Progress and Challenges in Machine Learning Geomodel Development at LANL for the New Mexico CORE-CM Effort

Introduction

As part of Task 2.0, Los Alamos National Lab was tasked with developing a machine learning based model of rare earth and critical mineral elemental distribution within the New Mexico CORE-CM region. A traditional approach to a elemental distribution map would rely on aggregation of known elemental concentrations through field sampling and subsequent geochemical analyses, followed by integration into a geographic information system and some sort of interpolation algorithm applied (e.g., Kriging). In cases where there is abundant data, and where predictive capabilities are not needed (e.g., measuring a set of elements in the field to predict others), this can often be a fruitful approach. However, in the case of New Mexico as relates to REE/CM distribution, we have a twofold challenge of often missing data fields within historical data (e.g., missing elements in USGS Coalqual database) and/or sparse data concentrated in just a few regions. The LANL developed SmartTensors packages is ideal for this task. Below, we detail progress made in developing a predicative geospatial model of REE/CM in the New Mexico CORE-CM basin. Finally, we detail key successes, key challenges, and goals for future work.

Methods and Challenge

We used the package SmartTensors to develop the set of signals corresponding to REE enrichment. As implemented here, SmartTensors is general high-performance unsupervised machine learning algorithm. Here, we employ non-negative matrix-factorization + k-means clustering. Algorithms such as these find patterns in the available data, and in this case is done so in an unsupervised manner on the entire database. The advantage here, is that the resulting patterns yield an output with many fewer variables of the most importance and is generally insensitive to missing values. Table 1 shows an example of missing data from the New Mexico dataset.

Table 1. Example of input data file, highlighting lack of information for many elements and many samples and the complex need to account for missing data.

Sample	**Lat**	**Lon**	**Datum**	**Depth**	**Cr**	**Cs**	**Cu**	**Ga**	**Ge**	**…**
A	36.803	-108.4407	NAD27	0	49.073	NA	50.525	NA	NA	…
B	35.484	-107.6634	NAD27	-313.08	49	10.11	NA	NA	17.1	…
C	35.484	-107.6634	NAD27	-308.08	7	NA	NA	NA	NA	…
D	35.484	-107.6634	NAD27	-269.93	5	NA	NA	5.45	NA	…
E	35.484	-107.6634	NAD27	-240.46	3	NA	NA	NA	NA	…

The data was provided by the resource characterization team developed as part of Task 2.0 and includes new analyses performed as part of CORE-CM plus a compilation of data from the USGS Coalqual database. A total of 252 samples were used, with approximately 100 input variables. Bad data was removed, and unusable (e.g., textual data) was removed for this initial

effort. We will elaborate on needs for future work around textual data more below. Samples cover an area of approximately 20,000 mi^2, mostly from the San Juan Basin coal fields (Figure 1). Also included however are samples from other potential REE deposits, some of which are higher in concentration. This sparsity of data at multiple scales, and wide range in observed concentrations will important impacts on result map products.

Fundamentally, the initial output will be signals associated with only points of known elemental concentration, in other words and map of signals for our training data. IN a perfect world, the data would be abundant and evenly distributed across the domain of interest. However, in our case here, we need to spatially interpolate resulting signals. We explored this using a variety of Kriging methods.

