Final Project MGNG599

Jaime F. Moraga

Colorado School of Mines

Author Note

This report corresponds to the final project for MNGN599

Abstract

This report summarizes the results of the analyses performed in a new dataset from the Brady Geothermal zone (OpenEI, 2021a), containing 8 data layers: Geothermal, Temperature, Faults, Slope, Chalcedony, Kaolinite, Gypsum and Hematite. These layers are first shown to present low collinearity, and then different Machine Learning methods (Linear regression, Neural networks, random forests, and support vector machine) are applied to them and compared against results of running the Geothermal AI (Moraga et al., 2022) on the same datasets.

The results show that a neural network with 7 hidden neurons achieves results comparable to the Geothermal AI for this dataset.

Keywords: Geothermal exploration, Remote sensing, Machine Learning, Artificial Intelligence, Spatial Statistics, Geospatial data

Final Project MGNG599

Geospatial data are a special case of data where the source relates to specific locations on earth at a specific time. These types of data are then usually correlated in both space and time, and time has also the characteristic of being seasonal and cyclical for certain phenomena.

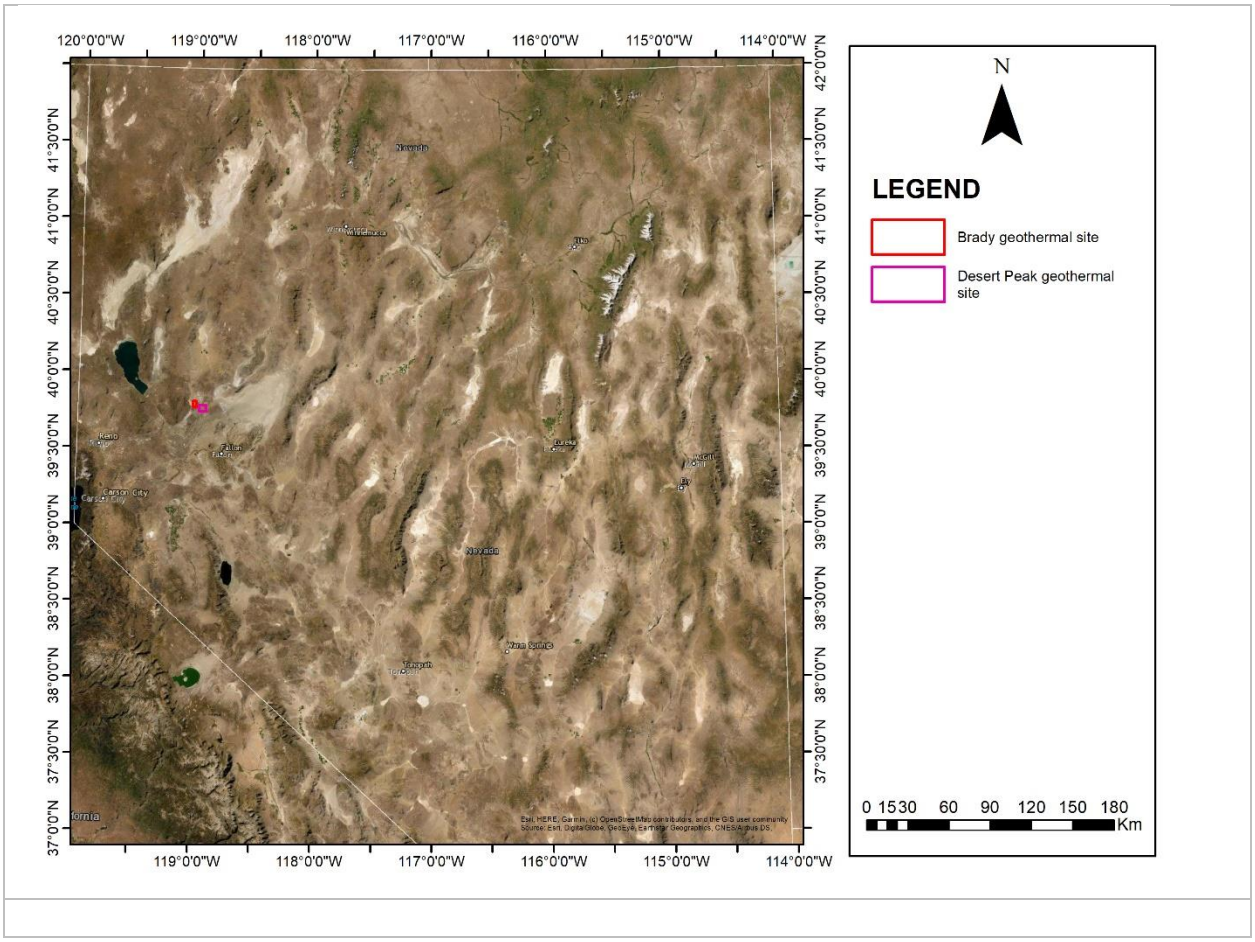
When analyzing statistically these type of data sets, care must be taken to properly take into account for these characteristics, both to improve the modelling and to prevent bias in the analyses.

