Estimating Corrosion Growth Rate for Underground Pipeline: A Machine Learning Based Approach

Joseph Mazzella
Engineering Director, Inc.
807 Davis Street, Unit 314
Evanston, IL 60201
USA

Len Krissa
Enbridge Employee Services Canada Inc.
10175-101 Street
Edmonton, Alberta, T5J 0H3
Canada

Thomas Hayden
Northwestern University
803 Hinman Ave
Evanston, IL 60202
USA

Haralampos Tsaprailis
Enbridge Employee Services Canada Inc.
10175-101 Street
Edmonton, Alberta, T5J 0H3
Canada

ABSTRACT

Estimating corrosion growth rate for underground pipelines is a non-linear, multivariate problem. There are many potential confounding variables such as soil parameters, cathodic protection, AC/DC interference, seasonal / climate conditions, and proximity to unique geographic features such as wetlands or polluted environments. The work presented provides an approach for estimating underground corrosion growth rates using a dataset of observations from a North American pipeline operator. Extensive geospatial data is utilized that has been obtained from public and private sources and extrapolated using Inverse distance weighted (IDW) interpolation. This work presents a model using IDW to estimate parameters involving soil, interference, geography, and climate factors for any location in North America.

Using this data, this work then presents several different machine learning approaches, including Generalized Linear Models, eXtreme Boosted Trees, and Neural Networks. All three provide an accurate estimation for corrosion growth rates for an underground asset at any latitude and longitude pair in North America. Each method comes with potential benefits and pitfalls, specifically; trade-offs between model accuracy and transparency. This work presents a framework for comparing geo-spatial and machine learning estimates.

Key words: Underground Pipeline, Machine Learning, Neural Networks, IDW, AC Interference, Soil, Geology

INTRODUCTION

The business case for this paper involves cost-effectively and efficiently assessing environmental conditions and the related impact of corrosion on underground pipeline using geographical information systems (GIS) and spatial data, with limited excavation. The objective is to proactively target those areas that have the highest likelihood of advanced corrosion (based on rate and degree of corrosion) and thus reduce risk of failure, while maximizing both capacity and related cost of inspection.

Geostatistical Analyst tools are used to emulate a phenomenon occurring in the landscape that is of interest such as pH and electrical conductivity in the soil, powerlines that contribute to alternating current (AC) interference of rectifiers, magnetic anomaly, road salts, and other known contributors to corrosion of underground pipeline.

By using the geospatial tools to generate data for inputs into machine learning, this paper proposes a tool which estimates corrosion growth rates from a large range of environmental variables. This isn't an uncommon approach, estimating corrosion growth rates using machine learning is an active area of research for at least two decades now. This approach is similar to work done by others but by including a wider range of spatial variables, the models are specifically designed for high dimensionality geo-spatial input, representing the wide array of environmental risk factors for an underground pipeline.[1 2 3]

DATA TRANSFORMATIONS

Data is collected from public and proprietary sources detailed below. Where relevant, any transformations or changes are noted. There are three main transformations made to the data, based on the type:

- Categorical Data - Unlike data that has a continuous or discrete range, categorical data is based on text-based attributes of data and retains no ordering. One of the most common examples is pipeline coating manufacturer. The coating is a strongly correlated independent variable but it is not in a form mappable to an integer or continuous value.
- One-Hot Encoding - One common method for dealing with categorical variables is to map the values to a table of values, where the value is 1 if the value is present and 0 otherwise. In the case of coating manufacturer, it may result in a matrix of a dozen columns (one for each manufacturer). This is a very common technique for mapping categorical data to an ordinal mathematical value.
- Binarization - Another common method used in feature engineering is to take a continuous variable and "binarize" it by converting it into a (0, 1) value. Usually this is done by looking at some threshold and creating a feature where the value is less than or greater than some threshold.

TRAINING DATA – HISTORIC CORROSION GROWTH RATES

The primary dependent variable comes from prior work by Ping et al.[4] This study utilizes In-line Inspection (ILI) back-to-back measurements collecting using Magnetic Flux Leakage measurements (MFL). The data, from a North American operator, includes measurements in the United States and Canada at the resolution of each GirthWeldAddress (GWA) along the pipeline. Below are the overall counts along with the number of available CGR measurements (Mean ILI Back2Back) per country.

This measurement (Mean ILI Back2Back) forms the basis of all the estimations in this work as our primary dependent variable.

