

Quarterly Report – Public Page

Date of Report: 3rd Quarterly Report-June 30th, 2020

Contract Number: 693JK31910018POTA

Prepared for: DOT PHMSA

Project Title: Mapping Indication Severity Using Bayesian Machine Learning from Indirect Inspection Data into Corrosion Severity for Decision-Making in Pipeline Maintenance

Prepared by: TEES (Texas A&M Engineering Experiment Station) and University of Dayton

Contact Information: Homero Castaneda, hcastaneda@tamu.edu, 979 458 9844.

For quarterly period ending: June 30th, 2020

1: Items Completed During this Quarterly Period:

<i>Item</i>	<i>Task</i>	<i>Activity/Deliverable</i>	<i>Title</i>	<i>Federal Cost</i>	<i>Cost Share</i>
#7	#1, 2	Report with current results based on Task1 and 2: Establishing a database and Experiments and analysis to bridge the gap in uncertainty quantification.	3 rd Quarterly Report	4,000.00	0.00

The title of the table is based on the file Technical and Deliverable Payable Milestone

2: Items Not-Completed During this Quarterly Period:

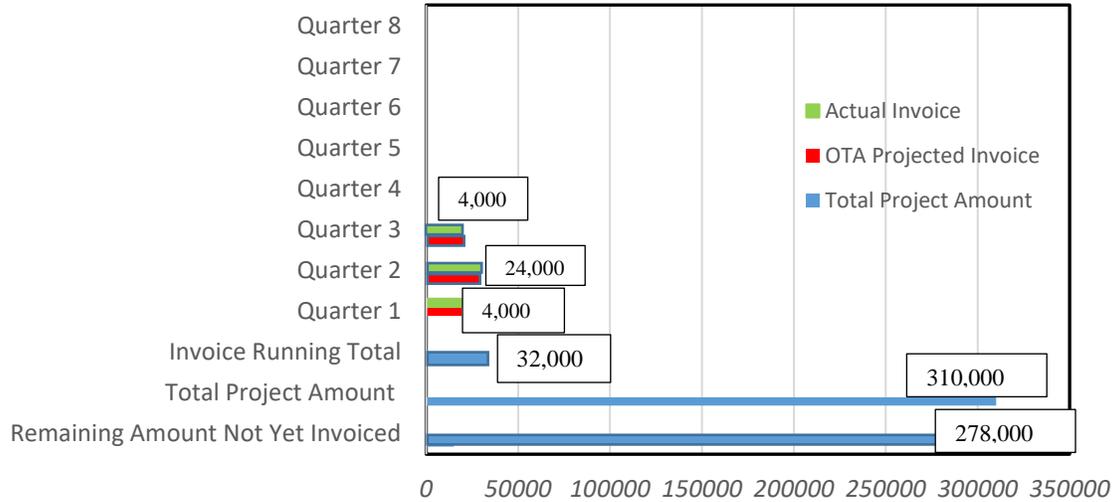
Task number 2, laboratory set up and extract basic corrosion model parameters started during the previous quarter report. Part of Task 2 will be cover in the following partial report. The following activities will be ready in the following quarter report based on the Technical and Deliverable Payable Milestone

<i>Item #</i>	<i>Task #</i>	<i>Activity/Deliverable</i>	<i>Title</i>	<i>Federal Cost</i>	<i>Cost Share</i>
4	1	Mapping available data via GIS tools and geographically co-register all datasets.	Establishing a database	20,000.00	10,000.00
6	2	Laboratory set up and electrochemistry mechanisms and corrosion assessment	Experiments and analyses to bridge gaps in prior knowledge	55,000.00	0.00
8	2	Extract basic corrosion model and embed into the previously developed stochastic corrosion rate model framework	Experiments and analyses to bridge gaps in prior knowledge	24,000.00	0.00

3: Project Financial Tracking during this Quarterly Period:

The table has been updated based on the deliverables and corrected attachment No5 *Technical and Deliverable Payable Milestone*.

Quarterly Payable Milestones/Invoices 693JK31910018POTA



4: Project Technical Status –

The following tasks are included in the project:

- **Task 1: Establishing a database**
- **Task 2: Experiments and analyses to bridge gaps in prior knowledge**
- **Task 3: Bayesian machine learning to bridge gaps in uncertainty quantification.**
- **Task 4 Finalize and evaluate/validate the model.**

During the third quarter, the team members from Texas A&M University (TAMU) and the University of Dayton (UD) had different meetings and an informal workshop conducted to deepen the understanding of laboratory results and how we are linking the data with filed measurements. In addition, the sponsor team gave a feedback of the analysis and experimental procedure for this project.

The team organized an internal Workshop entitled “*Indirect tools outcome as a parameter for severity and how we link the parameters with laboratory results*”

The outcomes of the workshop will help the PhD students in both TAMU and UD teams with different knowledge backgrounds to understand the corrosion mechanism and which parameters are used as primary precursors in the field following the understanding in the well control environment of the laboratory.

During the internal team meeting, we discussed different actions to cover task 1 and task 2.

Some of the results and highlights are summarized in this report as follows:

Tasks 1 and 3 – Establishing a database and Bayesian machine learning to bridge the gaps in uncertainty quantification.

Large scale environmental Database: In order to incorporate more features for the Bayesian Machine learning, a broader database with environmental data was established for better characterizing large-scale soil environment. With a number of open source data available, many databases were explored and data was analyzed using Python, QGIS and Google Earth Engine. Raster files of vegetation cover, precipitation and topology for the region of interest were processed.

Vegetation Cover: Normalized Differential Vegetation Index (NDVI) gives information corresponding to vegetation cover and reflecting the relevant level of organic content in the soil. NDVI is calculated as the difference between near-infrared (reflected strongly by vegetation) and red (absorbed by vegetation). In this project LANDSAT 8 data was used, and in LANDSAT 8 dataset, band 5 is near-infrared(0.85- 088µm) and band 4 is red (0.64 -0.67µm). Hence NDVI can be calculated as:

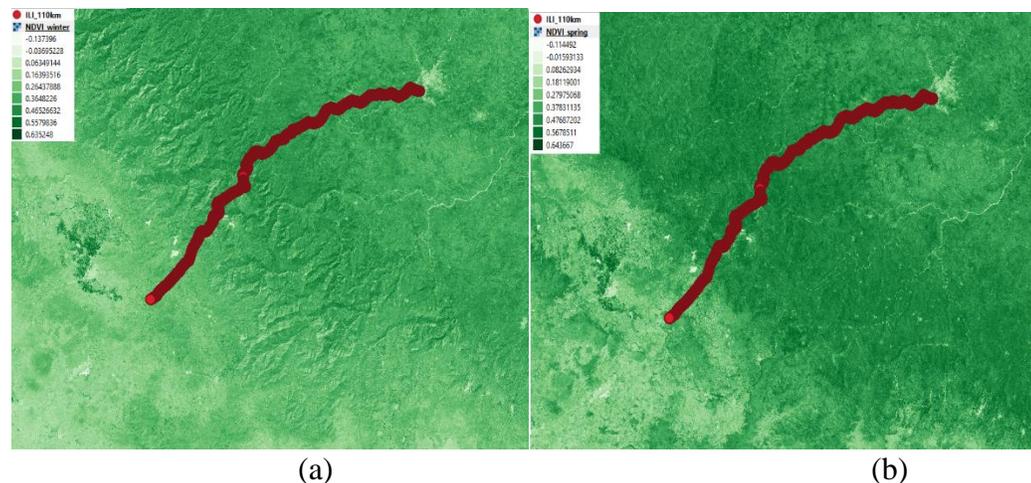
$$NDVI = \frac{(NIR - RED)}{(NIR + RED)} = \frac{(BAND 5 - BAND 4)}{(BAND 5 + BAND 4)}$$

In order to understand seasonal effects to the corrosion severity, NDVI was processed seasonally (winter, spring, summer and fall).

Resolution:30m

Processing Tools: Google Earth Engine (GEE) and QGIS

GEE combines multiple catalog of satellite imagery and geospatial datasets which can be accessed and manipulated on the cloud computing platform. The required dataset can be searched and imported to the users customized programming script. For this project LANDSAT 8 data was imported. Once imported, the dataset can be further filtered and processed to obtain the NDVI raster map. The seasonal NDVI raster maps are shown below in Figure 1.



