PROCEEDINGS, 50th Workshop on Geothermal Reservoir Engineering
Stanford University, Stanford, California, February 10-12, 2025
SGP-TR-229

GeoDAWN To GeoTGo: From Complex Data To Decisions Related To Geothermal Prospectivity

Tracy Kliphuis, Ari Markowitz, Rishi Yang, Velimir ("monty) Vesselinov

EnviTrace LLC, Santa Fe, New Mexico, USA

trais@envitrace.com, ari@envitrace.com, rishi@envitrace.com, monty@envitrace.com

Keywords: machine learning, data imputation, feature extraction, hidden geothermal resources, geothermal exploration, prospectivity.

ABSTRACT

GeoTGo is a cloud-based application designed to streamline geothermal exploration, monitoring, and resource utilization. The platform leverages advanced data analytics, cloud/high-performance computing, machine learning, and GIS technology to optimize the entire geothermal energy lifecycle—from exploration and site assessment to real-time well monitoring and energy output optimization. Key features include AI-powered geothermal mapping, thermal gradient analysis, and reservoir simulations. The app is also designed to support near-real-time data analytics integrated with remote sensors. GeoTGo will also offer financial analysis tools for energy yield estimation, cost-benefit analysis, and scenario modeling, helping developers and investors evaluate the economic feasibility of geothermal projects. The application will also provide collaboration and project management capabilities through a centralized cloud dashboard, enabling teams to work efficiently on geothermal projects. GeoTGo's environmental monitoring tools will track sustainability metrics and link users with compliance methods with regulations. By leveraging cloud infrastructure and HPC, GeoTGo significantly reduces exploration time, optimizes geothermal resource utilization, and enhances decision-making. The app is targeted at geothermal exploration companies, energy providers, investors, regulators, and research institutions, aiming to promote sustainable energy development and maximize the potential of geothermal resources. Here, we demonstrate how GeoTGO is applied to process the GeoDAWN airborne datasets collected over Nevada. Our ML analyses extracted features (signals) in the data that are potentially important for the evaluation of geothermal prospectivity. We also demonstrated the applicability of our ML techniques to impute and blindly predict missing data.

1. INTRODUCTION

The reconstruction of sparse geospatial datasets has gained significant attention in recent years, driven by the increasing demand for high-resolution models of physical and chemical properties over extensive spatial domains. The GeoTGo ML algorithms can be applied to address these challenges. Here, we also explore expanding the available techniques for spatiotemporal data imputation by comparing and contrasting different imputation techniques. This study addresses the challenges of predicting values in sparsely sampled variables by leveraging alternative higher-resolution datasets such as magnetics and gravity from airborne geophysics surveys. These surrogate datasets and our study serve as proxies to infer information about harder-to-measure information, such as heat flow and geothermal prospectivity.

A significant challenge in geothermal exploration lies in extracting and interpreting meaningful "hidden" features present within the available data. These concealed features, often overlooked or underestimated, can hold substantial potential in accurately assessing and evaluating geothermal prospectivity. Uncovering these hidden indicators requires advanced ML analysis techniques and a deep understanding of the geological, geophysical, and geochemical processes that govern geothermal systems. By effectively identifying and integrating these "hidden" features into the evaluation process, we can enhance the accuracy and reliability of geothermal resource assessments. This leads to more informed decision-making and, ultimately, a greater success rate in geothermal exploration and development projects.

A particular focus area of our research is the GeoDAWN region. GeoDAWN (Geoscience Data Acquisition for Western Nevada) project is supported by the U.S. Geological Survey (USGS) and the Department of Energy (DOE) to acquire high-resolution airborne magnetic and radiometric data over northern and western Nevada and eastern California. The collected data aims to support geologic and geophysical mapping and modeling to assist geothermal and critical mineral studies. The survey spans areas of major resource potential associated with the Walker Lane and the Western Great Basin. This region is characterized by its geological complexity and heterogeneous spatial patterns, making it an ideal testbed for evaluating advanced imputation methodologies.

Our research builds on past work by us and others that emphasized the need for advanced imputation strategies to address the limitations of existing methods when applied to diverse and complex datasets, especially those related to geosciences (Jarrett *et al.*, 2022; V. Vesselinov *et al.*, 2022; V. V. Vesselinov *et al.*, 2022; Vesselinov, 2023). By comparing a range of state-of-the-art techniques, we aim to evaluate their accuracy and computational efficiency performance systematically. There are several key challenges associated with geoscience data:

- **Uncertainties** arise when there is a lack of confidence or precision in data.
  - ***Aleatoric (irreducible) uncertainties*** are caused by inherent randomness in the geologic system (e.g., pressure fluctuations) that cannot be reduced but can be modeled probabilistically.

  - ***Epistemic (reducible) uncertainties*** are pervasive in geoscience and are caused by incomplete knowledge or a lack of data (e.g., limited samples). However, they can be reduced with better data, improved models, or additional research.
  - ***Systematic (conceptual) uncertainties*** result from system biases, such as errors in measurement devices or methodological errors in the applied models for data interpretation or prediction.
- **Measurement errors** occur when observed data deviates from the actual value; they can be random (unpredictable variations caused by noise) or systematic (consistent bias due to flawed instruments or experimental design).
- **Inaccuracies** are caused by data-entry mistakes and typos.
- **Differences in spatiotemporal support scales of data** can significantly impact analysis, modeling, and decision-making. These differences occur because datasets represent different spatial and temporal scales.
- **Differences in spatiotemporal scales of the characterized geologic features**: Geologic features exhibit distinct spatial, and sometimes temporal, scales, depending on the governing processes. A key aspect is geologic heterogeneity, which occurs across broad scales (from pore to basin, microns to kilometers). These differences are crucial for understanding resource and hazard assessment.

2. METHODOLOGY

In this study, we evaluated five state-of-the-art imputation methods—NMFk (Nonnegative Matrix Factorization with k-means clustering), SVR (Support Vector Regression), XGBoost (eXtreme Gradient Boosting), ICE (Iterative Conditional Expectation), HyperImpute, and Sinkhorn.

We apply these algorithms to reconstruct missing data introduced in contiguous spatial blocks in one of the geospatial attributes (layer). Each algorithm utilized complementary information from the remaining portion of the same layer and from other layers. In this way, the imputation analyses also provide testing and validation of the algorithms to blindly predict unseen data.

