PROCEEDINGS, 45th Workshop on Geothermal Reservoir Engineering
Stanford University, Stanford, California, February 10-12, 2020
SGP-TR-216

Discovering Signatures of Hidden Geothermal Resources based on Unsupervised Learning

V. V. Vesselinov, M. K. Mudunuru, B. Ahmmed, S. Karra, and R. S. Middleton

[1]Earth and Environmental Sciences Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA

Correspondence: vvv@lanl.gov, maruti@lanl.gov

Keywords: Geothermal energy, unsupervised machine learning, non-negative matrix factorization, clustering, hidden signatures.

ABSTRACT

Rapid advancements in the field of machine learning (ML) offer a substantial opportunity to accelerate discovery and reduce the costs associated with geothermal exploration, development, and production lifecycle. Application of new and innovative ML methods to multi-source and multi-physics datasets may lead to the discovery of new signatures or play fairway types for the detection of hidden geothermal resources. In our ML for geothermal exploration research, one of our goals is to discover the signatures (features) characterizing geothermal resources and favorable exploration sites from the regional-scale geothermal datasets. To achieve this goal, we have applied an unsupervised ML method to extract latent/hidden features or signals from the regional-scale geothermal data for geothermal resource exploration. The data describe the known-geothermal resources in southwest New Mexico. The unsupervised ML is based on a non-negative matrix factorization (NMF) method coupled with a custom semi-supervised *k*-means clustering algorithm. Our methodology, called NMF*k*, is capable of identifying latent/hidden signals, an optimal number of clusters, and a dominant set of features hidden in the large-scale geothermal datasets. Based on our NMF*k* analyses and associating the obtained ML results with site information (e.g., regional physiographic provinces), the optimal number of clusters identified is equal to 4. The dominant set of attributes, among a total of 22 geothermal attributes, that were identified through NMF*k* analysis include air temperature, gravity, depth to water table, elevation, crustal thickness, drainage, and lithium concentration. These dominant attributes in the geothermal data indicate favorable sources of data collection to explore geothermal resources in each province (e.g., The Rio Grande Rift, the Mogollon-Dalit volcanic field). Moreover, the proposed NMF*k* method is widely applicable to extract features/signals from large-scale geothermal data (including observational and simulation outputs). This broad applicability of our ML tools makes it attractive to discover, quantify, and assess hidden geothermal energy resources (e.g., meeting DOE-EERE GTO's mission).

1. INTRODUCTION AND MOTIVATION

Enhanced Geothermal Systems (EGS) utilize engineered subsurface reservoirs that are created where there is hot rock but insufficient natural permeability (e.g., see Brown et al. (2012), Kelkar et al. (2015), McClure and Horne (2014)). EGS provide potential for dramatically expanding the use of geothermal resources and represents a domestic energy source that is clean and reliable. Through efficient utilization of EGS, more than 100 GWe of economically viable capacity may be available in the United States. However, to fully develop EGS, we need to discover locations that favor permeability enhancement (e.g., Bielicki et al. (2015), Mudunuru et al. (2015)). An aim of this project is to identify site/regions that are both suitably hot and have suitable permeability to access untapped geothermal energy. To achieve this research goal, our first step is to focus on discovering hidden features/signals from an existing geothermal data that is representative of well temperatures and rock permeability (e.g., attributes related to faults and fluid flow).

To this end, we develop a new way to analyze available geologic, geochemical, and geophysical data (Bielicki et al. (2015); Pepin (2019)) to reduce the risk of geothermal resource exploration and increase success rates associated with EGS development. Here, we present an unsupervised machine learning (ML) method (Cichocki et al. 2009, Vesselinov et al. 2019) based on non-negative matrix factorization (NMF) coupled with a semi-supervised clustering algorithm to perform exploratory data analysis on site-scale and regional-scale geothermal data (Bielicki et.al, 2015 GDR; Pepin 2019) from New Mexico. Our unsupervised ML methodology, called NMF*k* (Alexandrov and Vesselinov 2014, Vesselinov et al. 2018), is capable of identifying (a) optimal number of hidden signals in geothermal data, and (b) the dominant set of attributes/parameters that correspond to these hidden signals. We note that our NMF*k* algorithm has been tested on a wide variety of synthetic and real-world site data other than geothermal (e.g., see Alexandrov and Vesselinov 2014, Vesselinov et al. 2018, 2019, TensorDecompositions, NMFk.jl, NTFk.jl).

The goal of our work is to discover the signatures (features) characterizing geothermal resources and favorable EGS sites from the regional-scale geothermal datasets. ML is used to clean, preprocess, and combine independent data streams (e.g., different geothermal data attributes) to analyze geothermal resources in southwest New Mexico. To discover latent/hidden features along with optimal number of clusters in large geothermal datasets, NMF*k* is at the forefront among various unsupervised ML methods such as PCA, ICA, SVD and its variants, *k*-means clustering, Gaussian mixture models (Friedman et al. (2001)). Since the geothermal data attributes analyzed here are non-negative (e.g., lithium and boron concentrations, fault density, heat flow, silica geothermometer temperature), a ML methodology that preserves the non-negativity when extracting hidden signals from these attributes is preferred, making NMF*k* a natural choice to analyze this non-negative geothermal data. Note that there may some instances of geothermal data that can be non-negative. For example, amplitude values of seismic or acoustic signals are not always non-negative. In such a case, the NMF*k* workflow allows for two alternative approaches. The first one is to preprocess the data by applying mathematical transformations to make it non-negative and amenable for classical NMF*k* analysis. An alternative approach allowed in NMF*k* is to relax some of the non-negative constraints in the matrix decomposition process.