Date of Report: 3rd Quarterly Report –

Prepared by Homero Castaneda, TEES
Hui Wang, U. Dayton

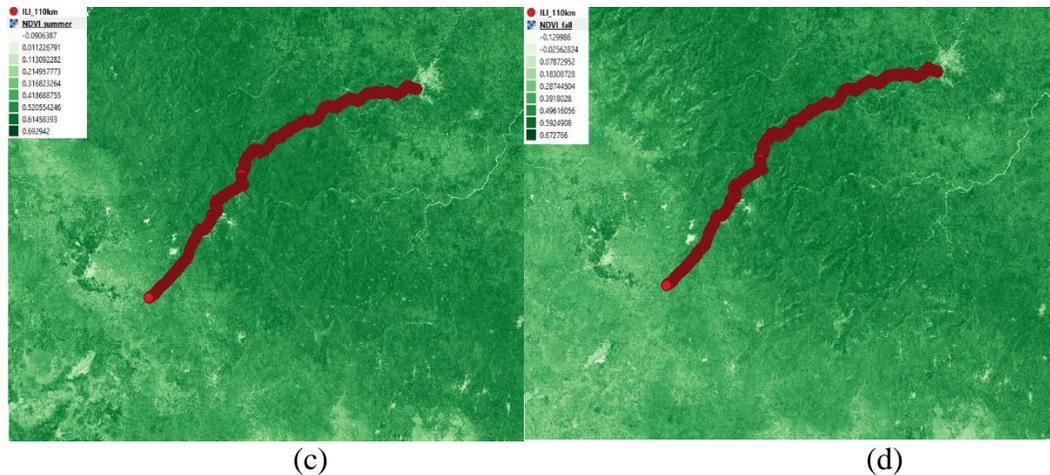


Figure 1: (a) Winter (b) Spring (c) Summer (d) Fall NDVI overplotted with 110km pipeline RoW

Precipitation: Precipitation describes the amount of water input in a given area and hence one of the environmental factors which can contribute to pipeline external corrosion. In this report, Climate Hazards Group InfraRed Precipitation with Station Data (CHIRPS) was used to create rainfall time series in the given region of interest. CHIRPS incorporates satellite imagery and in-situ station data to create gridded rainfall time series.

Resolution: 0:05 arc degrees.

Processing Tools: Google Earth Engine (GEE)

CHIRPS raster database has only a single precipitation band with units mm/pentad (pentad means 5 day period). Since in our project we are interested in the years 2005 to 2010. The data is filtered and processed for these years. Figure 2 shows the mean yearly precipitation for 5 years and Figure 3 shows the mean yearly precipitation for 2009. From the plots we see that the region gets maximum precipitation in the months of September and November and clear periodical trend can be noticed. The external corrosion processes can be affected by this seasonal fluctuation of precipitation in terms of water level condition (i.e., submerged, semi-submerged, or above water) and combined with elevation effect. Figure 4 shows the mean precipitation distribution within the area of interest overlaid with 110 km pipeline RoW. As expected we see more precipitation nearer to the coast and decreases as we move towards the mountain.

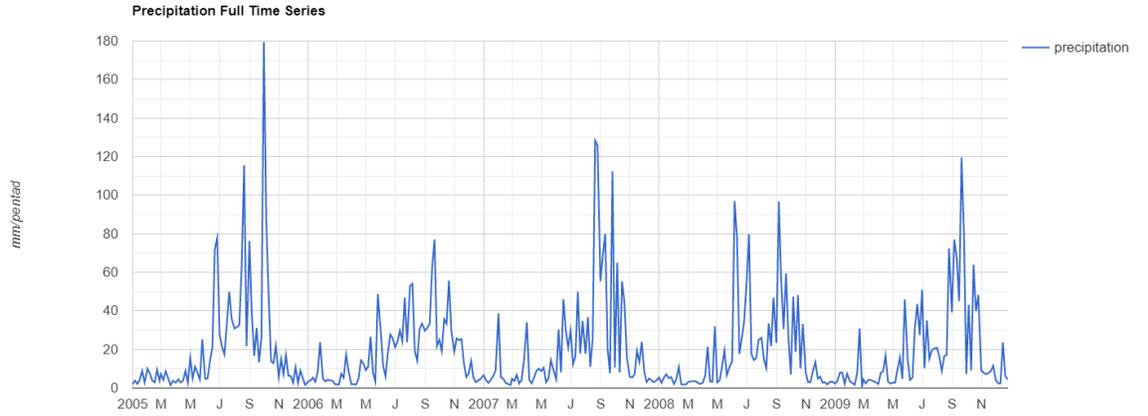


Figure 2: Five year mean precipitation.

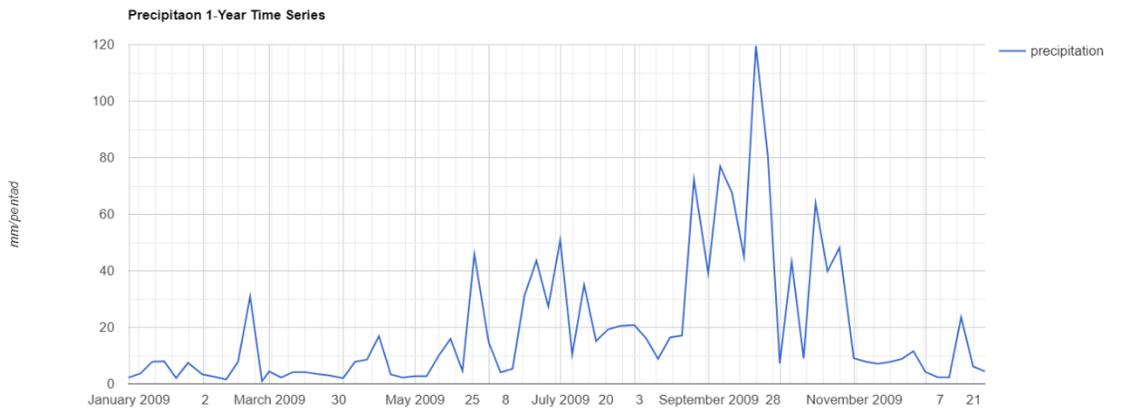


Figure 3: Yearly mean precipitation

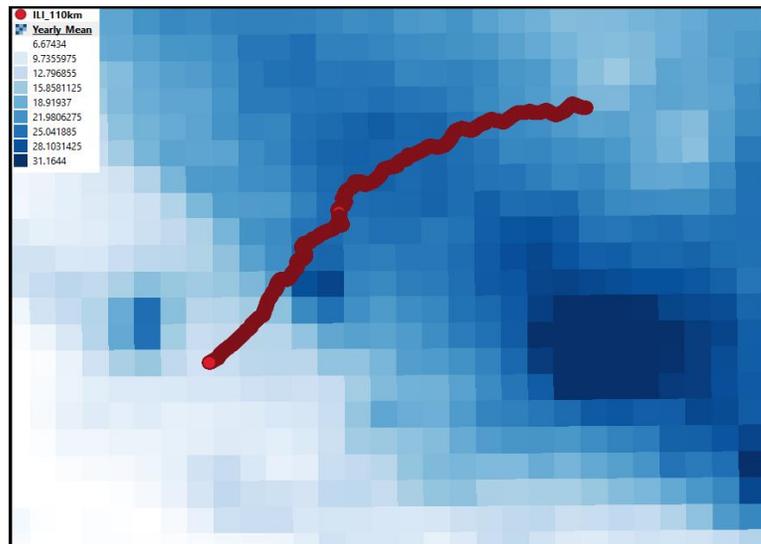


Figure 4: Mean yearly precipitation in the region of interest.

1. Soil physical properties

To study the effects of soil physical properties on corrosion of underground pipeline, different soil properties like bulk density of the fine earth fraction, proportion of clay particles (< 0.002 mm) in the fine earth fraction, volumetric fraction of coarse fragments, proportion of sand particles (> 0.05 mm) in the fine earth fraction and proportion of silt particles (≥ 0.002 mm and ≤ 0.05 mm) in the fine earth fraction were studied.

Bulk density: It is an indicator of soil compaction. It is calculated as the dry weight of soil divided by its volume. This volume includes the volume of soil particles and the volume of pores among soil particles.

Coarse Fragment: Particles that are more than 2 millimeters in diameter are not included in chemical, mineralogical, and some physical analyses, and they are called coarse fragments.

Clay: Clays are made up of secondary clay minerals and oxides/oxyhydroxides of iron and aluminum, and are less than 2 microns in diameter.

Sand: Comprise quartz and resistant primary minerals such as mica. Sand particles are between 2 mm and 20 microns in size.

Silts: These are typically composed of quartz and small mineral particles such as feldspars and mica, and are between 2 and 20 microns in diameter.

Data Source: SoilGrids- <https://soilgrids.org/> . SoilGrids is a system for global digital soil mapping that makes use of global soil profile information and covariate data to model the spatial distribution of soil properties across the globe. SoilGrids maps are a global soil data product generated at ISRIC — World Soil Information as a result of international collaboration. Data at 5 depths are available. Figures 5-9 shows each of these property data at 60cm depth overlaid with pipeline right of way.

Resolution: 250 m.