This report chronicles the analyses performed in datasets related to geothermal regions in this course (MNGN599).

Dataset

The Brady Hot Springs Geothermal area is located in Churchill County, Nevada. It sits northeast from Fernley, NV, and is part of the Northwest Basin and Range Geothermal Region. In this area, there are geothermal operations in two sites, Brady (39.79°N, 119.02°W) and Desert Peak (39.75°N, 118.95°W) (OpenEI, 2021a).

The area has been extensively studied and explored, and there is a broad amount of data available through the Open Energy Information's Geothermal Data Repository (OpenEI, 2021b).



A new dataset was compiled from this area by first reusing data from the Geothermal AI project (Geothermal, Minerals, Temperature, and Faults layers from (Moraga et al., 2022)), by removing the Geothermal layer and adding 7 new layers: slope, deformation, Chalcedony,

Kaolinite, Gypsum, Hematite and Epsomite. The last 5 layers were the component minerals from the original “Minerals” layer (Moraga, 2021).

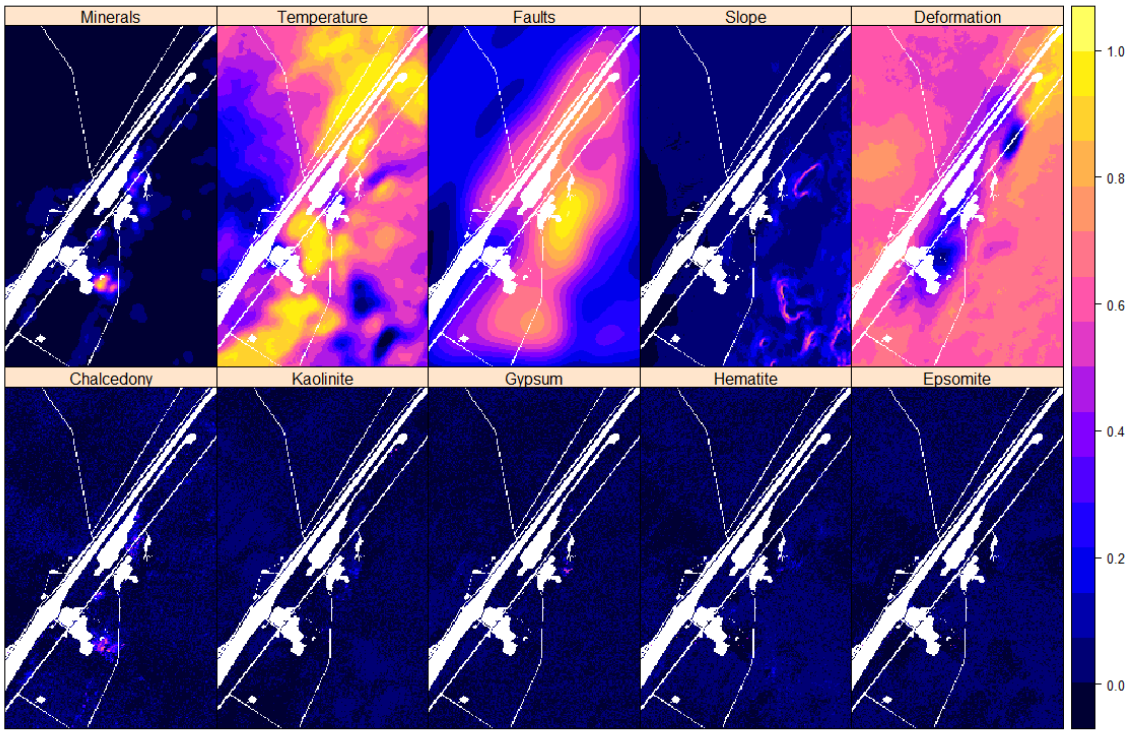


Figure 1 – Dataset 1: Base dataset used for collinearity analysis in this report

The original layers are described in detail in (Moraga et al., 2022), they are:

- Minerals: This is the result of voting using the layers below, all constructed by using ENVI’s Target detection wizard on a HyMap scan of the area of interest done in 2003, against relevant mineral markers associated with hydrothermal activity in the zone.
    - Chalcedony: The values matching the chalcedony spectrum, with values greater or equal to zero.