The outline of the paper is as follows: In Sec.2, we briefly review our NMFk methodology. We provide details on the regional-scale data and associated attributes used in our NMFk analysis in Sec.3. The regional-scale data consists of well records obtained from our previous Geothermal Play Fairways Analysis project (e.g., Bielicki et al. (2015); Pepin (2019)). This dataset contains hidden information on presence of heat, fluid, chemical concentration, and permeability in southwest NM through data attributes such as silica geothermometer measurements, lithium and boron concentrations, depth to water table, and fault density. Results are presented in Sec.4 and conclusions are drawn in Sec.4. The ML analysis and the work presented here is a part of the ongoing initiative supported by the U.S. Department of Energy – Geothermal Technologies Office to apply machine learning for geothermal energy exploration.

2. UNSUPERVISED MACHINE LEARNING METHODOLOGY BASED ON NMFK

Given observational data X of size *(n,m)* with non-negative values, where n is the number of observational wells from which data are sampled and m is the number of geothermal attributes. The goal of NMFk is to (1) decompose this matrix X into a non-negative feature matrix W of size *(n,k)* and non-negative mixing matrix H of size *(k,m)* and (2) find the optimal number of hidden signals k. This is achieved by minimizing the following objective function O based on Euclidean norm for all possible values of $\boldsymbol{k = 1, 2, \cdots, d}$. The maximum value of d that can be explored is user-defined and cannot exceed m (the number of geothermal attributes):

$$\mathcal{E}^{(k)} = \|X - W \times H\|_2^2 \quad \text{subject to } W_{pi} \geq 0, H_{iq} \geq 0, \text{and} \sum_{i=1}^{k} W_{pi} = 1 \quad \forall\, p = 1, 2, \cdots, n; \;\; q = 1, 2, \cdots, m \qquad (1)$$

where $\mathcal{E}^{(k)}$ is the reconstruction error for a given value of k. For each value of k in $1, 2, \cdots, d$, Eq.1 is solved for 1000 times based on random initial guesses for W and H matrices. The resulting 1000 solutions of H are clustered into k clusters using a customized semi-supervised clustering. During clustering, we enforce the condition that each of the k clusters contain equal number of solutions (i.e, 1000 solutions). After clustering, the average silhouette width *S(k)* is computed. This metric, *S(k)* (see Vesselinov et al. 2018), measures how well the 1000 solutions are clustered for given value of k. The value of *S(k)* varies from -1 to 1. Typically, *S(k)* declines sharply after an optimal number, k_{opt}, is reached. k_{opt} value can be select to be equal to the minimum number of signals that accurately reconstruct the observational data matrix X and the average silhouette width $S(k_{opt})$ is bigger than 0.8. More details on the NMFk algorithm and its implantation are discussed in Alexandrov and Vesselinov 2014, Vesselinov et al. 2018.

3. REGIONAL-SCALE KNOWN-GEOTHERMAL RESOURCES (KGR) DATA

To illustrate our NMFk method, we use the regional-scale known-geothermal resources (KGR) data (Bielicki et al. (2015); Pepin (2019)) from the OpenEI's Geothermal Data Repository (GDR) (see Bielicki et al. (2015)). The KGR data consists of wells with known temperatures. The data was collected during the Geothermal Play Fairways Analysis project (Bielicki et al. (2015); Pepin (2019)). These temperature data are developed from the USGS Identified Hydrothermal, USGS Isolated Geothermal Systems, and USGS Identified Delineated-Area Geothermal Systems (Pepin (2019)). The raw data from the U.S. Department of Energy National Renewable Energy Laboratory Geothermal Prospector tool (NREL (2018)) was processed to create this well temperature data (e.g., Courtesy of Dr. J. D. Pepin, USGS). The well temperature range between 22^{o}C (low temperature resource) to 130^{o}C (moderate/high temperature resource). There are 22 attributes that are associated with this well temperature data. These attributes include boron concentration, gravity anomaly, magnetic intensity, volcanic dike density, drainage density, fault intersection density, quaternary fault density, seismicity, New Mexico state map fault density, springs density, volcanic vent density, lithium concentration, precipitation, air temperature, silica geothermometer temperature, subcrop permeability, hydraulic gradient, water-table elevation, heat flow, elevation, depth to water table, crust thickness, and depth to basement. These data attributes are pre-processed (e.g., boron and lithium concentration) and transformed in to a log-scale. Then, they are rescaled within the range 0.0 to 1.0 using unit range transformation for NMFk analysis.