Mapped Units:

Name	Mapped Units
Bulk Density	cg/m ³
Coarse Fragment	cm ³ /dm ³ (vol%)
Clay Content	g/kg
Sand Content	g/kg
Silt Content	g/kg

Processing Tools: QGIS 3.1

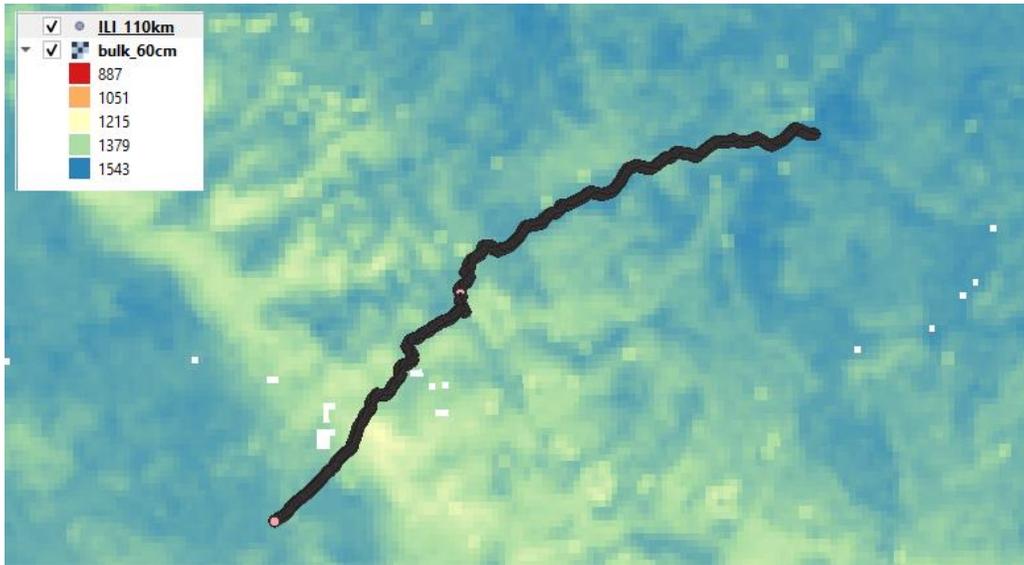


Figure 5: Soil Bulk Density

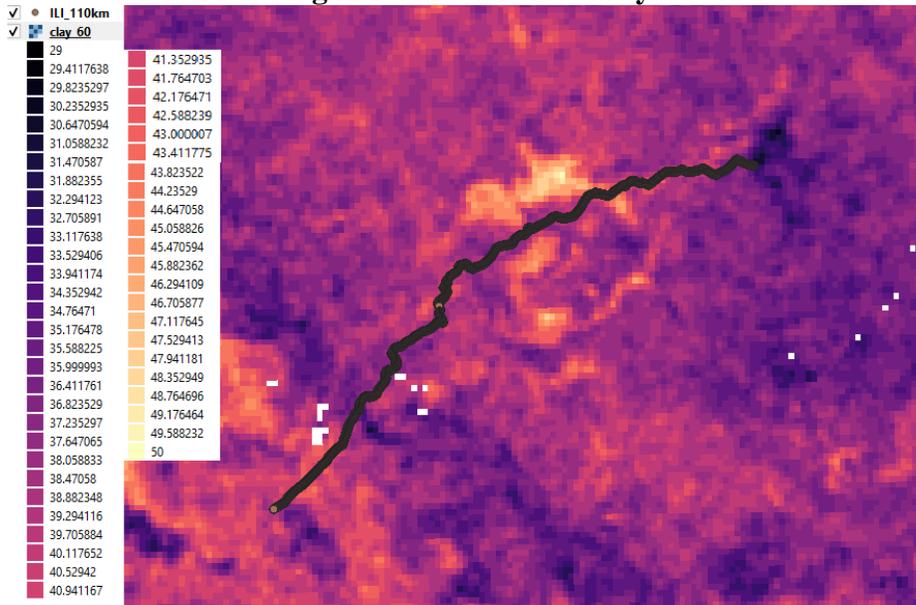


Figure 6: Soil Clay Content

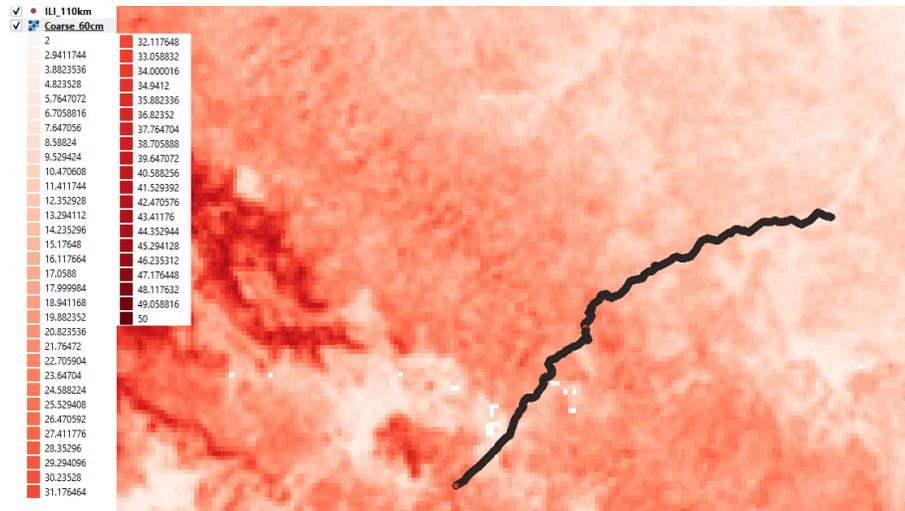


Figure 7: Soil Coarse Fragment

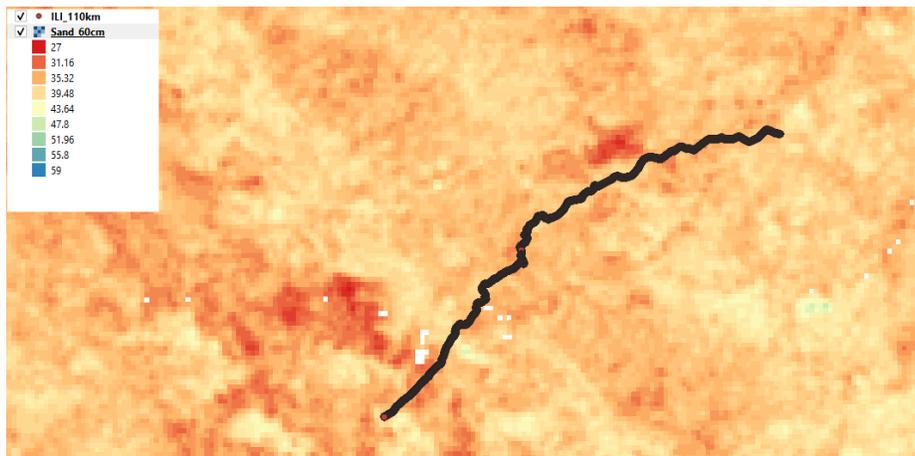


Figure 8: Soil Sand Content

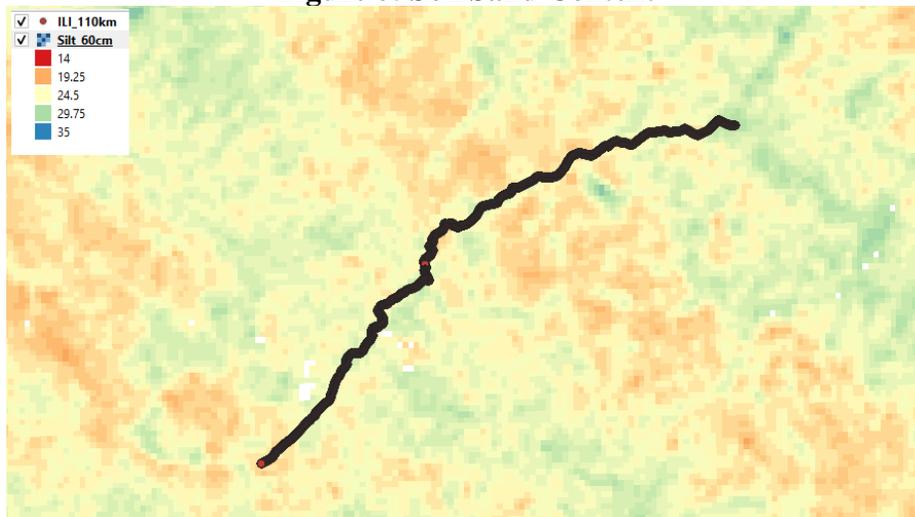


Figure 9: Soil Silt Content

Task 2: Experiments and analyses to bridge gaps in prior knowledge

During this quarter, we started to identify critical gaps in prior knowledge (i.e., current indirect survey, environmental conditions and other databases) and coded (or related) to deterministic and probabilistic modeling by following the corrosion mechanism. The correlation between primary and critical parameters for corrosion precursors has been considered by theoretical approach and find the possible gaps in experimental conditions.

Experiments set up. Gaps in prior knowledge relating coating conditions and corrosion severity under controlled environmental factors will be addressed through laboratory experiments.

The first set of laboratory experiments primarily consider the effects of soil resistivity and the metallic surface condition in the presence of holidays (specifically active and passive state) under cathodic protection. The experimental design is presented in Table 1. Buffer solution (with defined conductivity, pH and TDS Standards) is applied to adjust solution resistivity. The passive holiday can be realized by external anodic current via potentiostat (Gamry, The Interface 600plus™). NS4 solution with composition (unit: g/L) of KCl: 0.122, NaHCO₃: 0.483, CaCl₂: 0.093 and MgSO₄: 0.131 is used to simulate soil conditions. Figure 10 shows the carbon steel polarized in a sodium carbonate-bicarbonate buffer solution with a pH of 10 at a scan rate of 0.167 mV/s. Figure 10b, shows constant potential of 0.3 V/SCE was applied on the metal surface for 3 hours to ensure the sufficient low and stable current density. A final current density of around 150 nA/cm² was observed.