The HyperImpute algorithm was implemented in Python, whereas SVR, XGBoost, ICE, and Sinkhorn were implemented in Julia. SmartML was employed to facilitate automated hyperparameter tuning for SVR and XGBoost.

2.1. GeoDAWN dataset

The United States Geological Survey (USGS), in partnership with the Department of Energy (DOE), has amassed a vast airborne dataset as part of its research initiatives. This dataset encompasses 149,030 line-kilometers of flight lines, covering an expansive area of 51,857 square kilometers. Data was gathered using airborne platforms flying at a spacing of 200 to 2000 meters apart, facilitating broad coverage while maintaining a reasonable resolution. The dataset is rich in detail, incorporating more than 20 distinct measured attributes, providing a multifaceted perspective on the surveyed region. This comprehensive dataset amounts to a substantial volume of data, exceeding 5 gigabytes. The collected information holds significant potential for advancing our understanding of subsurface geological structures and resources, particularly those relevant to geothermal energy exploration and development as well as in-situ mining of rare-earth elements.

Our analysis incorporates data from the GeoDAWN region, supplemented with geophysical, geochemical, and hydrochemical measurements from sources such as the INGENIOUS project (Ayling, 2022) and state geologic institutions. The final pre-processed dataset includes 19 metrics, including Magnetics, Gravity, and a range of geochemical and hydrochemical elements (K, U, Th, Li, Mg, Na, Fe, Ba, B, Ca, As, HCO_3, SiO_2, F, SO_4, Cl, Tc). This integrated dataset provides a comprehensive spatial representation of geological and hydrological characteristics across the study area. **Figure 1** shows some of these datasets.

Our team has already processed and gridded many additional features for this region that would benefit this analysis further. These include 500k-scale geologic maps, geologic structures, quaternary faults, well and spring chemistry, aquifers, and subsurface fluid flow.

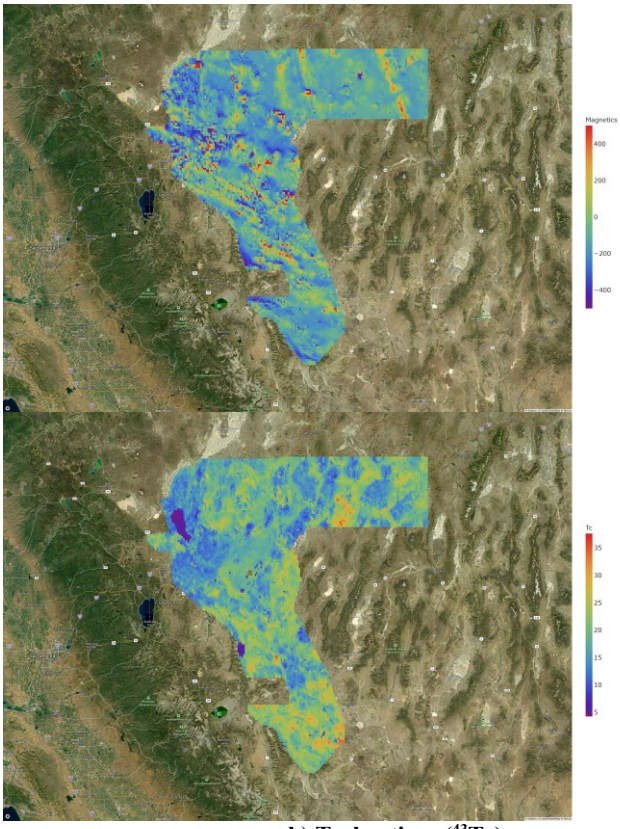
2.1 Overview of the Imputation Framework

Data Preparation:

Test 1: This test focused on data interpolation. A geospatial layer in the dataset was selected for imputation, and 30% of its data values were removed in four spatially contiguous squares within the boundary of the covered area. Other layers in the dataset remained intact and served as auxiliary information.

Test 2: This test focused on data extrapolation. A geospatial layer in the dataset was selected for imputation, and the entire northwest region (about 30%) of its data values were removed in four spatially contiguous squares within the boundary of the covered area. Other layers in the dataset remained intact and served as auxiliary information.

We have developed a workflow for future exploration of the behavior of imputation methods on various types of missingness in geologic datasets by simulating the following: MCAR (Missing Completely at Random), MAR (Missing at Random), and MNAR (Missing Not at Random) (Barnett and Deutsch, 2015). The implications and issues associated with these types of missingness are summarized in **Table 1**. Our goal is to be able to handle all these types of data issues with our ML algorithms implemented in GeoTGo. Accounting for these types of data gaps is critical for accurately evaluating geothermal prospectivity.



a) Magnetics **b) Technetium (^{43}Tc)**

c) Thorium (^{232}Th) **d) Uranium (^{238}U)**

Figure 1: Different attributes available from the GeoDAWN dataset.

Table 1: Types of missingness and their geologic implications.

Aspect	MCAR (Missing Completely at Random)	MAR (Missing at Random)	MNAR (Missing Not at Random)
Dependence	Missingness is unrelated to data.	Missingness depends on observed data but is not related to unobserved data.	Missingness depends on unobserved data.
Bias/Risk	No bias is introduced.	Low risk if adequately modeled.	High risk without external correction.
Ease of Handling	Easy (e.g., deletion or imputation).	Moderate (requires imputation models).	Difficult (requires assumptions or extra data).
Example	Equipment failure.	Data is missing because borehole sections with known loose sediments cannot be adequately sampled.	Geothermal regions may not be sampled/analyzed because they are deemed uneconomical/non-prospective.

Imputation Pipeline: Each algorithm received the partially observed layer and the full auxiliary layers as inputs. The aim was to reconstruct the missing regions using both within-layer and cross-layer correlations.

Evaluation Metrics: Quantified by Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) for prediction accuracy. Measured computation time as the average time (in seconds) required for a single imputation iteration on the same hardware configuration.