Figure 1 shows the study area is southwest New Mexico. This region is highlighted in red and consists of known-geothermal resources. The right figure shows the well locations (e.g., latitude, longitude) where temperature is sampled. These well samples are clustered based on temperature and other attributes described above (e.g., concentration of lithium and boron, depth-to-water table) using the NMFk algorithm. Figure 2 shows the spatial plots of some important features used in ML analysis. The left figure shows the concentration of boron (~8800 records), the middle figure shows the concentration of lithium (~5800 records), and temperature from silica geothermometers (~8260 records). These data records are preprocessed in ArcGIS to specify feature values at the locations (Pepin (2019)) specified in Fig.1. Figure 3 shows the preprocessed data feature maps. Data preprocessing is performed using the records from Fig.2 based on ArcGIS's inverse distance weighted (IDW) interpolation scheme. The IDW interpolation is used to specify feature values at the well locations based on Fig.1. Important feature maps include boron concentration (top left figure), lithium concentration (top right figure), temperature from silica geothermometers (bottom left figure), and water table gradient (bottom right figure).

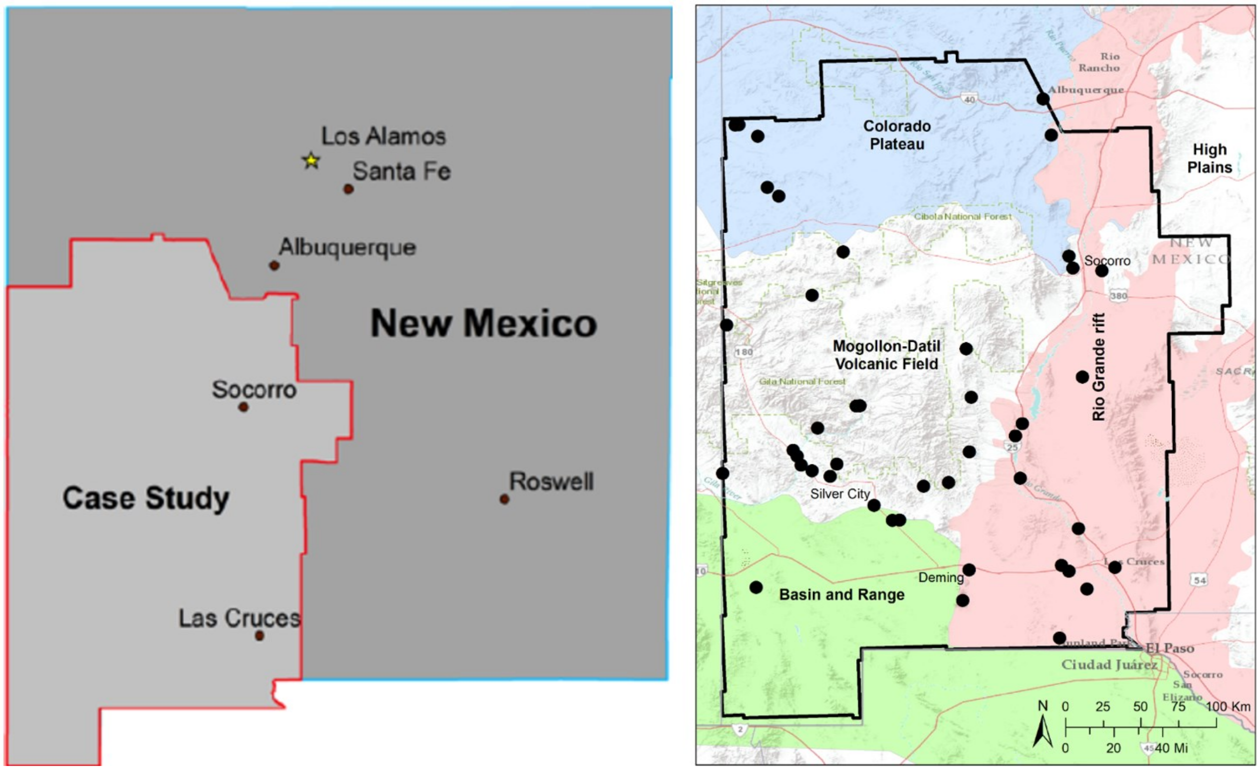

Figure 1: The study area is southwest New Mexico as shown in the left figure. Well locations where temperature is sampled is shown in right figure with black solid dots.