Table 1: Quantities of Data with and without ILI Measurements

Country	GirthWeldAddress w/ILI Measurements	Total GirthWeldAddresses
United States	139,024	882,034
Canada	201,950	976,340
Total	340,974	1,858,137

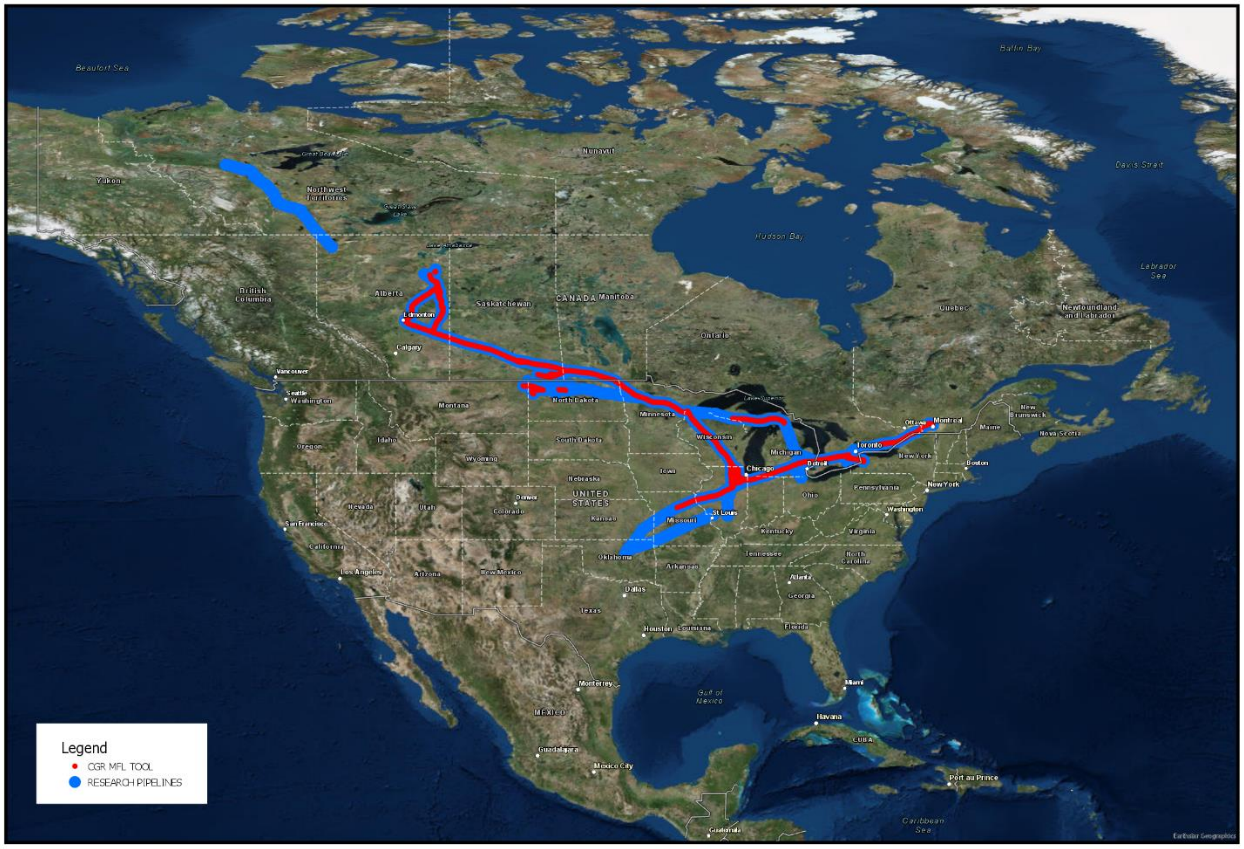

Figure 1: North American Pipelines for Model Training and Evaluation. Existing ILI CGR available in Red

DATASETS USED FOR MODELING

To train the Machine Learning models, a large array of different dependent variables including atmospheric conditions, human activity, and alternating current (AC) interference are included.

Atmospheric Conditions

From prior work on ISO9223[1] long term climatological and atmospheric conditions were included. In particular, the following variables: [5 6]

- Time of Wetness (TOW) - Time of Wetness is defined as the number of hours per year where the temperature is above freezing and humidity is greater than or equal to 80 percent.
- Mean Average Temperature - The mean average temperature (averaged hourly) over the course of a year.
- Total Number of Days Below Freezing - Number of days observed where at least one hour of the day was below zero degrees Celsius.