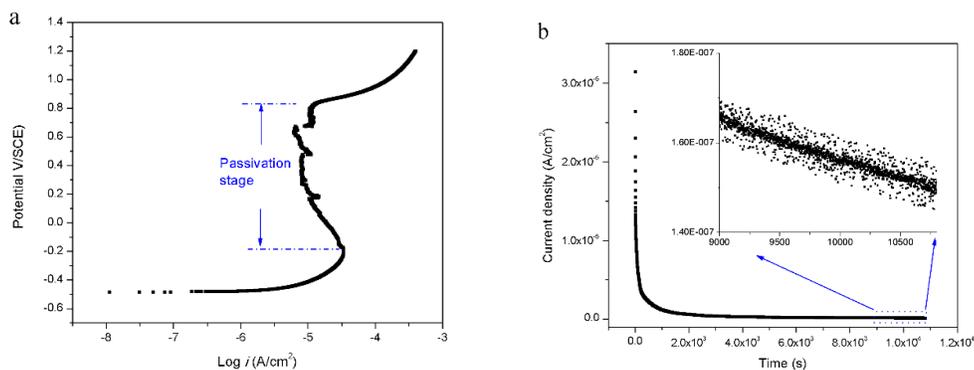


Fig.10 (a) Potentiodynamic polarization curves and (b) current density response to potentiostatic polarization of carbon steel for 3 hours in a carbonate-bicarbonate buffer solution with a pH of 10

Sequence of Non-destructive methods for Close interval survey simulation; the methods include: the evolution of open circuit potential (OCP), Electrochemical impedance spectroscopy (EIS), linear polarization resistance (LPR), and surface analysis will be carried out in sequence to measure the response of experimental set-ups for the analysis of CIS database. Also the parameters that will simulate the DCVG database will be included

in the sequence of techniques and outcomes. For the electrochemical measurement system, the three-electrode method will be used. The frequency range of EIS starts from 100 kHz to 10 mHz with an amplitude of 10 mV. A small voltage variation (± 20 mV) based on its corrosion potential is applied during LPR tests. Microscopic or direct observation will be used to characterize the corrosion severity of holidays. All measurements will include duplicates to ensure reproducibility.

Table 1 Experimental design matrix for CIS and DCVG analysis

Samples	Coatings Thickness	Soil Composition	Distribution of soil (resistivity)	Severity based on active-passive concept	pH
AISI 1008/API X52	10-20 mils	NS4	Conductivity 1	Active Holiday	4
AISI 1008/API X52	10-20 mils	NS4	Conductivity 1	Active Holiday	7
AISI 1008/API X52	10-20 mils	NS4	Conductivity 1	Active Holiday	9
AISI 1008/API X52	10-20 mils	NS4	Conductivity 1	Passive Holiday	4
AISI 1008/API X52	10-20 mils	NS4	Conductivity 1	Passive Holiday	7
AISI 1008/API X52	10-20 mils	NS4	Conductivity 1	Passive Holiday	9
AISI 1008/API X52	10-20 mils	NS4	Conductivity 2	Active Holiday	4
AISI 1008/API X52	10-20 mils	NS4	Conductivity 2	Active Holiday	7
AISI 1008/API X52	10-20 mils	NS4	Conductivity 2	Active Holiday	9
AISI 1008/API X52	10-20 mils	NS4	Conductivity 2	Passive Holiday	4
AISI 1008/API X52	10-20 mils	NS4	Conductivity 2	Passive Holiday	7
AISI 1008/API X52	10-20 mils	NS4	Conductivity 2	Passive Holiday	9

Table 2 Technique to be used and parameter to quantify for CIS functional expressions

Experimental technique	Outcome parameters	Correlation to the field interpretation
OCP and Ecorr	Potential at No IR Current decay profile vs. time	Pipe to soil-potential
EIS	Surface mechanisms, corrosion rate Passive state or active state magnitudes	Severity and criterion Ranking for correlations
LPR	Corrosion rate MPY	Severity and ranking for correlations

Table 3 Experimental design matrix for DCVG analysis

If concentration of ions are considered, we will include more tests on concentration of HCO_3 , Cl and SO_4 at fixed conductivity and pH

Sample	Soil Composition	Severity based on active-passive concept	HCO_3	Cl^-	SO_4
API X52	NS4	Active Holiday	#1		
API X52	NS4	Passive Holiday	#2		
API X52	NS4	Active Holiday		#1	

API X52	NS4	Passive Holiday		#2	
API X52	NS4	Active Holiday			#1
API X52	NS4	Passive Holiday			#2

Table 4 Technique to be used and parameter to quantify for DCVG and assist in the generation of functional expressions

Experimental technique	Outcome parameters	Correlation to the field interpretation
Resistivity of the soil	Soil conditions Resistivity per element to isolate the meaningful resistance	Resistivity vs. IR
OCP and EIS	Surface mechanisms, corrosion rate Current decay and IR due to severity	Severity and criterion Ranking for correlations
Potential gradients by using DCVG technique approach	Corrosion rate MPY	Potential gradient vs severity

Field data analysis

Upon completion of compiling the data to create the master files for the 110 km pipeline and 60 km pipeline data, certain discrepancies were found when attempting to map the data. Issues regarding soil moisture and water distribution information are being resolved. For ease in designing the simulation strategy and setting up experiments, the closed-interval survey (CIS) data has been considered against other environment-related factors such as pH and ionic concentrations of species to generate probability density functions (PDFs). These were plotted with the use of the open-circuit potential (off potential) measured via CIS and other soil parameters that are likely to affect this magnitude. These will be used further to determine functional expressions and how will be characterized in the laboratory.

Mechanistic analysis of the soil and corrosive environment in the field included thermodynamic fundamentals by means of Nernst Equation and semi-empirical relations based on corrosion potential, open circuit potential and/or Off potential concepts with the soil parameters will be performed after having the full set of results from the laboratory experiments.

To perform the Bayesian machine learning analysis, the geo-referenced database is converted to an object based database such that each pipeline segment (with a predefined length, say 0.06 miles) with associated features (i.e., measured soil environment, direct and indirect inspections). Various data collected as soil chemical and physical properties, direct inspection data and indirect inspection data are converted into object based data-frames indexed by the pipeline segment ID. Hence in the global data-frame, each pipeline segment will have the data structure shown in Figure 11.

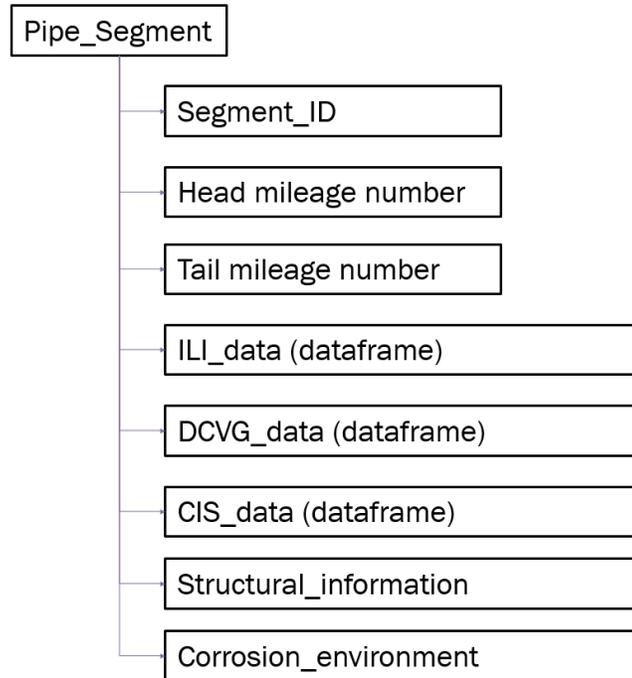


Figure 11: Data structure for Bayesian machine learning

Now the feature space comprises of all the features associated to each segment. Referring to the proposed work flow, all the features can be categorized into two groups: 1) Indicative features and 2) target features. The indicative features include all the indirect information from the environment of the pipeline. These are the chemical and physical properties of the soil environment as well as indirect inspection results. In this report and also according to the proposed work, the soil chemical and physical properties are used for clustering the pipeline segments into different groups in order to detect possible heterogeneity of the soil corrosion environment along the pipeline right-of-way. Pipeline segments are categorized into multiple groups such that items in the same cluster will have similar soil environment and hence are assumed to be exposed to the similar corrosivity level. The inter cluster difference is assumed to be a part of the driving force that leads to different corrosion severities.

Task 3: Bayesian machine learning to bridge the gaps in uncertainty quantification

The entire machine-learning framework comprises of three layers:

- Unsupervised clustering on regional environment and soil information for corrosivity similarity analysis.
- Supervised classification for identifying corrosion defects and severities.
- Regression and projection for spatial and temporal characterization of active corrosion process.

The first layer of unsupervised clustering is performed by using Hidden Markov Random Field (HMRF) model. This model can extract similarity of corrosion environment among multiple pipeline segments in both physical and feature space. In order to use HMRF model, we need to select relevant features (i.e., measured corrosion environment data)

which contributes to the inter cluster separability and identify the ideal number of clusters to be made.

Feature Selection

Clustering analysis is the process of grouping data points based on some similarity in properties or features. For supervised learning feature selection algorithms, this can be achieved by maximizing some target function of predictive accuracy, because we are given class labels and we want to keep only the features that are related to or lead to good prediction performance. But in unsupervised learning we are not given class labels. In such case we will have to perform feature selection because:

- Finding the ones that contribute to the prediction since some features may be redundant, some may be irrelevant, and some can misguide clustering results.
- Curse of dimensionality: If there are N features we need at least 2^N data points [1,2].