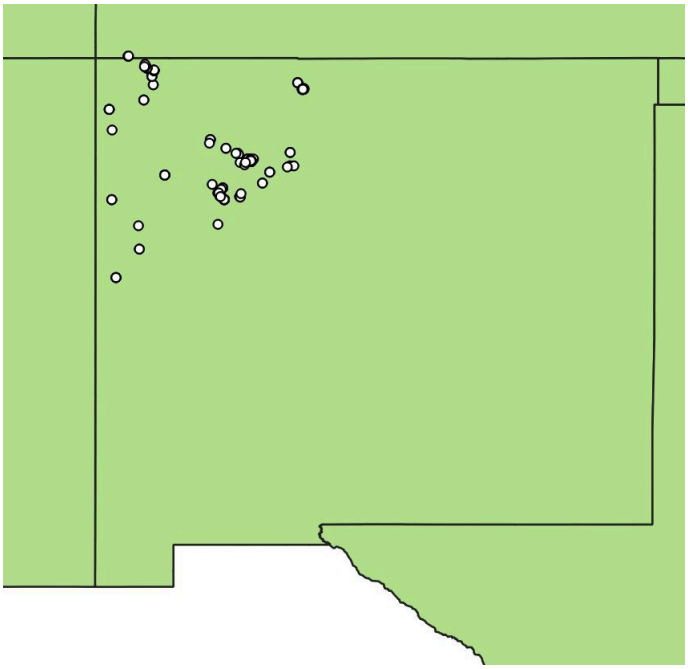

Figure 1. Distribution of samples for initial model development.

Results

Initial results are promising but will require refinement. Example signals are shown in table 2, and for each location, there is a unique set of matrix factorizations for a given signal. These signals are then Kriged to generate a spatial representation of the various signals. For the 252 test samples considered here, there were three resulting signals. The number of signals is not initially set by the model, and instead is part of the learning process. Table 2 and Figure 2 shows the results for the three signals determined by the model. It is important to note that there are not any specific REE nor a concentration, but rather a heat map of the importance spatially of a various signal.

Table 2. Example output of NMFk signals from our 252 sample test dataset.

X	Y	Depth	Signal 1	Signal 2	Signal 3
35.4	-107.6	-82.18	19.858	11.091	157.884
36.8	-104.9	-35.8	1.225	11.537	143.244
35.6	-107.6	-290	122.386	53.054	2.083
35.9	-107.4	0	108.681	9.343	0.000
…	…	…	…	…	...

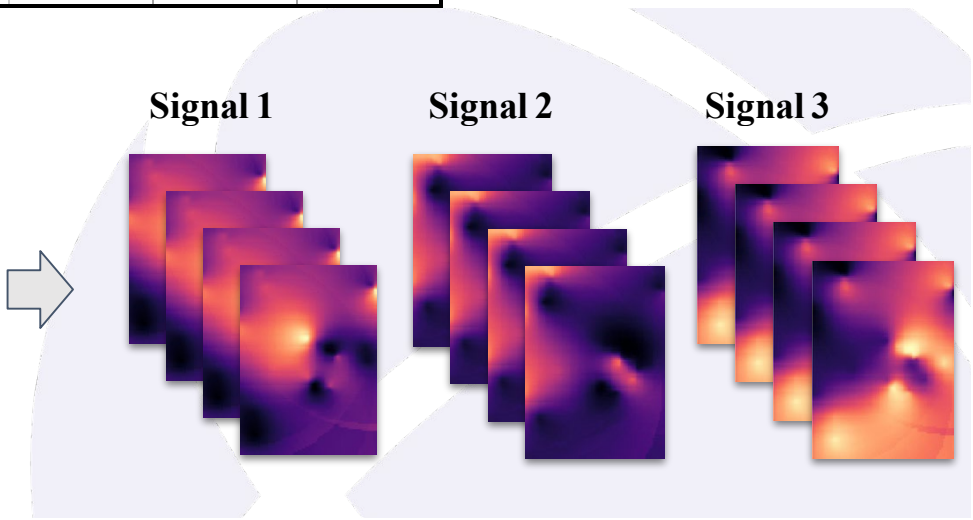

Figure 2. Example of NMFk signals as a stack to represent the vertical direction. Shown are unique signals for the different signals, which are influenced by different combinations of geochemical parameters.

Figure 3 shows the conversion of the 3 NMFk signals when used to determine the total REE for the New Mexico domain. Two main takeaways from this figure, the first is strong anisotropy between the horizontal and vertical domains. While there may exist anisotropy as a result of variation in total REE concentration, we are most likely observing strong biases in the distribution of data points. Second, for the surface concentrations the relatively large spread in concentration means that data representation is currently poor with little resolution to refine variation in REE concentration.