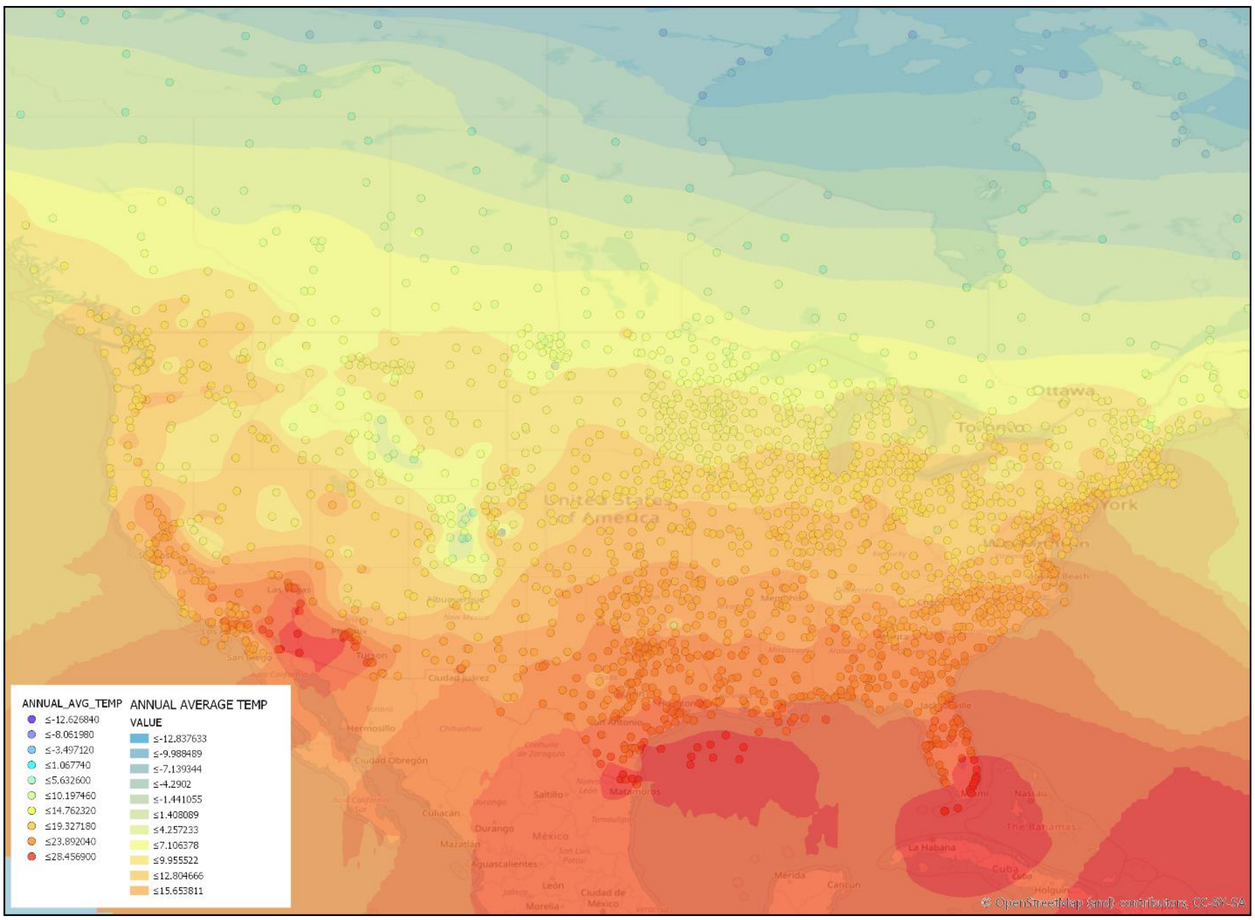

Figure 2: North American Average Mean Temperature

IDW was used to extrapolate 3,400 stations reporting hourly in the United States to annual averages. In addition, a similar process was done for pollutant data using approximately 300 stations from the EPA's NADP[(2)] program.[7]

[(1)] International Organization for Standardization. BIBC II Chemin de Blandonnet 8 CP 401 1214 Vernier, Geneva Switzerland. http://www.iso.org

[(2)] National Atmospheric Deposition Program. Wisconsin State Laboratory of Hygiene 465 Henry Mall University of Wisconsin Madison, WI 53706. http://nadp.slh.wisc.edu/nadp/contacts.aspx

- Mean Annual SO_2 Dry Deposition - Estimated annual SO_2 dry deposition amounts measured in mg/m^3
- Mean Annual Chloride Dry Deposition - Estimated annual Cl dry deposition measured in mg/m^3

AC Interference

There are decades of research on the subject of the role of alternating current's (AC) role in corrosion growth.[8] [9] Two sources of potential interference were included in the training dataset:

- Proximity to High Voltage Powerlines - For each GWA, the number and maximum voltage of nearby powerlines within a 300m and 100m radius at points greater than 300 maximum volts.
- Proximity to Power Substations - For each GWA, the number and maximum voltage of power substations within a 500m radius.