    - Kaolinite: The values matching the kaolinite spectrum, with values greater or equal to zero.
    - Gypsum: The values matching the gypsum spectrum, with values greater or equal to zero.
    - Hematite: The values matching the hematite spectrum, with values greater or equal to zero.
    - Epsomite: The values matching the epsomite spectrum, with values greater or equal to zero.
- Temperature: The result of k-mean clustering and selection of the highest temperature areas throughout a year, using LANDSAT-8's ADR LST dataset.
- Faults: The result of applying to a shapefile with the faults in the area a density function with a 1500 m radius.
- Slope: The result of measuring slope in a high-resolution digital elevation model (DEM) of the area of interest. The values are in radians from 0 to 1
- Deformation: The result of a PSInSAR analysis using SENTINEL-1A data from 72 datapoints captured between Dec 2017 and Dec 2019 in the area of interest (Çavur et al., 2021). The results are normalized using min max normalization to values from 0-1, after adjusting to the mean and scaling using standard deviation.

A second dataset is derived using the same initial dataset but adjusting to remove collinearity (thus dropping both Minerals and Epsomite as explained in the Mineral Marker collinearity sub-section).

To use this dataset for geospatial analysis, a new layer is built as a derivative of the Deformation layer. The deformation layer is used as a proxy for Geothermal activity, by determining the zones that have suffered subsidence as a result of the operation of the Brady Geothermal Plant. The resulting layer is thus the deformation layer where subsidence is 1mm or higher (Çavur et al., 2021). The final dataset, used for regression and other Machine learning algorithm is thus:

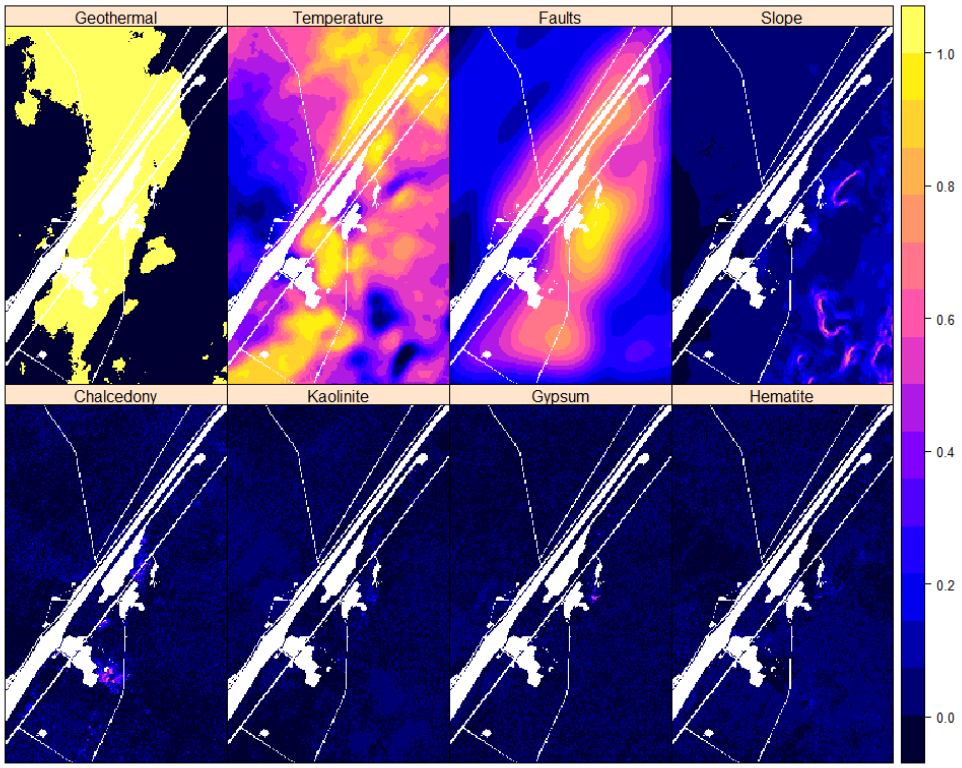


Figure 2 – Dataset 2: Base dataset used for machine learning in this report

Multicollinearity analysis

Given a linear regression model (McCullagh & Nelder, 2019):

$$Y = X\beta + \varepsilon$$

Where the vector Y is the dependent variable or response, X is a matrix that contains the independent variables (regressors), β is the vector of coefficients, and ε is a vector of random errors that are assumed to be independent normally distributed random values with mean 0 and a constant variance.