Therefore choosing a subset of features leads to better performance.

Feature selection are of three types:

1. Filter Model: Evaluates the relevance of a feature by studying its characteristics using certain statistical criteria.
2. Wrapper Model: Utilizes quality of clustering as a selection criteria.
 - a. Computationally expensive and biased on method used.
 - b. Better accuracy
3. Hybrid Model: Bridges gap between filter and wrapper models. Uses statistical criteria to select candidate features and then chooses the subset with the highest classification accuracy.

In this project we have three main categories of features:

1. Physical soil properties like bulk density of soil, percentage sand content, clay content, coarse fragment and silt content.
2. Large scale environmental features like elevation, yearly mean precipitation and NDVI,
3. Soil electrochemical properties from site samples like pH, ion concentration, half-cell potential and resistivity.

Among these features soil electrochemical features are onsite data and are the most wisely studied features and they are directly related to corrosivity. Hence the aim is to select relevant features from the other two subsets that will add to the predictive analysis. For this project feature selection we first start with finding the correlation across different features, correlation between the potential features and the maximum depth of corrosion defects within each pipeline segment and the number of corrosion defects found in each pipeline segment. Further we check the heterogeneity of the features by performing Kernel density fitting on each feature. Then, we apply and compare other unsupervised feature selection methods to determine the most important features.

Correlation Analysis

Features with high correlation are more linearly dependent and hence have almost the same effect on the dependent variable. So, when two features have high correlation, we can drop one of the two features. Figure 12 gives the correlation matrix of each feature with each other, Figure 13 shows the sum of correlation values of each feature, Figure 14 and 15 shows the correlation of each of the features with maximum corrosion depth and maximum number of corrosion defects in a 100 m segment.

As seen in Figure 12 Bulk density of the soil has good correlation between pH and Elevation which means information of bulk density is contained in these features and hence can be removed. Among pH and Elevation also has a correlation of 0.68. Elevation also has the maximum total correlation value as shown in Figure 13. Hence among these three features pH can be selected as its has been shown in number of literature that pH variation is a major indicator of corrosivity. Similarly, clay content also has correlation with sand and silt content and among these two according to figure 13 sand content has higher correlation value. Therefore, in physical soil properties silt content can give more information for the clustering analysis.

In the case of large-scale environmental features elevation has the highest correlation with all features and can be eliminated. Among electrochemical properties all features have correlation below 0.5.