2.2 Imputation Methods

2.2.1. NMFk

NMFk is at the forefront among various unsupervised ML methods such as NMF, PCA, ICA, SVD and its variants, k-means clustering, and Gaussian mixture models. In contrast with traditional NMF (Lee and Seung, 1999), NMFk allows for the automatic identification of the optimal number of signatures (features) present in the data (Vesselinov, Alexandrov and O'Malley, 2018b, 2018b). The non-negativity constraint makes the decomposed matrices easier to interpret than PCA, SVD, and ICA because the extracted signatures are additive (Lee and Seung, 1999). Moreover, our version of NMF (implemented in NMFk) can also handle categorical and missing data. Missing data is challenging or impossible to address with other supervised and unsupervised ML methods (Vesselinov, Alexandrov and O'Malley, 2018a; Vesselinov *et al.*, 2019; Siler *et al.*, 2021). Even more importantly, the missing data can be reconstructed from available data based on the estimated matrix factorization. NMFk is part of our SmartTensors ML framework.

2.2.2 Support Vector Regression (SVR)

SVR is a kernel-based supervised learning algorithm that projects data into a higher-dimensional feature space to capture nonlinear relationships (Drucker *et al.*, 1996; Chang and Lin, 2013). By employing a suitable kernel (such as a Radial Basis Function), SVR can model complex interactions among the variables and the spatial structure of the missing data. One principal advantage of SVR lies in its robustness against overfitting, stemming from a well-defined regularization framework. Its capacity to accommodate moderate-sized datasets with meaningful complexity often leads to accurate imputations in scenarios with relatively smooth or structured patterns. However, tuning kernel parameters, regularization terms, and insensitivity thresholds can be computationally intensive and highly sensitive to data scale. Moreover, performance may degrade on huge datasets because of the time and memory required to compute and store kernel transformations.

2.2.3 XGBoost

XGBoost (eXtreme Gradient Boosting) is an ensemble method that builds decision trees in sequence, with each tree attempting to reduce the residual errors from the preceding iteration (Chen and Guestrin, 2016). Thanks to a combination of efficient tree construction, sparse data handling, and integrated regularization, XGBoost has gained attention for its strong predictive performance across diverse domains. The main strength of XGBoost is its flexibility in modeling complex interactions and scalability on large datasets when suitably optimized. Regularization techniques help control overfitting, and parallelization speeds up training. However, this flexibility entails a large hyperparameter search space, including learning rates, tree depths, and the number of boosting rounds. Prolonged tuning procedures can significantly increase computational costs, potentially making XGBoost less appealing if time or computational resources are limited.

2.2.4 Iterative Conditional Expectation (ICE)

ICE imputes missing values by iteratively updating each variable based on all others; it is likened to Multiple Imputation by Chained Equations (MICE) (Buuren and Groothuis-Oudshoorn, 2011). It systematically cycles through variables with missing data, fitting a predictive model for each variable in turn and refining imputed values until convergence. The straightforward conceptual framework of ICE allows it to capitalize on inter-variable correlations, making it particularly appropriate when multiple auxiliary layers exhibit strong relationships with the partially observed layer. Its primary downside is the risk of slow or erratic convergence, especially in high-dimensional or noisy datasets that require a large number of iterations. Performance can also vary depending on the choice of model within each iteration; more sophisticated models may deliver better imputation results but raise computational overhead.

2.2.5 HyperImpute

HyperImpute is a matrix-completion-based method that integrates statistical modeling, factorization techniques, and iterative refinement (Jarrett *et al.*, 2022). It is well-suited to datasets expected to exhibit lower-rank or smooth latent structures. By approximating the underlying matrix in a reduced dimension, HyperImpute can effectively capture shared patterns among different variables or geospatial layers. The principal advantage of HyperImpute lies in its ability to exploit global data structure, leading to accurate imputations for variables that vary smoothly in space or share linear and low-rank dependencies. Its effectiveness depends, however, on whether the data genuinely conforms to these assumptions. Selecting appropriate factorization parameters or regularization terms can also be challenging, and inadequate tuning may lead to slower convergence or suboptimal imputations.

2.2.6 Sinkhorn

Sinkhorn imputation is inspired by entropy-regularized optimal transport (Muzellec *et al.*, 2020). It aligns observed and target distributions by iteratively adjusting a transport plan that displaces the "mass" of known data to fill in the gaps. This framework naturally respects global distributional constraints and marginal properties, making it valuable for preserving domain-specific properties in the imputed layer. A key merit of Sinkhorn is its capacity to preserve overall data distributions, which can be essential in geospatial tasks where physical or chemical constraints must remain intact. The computational load is higher than for more straightforward regression-based methods, as the iterative procedure to solve the optimal transport problem is relatively expensive. Additionally, choosing an appropriate entropy regularization parameter is often nontrivial and can drastically influence performance or stability.

2.3 Implementation Details

Experiments were run on a standard multicore CPU workstation with the following specifications: Intel(R) Xeon(R) W-2255 CPU @ 3.70GHz, 128 GB RAM, and a 64-bit operating system. Timings reported in **Section 3.3** reflect the computational resource constraints of this hardware setup. The software used for this project was Julia (for SVR, XGBoost, ICE, and Sinkhorn), Python (for HyperImpute), and SmartML (for automated hyperparameter tuning of SVR and XGBoost). Parameter tuning was performed using ICE (with convergence criteria based on changes in predicted values), HyperImpute (with a grid search for λ to balance accuracy and runtime), and Sinkhorn (with a preliminary search over ϵ values to ensure stable convergence). Overall, this multi-method approach allows for a broad benchmarking of state-of-the-art imputation strategies in geospatial applications, with particular attention to accuracy, spatial coherence, and computational scalability.

3. Results

3.1 Imputation

In our "Test 1" (interpolation) scenario, the comparative analysis of the methods revealed substantial differences in imputation accuracy and computational efficiency. **Figure 2** demonstrates the RMSE for imputing a given layer from the rest using the different imputation methods.