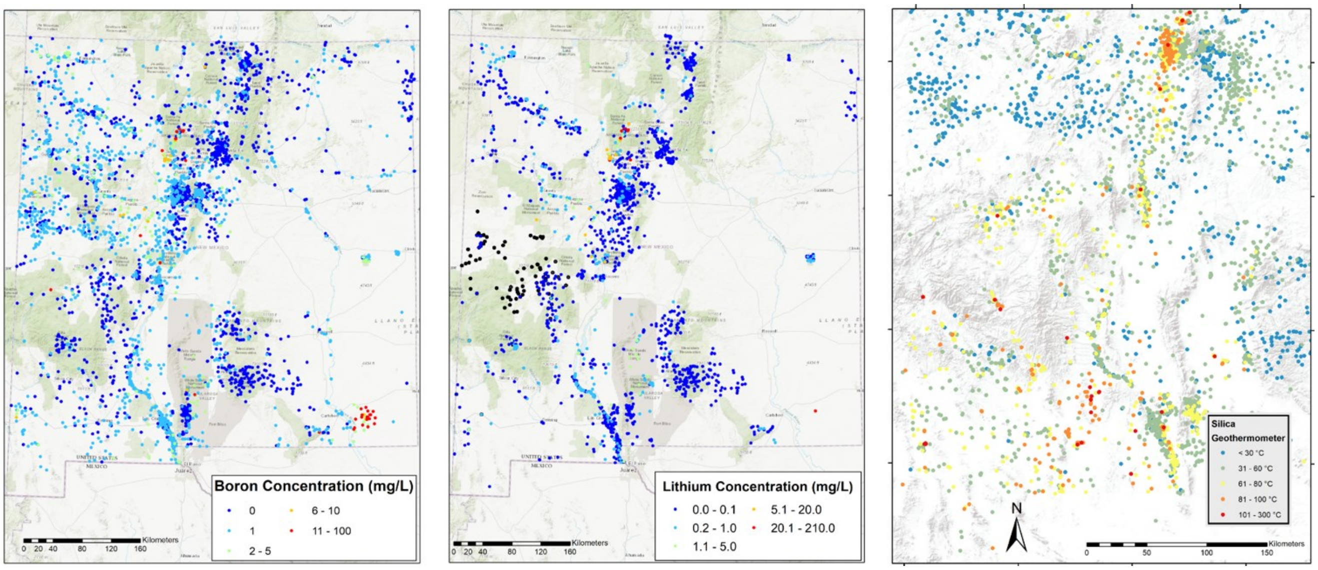

Figure 2: Spatial plots of some important features used in NMF*k* analysis.

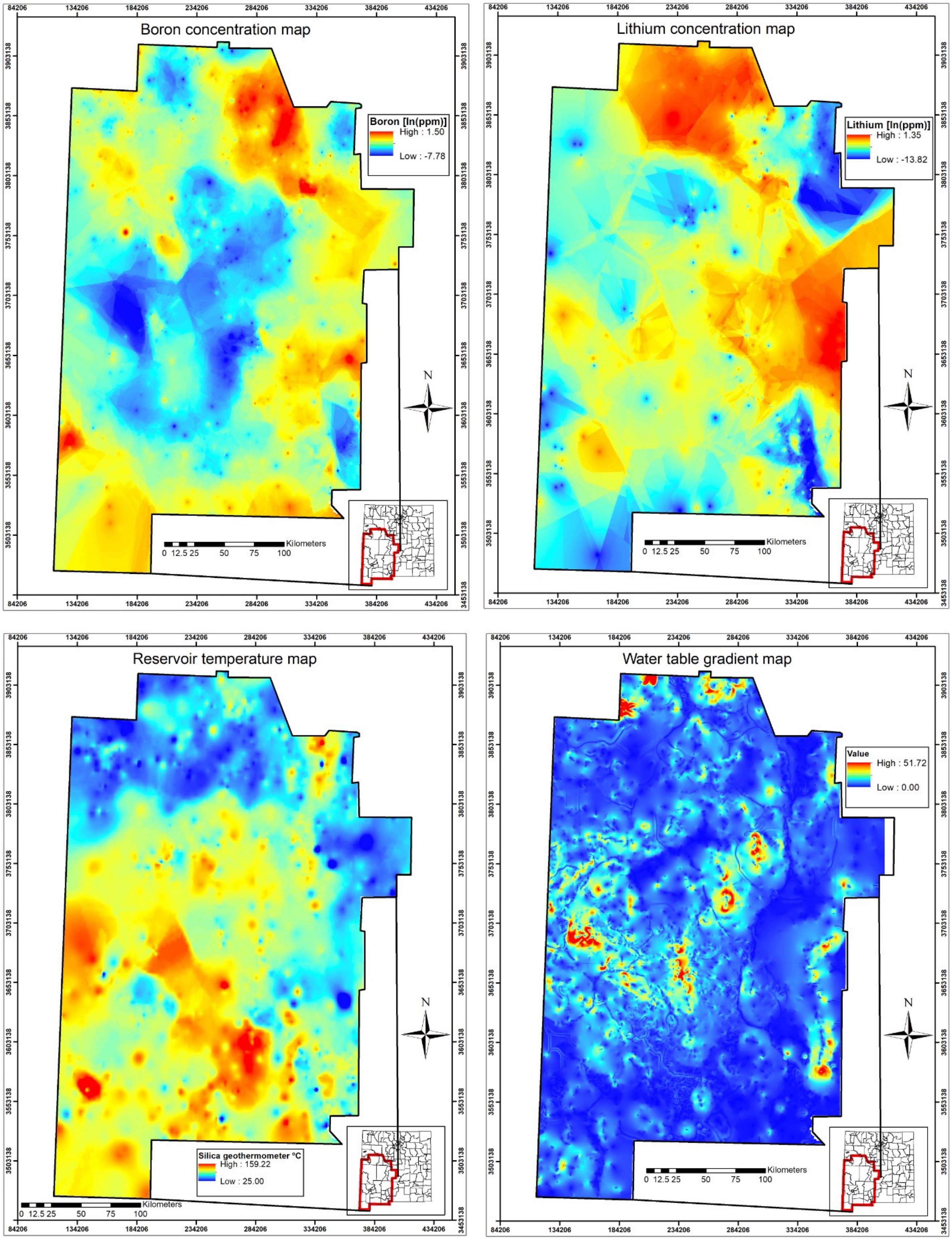

Figure 3: Preprocessed data (e.g., lithium and boron concentration, silica geothermometer temperature, depth to water table) feature maps using ArcGIS's inverse distance weighted (IDW) interpolation scheme.