Result: Physical Soil Properties: Silt content

Large Scale Environmental: Yearly Mean Precipitation and Avg NDVI

Electrochemical Properties: Eh, Resistivity, pH, CO₃, HCO₃, Cl, SO₄

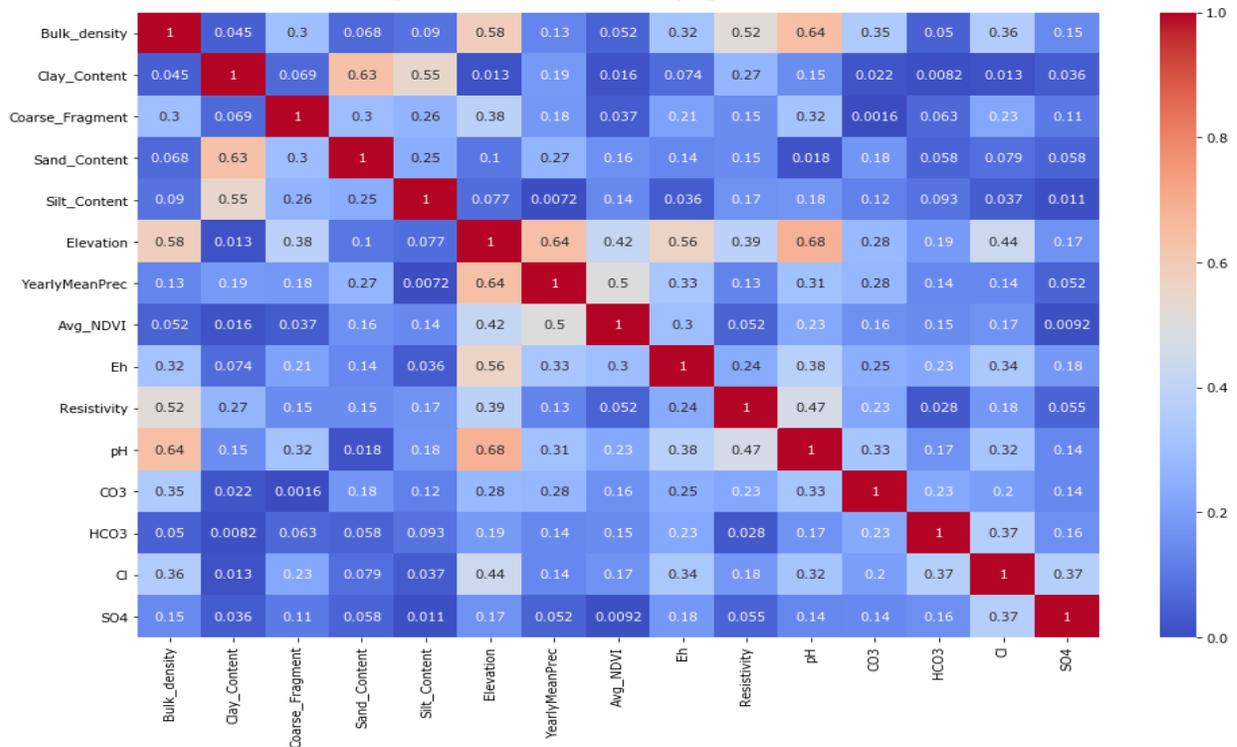


Figure 12: Feature Correlation Matrix

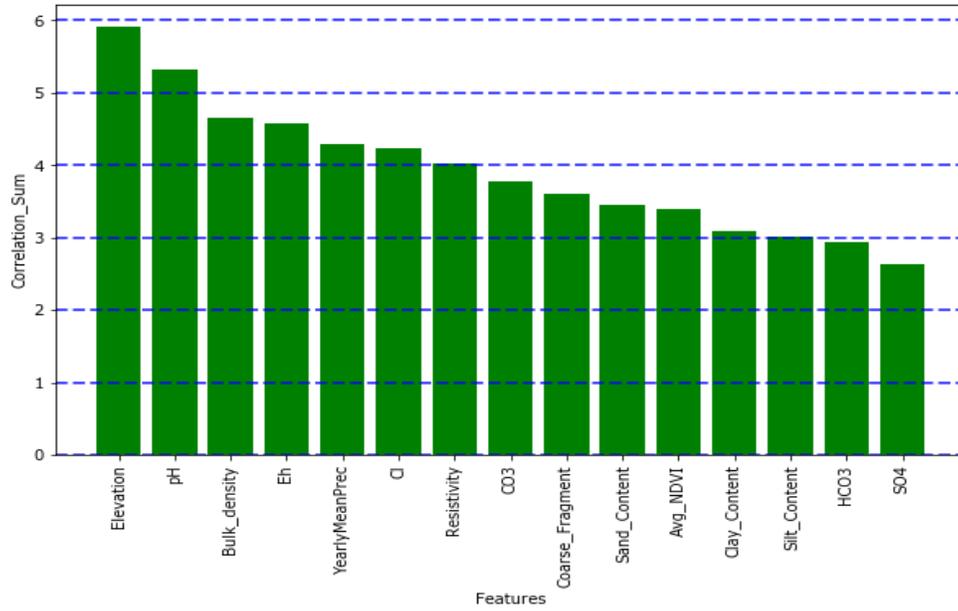


Figure 13: Order of correlation sum

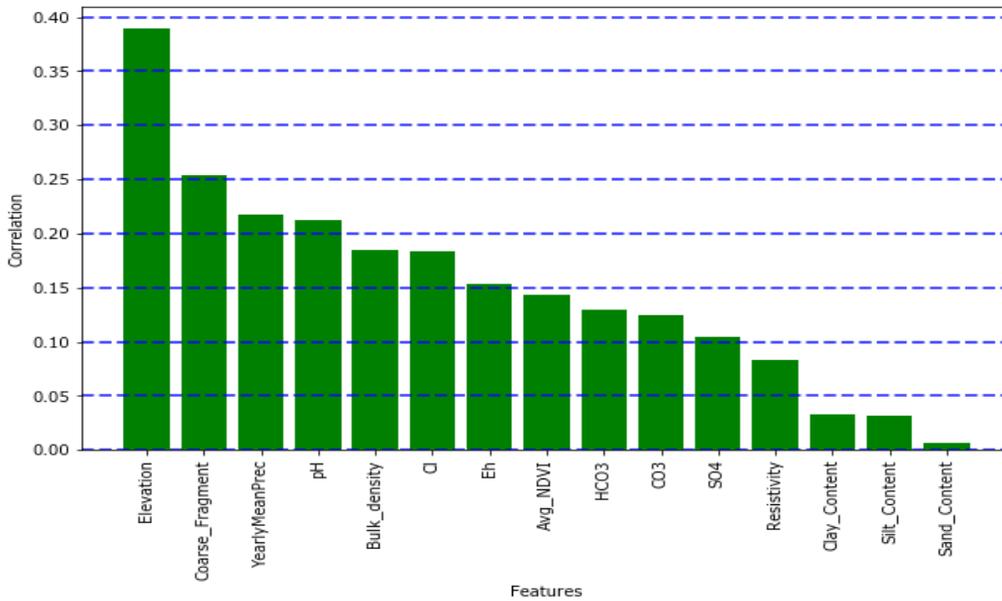


Figure 14: Correlation with maximum corrosion depth

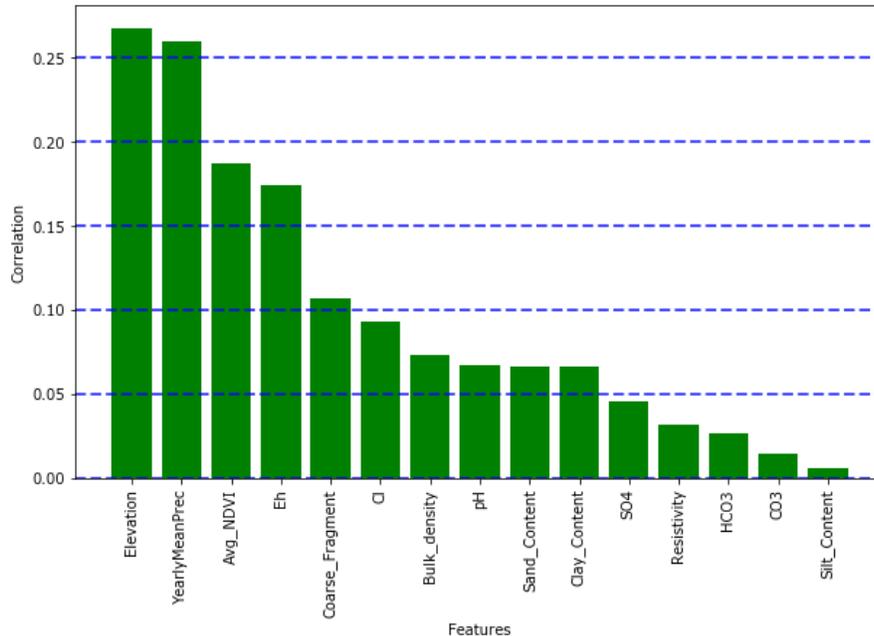


Figure 15: Correlation with number of corrosive spots.

Probability Density Analysis

Probability density function(PDF) provides a relative likelihood of the observed feature in the sample space. The methods for estimating probability density functions can be categorized into parametric and nonparametric approaches. The parametric methods assume a functional form for the density, such as Gaussian and hence only the distribution parameters need to be estimated. Non-parametric methods do not assume any knowledge about the density of the data in priori and computes the density directly from the instances and because of this reason they are in more of interest. Non-parametric density estimation is given as:

$$p(x) \cong \frac{k}{N*V} \quad (1)$$

where $p(x)$ is the estimated probability density distribution function for feature x , V volume surrounding x , N is the total number of instances and k is the number of instances inside V . There are two approaches for non-parametric density estimation, one fixing value of k and find corresponding volume V called k Nearest Neighbor (kNN) Method and other volume V is fixed and k is determined called Kernel Density Estimation(KDE). kNN PDF estimation methods are prone to local noise and since its integral over all the instance space diverges the probability density is less accurate. Therefore KDE with Gaussian kernel was used. The KDE plots for each feature is shown in Figures 16,17,18. These plots are a visualization of the features in the given space. Single peak indicates an overall presence and more the number of modes more heterogeneity and hence multimode features can contribute distinct information to the predictive analysis.

Result: *Physical Soil Properties:* Coarse Fragment Content
Large Scale Environmental: Elevation and yearly mean precipitation
Electro Chemical Properties: CO₃, Eh, pH

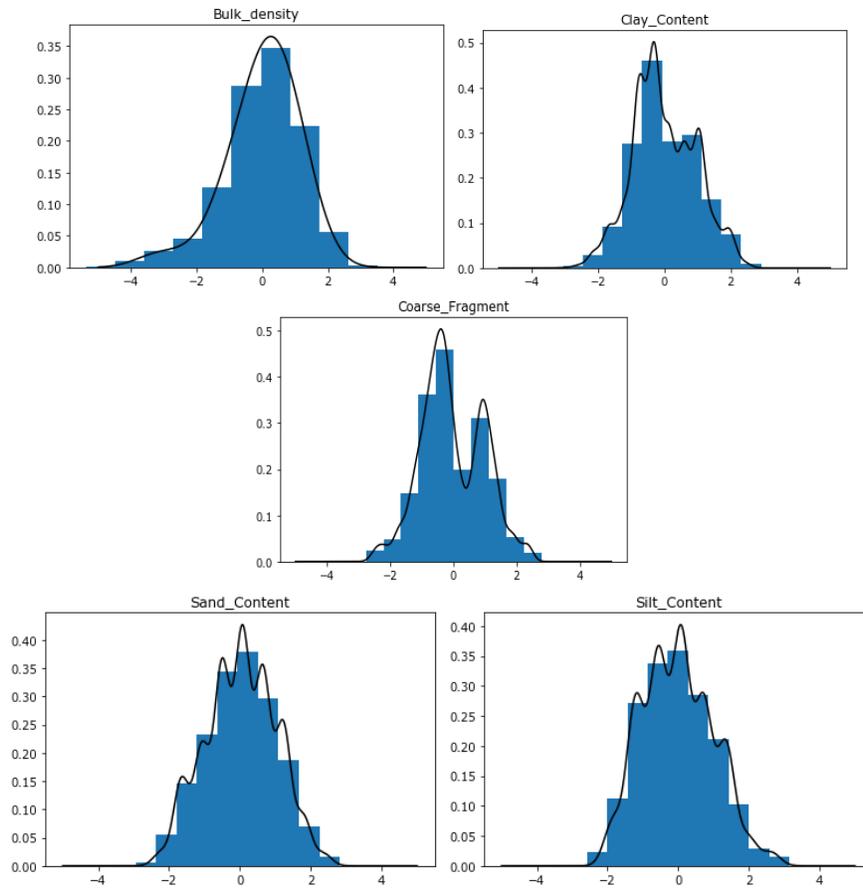


Figure 16: KDE for physical soil properties

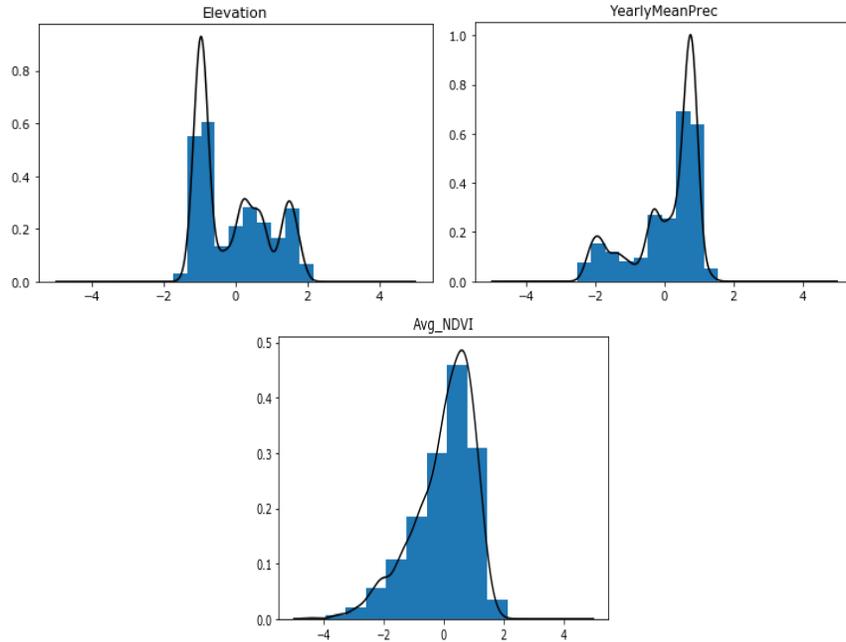
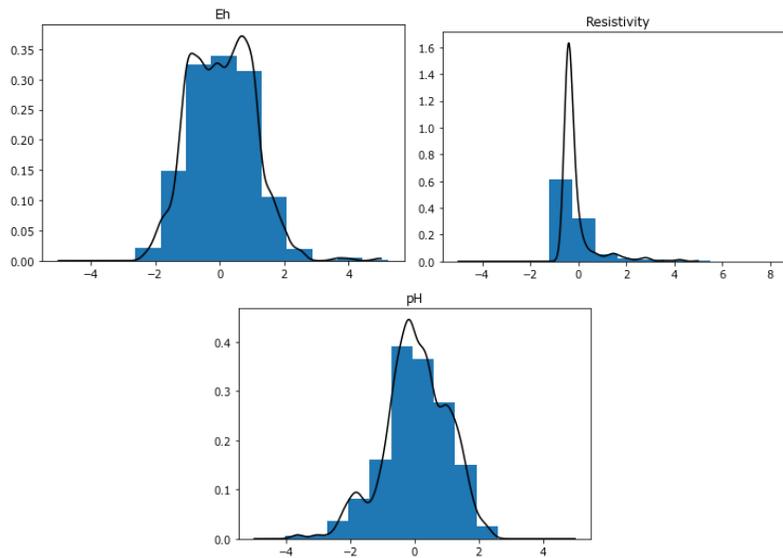