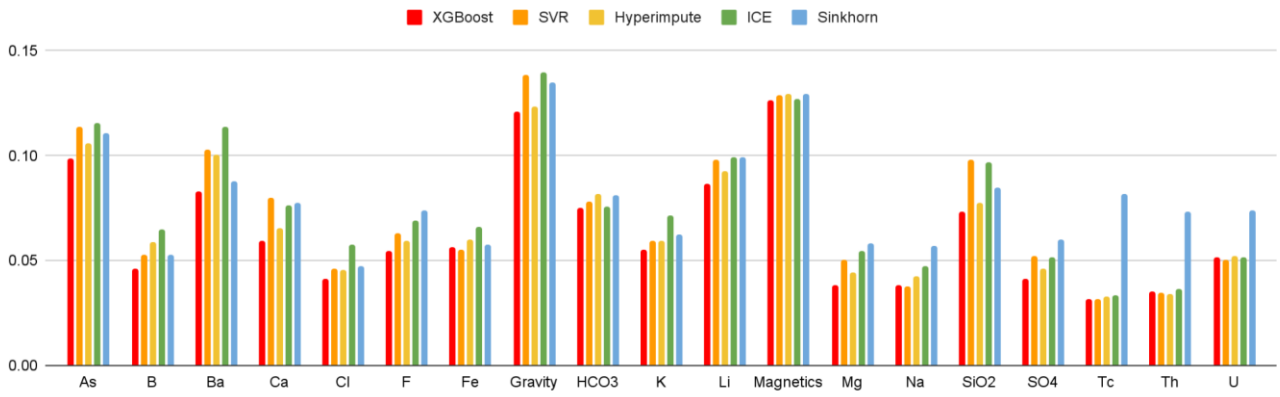


Figure 2: Imputation results for imputation Test Case #1 in terms of root mean squared errors. XGBoost is outperforming all the other tested techniques.

XGBoost emerged as the best-performing method overall, achieving highly competitive RMSE values across most layers. Its self-tuning capability ensured consistently strong performance without requiring extensive manual intervention, making it particularly effective in chemically complex layers such as Cl, B, and SO_4. On smoother layers such as Tc and Th, XGBoost matched HyperImpute and SVR in performance while excelling in spatially and chemically variable layers.HyperImpute and SVR also demonstrated excellent performance. HyperImpute's matrix factorization approach excelled in capturing smooth spatial patterns, particularly in Tc and Th. SVR's nonlinear modeling capabilities allowed it to handle gradual spatial variations effectively. In most cases, ICE showed reliable performance on several

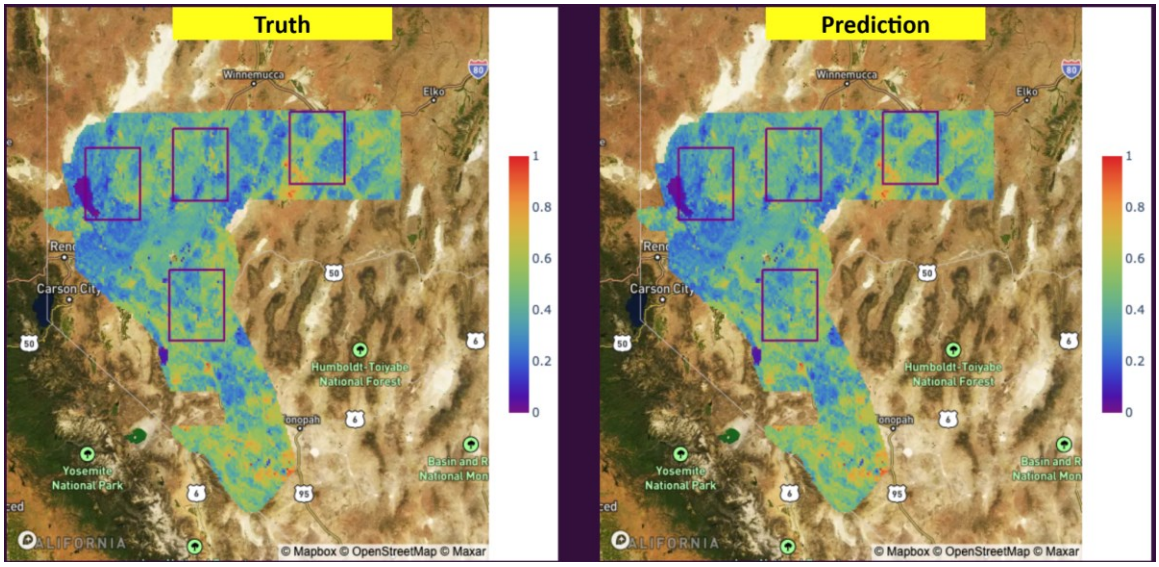
layers but was slightly less accurate than XGBoost and HyperImpute. Sinkhorn, while conceptually robust, exhibited moderate accuracy and struggled with computational efficiency, reflecting the trade-offs inherent in its optimization-based approach.

Table 2: Comparison of execution times for each imputation method applied to a given layer in the "Test 1" scenario (imputation). *XGBoost was executed using auto-hyperparameter tuning, yielding considerably higher execution times.

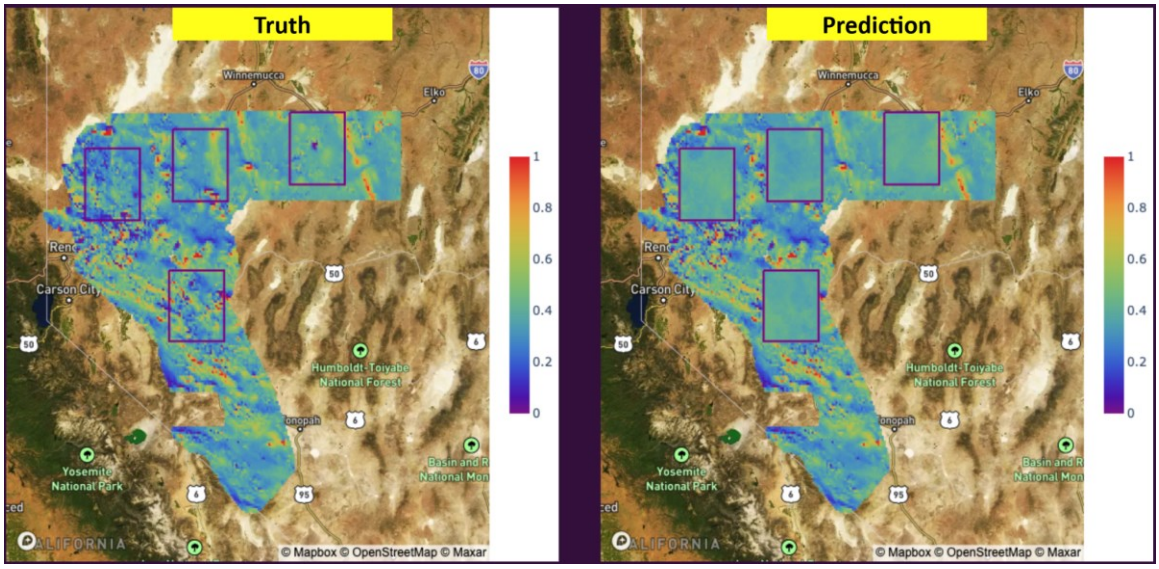
Method	Single Imputation Time (Seconds)
ICE	1.35
SVR	12.54
Hyperimpute	21.48
Sinkhorn	458.25
*XGBoost**	795.10