4. RESULTS

In this section, we summarize the results and analyses of NMF*k* algorithm on regional geothermal data described in Sec.3. Table 1 provides the values of reconstruction quality ($\mathcal{O}$) and average silhouette width (*S*) for different number of hidden features/signals in the data. From the average silhouette width values, it is clear that the possible number of hidden signals can be 2, 4, and 5; note that the solution for 3 signals has been rejected by the algorithm. For these number of hidden signals *k*, the values of *S(k)* is close to 1. By correlating these hidden features with regional physiography (e.g., see Fig.1 which provides details on wells in the Colorado Plateau, the Mogollon-Datil volcanic field, Basin and Range, the Rio Grande rift), we conclude that the optimal number of signals is equal to four (i.e., $k_{opt} = 4$) based on KGR. Fig.4–6 show the dominant attributes, mixing of these attributes, and spatial clustering of geothermal wells based on NMF*k* method.

Figure 4 shows the plots of feature matrix (W), mixing matrix (H), and spatial clustering of wells based on the values of mixing matrix for *k* = 2. From this figure, it is evident that for hidden signal S1, the dominant attributes are crustal thickness, elevation, water table gradient, drainage, precipitation, lithium concentration, fault density, magnetic intensity, and boron concentration. For hidden signal S2, the dominant attributes are air temperature, gravity, boron concentration, magnetic intensity, quaternary fault density, lithium concentration, and temperature from silica geothermometers. Hidden signal S1 mainly corresponds to the 27 wells in cluster-1 (denoted by A and red color dots in cluster map) and hidden signal S2 mainly corresponds to the 17 wells in cluster-2 (denoted by B and blue color dots in cluster map). Correlating these well clusters with the physiographic provinces (e.g., Fig.1), wells representing cluster-1 (denoted by A and represented by red color dots in cluster map) are predominantly located in the Colorado Plateau and the Mogollon-Datil volcanic field. The mixing matrix *H* illustrates how the extracted signals are mixed (represented) in each observational well. Some of the wells in cluster-1 are located in the Rio Grande rift towards Rio Rancho, Albuquerque (e.g., see wells with coordinates in the range of ~34–35° latitude and ~ -107° longitude). The wells representing cluster-2 (denoted by B and blue color dots in cluster map) are mainly located in the Basin and Range province and the Rio Grande rift towards Las Cruces.

Figure 5 shows the plots of feature matrix (W), mixing matrix (H), and spatial clustering of wells based on the values of mixing matrix for *k* = 4**.** The wells are divided in to four clusters. NMF*k* analysis provides us with four different hidden signals, which are S1, S2, S3, and S4. The dominant attributes characterizing hidden signal S1 are drainage, quaternary fault intersection density, quaternary fault density, seismicity, fault density, and spring density. Hidden signal S1 mainly corresponds to the 7 wells in cluster-4 (denoted by D and orange color dots in the cluster map). The dominant attributes characterizing hidden signal S2 include boron concentration, gravity, magnetic intensity, quaternary fault density, lithium concentration, air temperature, temperature from silica geothermometers, heat flow, and depth to basement. Hidden signal S2 mainly corresponds to the 12 wells in cluster-2 (denoted by B and blue color dots in the cluster map). The dominant attributes characterizing hidden signal S3 include boron concentration, magnetic intensity, drainage, lithium concentration, depth to water table, elevation, and crustal thickness. Hidden signal S3 mainly corresponds to the 10 wells in cluster-3 (denoted by C and green color dots in the cluster map). The dominant attributes characterizing hidden signal S4 include magnetic intensity, volcanic dike density, fault density, lithium concentration, precipitation, air temperature, silica geothermometers, water table gradient, depth to water table, and elevation. Hidden signal S4 mainly corresponds to the 15 wells in cluster-1 (denoted by A and red color dots in the cluster map). Correlating the well clusters with the physiographic provinces (e.g., Fig.1), wells representing cluster-1 (denoted by A and represented by red color in the cluster map) are predominantly located in the Mogollon-Datil volcanic field. Wells representing cluster-2 (denoted by B and represented by blue color in the cluster map) are predominantly located in the Basin and Range province, and southeast region of the Rio Grande rift (e.g., closer to Las Cruces). Wells representing cluster-3 (denoted by C and represented by green color in the cluster map) are predominantly located in the Colorado Plateau. Some of these wells in cluster-3 are also located in northwest region of the Mogollon-Datil volcanic field (e.g., see wells with coordinates in the range of ~33.5-34.5° latitude and ~ -109-108° longitude). Wells representing cluster-4 (denoted by D and represented by orange color in the cluster map) are predominantly located in northeast region of the Rio Grande rift (e.g., closer to Socorro).