Proximity was computed between a given GWA and a particular power line, not a specific tower. For further discussion on this subject, please see the future work portion at the end of this paper.

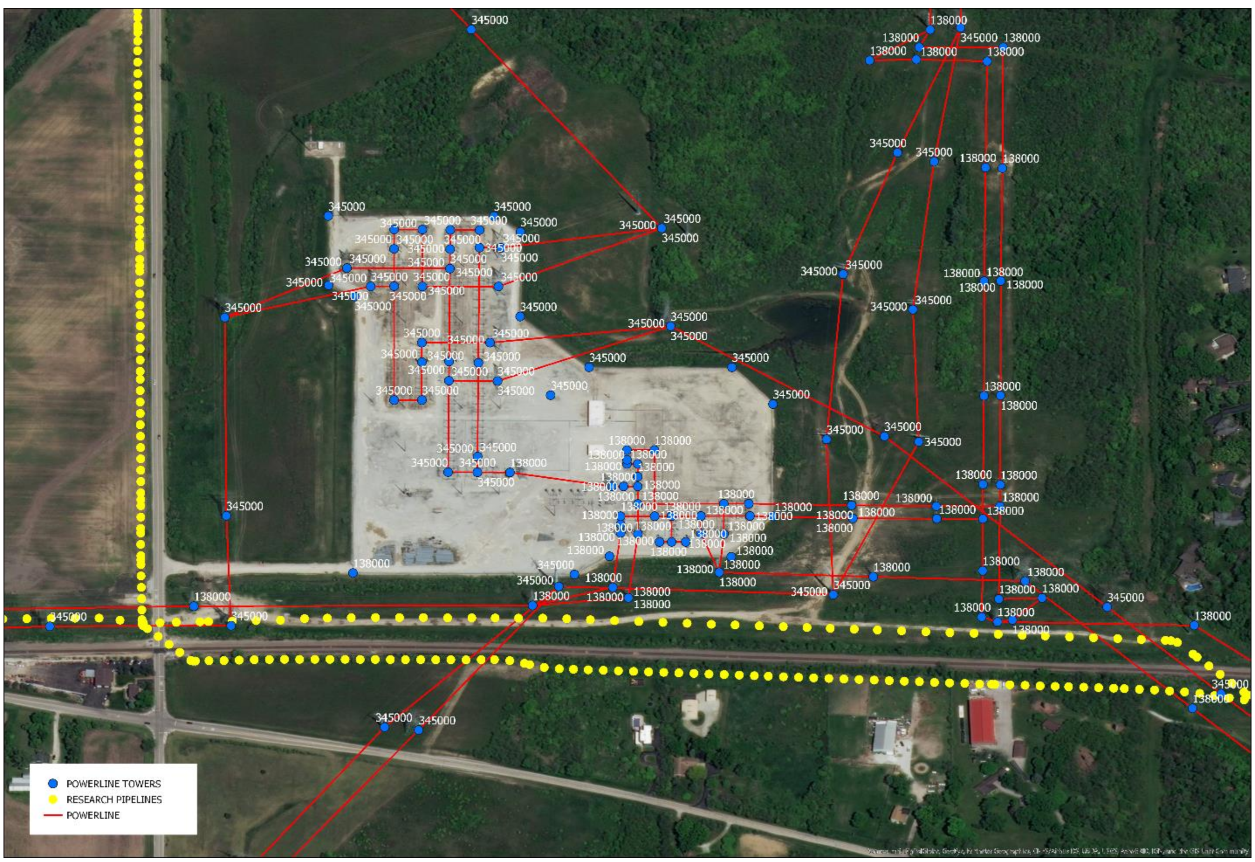

Figure 3: Example of AC Interference – Pipeline Intersecting with High Tension Power Lines and Substations

Road & Railways

Proximity to roadways[10] and railways[11] is included because in snowier parts of the country, road salt is a major potential contributor to corrosion. Therefore, roads with mean and maximum speeds above 40 miles-per-hour are included.

Bodies of Water

For identifying when a pipeline is near or intersecting with a body of water, shapefiles[9] are binarized by determining whether a given GWA is directly in contact with land or not. In the case of the pipeline being on land a dummy variable of 1 is assigned and 0 otherwise.