Figure 17: KDE for Large scale environmental features



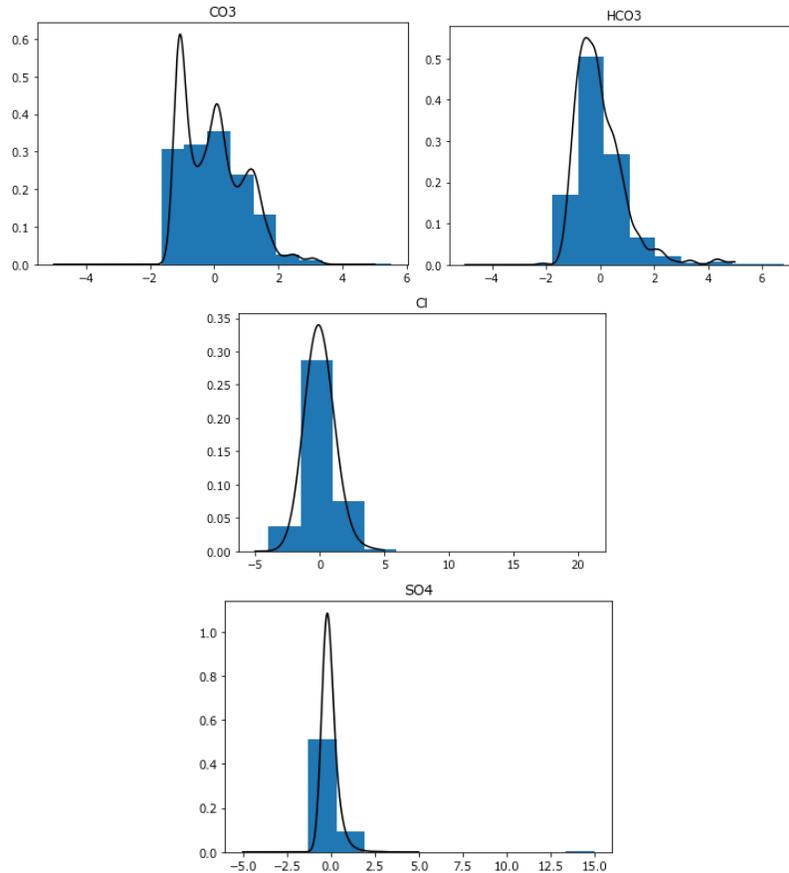


Figure 18: KDE for Soil Electrochemical features

Unsupervised Feature Selection

Spectral Feature Selection (SPEC): SPEC estimates the feature relevance by estimating feature consistency with the spectrum of a matrix derived from a similarity matrix S using Radial-Bases Function (RBF) as similarity function. Based on S , a graph structure is constructed and uses Spectral Graph Theory to find features with the best separability in assumption that points are separated to the predefined number of clusters [2]. SPEC algorithm selects k features from the given set having highest scores. Figure 19 shows the SPEC scores.

Result: *Physical Soil Properties:* Bulk density ,Coarse Fragment
Large Scale Environmental: Yearly Mean Precipitation
Electro Chemical Properties: Eh,Resistivity,CO3

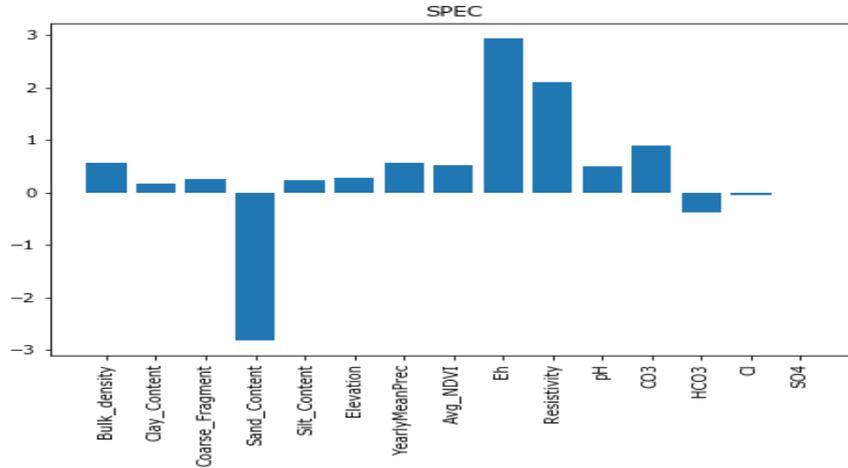


Figure 19: SPEC scores

Laplacian Score: is a special case of SPEC based on the observation that data from the same class is often close to each other and thus we can evaluate the importance of a feature by its power of locality preserving [3]. A Laplacian score is then calculated for each feature and will have the property that **smallest values** correspond to the most important dimensions. Figure 20 shows the Laplacian scores for each feature.

Result: *Physical Soil Properties:* Bulk density

Large Scale Environmental: Elevation

Electro Chemical Properties: SO4, Resistivity

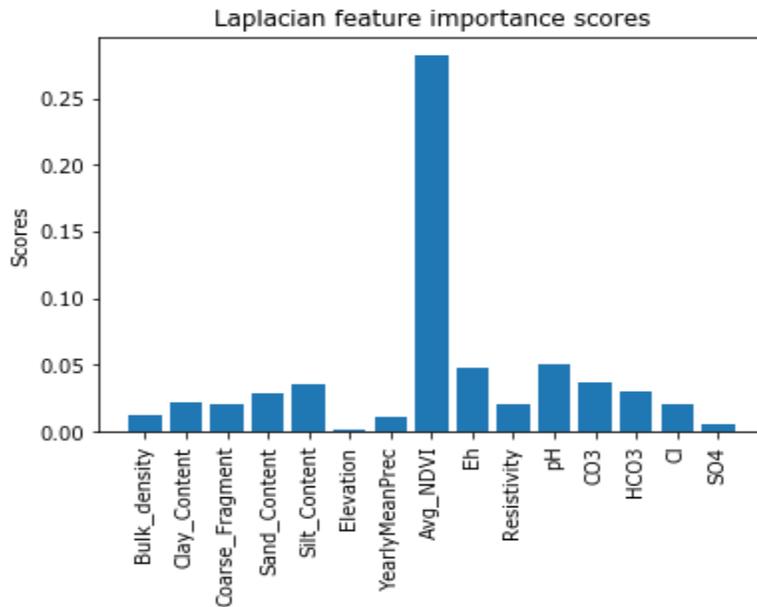


Figure 20: Laplacian scores

Multi-Cluster Feature Selection: Selects the set of features that can cover all the possible clustering in the data. In MCFS a spectral analysis is performed to measure the correlation between different features. The top eigenvectors of the graph Laplacian are

used to cluster the data and a feature score is being computed. The algorithm returns top k features from the set of features passed. Scores calculated are shown in Figure 21.

Result: *Physical Soil Properties:* Bulk density
Large Scale Environmental: Elevation
Electro Chemical Properties: Resistivity,pH

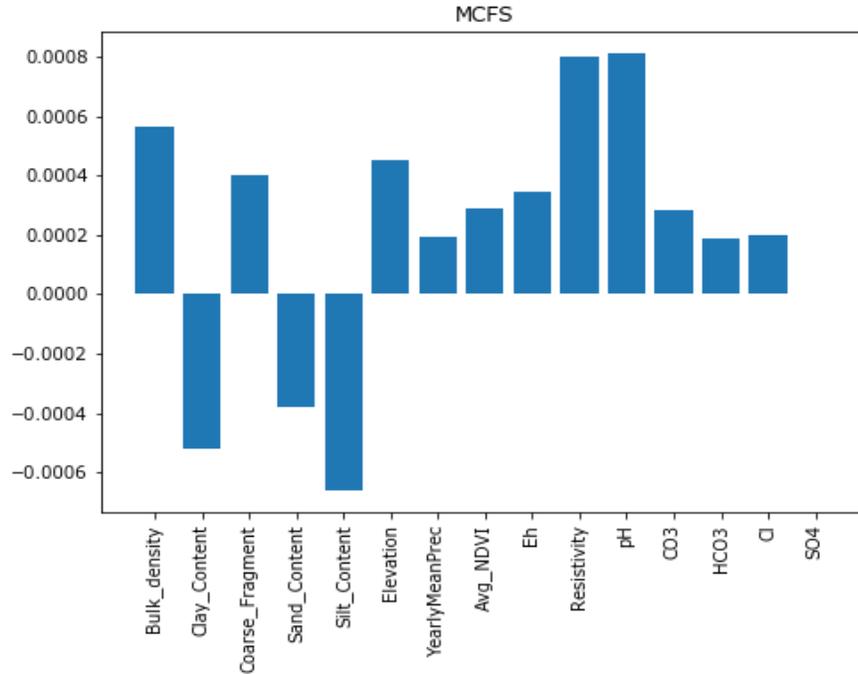


Figure 21: MCFS score plots

Summary:

On comparing all the above methods bulk density and coarse fragment content can be chosen soil physical features, elevation and yearly mean precipitation from large scale environmental features and keeping all the electrochemical properties we have a total of 11 features. Among the ion concentrations the least preferred from the scores is HCO3.