The computational efficiency analysis highlighted in **Table 2** significant differences among the methods. ICE was the fastest, requiring only 1.35 seconds per iteration. SVR followed with an average of 12.54 seconds per iteration, striking a balance between speed and accuracy. HyperImpute, while slightly slower at 21.48 seconds per iteration, consistently produced highly accurate results. Sinkhorn exhibited significantly longer runtimes, averaging 458.25 seconds per iteration, while XGBoost, with self-tuning, required the most computational time, averaging 795.10 seconds.

Figure 3 compares predicted and original values for Technetium (^{99}Tc) and Magnetics, respectively. These figures highlight the effectiveness of XGBoost in predicting Tc compared to its considerably worse ability to predict Magnetics.



a): Technetium (Tc)



b) Magnetics

Figure 3: Map of the 4 imputation regions (purple squares) applied to test the selected imputation algorithms (Test Case #1). The results above are obtained using XGBoost for Tc (a) and Magnetics (b).

In our "Test 2" (extrapolation) scenario, the comparative analysis of the methods again revealed substantial differences in imputation accuracy. **Figure 4** demonstrates the RMSE for imputing a given layer from the rest using the different imputation methods. As expected, the RMSEs are slightly higher overall for XGBoost and SVR. Still, the newly introduced method NMFk performs very well, with consistently the lowest RMSE values across all variables except Th. In particular, NMFk produces reasonable predictions for Magnetics, especially when compared to XGBoost and SVR. Overall, the method demonstrates robustness in handling diverse variables.

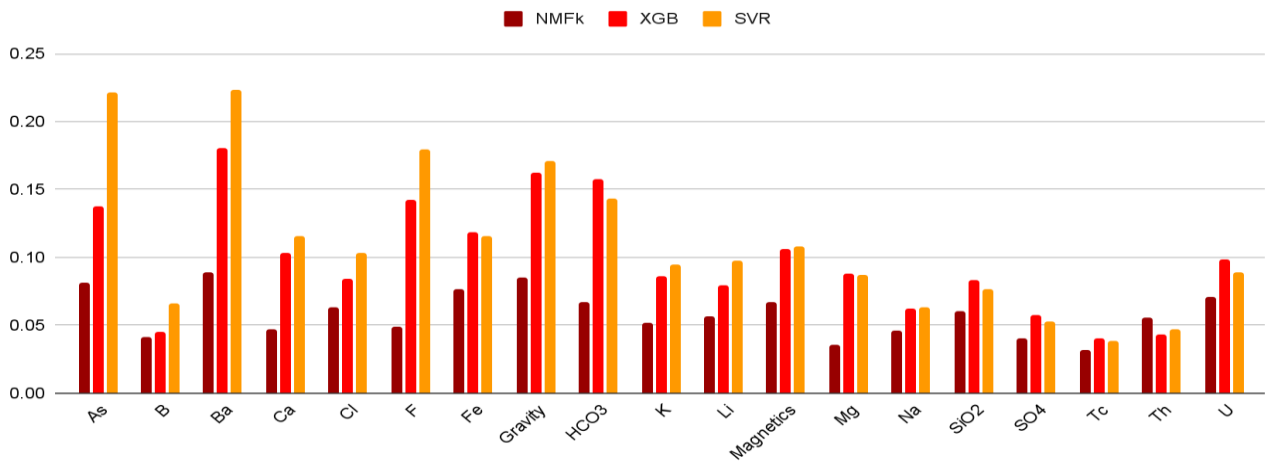
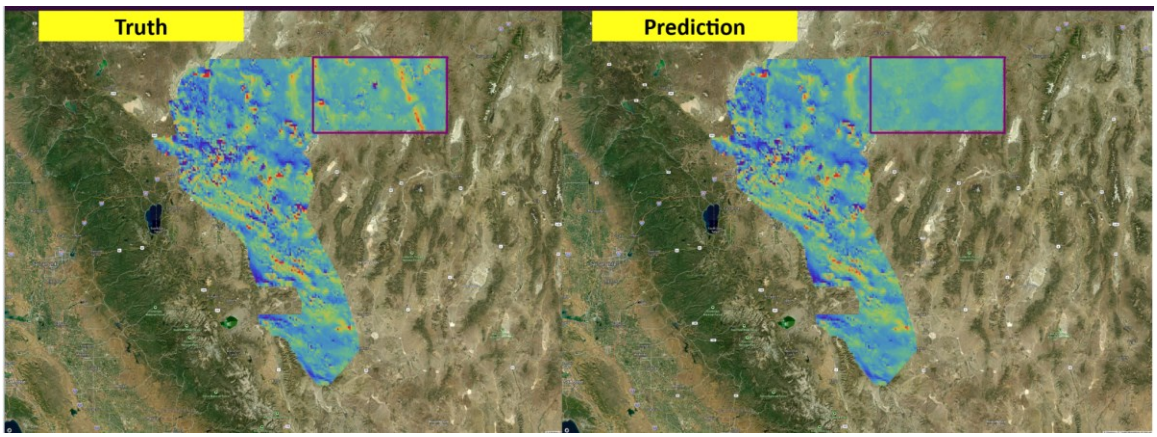
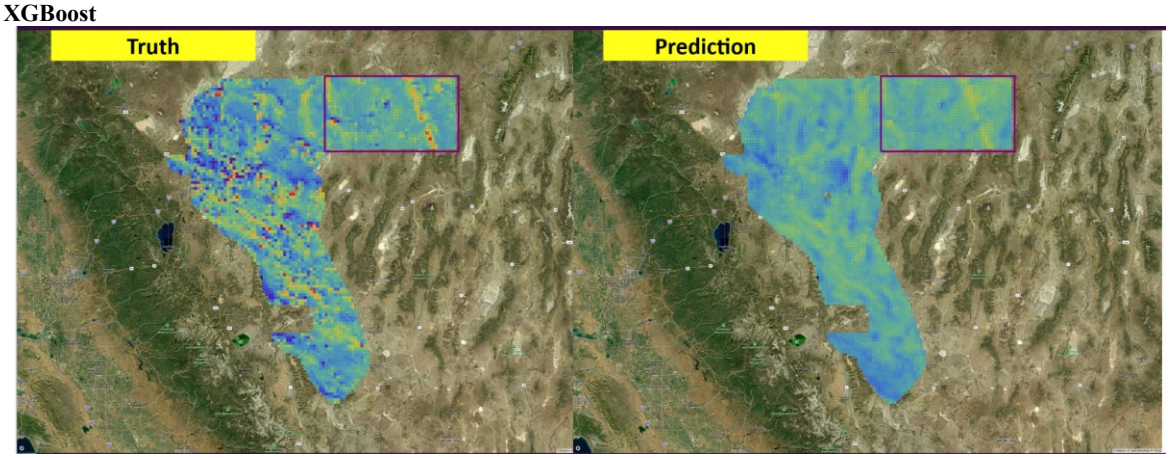