Table 2: Examples of Binarized Water Variable

Point Location	Type	Binarized Water Variable
53.5248, -113.342693	Land	1
39.980854,-91.454796	Mississippi River	0
41.649921,-88.066504	Land	1
41.648017,-88.065707	Des Plaines River	0

Soil Database

The most practical dataset for working with underground assets comes from the ISRIC-WISE[(3)] soil database.[13] [14] It is an extensive shapefile-based database focused on a variety of soil properties. Initially, the entire dataset was fed into the machine learning training algorithm but, many of the variables are correlated and often derived from each other. Therefore, the following variables are retained:

Table 3: Soil Parameters from ISRIC Data Used in Models

Soil Parameter	Units
Organic Carbon	g Carbon
Total Available Water Capacity	-33 to –1500 kPa
Soil pH	pH
Silt Mass %	% Percentage
Sand Mass %	% Percentage
Clay Mass %	% Percentage

[(3)] International Soil Reference and Information Center. Droevendaalsesteeg 3 6708 PB Wageningen, The Netherlands. https://www.isric.org

Magnetic Anomalies

The World Magnetic Anomaly Map (WDMAM) is a project to aggregate ground-based magnetic anomalies using primary satellite data. The data is published in a single banded raster file, representing the magnetic anomaly at a 2-degree arc level.

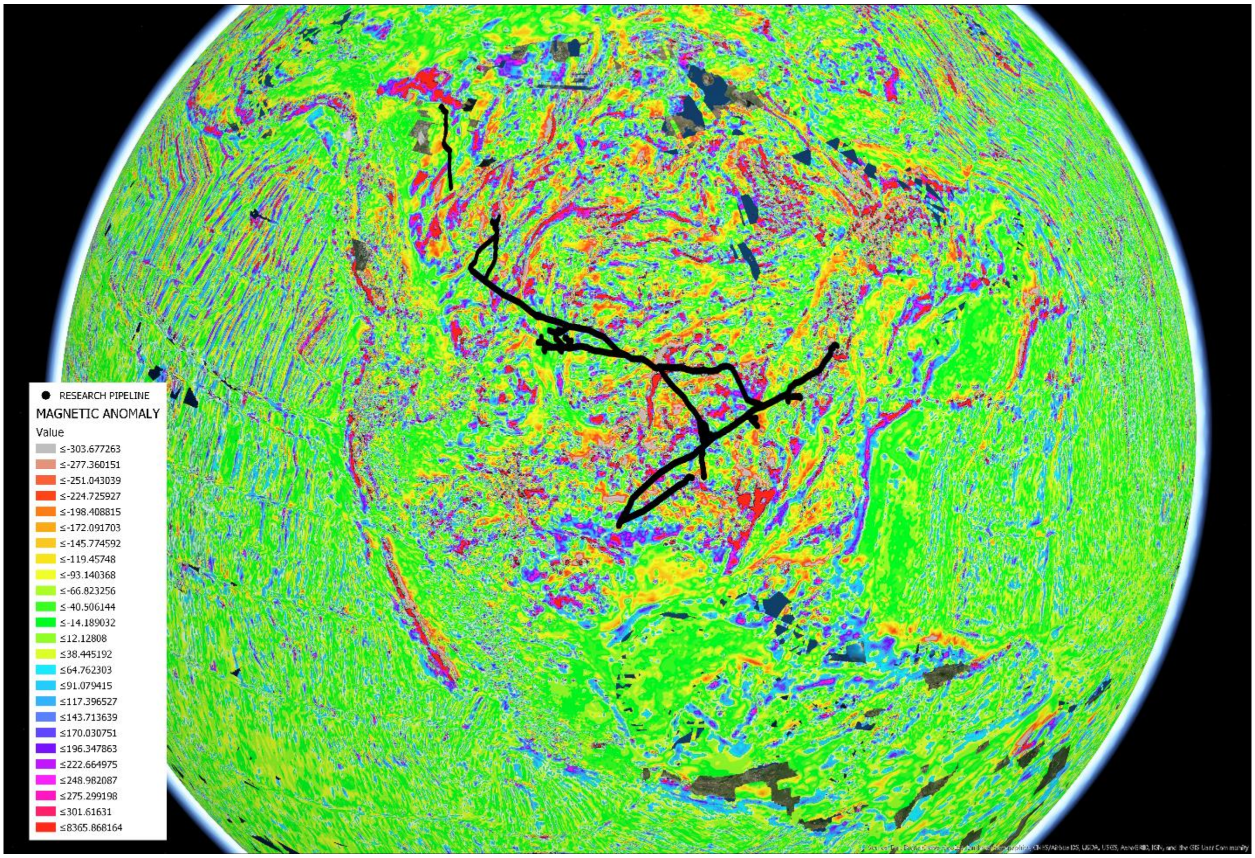

Figure 4: Global Magnetic Anomaly Map – showing North America, with pipeline visible in black