Table 5: Features selected using different methods

Feature	Correlation Matrix	KDE	SPEC	Laplacian	MCFS
Bulk density			Y	Y	Y
Clay Content					
Coarse Fragment		Y	Y		
Sand Content					
Silt Content	Y				
Elevation		Y		Y	Y
Yearly Mean Precipitation	Y	Y	Y		
Avg. NDVI	Y				
Eh	Y	Y	Y		
Resistivity	Y		Y	y	Y

pH	Y	Y			Y
CO ₃	Y	Y	Y		
HCO ₃	Y				
Cl	Y				
SO ₄	Y			Y	

Number of clusters

An essential input parameter for clustering is the number of clusters that best fits a given dataset. Number of measures have been well developed for this problem in literature. In general, they can be categorized into three types: external criteria, internal criteria and relative criteria. An external criterion evaluates the result of clustering based on a pre-specified structure [5]. Meanwhile an internal criterion is based on quantities that involve the vectors of the data set themselves. The basis of external and internal criteria is statistical testing and is complex. The relative criterion is more often used.

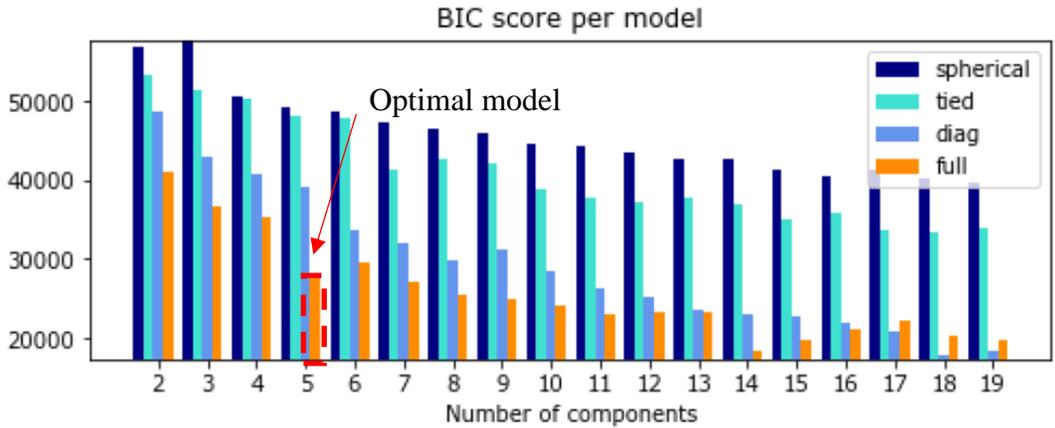
Since each clustering use different properties there are different types of measures which has been proposed. For model based clustering the “Approximate Weight of Evidence Criterion (AWE)” is used to determine the number of clusters. When EM is used to find the maximum mixture likelihood an approximation to AWE called Bayesian Information Criterion (BIC) is applicable.

$$BIC = L(\theta) - \frac{1}{2}m * \log(n)$$

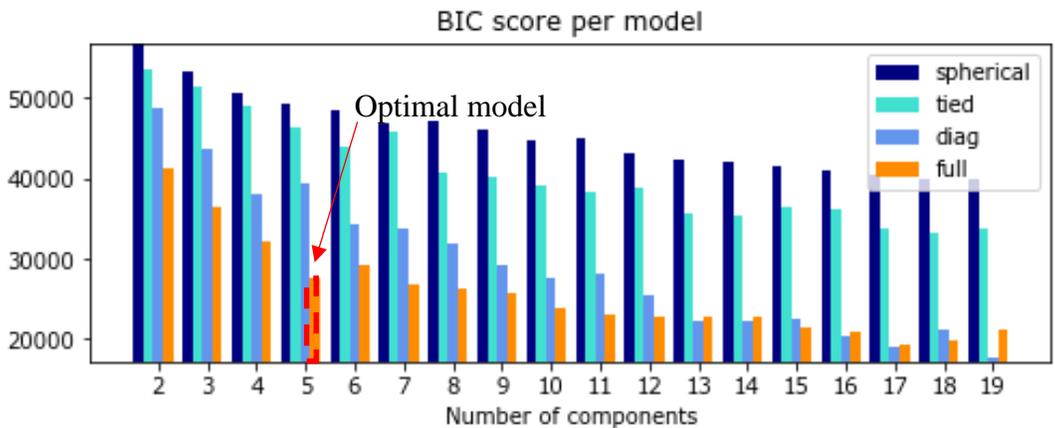
(2)

where, $L(\theta)$ loglikelihood function, m=no. of free parameters to be estimated , n= number of observations

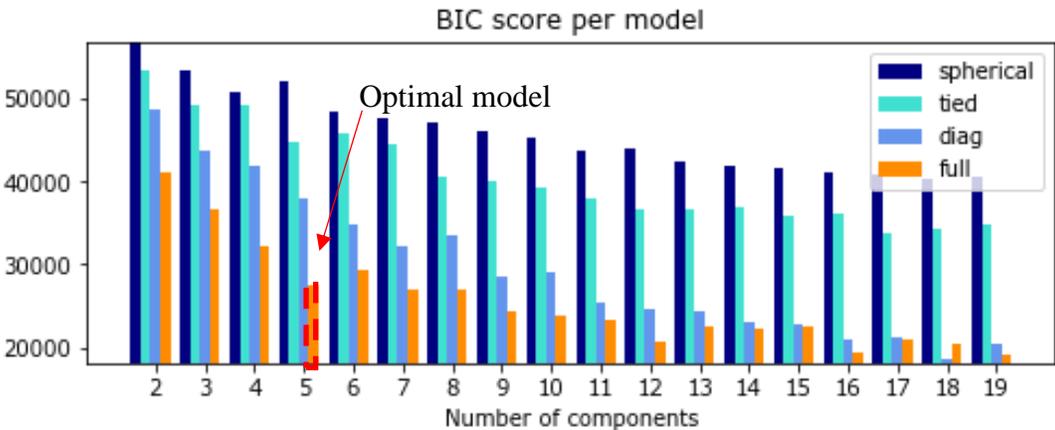
A model with a higher BIC is a better model, since if data fits well to the model, the log likelihood would be higher and m should be a minimum. Based on the degree of freedom in the shape of the cluster there can be four covariance types. Spherical corresponds to equal volume, no orientation for the cluster shape. Diagonal which means that the size of the cluster along each dimension can be set independently, with the resulting ellipse constrained to align with the axes. Tied means the clusters have same shape but the shape can be anything. Full means the components can have any shape or orientation independently(unconstrained). With the selected 11 features the number of clusters were calculated with assuming k mixed multivariate Gaussian distribution. k varied from k=2 to k=20 and the result is as shown in figure 22 with different dataset. The optimal number of clusters is considered at the number where the minimum of the BIC is achieved. In both datasets k=5 components for full (unconstrained) model shows the knee point which means 5 clusters will be preferable for the dataset.



(a) Result of test set 1



(b) Result of test set 2



(c) Result of test set 3

Figure 22: BIC score for different test sets.

References:

1. T. Hastie, R. Tibshirani, and J. Friedman. The Elements of Statistical Learning, Springer, 2001.

2. Alelyani, Salem, Jiliang Tang, and Huan Liu. "Feature Selection for Clustering: A Review." *Data Clustering: Algorithms and Applications* 29 (2013): 110-121.
3. M. Belkin and P. Niyogi, "Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering," *Advances in Neural Information Processing Systems*, Vol. 14, 2001.
4. Cai, Deng, Chiyuan Zhang, and Xiaofei He. "Unsupervised feature selection for multi-cluster data". *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2010.
5. Q. Zhao, V. Hautamaki and P. Fränti, "Knee Point Detection in BIC for Detecting the Number of Clusters", *ACIVS*, 2008.

5: Project Schedule –

The project is on-schedule as originally-proposed.

During the following quarter, the team will perform and cover the experimental laboratory testing planned. During the analysis of experimental laboratory data we will consider gaps in prior knowledge relating coating conditions and corrosion severity under controlled environmental factors. We also will define parameters in the field that has not been considered for severity. We also are using a methodology that considers a new interpretation for CIS and DCVG technologies.

Fieldwork will involve pipelines whose RoW reflects conditions of different soil scenarios, and a host of topographic conditions will be included, to cover the range of typical US conditions. Trends in the outcomes will be examined and/or deterministic or semi-empirical models or expressions will be developed to quantify the damage evolution in the pipeline/soil system.

The activity for mapping available data via GIS tools and geographically co-register all datasets will be continue to have the task completed.

In the following quarter, we will perform model selection in order to determine the number of clusters regarding the heterogeneity of the soil environment. The Bayesian information criteria will be adopted as a quantitative measure of the preference as it systematically takes both model complexity and fitting error into consideration and hence possible overfitting can be avoided.

Another work will be finished in the following month should be the detailed clustering analysis as scheduled in the original proposed tasks. The optimal number of clusters will be used and the spatial distribution of different clusters (i.e., soil corrosion environment) will be visualized and discussed.

6. Publication

On May 20th 2020, a conference abstract was submitted to the NACE 2021 Conference to be held in Salt Lake City Utah USA. Currently, the abstract is under review.