negative sites. The result of our PCAk analysis of training sites indicates that the clustering of a large proportion of our positive sites have a strong positive correlation to these same features (Figure. 8). PCAk can also inform about feature correlation. The training site biplot identifies earthquakes and geodetic strain as strongly correlated, which may explain why our permutation analysis did not present earthquakes as a high scoring feature. It is commonly observed that with two collinear features, supervised models depend more on the one that best explains variance (Pedregosa et al., 2011), and thus a supervised model may be de-emphasizing the earthquakes feature in this study if both features are accessible.

Knowledge of the influence of individual features can help in the assessment of potential overfitting with any of our ML efforts. PCA and other linear models/dimensionality reduction methods (e.g., non-negative matrix factorization) give clues into variance control of features. Alongside dimensionality reduction, clustering provides clues into training site grouping/outliers with respect to a feature set. This information is helpful not only for deciding which features to use in our model, but also for performing de-correlation and/or identifying extreme values to reduce the effect of spurious outlier samples and features.

3.1 Constraining Our Problem

Although our PCAk method is better constrained than the supervised feature selection method, it too suffers from non-uniqueness associated with the selection of variables to include in the analysis, the method of handling outliers, and the employed clustering algorithm. For example, the geodetic strain feature (Figure. 4) in our study uses non-corrected fields (that include the transient part of the relaxation strain rate field) and was chosen over the corrected field because part of this signal may reflect the viscoelastic transient effects associated with earthquakes in the region over the past ~100 years (Faulds et al., 2015). This feature may provide new insight or affect the distribution of clusters if the alternative corrected field is used. Removing the geodetic strain feature all together does not appreciably change the training site (Figure. 8) or structure (Figure. 9) PC1 vs. PC2 biplot, but this is likely a product of its strong correlation to the earthquake feature and the number of features considered in this study; analyses that consider fewer variables will be more susceptible to larger sensitivities associated with variable selection and handling. In practice, determining which datasets to include and how to represent them in the analysis will likely depend on identifying regional constraints and the level of detail in PFA input data. Additionally, the type of k-means clustering algorithm used, or the use of an entirely different clustering method (e.g. density based spatial clustering of applications with noise (DBSCAN)) may alter the results. A significant advancement to our approach would be to employ ensemble techniques, such as Monte Carlo methods, that consider different combinations and representations of variables and outlier criteria to gain a better understanding of these non-uniqueness issues and find best-fit solutions (Pepin, 2019). Also, there are cases where discriminative information actually resides in components with smaller variances, such that PCA could greatly hurt classification performance. If most discriminative information is in smaller eigenvectors, we will want to explore isolating these vectors or introducing alternative matrix decomposition methods (e.g., non-negative matrix factorization) to our problem (e.g., Vesselinov et al., 2014, 2020).

Generally, our initial PCAk results yield a more unique solution if focused on how our groupings are controlled by structural domains of the region. Thus far, the information from these results agrees well with previous work in the PFA study, while utilizing a relatively straightforward methodology. This approach also allows investigation of both the co-location and correlation of variables and increase in the level of detail and quality of PFA input data by constraining the influence of strongly correlated anti-correlated features to each principal domain. Overall, we were surprised by how well organized our clustering results turned out in the context of major structural domains. This was especially evident when we trained our PCAk model on the 375 structural settings. The comparison of our structural setting modeling to our training site modeling indicated similar feature influences, but also the variation in our results appears to highlight which features act better for generalizing solutions and which are more tailored to specific solutions. For example, the horizontal gravity gradient is a feature which has a lot of discriminatory power in the supervised feature selection (sec 2.3.1) and PCAk analysis (sec 2.4) of training sites, but this feature becomes less significant in the PCAk modeling of the mean data within a structure. By nature we are smoothing this and other features by taking the mean values within each structure, and also resolving a more generalized spatial pattern in our domains with a larger sampling population. It is worth exploring more representative approaches with structural settings, such as constraining which grid blocks within structural ellipses best represent where known geothermal sites reside and/or where areas of geothermal upwelling occur. Even still, the PCAk result from the mean numerical representation of structures in this study, as compared to training site samples, helps to identify favorability criteria in geothermal exploration efforts and indicate which features may act as the best proxies for permeability/heat.