Collinearity is a problem that arises in statistical analysis when the independent variables are not orthogonal. That is, there is a linear relationship between subsets of the independent variables. This is a particularly important consideration in geospatial data because physical and natural phenomena can be related spatially among each other (for example, plant growth can be related to amount of light, type of soil and humidity in a zone).

The variance inflation factor (VIF) is a tool used to determine the collinearity for each of the independent variables.

VIF is defined as:

$$VIF_i = \frac{1}{1 - r_i^2}$$

For each independent variable i, where r_i^2 is the coefficient of determination obtained by firring a regression model of the variable i against all the other independent variables. That is, the results of solving

$$X_i = Z_i\delta_i + \varepsilon_i$$

Where X_i is the column from X, Z_i is X without column X_i, δ_i are the coefficients, and ε_i vectors of independent random variables with mean 0, constant variance and normally

distributed. When there is no correlation, VIF is 1, and when a variable is perfectly correlated to the other variables VIF is infinite.

There are no formal criteria to decided when a VIF is too large, but in practice VIF values greater or equal to 5 or 10 are commonly used as cutoff points for cases where the collinearity is high enough to require remedial measures (for example, eliminating one or more of the high VIF variables from the analysis) (Craney & Surles, 2002).

In our case, we will use a more conservative cutoff value of 2.5 and 2 as a warning sign.

Geothermal dataset collinearity

For the geothermal dataset, we look for the relationship between Geothermal presence in an area as a function of 3 variables: Temperature, Faults and Mineral markers.

A generalized linear model is generated in R with coefficients:

Table 1 – Generalized linear model parameters

Intercept	Minerals	Temperature	Faults
-14.698	40.960	7.405	31.772

The VIF results calculated are:

Table 2– Generalized linear model VIF

Minerals	Temperature	Faults
1.009278	1.704965	1.704983

In all three cases, the VIF values are significantly lower than 2.5. In particular, Mineral markers seems almost completely independent of the other independent variables.

From a physical standpoint, faults and temperature can be correlated because in a geothermal area with surface manifestations, near the faults there are hot zones from fumaroles, hot springs and mud pots.

Mineral marker collinearity

Instead of using the Minerals band, an alternative is using the component minerals, and eliminate those minerals with high collinearity.

Using bands Geothermal, Temperature and Faults, and breaking the mineral markers layer into its components: Chalcedony, Kaolinite, Gypsum, Hematite and Epsomite, the analysis results are the following.

Table 3– Mineral markers generalized linear model VIF

	VIF						
AIC	Temperature	Faults	Chalcedony	Kaolinite	Gypsum	Hematite	Epsomite
144000	1.8405	1.7695	1.3331	1.3808	2.0482	2.2021	2.4437
144000	1.8363	1.7686	1.2979	1.3358	1.0956	1.1450	
144100	1.8096	1.7728	1.0273		1.0925	1.1112	
144500	1.7566	1.7484	1.0195		1.0186		
144000	1.8074	1.7652		1.0612	1.0957	1.1522	
144100	1.7916	1.7588		1.0245	1.2667		1.2576

The Akaike information criterion (AIC) results allow us to compare the relative accuracy of the models, by using the formula (Bozdogan, 1987):

$\boldsymbol{AIC = 2k - 2\ln(\mathcal{L})}$ **,** where k is the number of estimated parameters in the model, and $\mathcal{L}$ is the maximum value of the likelihood function of the model.

That is, a lower AIC is better, penalizes the addition of parameters to the model, and improves logarithmically as the likelihood increases.

Three of the models in the table are equivalent based on AIC, with the 5th case having the lowest number of parameters. Nevertheless, the 2nd case has the same AIC with all columns having a VIF significantly lower than 2.5. This is then the dataset we will use from now on.

The base case: Deep learning model

To compare models, we will use as base case the state-of-the-art Geothermal AI. The Geothermal AI is a deep learning model (DLM) that uses as the ground truth the results of a self-organizing map (SOM) (Kohonen, 1982) created with a superset of the layers used for the training and testing of the AI (Moraga et al., 2022). This approach has advantages in the creating of a "ground-truth" layer without the intervention of experts in geothermal exploration, but has the drawback of generating a "ground-truth" biased to a linear combination of the inputs, as a result of using SOM. Although the linear combination may not be perfectly accurate given the additional information contained in the input layers to the SOM not used in the Geothermal AI's DLM. By using deformation as ground-truth, this bias is eliminated, but there is a possibility that no relationship exists between the inputs and the new ground-truth.