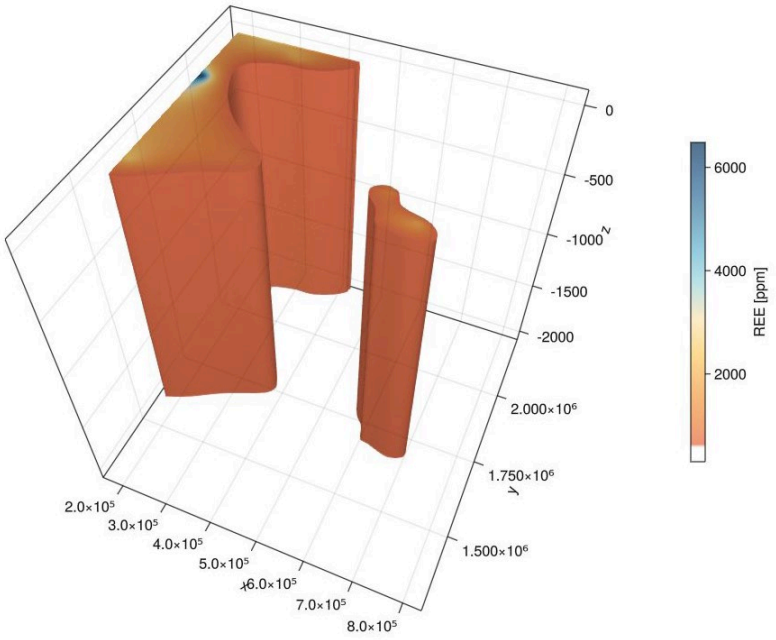

Figure 3. Heat map of predicted total REE concentration based on three NMFk signals

Figure 4 further highlights the issue of anisotropic variation in the horizontal and vertical domains. In particular, it demonstrates the sensitivity of the resulting data to the application of scale factors. All units in feet.

Figure 4. See next page.

Finally, we examined the correlation between NMFk signals (weights) and the parameters in source data set. Such an analysis starts to examine which measurable/observable parameters control a given signal and its weight. For example, Signal 1 has high weight lithologically with "coal", it also has high weight elementally with carbon, along with a few other elements. One caveat here however is that the model currently handles textual data so common in geologic literature as binary. That is, for any given variable such as coal, it is either coal and converted to a value of 1, or if false, 0.

Conclusions and Key takeaways

We have made good initial progress on machine learning based modeling of REE/CM distribution in the New Mexico CORE-CM domain. We tested model development using 252 legacy and new samples developed by Task 2 and implemented using the SmartTensors machine learning package. Anisotropy is the norm within the data and highlights the need for additional data much as possible; however, we can also explore potential solutions in data presentation. Finally, additional work is needed in handling textual geo data, for example lithologic descriptions, formation names, etc.

Successes

- Built a proof-of-concept model that can predict concentrations of REE/CMs across the region;

- Combined unsupervised ML techniques with geostatistics to learn signatures of REE/CMs and map them spatially;
- The current set of data provide useful information for horizontal variation in REE concentration;

Key Challenges

- Lack of data in the vertical direction sets up strong anisotropy in resulting data models;
- Overall lack of samples to generate tall skinny matrices, which are preferred in ML models. Currently have too many variables for the number of samples, resulting in model runs not always replicating adequately;
- Incorporation of geologic textual information needs advancement (e.g., natural language models);

Work remaining that might constitute a Phase 2

- Incorporating detailed stratigraphic information and improving the way we represent geologic information in the model;
- How much data is too much, in this case it isn't too many samples but rather too many variables as possible signals;
- Connecting the existing model with a GIS framework. We have some initial, promising progress on the GIS front, but there's a lot more to do;
- Improving the way we deal with anisotropy in the vertical direction.

Figure 4. Heat map of NMFk signals for a given geologic or geochemical parameter.