Figure 6 shows the plots of feature matrix (W), mixing matrix (H), and spatial clustering of wells based on the values of mixing matrix for *k* = 5. Similar to Fig.5, the dominant parameters are precipitation, air temperature, quaternary fault density, lithium concentration, and crustal thickness for S1-S5, respectively. The signals S1-S5 are mainly associated with clusters labelled by D, B, E, C, and A. Correlating the well clusters with the physiographic provinces, we can see that wells in A are located in the Colorado Plateau, the Mogollon-Datil volcanic field, and northeast of the Rio Grande rift. Wells in B and C are located mainly in southwest of the Rio Grande rift and the Mogollon-Datil volcanic field. Wells in D and E are predominantly in the Mogollon-Datil volcanic field and the Rio Grande rift.

Table 1: NMF*k* results for regional-scale geothermal data described in Sec.3.

Number of hidden (latent) features/signals (*k*)	Reconstruction quality, ($\mathcal{O}$)	Average silhouette width of the clusters (*S*)
2	36.52	1.0
3	28.20	0.291
4	21.25	0.999
5	16.76	0.998
6	14.10	0.0928

7	12.02	-0.164
8	9.91	0.217
9	8.08	0.316
10	6.65	0.033

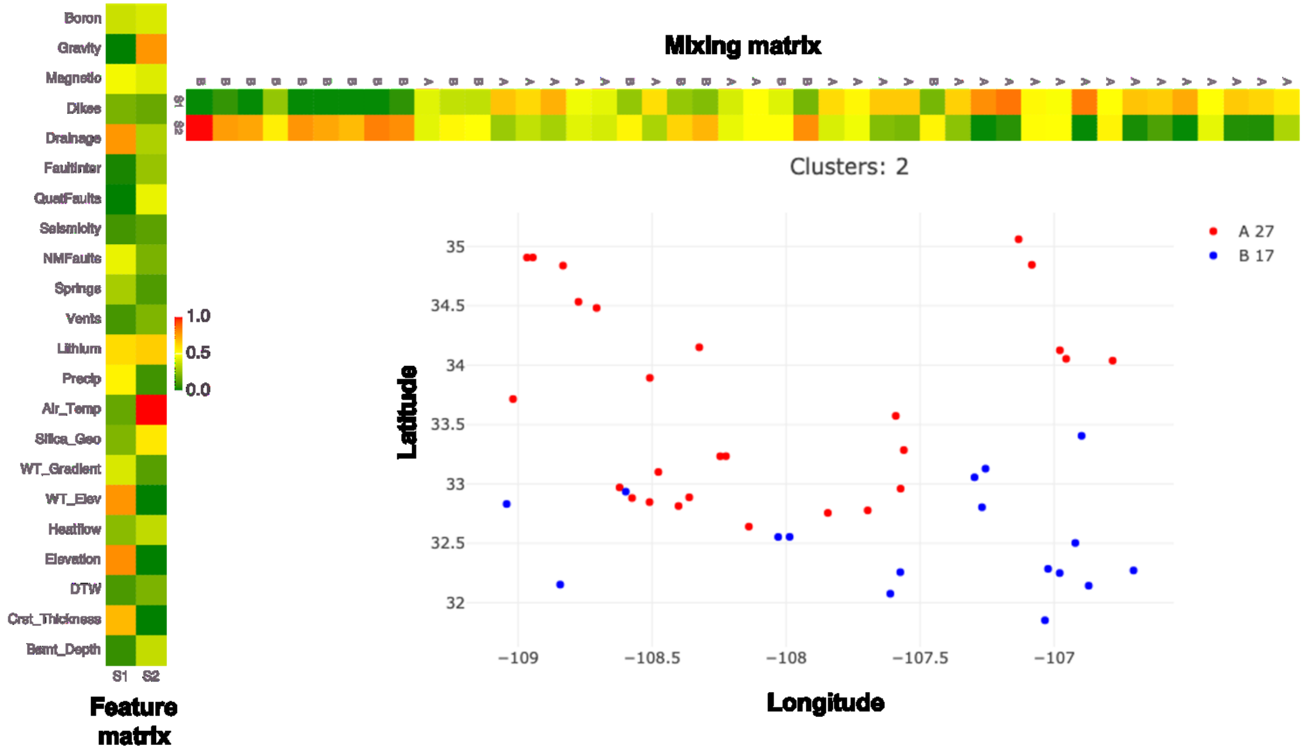

Figure 4: NMF*k* analysis for *k* = 2.