In the future, it should be possible to expand our PCAk analysis (or a similar unsupervised approach) to evaluate a larger sample population (e.g., the entire fairway) and use spatial patterns from training sites and favorable structural settings to evaluate unexplored regions (similar to Pepin, 2019). PCAk and similar methods also offer the opportunity to introduce new reduced feature inputs, where we can combine features that present a strong correlation in a principal component (e.g., combining earthquakes, geodetic strain rate, and fault density) as a preliminary step before our supervised modeling, thus simplifying the dimensionality of our inputs and structure of a model.

ACKNOWLEDGEMENTS

This material is based on work supported by the U.S. Department of Energy's Office of Energy Efficiency and Renewable Energy (EERE) under the Geothermal Technologies Program and Machine Learning initiative award number DE-EE0008762. Support from Jake DeAngelo, Jonathan Glen, and Erick Burns was provided by the U.S. Geological Survey Energy Resources Program. Additional support was provided by John Queen and Hi-Q Geophysical, Inc., Ponca City, OK. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

REFERENCES

Alexandrov, B. S., & Vesselinov, V. V.: Blind source separation for groundwater pressure analysis based on nonnegative matrix factorization, *Water Resources Research*, v.50(9), (2014), p. 7332-7347.

Bell, J.W., and Ramelli, A.R.: Active faults and neotectonics at geothermal sites in the western Basin and Range: Preliminary results, *Geothermal Resources Council Transactions*, v. 31, (2007), p. 375-378.

Bell, J.W., and Ramelli, A.R.: Active fault controls at high-temperature geothermal sites: Prospecting for new faults, *Geothermal Resources Council Transactions*, v. 33, (2009), p. 425–429.

Brooks, L.E., Masbruch, M.D., Sweetkind, D.S., and Buto, S.G.: Steady-state numerical groundwater flow model of the Great Basin carbonate and alluvial aquifer system, *U.S. Geological Survey Scientific Investigations Report,* 5213, (2014), 86 p.

Brown, S., Coolbaugh, M., DeAngelo, J., Faulds, J., Fehler, M., Gu, C., Queen, J., Treitel, S., Smith, C., and Mlawsky, E.: Machine learning for natural resource assessment: An application to the blind geothermal systems of Nevada: *Geothermal Resources Council Transactions*, v. 44, (2020),14 p.

Caskey, S. J., Bell J. W., and Wesnousky S. G.: Historic surface faulting and paleoseismicity in the area of the 1954 Rainbow Mountain–Stillwater earthquake sequence, *Bull. Seismol. Soc. Am.*, **94**, (2004), p. 1255–1275.

Coolbaugh, M.F., Raines, G.L., and Zehner, R.E.: Assessment of exploration bias in data driven predictive models and the estimation of undiscovered resources: *Natural Resources Research*, v. 16, no. 2, (2007), p. 199-207.

Craig, Jason W.: Discovery and analysis of a blind geothermal system in southeastern Gabbs Valley, western Nevada [M.S. Thesis]: *University of Nevada, Reno,* (2018),111 p.

Curewitz, D. and Karson, J.A.: Structural settings of hydrothermal outflow: Fracture permeability maintained by fault propagation and interaction, *Journal of Volcanology and Geothermal Research*, v. 79, (1997), p. 149-168.

Dixon, T.H., Miller, M., Farina, F., Wang, H., and Johnson, D.: Present-day motion of the Sierra Nevada block and some tectonic implications for the Basin and Range province, North American Cordillera, *Tectonics*, v. 19, (2000), p. 1-24.

Dixon, T.H., Robaudo, S., Lee, J., and Reheis, M.C.: Constraints on present-day Basin and Range deformation from space geodesy, *Tectonics*, v. 14, (1995), p. 755-772.

Doust, H.: The exploration play: What do we mean by it?, *American Association of Petroleum Geologists Bulletin*, v. 94, (2010), p. 1657-1672.

Everitt BS, Landau S, Leese M, Stahl D.: Cluster Analysis. *JohnWiley & Sons, Ltd., West Sussex, United Kingdom,* (2011), 348 p.

Faulds, J.E., Coolbaugh, M.F., Vice, G.S., and Edwards, M.L.: Characterizing structural controls of geothermal fields in the northwestern Great Basin: A progress report: *Geothermal Resources Council Transactions*, v. 30, (2006), p. 69-76.

Faulds, J.E., and Henry, C.D.: Tectonic influences on the spatial and temporal evolution of the Walker Lane: An incipient transform fault along the evolving Pacific – North American plate boundary, *in Spencer, J.E., and Titley, S.R., eds., Circum-Pacific Tectonics, Geologic Evolution, and Ore Deposits: Tucson, Arizona Geological Society*, *Digest* 22, (2008), p. 437-470.