Figure 4: Imputation results for imputation Test Case #2 in terms of root mean squared errors. SmartTensorts NMFk outperforms all the other tested techniques in most cases (except Th).



(a) XGBoost



(b) SmartTensors' NMFk

Figure 5: Maps of the XGBoost vs. SmartTensors' NMFk imputation (blind prediction) of the Magnetics GeoDAWN dataset (Test Case #2). NMFk (b) was capable of capturing the spatial patterns in the imputed region. The XGBoost performed very poorly in this case (a), producing unrealistic spatial patterns in the test region (purple square).

In terms of spatial accuracy, NMFk captures realistic and coherent patterns, closely matching the ground truth, as shown in **Figure 5**. In contrast, XGBoost produces disjointed and unrealistic spatial predictions, particularly in the test region. These results indicate that NMFk not only improves accuracy but also preserves spatial consistency, making it more effective for geospatial imputation tasks.

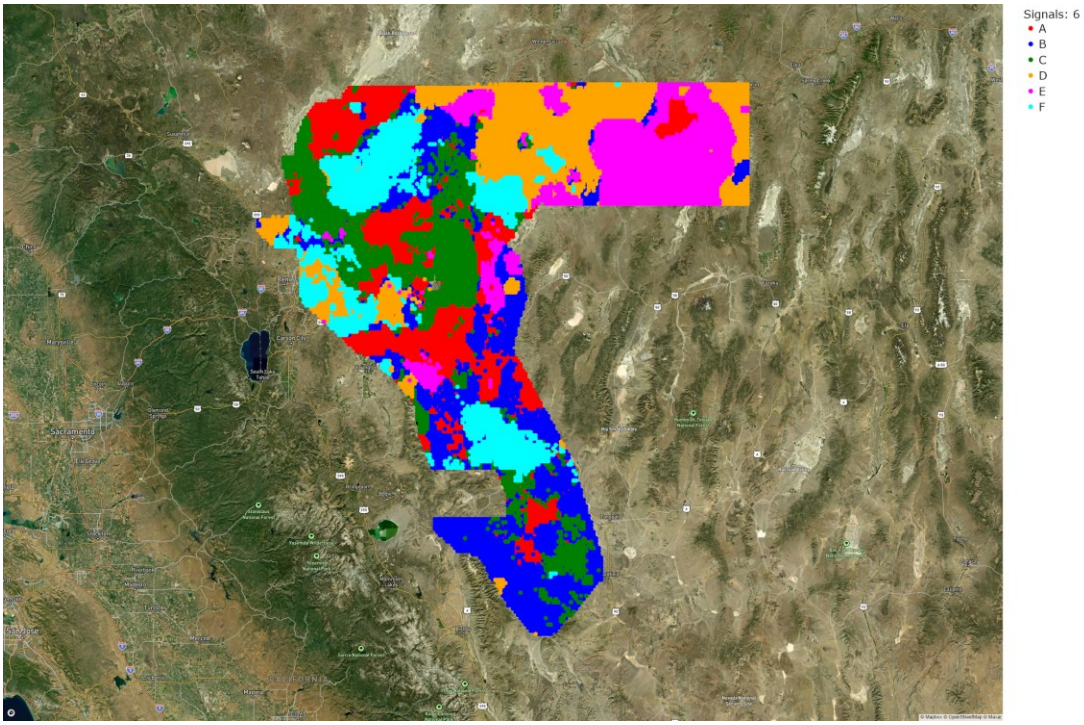


Figure 6: Map ML-identified provinces (spatial domain) in the GeoDAWN dataset. The 6 provinces are associated with the following 6 signals with their respective data attributes: A: F (Na, Li), B: Tc (Th, U), C: Cl (SiO_2, B, …), D: Gravity, E: Magnetics, and F: HCO_3 (Ba, Ca, Mg…). See also the attribute bi-plots in Figure 7.