Nevertheless, by running the Geothermal AI with this report's dataset 2 (Figure 2), the following results were obtained:

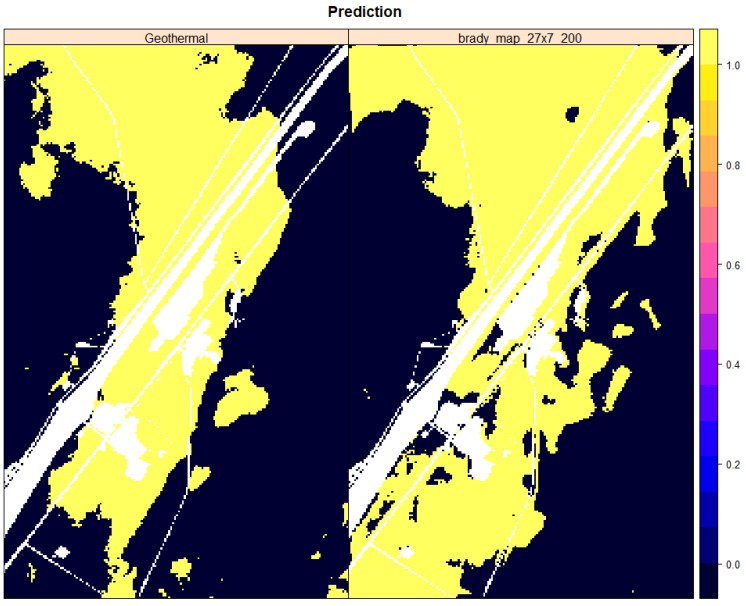


Figure 3 – Results of the Geothermal AI with dataset 2

The Geothermal AI obtains an accuracy of 80.7% in the test set. By running the AI and mapping the whole area of interest, an overall accuracy of 77.4% is obtained.

The confusion matrix is the following, with '1' being the positive case (Geothermal):

Table 4 – confusion matrix for Geothermal AI's Brady prediction

		Prediction	
		0	1
Ground truth	0	504,681	190,217
	1	64,661	344,782

In this case there is an imbalance in the number of positive and negative cases, therefore a more accurate measure of accuracy is balanced accuracy (BAcc), defined as:

Equation 1 – Balanced accuracy equation

$BAcc = \frac{Sensitivity+Specificity}{2}$, where sensitivity and specificity are defined as

Equation 2 – Sensitivity equation

$Sensitivity = \frac{TP}{(TP + FN)}$, and

Equation 3 – Specificity equation

$Specificity = \frac{TN}{(TN + FP)}$

Therefore:

$$BAcc = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right) = \frac{1}{2}\left(\frac{344{,}782}{344{,}782+64{,}661} + \frac{504{,}681}{504{,}681+190{,}217}\right) = \mathbf{78.4}\%$$

Additional measures are:

Table 5 – Precision, recall and f1-score for geothermal AI

	Non-geothermal	Geothermal	Aggregate
Precision	88.6%	64.4%	76.5%
Recall	72.6%	84.2%	78.4%
f-1	79.8%	73.0%	77.5%

This approach uses area-based analysis, so the expectation is that pixel-based analyses will obtain lower values than these.

Regressions

Statistical regression is a method to determine the strength and type of relationship between inputs and outputs in a dataset. Linear regressions are defined in the Multicollinearity analysis section.

To train and test regressions and reduce bias, one technique is to use k-fold cross-validation. That is, split the dataset in k subsets (folds) and using one of the sunsets for testing and the rest for training. Repeated k-fold cross-validation corresponds to the same operation repeated a certain number of times.

In this report, cross-validation (CV) will imply the use of 5-fold cross-validation repeated twice for a total of 10 cross-validation iterations.

Linear Regression with cross-validation

Using a generalized linear model and 5-fold cross-validation repeated twice, the values obtained are:

Table 6 – Linear regression with cross-validation

Iteration	Accuracy	Balanced accuracy	F1-score
1	68.2%	63.5%	51.5%
2	68.2%	63.6%	51.6%
3	68.1%	63.5%	51.6%
4	67.9%	63.3%	51.3%
5	68.0%	63.4%	51.4%
6	68.0%	63.4%	51.4%
7	68.0%	63.4%	51.3%
8	67.9%	63.4%	51.4%
9	68.3%	63.7%	51.8%
10	68.0%	63.4%	51.4%
Aggregate	**68.1%**	**63.5%**	**51.5%**

The aggregated model achieves a balanced accuracy of 63.5%, with the best model achieving at most 63.7% accuracy.