Faulds, N.H., and Hinz, N.H.: Favorable tectonic and structural settings of geothermal settings in the Great Basin Region, western USA: Proxies for discovering blind geothermal systems, *Proceedings, World Geothermal Congress 2015, Melbourne, Australia*, (2015), 6 p.

Faulds, J.E., Hinz, N.H., Coolbaugh, M.F., Siler, D.L., Shevenell, L.A., Queen, J.H., dePolo, C.M., Hammond, W.C., and Kreemer, C.: Discovering blind geothermal systems in the Great Basin region: an integrated geologic and geophysical approach for establishing geothermal play fairways, *Final report submitted to the Department of Energy*, (2015), 245 p.

Faulds, J.E., Hinz, N.H., Coolbaugh, M.F., Shevenell, L.A., Sadowski, A.J., Shevenell, L.A., McConville, E., Craig, J., Sladek, C., and Siler D.L.: Progress report on the Nevada play fairway project: Integrated geological, geochemical, and geophysical analyses of possible new geothermal systems in the Great Basin region, *Proceedings, 42nd Workshop on Geothermal Reservoir Engineering, Stanford University, Stanford, California*, SGP-TR-212, (2017), 11 p.

Faulds, J.E., Craig, J.W., Coolbaugh, M.F., Hinz, N.H., Glen, J.M., Deoreo, S.: Searching for blind geothermal systems utilizing play fairway analysis, western Nevada, *Geothermal Resources Council Bulletin*, v. 47, (2018), p. 34-42.

Faulds, J.E., Hinz, N.H., Coolbaugh, M.F., Ramelli, A., Glen, J.M, Ayling, B.A., Wannamaker, P.E., Deoreo, S.B., Siler, D.L., and Craig, J.W.: Vectoring into potential blind geothermal systems in the Granite Springs Valley area, western Nevada: Application of the play fairway analysis at multiple scales, *Proceedings 44th Workshop on Geothermal Reservoir Engineering, Stanford University, Stanford, California*, SGP-TR-214, (2019), p. 74-84.

Goodfellow, I., Bengio, Y., and Courville, A.: Deep Learning, MIT Press, (2016).

Kohavi, R., G. John.: Wrappers for feature selection, *Artificial intelligence*. 97(1- 2), (1997), p. 273-324.

Otero, N., Tolosana-Delgado, R., Soler, A., Pawlowsky-Glahn, V., Canals, A.: Relative vs. absolute statistical analysis of compositions: a comparative study of surface waters of a Mediterranean river, *Water Resources*. 39, (2005), p. 1404–1414.

Pedregosa *et al.*: , *Scikit-learn: Machine Learning in Python* *JMLR* 12, (2011), pp. 2825-2830

Pepin, J. D.: New approaches and insights to geothermal resource exploration and characterization, [Ph.D. Dissertation] *New Mexico Institute of Mining and Technology, Socorro, New Mexico*, (2019),186 p.

Radivojac P., Obradovic Z., Dunker A.K., Vucetic S.: Feature Selection Filters Based on the Permutation Test. *Machine Learning, ECML Lecture Notes in Computer Science, Springer, Berlin, Heidelberg*. v. 3201, (2004), p. 334–346.

Richards, M., and Blackwell, D.: A difficult search: Why Basin and Range systems are hard to find, *Geothermal Resources Council Bulletin*, v. 31, (2002), p. 143-146.

Stewart, J.H.: Tectonics of the Walker Lane belt, western Great Basin: Mesozoic and Cenozoic deformation in a zone of shear, *in Ernst, W. G., ed., Metamorphism and crustal evolution of the western United States, Prentice Hall, Englewood Cliffs, New Jersey*, (1988), p. 681-713.

Vesselinov, V.V., Mudunuru, M.K., Ahmmed, B., Karra, S., Middleton, R.S.: Discovering signatures of hidden geothermal resources based on unsupervised learning, *45th Workshop on Geothermal Reservoir Engineering, Stanford University, Stanford, California,* SGR-TR-216, (2020), 11 p.

Williams, C.F., Reed, M.J., DeAngelo, J., and Galanis, S.P. Jr.: Quantifying the undiscovered geothermal resources of the United States, *Geothermal Resources Council Transactions*, v. 33, (2009), p. 995-1002.