Table 4: Summary of All Variables

Variable ID	Type	Description
SO_2 per m2/d	Atmospheric	Sulfide Pollution
Cl per m2/d	Atmospheric	Chloride Pollution
Avg TOW	Atmospheric	Time of Wetness
ANNUAL_AVG_TEMP	Atmospheric	Annual Average Temperature
AVG_DAYS_BELOW_0	Atmospheric	No of Days per Year < 0 C
is_Water	Geography	Over Land (1) or Water (0)
SUBSTN_FREQ	AC Interference	Qty of Nearby Substations
ORGCSoilOrganic	Soil	Organic Carbon
PHAQSoilSoil	Soil	pH
STPCSoilSilt	Soil	Mass %
SDTOSoilSand	Soil	Mass %
CLPCSoilClay	Soil	Mass %
TAWCSoilTotal	Soil	Available Water Content
Is_ELCO_gt_2	Soil	Electrical Conductivity > 2
POWERL_FREQ	AC Interference	No of AC Powerlines within 300m
Is_MPH_MAX_gt_40	Human Activity	Roads within 100m with Max MPH > 40
Is_MPH_MEAN_gt_40	Human Activity	Roads within 100m with Avg MPH > 40
MPH_FREQUENCY	Human Activity	Number Roads within 100m with MPH > 40
RAIL_OPERATIONAL	Human Activity	Is Nearby Operational Railway
MAGANOM	Magnetism	Magnetic Anomaly Value
PIPELINE_COUNT	Human Activity	Number of Nearby Pipelines
Is_MAX_VOLTS_gt_300	AC Interference	Is Line within 300m, Voltage > 300v
Is_POWERL_LT_100	AC Interference	Is Powerline within 100m
Is_SUBSTN_NEAR_500	AC Interference	Is AC Substation within 500m
Is_POWERL_MAX_VOLTS_GT_100	AC Interference	Is nearby powerline > 100V

When included, these are the pipeline-based variables used:

Table 5: Pipeline Variables included

Variable ID	Type	Description
PipeManufacturerPipeline	Pipeline	One-Hot Encoded Name
YearofMillRun	Pipeline	Year of Mill Run
ActualOuterDiameter	Pipeline	mm Diameter of Pipeline

Data Exclusions

The overall data set includes 1.8 million GWA values. In this set, there are 5,798 GWA that are intersecting or contained within a body of water polygon.[13] These GWA are excluded from both training the machine learning algorithms and the process of inferring values for all records. The underlying conditions for these GWA differ significantly from the rest of the dataset. Specifically, the soil conditions which may impact a GWA 150cm underground may have little relevance to a GWA 150cm underwater. This applies to almost all variables in the model.

Inferring values from a model trained on one set of conditions to a point that exists in another operation condition increases risk of type one errors. Water-based assets should have their own statistical models trained with a completely different set of parameters, analysis and model fit.

MACHINE LEARNING

Three main approaches to estimating ILI Back-to-Back CGR were used. In the first approach, a log-linear regression with transparent feature mappings. In the second approach, a modern machine learning toolset called eXtreme Gradient Boosting (xgboost) and in the third approach, an artificial neural net was training on the same data.[16]

Model Evaluation Methodology

For all models, the following methodology was used to evaluate performance. First, the data was split using a standard 90/10 training/validation split evenly with respect to the CGR back-to-back values. This ensures an even distribution of true values to train and evaluate on. The training set is 340,974 GirthWeldAddresses with ILI values, and thus the validation set will be around 34,000 GirthWeldAddresses. Then the following evaluation metrics are used:

- Root Mean Squared Error (RMSE) is defined as the square root of the average squared error. Intuitively, RMSE is a measurement of the error of a predictor. RMSE, along with Mean Absolute Error (MAE) are the most commonly used metrics for evaluating model performance.
- Correlation between Predicted Values and True Values. While only a linear predictor, the correlation between predicted and true value gives an approximate estimate of how closely aligned the values are. Ideally, a correlation of 1.0 would indicate that a model and the true values are perfectly aligned while a value of 0.0 would indicate the model has no power.

The reader is referred the literature for a more comprehensive coverage on measuring error in machine learning.[17] [18]

Log Linear Models

The log-linear model optimizes to solve the following mathematical equation:

$$\ln(Y) = X_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_N X_N$$

Where Y is the dependent variable, X_i is an independent variable and β_i is the fitted predictor weight. In the model, the dependent variable Y is back-to-back ILI CGR. A log-linear model is chosen because it is common in corrosion literature[1] and outperforms traditional regression in all model performance measures; R^2 and RMSE. The intercept value X_0 is omitted from the log-linear models, thus giving the following simplified form. This is a common practice, when the generating process (in this case corrosion) has a true intercept at zero.

$$\ln(Y) = \beta_1 X_1 + \beta_2 X_2 + ... + \beta_N X_N$$

eXtreme Gradient Boosting

eXtreme Gradient Boosting (xgboost)[14] is enjoying a renaissance in the machine learning space at the moment, winning competitions.[19] [20] In general, the algorithm is a continuation of prior work in the machine learning field around Gradient Boosting algorithms and tree-based models for computation. The libraries for xgboost are open-source and freely available on GitHub. Unlike the log-linear approach, xgboost is a black-box algorithm. This implies there is little visibility into understanding the particular logic the algorithm is following because the model.