Linear regression with spatial cross-validation

A higher bar is raised when using not just normal cross-validation but spatial cross-validation (SpCV) (Lovelace et al., 2019). In these cases, the training and testing subsets are spatially disjoint, thus reducing the effect of spatial correlation in the analysis. The results follow:

Table 7 – Linear regression with spatial cross-validation

Iteration	Accuracy	Balanced accuracy	F1-score
1	29.4%	50.0%	0.0%
2	59.9%	63.3%	62.6%
3	53.4%	53.0%	43.4%
4	86.0%	67.3%	44.1%
5	78.4%	64.9%	46.1%
6	58.6%	57.6%	50.7%
7	90.8%	50.0%	4.9%
8	57.1%	52.6%	9.9%
9	57.7%	51.9%	70.8%
10	71.8%	59.7%	34.8%
Aggregate	**64.3%**	**57.0%**	**36.7%**

The best case is comparable with the regular cross-validation, but the overall (aggregate) balanced accuracy achieves only 57%.

Machine Learning

To compare the traditional linear regression approach to other methods, supervised machine learning algorithms are selected, these include Neural Networks, Random Forests and Support vector machine.

Neural networks (Lu et al., 2017) are a machine learning paradigm that uses the brain as a model. In this method, neurons (nodes) are connected through each other by axons creating synapses (weights). Learning is achieved by comparing inputs to outputs by using feedforward and backwards approaches. The most traditional and basic module is called the perceptron, where inputs arrive to a single node, they are multiplied by weights, and a bias is added. The node then is triggered using an activation function; for example, a sigmoid, step or regularized linear unitor ReLU function. The results are compared with the expected output and corrections to the weights are made using an adjusting method; for example, backpropagation.

Random forests (Breiman, 2001), or random decision forests, are a classification or regression method that operates by building a tree, with the inputs at the root, and comparisons that split the inputs as branches of the tree. These branches become inputs for deeper trees until they achieve a final classification. By splitting the inputs this way, partitions are created that can be linear or nonlinear.

Support vector machines (SVM) (Noble, 2006), are another set of supervised learning methods that split the inputs using hyperplanes. The relationships are usually linear, but SVMs can also use polygonal and radial kernels to better model more complexity.

In all cases, the sampling is done by using spatial cross-validation (SpCV).

Neural network – one hidden neuron (nnet1)

A model was built using an input, an output and a single hidden neuron. The output neuron has a logistic (binary) activation function, that splits the result in two classes (Geothermal and non-geothermal).

The results are the following:

Table 8 – Neural network with one hidden neuron

Iteration	Accuracy	Balanced accuracy	F1-score
1	31.0%	51.1%	4.6%
2	58.0%	63.4%	64.5%
3	51.6%	59.2%	54.6%
4	73.9%	77.5%	45.4%
5	83.4%	77.1%	69.3%
6	60.1%	63.0%	61.6%
7	82.4%	51.8%	8.7%
8	65.2%	61.6%	37.7%
9	59.9%	52.4%	74.3%
10	63.9%	67.8%	60.0%
Aggregate	**63.0%**	**62.5%**	**48.1%**

Neural network – seven hidden neurons (nnet7)

A comparison model was built using an input, an output and seven hidden neurons. The output neuron has a logistic (binary) activation function, that splits the result in two classes (Geothermal and non-geothermal). The addition of neurons allows for polynomial and non-linear relationships to be captured by the neural network.

The results are the following:

Table 9 – Neural network with seven hidden neurons

Iteration	Accuracy	Balanced accuracy	F1-score
1	31.6%	51.5%	6.3%
2	54.2%	54.3%	49.4%
3	54.3%	52.8%	42.0%
4	82.9%	76.4%	51.0%
5	63.3%	54.7%	35.0%
6	57.3%	55.0%	44.6%
7	82.5%	47.7%	4.7%
8	79.8%	78.6%	74.9%
9	58.3%	57.4%	63.7%
10	65.9%	62.4%	50.6%
Aggregate	**63.0%**	**59.1%**	**42.2%**

Random Forest

A Random Forest algorithm was also used with the same dataset, with up to 100 trees limitation in the model. The results follow.