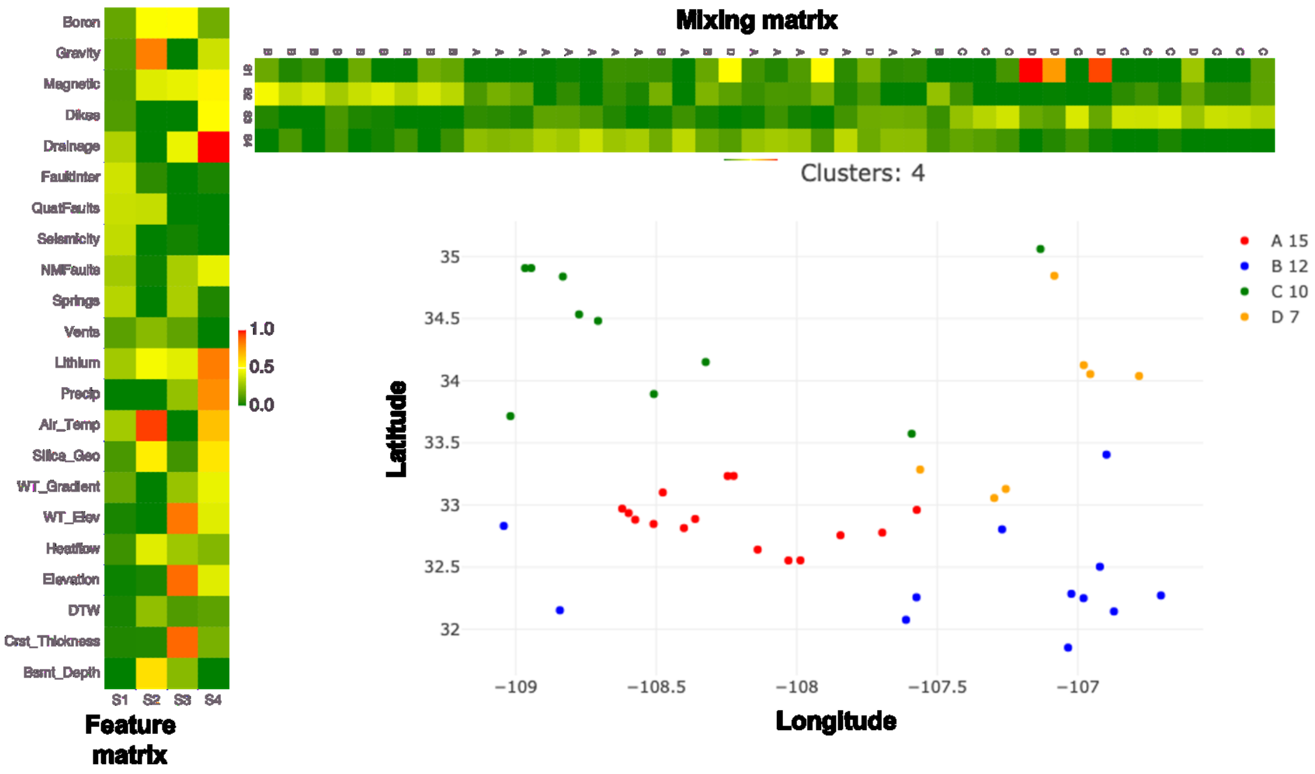

Figure 5: NMF*k* analysis for *k* = 4.

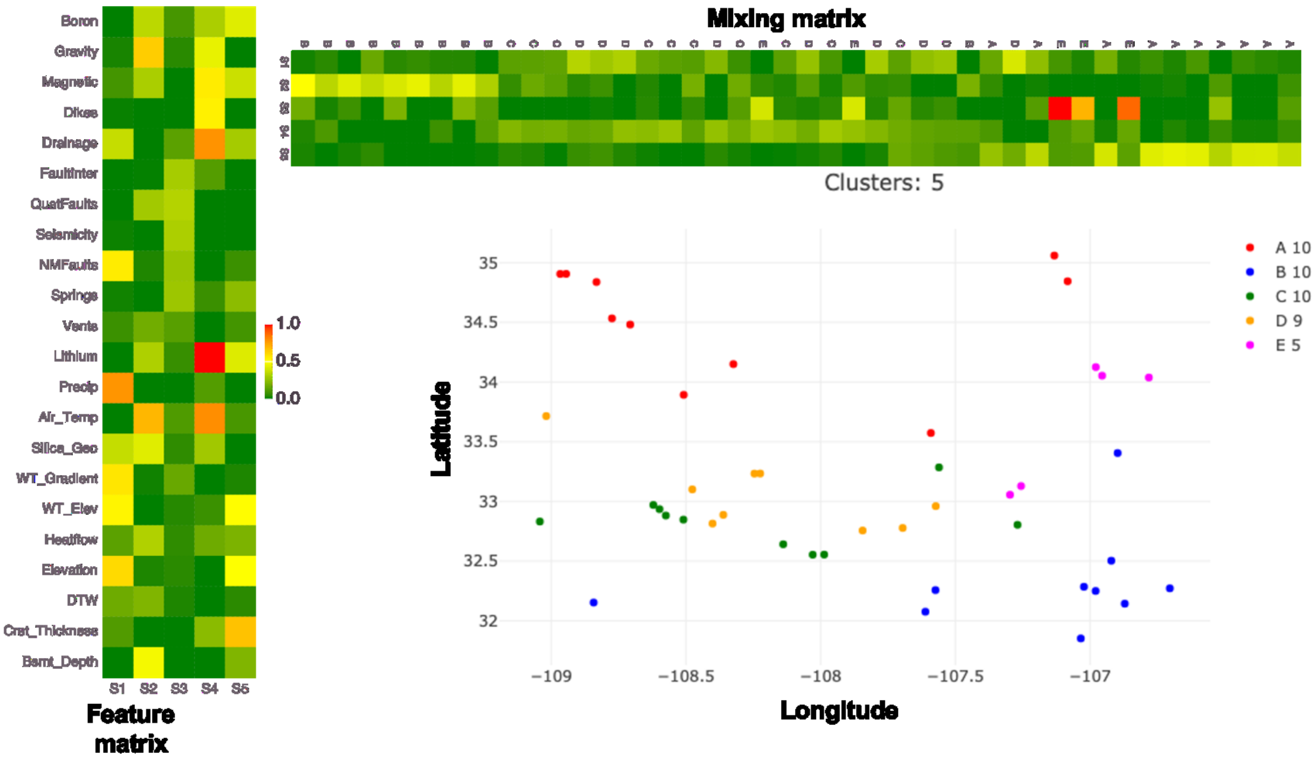

Figure 6: NMF*k* analysis for *k* = 5.