Neural Networks

Artificial Neural networks (ANN), also commonly known as deep learning or artificial intelligence (AI) are the hot computer science research at the moment. Neural networks excel in problem settings involving classification, particularly at tasks where humans do well. They're commonly used in many automation tasks just classifying images, identifying content of images, or working through complicated high dimensionality inputs such as sound or video files.[21] Fundamentally, ANNs can solve any type of problem that has a fixed input and a learnable out.

Like the xgboost algorithm, ANNs do not provide robust visibility or transparency into the algorithmic process. There has been some research in building models for understanding neural network decision making but currently this research is not yet ready for deployment in production environments and is an active area of theoretical computer science research.[22]

In the ANN for this paper, the model is a three layer artificial neural net, excluding the input and output layers. A tanh activation function is applied to the first two layers and a linear activation on the final embeddings layer. To prevent overfitting, a single dropout layer with p = 0.5 is used. A variety of network structures were tested and this network architecture had the best performance and conforms with prior ANN research in this field.[1]

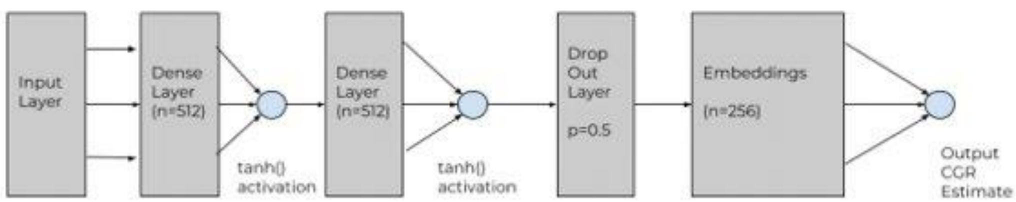

Figure 5: Three Layer Neural Net Structure with tanh() Activation and 256 Node Embeddings Layer

RESULTS

A total of six different models were trained and evaluated, three for environmental conditions only and three for a complete dataset, including pipeline parameters. The ideal model has a low RMSE and a high correlation scores, implying the mean error is low and correlates with true values. The authors have included notes as their own interpretations.

In general, as more data was provided to the model, in the form of pipeline specifications, the model performed more reliably.

Table 6: Model Performance

Algorithm	Dataset	RMSE	Correlation	R^2	Notes
Log-linear	Environmental	0.259	0.422	0.1150	Poor performing model
xgBoost	Environmental	0.120	0.888		Performs Suitably
Neural Net	Environmental	0.273			
Log-linear	Environment + Pipeline	0.126	0.890	0.0959	
xgBoost	Environment + Pipeline	0.069	0.956		Best Model By Far
Neural Net	Environment + Pipeline	0.139			

R^2 is only available for linear regression and is not the most reliable accuracy measure. The authors have left in for the sake of completeness. For the Neural Net models, the Pearson correlation co-efficient was not available for reporting.

Inference

After training the model on the 340,947 GirthWeldAddresses, it can then be inferred over the entire dataset of approximately 1.8 million GirthWeldAddresses.

CONCLUSIONS AND FUTURE WORK

Below are areas where the models can be improved, potentially quite dramatically:

- Identify and correct NULL data - In some cases, input data is missing from various sources, but primarily from the World Soil Organization. The NULL values in input result in NULL output values from the models.
- Incorporate data relating to cathodic protection (CP). Modeling for CP is notoriously difficult, CP is often located in places of increased corrosion rates and building the proper causal model is challenging.
- Incorporate improvements in AC Interference – datasets are now available which include distance between specific transmission tower sites and GWA locations. Including this data, will allow the models to capture not just the general AC interference but also the risks posed by proximity to metal lattice towers
- More and higher resolution data sources, especially with respect to localized emissions. It’s likely the atmospheric data would have a greater impact on the model by capturing localized emissions more directly. For instance, if the pipeline intersected directly with a coal fired powerplant or a Polyvinyl Chloride facility, the models should reflect localized levels of corrosion.
- Cross Validation of Models - All models were trained using a fixed training and validation set. Moving to a cross validation model will permit a wider range of model optimization.