Table 10 – Random Forest model

Iteration	Accuracy	Balanced accuracy	F1-score
1	48.4%	62.7%	43.3%
2	56.1%	57.1%	53.7%
3	55.9%	54.7%	44.3%
4	83.2%	70.3%	45.2%
5	68.9%	57.0%	34.0%
6	59.9%	56.2%	42.4%
7	85.5%	54.6%	11.6%
8	70.3%	67.8%	55.9%
9	48.1%	46.9%	54.9%
10	69.0%	64.1%	51.7%
Aggregate	**64.5%**	**59.1%**	**43.7%**

Support vector machine

Support vector machines have several implementations possible, including using neural networks. In this case, the SVM was created with a cache of 1GB, a linear kernel with a cost parameter of 100 and a tolerance of 1. The results follow:

Table 11 – Support vector machine (SVM)

Iteration	Accuracy	Balanced accuracy	F1-score
1	52.6%	54.4%	61.0%
2	25.3%	53.0%	12.2%
3	75.8%	76.2%	52.5%
4	58.8%	60.2%	67.2%
5	65.7%	62.6%	75.1%
6	58.7%	60.1%	66.9%
7	65.7%	62.6%	75.1%
8	75.8%	76.1%	52.4%
9	28.1%	54.7%	18.2%
10	53.3%	55.0%	61.3%
Aggregate	**56.0%**	**61.5%**	**54.2%**

Discussion and conclusions

The average results of the analyses show that, in general, the base model is the best and that, when averaged, the Linear regression with normal cross-validation achieves the best results. This is expected from the fact that by using normal cross-validation, the linear regression is also capturing information related to the spatial relationships of the data. Additionally, the linear regression model under spatial cross-validation is the worse performer, and, as expected, its results are worse than the linear regression under normal cross-validation.

Table 12 – Average results of all machine learning algorithms

Model	Accuracy	**Balanced accuracy**	F1-score
Base case	**77.4%**	**78.4%**	**77.5%**
LR w/CV	68.1%	**63.5%**	51.5%
LR w/SpCV	64.3%	**57.0%**	36.7%
Neural network – 1 hidden neuron SpCV	63.0%	**62.5%**	48.1%
Neural network – 7 hidden neurons SpCV	63.0%	**59.1%**	42.2%
Random forest SpCV	64.5%	**59.1%**	43.7%
Support vector machine SpCV	56.0%	**61.5%**	54.2%

On the other hand, the best results of the analyses (Table 13) show a different story.

Table 13 – Best results of all machine learning algorithms

Model	Accuracy	**Balanced accuracy**	F1-score
Base case	**77.4%**	**78.4%**	**77.5%**
LR w/CV	68.3%	**63.7%**	51.8%
LR w/SpCV	86.0%	**67.3%**	44.1%
Neural network – 1 hidden neuron SpCV	73.9%	**77.5%**	45.4%
Neural network – 7 hidden neurons SpCV	79.8%	**78.6%**	74.9%
Random forest SpCV	83.2%	**70.3%**	45.2%
Support vector machine SpCV	75.8%	**76.2%**	52.5%

The best pixel-based ML model is the neural network with 7 hidden neurons, and the results are competitive with the base case of the Geothermal AI.

When comparing the maps, the base case (Figure 3) shows more contiguous space and almost no salt and pepper errors. On the other hand, the prediction map (Figure 4) from the neural network with 7 hidden neurons (nnet7), although containing a lot of salt and pepper inside and at the edges of the geothermal region, does a very good job of highlighting the areas where there is geothermal potential.

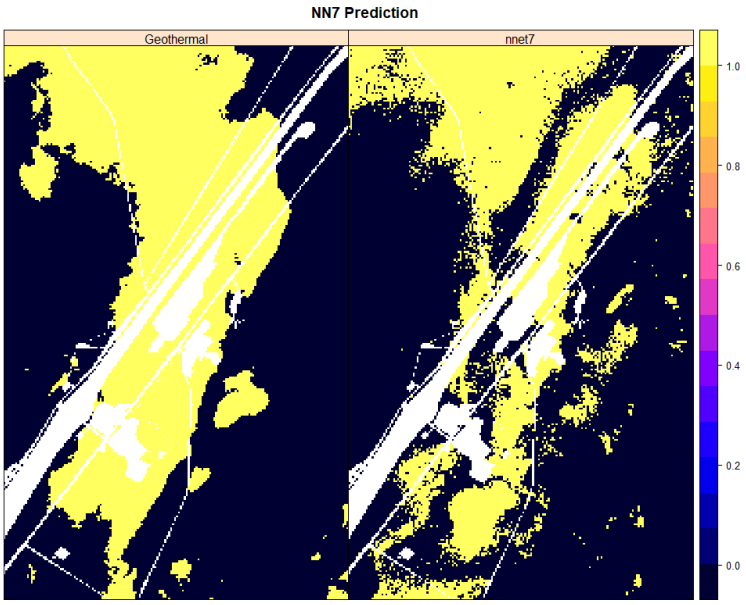


Figure 4 – Dataset 2 Geothermal layer and nnet7 prediction

Moreover, when comparing the statistics (Table 14) of the base case (Geothermal AI), against the best competing case (nnet7), we can conclude that the neural network with just 7 neurons is not only competitive overall but performs better at predicting positive values (76.88% vs 64.45%) by reducing the number of false positives to less than half (from 190,217 to 81,011).