3.2 Feature Extraction

Our NMFk analysis effectively identified and clustered six distinct geochemical and geophysical provinces within the GeoDAWN dataset. Each signal represents a unique combination of features, reflecting the diversity and complexity of the data. **Figure 6** shows the spatial extent of the extracted signals. **Figure 7** displays 2D projections of the extracted attributes within the identified latent space in the form of bi-plots. These projections allow for visualization and understanding the relationships between attributes and signals. Each signal is associated with a set of dominant attributes, which are colored and labeled; they are highlighted with ovals in the figure. The positioning of the dots representing these attributes reveals their significance in characterizing the corresponding signals. Dots (attributes) that are located closer to the primary axes and closer to 1 (the axes maximums) indicate a stronger association between the attribute and the signal represented by that axis. This visual representation allows for identifying the most influential attributes for each signal, providing insights into the underlying characteristics that define them. The six signals and their associated features are as follows (ordered based on their importance in reconstructing the original data):

1. **Signal A**: Moderately strong signal characterized by associations with **F, Na, and Li**, indicating a region with significant contributions from these elements. This feature (signal) might define areas where rare-earth elements might be predominantly located and suitable for in-situ mining (**Figure 6**).
2. **Signal B**: Strong signal predominantly linked to **Tc, Th, and U**, with contributions to the signal decreasing in that order. This signal highlights regions with elevated concentrations of these elements, potentially reflecting radioactive or heavy-metal-rich zones.
3. **Signal C**: This signal is essential for data reconstruction, but it is less discernible, which is inferred based on the relative absence of alignment of the associated attributes (**Cl, SiO_2, and B**) with the signal axes (the "green" dots associated with this signal are generally located in the center of the plots or close to 0). This may indicate a diffuse distribution or weaker clustering of these features.
4. **Signal D**: Strong signal that is almost exclusively linked to **Gravity**, suggesting a region dominated by gravitational anomalies, which could signify significant subsurface mass variations.
5. **Signal E**: Signal primarily associated with **Magnetics**, pointing to regions with magnetic anomalies, potentially linked to mineralized zones or specific rock types.
6. **Signal F**: Moderately strong signal associated with **HCO_3, Ba, Ca, and Mg**, indicative of regions with notable geochemical activity, such as carbonate-rich environments or hydrothermal systems.

These results demonstrate the robustness of the NMFk clustering approach in delineating geochemical and geophysical provinces, revealing clear and interpretable patterns in the GeoDAWN dataset.

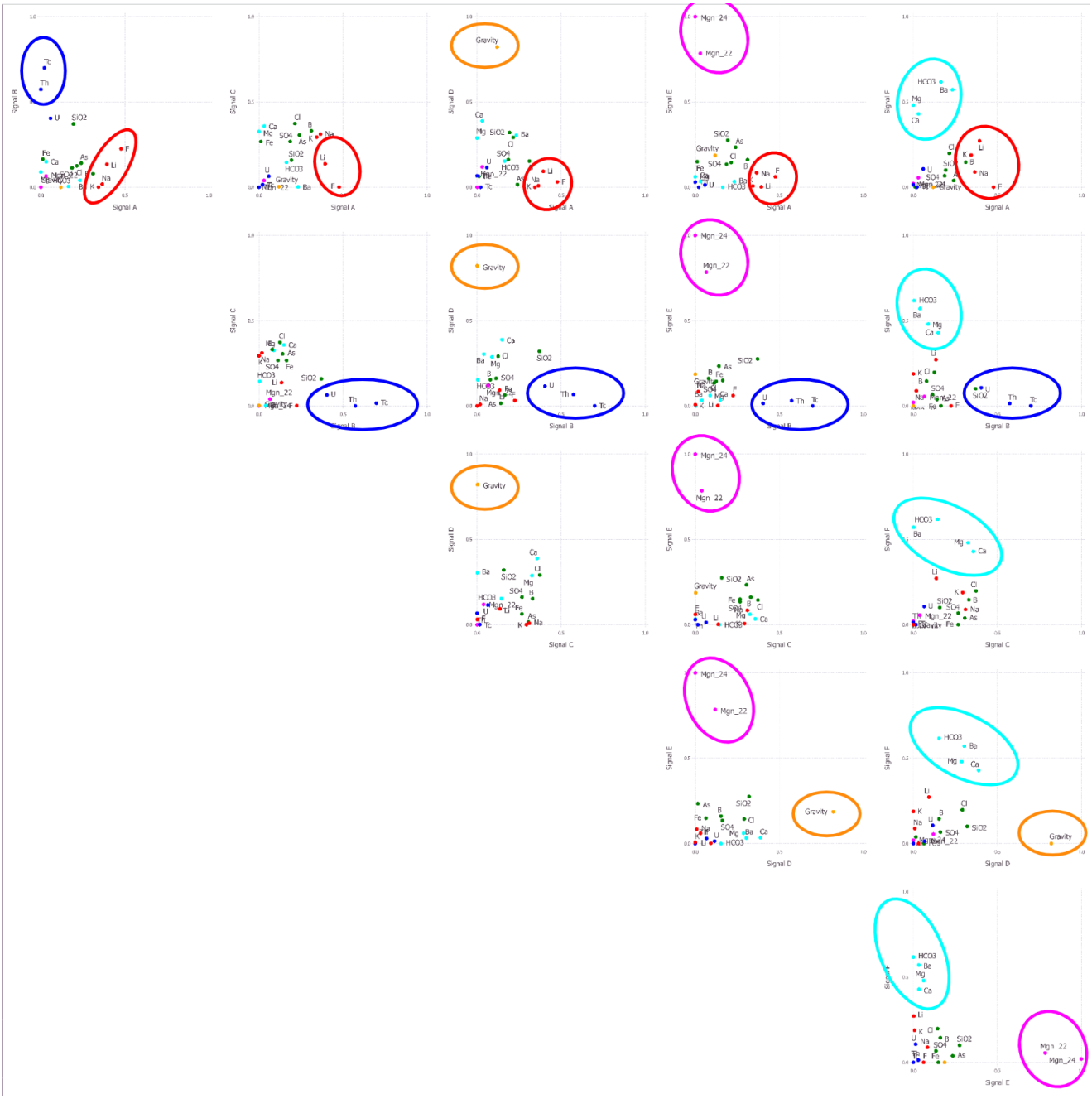


Figure 7: Map ML-identified 6 dominant signals (features) in the GeoDAWN dataset. Each of the 6 features is associated with the following data attributes: A: F (Na, Li), B: Tc (Th, U), C: Cl (SiO_2, B, …), D: Gravity, E: Magnetics, and F: HCO_3 (Ba, Ca, Mg…). The dots in the bi-plots are colored

4. CONCLUSIONS

Our findings thus far indicate that the SmartTensors' NMFk methodology for data imputation shows the most potential among the approaches we've explored. The initial results suggest that SmartTensors is capable of accurately and efficiently filling in missing data points within our geothermal datasets. This is crucial for improving the reliability and robustness of our ML models, as incomplete data can often lead to inaccurate predictions and biased results. We hypothesize that by continuing to refine and optimize our SmartTensors approach, we can further enhance its imputation capabilities and ultimately develop more accurate and effective predictive models for geothermal exploration and reservoir management.