5. CONCLUSIONS

In this paper, we have presented a robust unsupervised machine learning methodology to discover hidden signals in the data and extract dominant attributes corresponding to these signals. The unsupervised ML is based on a non-negative matrix factorization (NMF) method with a custom semi-supervised clustering. Through our results, we demonstrated the applicability of NMF*k* method to discover optimal number of features. By correlating the NMF*k* analyses with physiography of southwest New Mexico, we conclude that the optimal number of clusters is k_{opt} is equal to *4*. Quantitatively, through reconstruction quality and average silhouette width of the clusters, we also showed that $k = 4$ is an optimal cluster number. For this optimal cluster number, the dominant attributes among a total of 22 analyzed geothermal attributes include air temperature, gravity, depth to water table, elevation, crustal thickness, drainage, and lithium concentration. These identified attributes may indicate favorable data sources to prospect site-scale (e.g., Truth or Consequences) geothermal resources in each province (e.g., the Rio Grande rift, the Mogollon-Datil volcanic field). Our future work involves extracting hidden features from subset of the regional geothermal data in New Mexico, quantifying uncertainties and estimating confidence intervals on resource classification predictions (e.g., low, moderate/high temperatures), and analyzing ArcGIS preprocessed data (e.g., IDW interpolated geothermal data in Fig.3) to discover new geothermal locations. To conclude, the extracted dominant features using our unsupervised machine learning methods indicate favorable data sources to prospect geothermal resources in each province. Moreover, the proposed NMF*k* analyses is widely applicable to extract features/signals from large-scale geothermal data (including observational and simulation outputs). This broad applicability of our ML tools makes it attractive for researchers in geothermal industry and institutions to use our tools to discover, quantify, and assess hidden geothermal energy resources (e.g., meeting DOE-EERE GTO's mission).

ACKNOWLEDGEMENTS

This material is based upon work supported by the U.S. Department of Energy's Office of Energy Efficiency and Renewable Energy (EERE) under the Geothermal Technology Office (GTO) Machine Learning (ML) for Geothermal Energy funding opportunity, Award Number DE-EE-3.1.8.1.

Disclaimer: This paper was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

REFERENCES

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

Bielicki, J., Blackwell, D., Harp, D., Karra, S., Kelley, R., Kelly, S., Middleton, R., Pepin, J., Person, M., Sutula, G., and Witcher, J.,: Hydrogeolgic windows: Regional signature detection for blind and traditional geothermal play fairways, Final Reports, Los Alamos National Laboratory, LA-UR-15-28360, (2015).

Bielicki, J., Blackwell, D., Harp, D., Karra, S., Kelley, R., Kelly, S., Middleton, R., Pepin, J., Person, M., Sutula, G., and Witcher, J. (2015) https://gdr.openei.org/submissions/611

Brown, D.W., Duchane, D.V., Heiken, G., and Hriscu, V.T.: Mining the Earth's Heat: Hot Dry Rock Geothermal Energy, *Springer*, (2012).

Cichocki, Andrzej, Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation. John Wiley & Sons, 2009.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). The elements of statistical learning (Vol. 1, No. 10). New York: Springer series in statistics.

Kelkar, S., WoldeGabriel, G., and Rehfeldt, K.: Hot Dry Rock Final Report, Geothermal Energy Development at Los Alamos National Laboratory: 1970-1995, LA-UR-15-22668, April 2015.

McClure, M.W., and Horne, R.N.: An investigation of simulation mechanisms in Enhance Geothermal System, *International Journal of Rock Mechanics & Mining Sciences*, **72**, (2014), 242-260.

Mudunuru, M. K., Karra, S., Harp, D. R., Guthrie, G. D., & Viswanathan, H. S. (2017). Regression-based reduced-order models to predict transient thermal output for enhanced geothermal systems. *Geothermics*, 70, 192-205.

NREL (2018) Geothermal Prospector. National Renewable Energy Laboratory. https://maps.nrel.gov/geothermal-prospector/

Pepin, J. D., New approaches and insights to geothermal resource exploration and characterization, Ph.D Dissertation, Feb. 2019.

Vesselinov, V. V., Mudunuru, M. K., Karra, S., O'Malley, D., & Alexandrov, B. S. (2019). Unsupervised machine learning based on non-negative tensor factorization for analyzing reactive-mixing. *Journal of Computational Physics*, 395, 85-104.

Vesselinov, V. V., Alexandrov, B. S., & O'Malley, D. (2018). Contaminant source identification using semi-supervised machine learning. Journal of contaminant hydrology, 212, 134-142.

TensorDecompositions: https://github.com/TensorDecompositions and https://tensors.lanl.gov/

NMFk.jl https://github.com/TensorDecompositions/NMFk.jl

NTFk.jl https://github.com/TensorDecompositions/NTFk.jl