These results show that a simpler wide network can obtain results that are competitive with a more complex, area-based model.

Table 14 – Statistics and confusion matrix for (a) base case, and (b) nnet7

Confusion Matrix and Statistics Reference Prediction No Yes No 504,681 64,661 Yes 190,217 344,782 Accuracy: 0.7692 95% CI: (0.7684, 0.77) No Information Rate: 0.6292 P-Value [Acc > NIR]: < 2.2e-16 Kappa: 0.5347 Mcnemar's Test P-Value: < 2.2e-16 Sensitivity: 0.8421 Specificity: 0.7263 Pos Pred Value: 0.6445 Neg Pred Value: 0.8864 Prevalence: 0.3708 Detection Rate: 0.3122 Detection Prevalence: 0.4845 Balanced Accuracy: 0.7842 'Positive' Class: Yes	Confusion Matrix and Statistics Reference Prediction No Yes No 613,887 140,097 Yes 81,011 269,346 Accuracy: 0.7998 95% CI: (0.799, 0.8005) No Information Rate: 0.6292 P-Value [Acc > NIR]: < 2.2e-16 Kappa: 0.5578 Mcnemar's Test P-Value: < 2.2e-16 Sensitivity: 0.6578 Specificity: 0.8834 Pos Pred Value: 0.7688 Neg Pred Value: 0.8142 Prevalence: 0.3708 Detection Rate: 0.2439 Detection Prevalence: 0.3173 Balanced Accuracy: 0.7706 'Positive' Class: Yes
(a) Base case (Geothermal AI)	(b) Best competing model (nnet7)

This report show how different machine learning approaches can be applied to Geospatial Data, and follows the analysis of dataset 2 by first using the Geothermal AI and applying 4 different algorithms, two of them with different parameters or cross-validation methods.

It has been shown that a neural network with 7 neurons can achieve a performance similar to the Geothermal AI for this dataset, showing more salt and pepper error but obtaining a more conservative result that reduces false positives.

Further work can be done by analyzing different neural network architectures, including wide and deep learning models, and new sampling and cross-validation using the Geothermal AI to show with more certainty the power of deep versus wide approaches in this dataset.

References

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, *52*(3), 345–370.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Çavur, M., Moraga, J., Duzgun, H. S., Soydan, H., & Jin, G. (2021). Displacement Analysis of Geothermal Field Based on PSInSAR And SOM Clustering Algorithms A Case Study of Brady Field, Nevada—USA. *Remote Sensing*, *13*(3), 349. https://doi.org/10.3390/rs13030349

Craney, T. A., & Surles, J. G. (2002). Model-dependent variance inflation factor cutoff values. *Quality Engineering*, *14*(3), 391–403.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, *43*(1), 59–69. https://doi.org/10.1007/BF00337288

Lovelace, R., Nowosad, J., & Muenchow, J. (2019). *Geocomputation with R*. Chapman and Hall/CRC.

Lu, Z., Pu, H., Wang, F., Hu, Z., & Wang, L. (2017). The Expressive Power of Neural Networks: A View from the Width. *ArXiv:1709.02540 [Cs]*. http://arxiv.org/abs/1709.02540

McCullagh, P., & Nelder, J. A. (2019). *Generalized linear models*. Routledge.

Moraga, J. (2021). *Hydrothermal mineral alterations in the Brady and Desert Peak geothermal fields*. Colorado School of Mines; Geothermal Data Repository. https://doi.org/10.15121/1824162

Moraga, J., Duzgun, H. S., Cavur, M., & Soydan, H. (2022). The Geothermal Artificial Intelligence for geothermal exploration. *Renewable Energy*, *192*, 134–149. https://doi.org/10.1016/j.renene.2022.04.113

Noble, W. S. (2006). What is a support vector machine? *Nature Biotechnology*, *24*(12), 1565–1567.

OpenEI. (2021a). *Brady Hot Springs Geothermal Area*. Open Energy Information. https://openei.org/wiki/Brady_Hot_Springs_Geothermal_Area

OpenEI. (2021b). *Geothermal Energy | Open Energy Information*. https://openei.org/wiki/Gateway:Geothermal