SmartTensors possess the capability to identify and extract salient ("hidden") spatial and attribute features from within a given dataset. These features could include identifying patterns, trends, or anomalies in the spatial distribution of data or recognizing key characteristics or attributes associated with specific data points. This information can be leveraged to generate valuable insights and inform decision-making processes.

However, SmartTensors' NMFk technique is much more computationally expensive than other imputation approaches that we have tested. SmartTensors' XGBoost appears to be better than other specifically designed imputation methods we tested (ICE, HyperImpute, and Sinkhorn), especially when utilizing hyperparameter tuning.

Future work will aim to:

- Incorporate more data in the GeoDAWN domain, including the results from the Stanford Thermal Earth Model (M. Aljubran and Horne, 2024; M. J. Aljubran and Horne, 2024).
- Test against more missingness types (MAR, MCAR, MNAR).
- Explore more imputation techniques (including reinforcement ML), and further tune imputation methods we have already tested.
- Consider explicitly the spatial context of the data, including the support scale of the data and scale (size) of the analyzed features.
- Predict the presence of geologic features suggesting geothermal prospectivity.
- Estimate of geothermal productivity.

5. ACKNOWLEDGMENT

EnviTrace work is funded by DOE SBIR Grant DE-SC0022697 titled "GeoTGo: Equitable and inclusive tool for community-based geothermal development".

References

Aljubran, M. and Horne, R. (2024) 'Stanford Thermal Earth Model for the Conterminous United States'. DOE Geothermal Data Repository; Stanford University. Available at: https://doi.org/10.15121/2324793.

Aljubran, M.J. and Horne, R.N. (2024) 'Thermal Earth model for the conterminous United States using an interpolative physics-informed graph neural network', *Geothermal Energy*, 12(1), p. 25. Available at: https://doi.org/10.1186/s40517-024-00304-7.

Ayling, B. (2022) *INGENIOUS - Great Basin Regional Dataset Compilation*. 1391. USDOE Geothermal Data Repository (United States); GBCGE, NBMG, UNR. Available at: https://doi.org/10.15121/1881483.

Barnett, R.M. and Deutsch, C.V. (2015) 'Multivariate imputation of unequally sampled geological variables', *Mathematical Geosciences* [Preprint]. Available at: https://doi.org/10.1007/s11004-014-9580-8.

Buuren, S. van and Groothuis-Oudshoorn, K. (2011) 'MICE: Multivariate Imputation by Chained Equations in R', *Journal of Statistical Software* [Preprint]. Available at: https://doi.org/10.18637/jss.v045.i03.

Chang, C. and Lin, C. (2013) 'LIBSVM : A Library for Support Vector Machines', *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2, pp. 1–39. Available at: https://doi.org/10.1145/1961189.1961199.

Chen, T. and Guestrin, C. (2016) 'XGBoost: A Scalable Tree Boosting System', in *Proceedings of the 22nd ACM SIGKDD*. Available at: https://doi.org/10.1145/2939672.2939785.

Drucker, H. *et al.* (1996) 'Support Vector Regression Machines', in *Advances in Neural Information Processing Systems*. MIT Press. Available at: https://proceedings.neurips.cc/paper_files/paper/1996/hash/d38901788c533e8286cb6400b40b386d-Abstract.html (Accessed: 29 January 2025).

Jarrett, D. *et al.* (2022) 'HyperImpute: Generalized Iterative Imputation with Automatic Model Selection', in *Proceedings of the 39th International Conference on Machine Learning*. *International Conference on Machine Learning*, PMLR. Available at: https://proceedings.mlr.press/v162/jarrett22a.html (Accessed: 29 January 2025).

Lee, D.D. and Seung, H.S. (1999) 'Learning the parts of objects by non-negative matrix factorization.', *Nature*, 401(6755), pp. 788–91. Available at: https://doi.org/10.1038/44565.

Muzellec, B. *et al.* (2020) 'Missing Data Imputation using Optimal Transport', in *Proceedings of the 37th International Conference on Machine Learning*. *International Conference on Machine Learning*, PMLR. Available at: https://proceedings.mlr.press/v119/muzellec20a.html (Accessed: 29 January 2025).

Siler, D. *et al.* (2021) *Machine Learning to Identify Geologic Factors Associated with Production in Geothermal Fields: A Case-Study Using 3D Geologic Data from Brady Geothermal Field and NMFk*. 1344. USDOE Geothermal Data Repository (United States); United States Geological Survey. Available at: https://doi.org/10.15121/1832133.

Vesselinov, V. *et al.* (2022) 'GeoThermalCloud: Machine Learning for Discovery, Exploration, and Development of Hidden Geothermal Resources', in *Stanford Geothermal Workshop*. *Stanford Geothermal Workshop*, Stanford, CA. Available at: https://pangea.stanford.edu/ERE/db/GeoConf/papers/SGW/2022/Vesselinov.pdf.

Vesselinov, V.V. *et al.* (2019) 'Unsupervised machine learning based on non-negative tensor factorization for analyzing reactive-mixing', *Journal of Computational Physics* [Preprint]. Available at: https://doi.org/10.1016/j.jcp.2019.05.039.

Vesselinov, V.V. *et al.* (2022) 'Discovering hidden geothermal signatures using non-negative matrix factorization with customized k-means clustering', *Geotermics* [Preprint]. Available at: https://doi.org/10.1016/j.geothermics.2022.102576.

Vesselinov, V.V. (2023) *GeoThermalCloud.jl: Machine Learning for Geothermal Exploration*. Available at: https://github.com/SmartTensors/GeoThermalCloud.jl.

Vesselinov, V.V., Alexandrov, B.S. and O'Malley, D. (2018a) 'Contaminant source identification using semi-supervised machine learning', *Journal of Contaminant Hydrology*, 212(October), pp. 134–142. Available at: https://doi.org/10.1016/j.jconhyd.2017.11.002.

Vesselinov, V.V., Alexandrov, B.S. and O'Malley, D. (2018b) 'Nonnegative tensor factorization for contaminant source identification',

Journal of Contaminant Hydrology [Preprint]. Available at: https://doi.org/10.1016/j.jconhyd.2018.11.010.