In addition, there remains a lot of additional work to be done on building a high performing neural network model. In particular, designing the network structure is a combinatorial problem that doesn’t have a straightforward answer, beyond trying out different combinations. A three-layer tanh-activation ANN appears to be the most used standard[19] but there is probably a more optimal network available using either more layers, different activation functions, or vastly higher dimensionality input. The ideal outcome is to move past linear or tree-based modeling methods towards a comprehensive neural net approach that scales to any corrosion modeling situation.

DISCLAIMER

Any information or data pertaining to Enbridge Employee Services Canada Inc., or its affiliates, contained in this paper was provided to the authors with the express permission of Enbridge Employee Services Canada Inc., or its affiliates. However, this paper is the work and opinion of the authors and is not to be interpreted as Enbridge Employee Services Canada Inc., or its affiliates', position or procedure regarding matters referred to in this paper. Enbridge Employee Services Canada Inc. and its affiliates and their respective employees, officers, director and agents shall not be liable for any claims for loss, damage or costs, of any kind whatsoever, arising from the errors, inaccuracies or incompleteness of the information and data contained in this paper or for any loss, damage or costs that may arise from the use or interpretation of this paper.

ACKNOWLEDGEMENTS

Jerry Dewitt Enbridge Energy Partners, Alfonso Garcia Enbridge Energy Partners, Kim Hauer Enbridge Pipelines, Rob Floria Lake Superior Contracting, Niels Batjes ISRIC, Bob Dobos NRCS-NSSC, Steve Peaslee NSSC, Sharon Waltman USDA-NRCS, Jason Nemecek NRCS

REFERENCES

1. Engelhardt, G. R., D. D. Macdonald, and M. Urquidi-Macdonald. "Development of fast algorithms for estimating stress corrosion crack growth rate." *Corrosion Science* 41.12 (1999): 2267-2302.

2. Caleyo, F., et al. "Probability distribution of pitting corrosion depth and rate in underground pipelines: A Monte Carlo study." *Corrosion Science* 51.9 (2009): 1925-1934.

3. Yi-kun Cai, Yu Zhao, Xiao-bing Ma, Kun Zhou, and Hao Wang. “Long-Term Prediction of Atmospheric Corrosion Loss in Various Field Environments.” CORROSION. Vol 74. No 6. (2018)

4. Yanping Li, Len Krissa, Mona Abdolrazaghi and Gordon Fredine. “Validation of Corrosion Growth Rate Models.” NACE. (2017)

5. ISO 9223 “Corrosion of metals and alloys Corrosivity of Atmospheres Classification, Determination and Estimation” (2012)

6. Mazzella et al. “A Method for Extrapolating ISO9223 Response Functions. NACE International Nashville, TN. (2018)

7. National Atmospheric Deposition Program (NRSP-3). NADP Program Office, Illinois State Water Survey, University of Illinois. (2017)

8. Gummow, Robert A., Robert G. Wakelin, and Sorin Marius Segall. “AC corrosion–A new challenge to pipeline integrity.” No. CONF-980316–. NACE International, Houston, TX (United States), (1998)

9. Wakelin, Robert G., Robert A. Gummow, and Sorin Marius Segall. "AC corrosion-case histories, test procedures, & mitigation." *CORROSION 98*. NACE International. (1998)

10. Oak Ridge National Laboratories. *CTA Railroad Network*. https://www-cta.ornl.gov/transnet/RailRoads.html
11. ESRI. *North America Detailed Streets* 2018
12. TomTom North America and ESRI. *US and Canada Water Polygons* (2013).

13. Hengl, T., Mendes de Jesus, J., Heuvelink, G. B.M., Ruiperez Gonzalez, M., Kilibarda, M. et al. (2017) “SoilGrids250m: global gridded soil information based on Machine Learning.” PLoS ONE 12(2): e0169748. doi:10.1371/journal.pone.0169748.

14. Hengl T, de Jesus JM, MacMillan RA, Batjes NH, Heuvelink GBM, et al. (2014) “SoilGrids1km — Global Soil Information Based on Automated Mapping.” PLoS ONE 9(8): e105992. doi:10.1371/journal.pone.0105992

15. Maus, S., Barckhausen, U., Berkenbosch, H., Bournas, N., Brozena, J., Childers, V., Dostaler, F., Fairhead, J. D., Finn, C., von Frese, R. R. B., Gaina, C., Golynsky, S., Kucks, R., Luhr, H., Milligan, P., Mogren, S., Müller, R. D., Olesen, O., Pilkington, M., Saltus, R., Schreckenberger, B., Thebault, E. and Caratori Tontini, F. “EMAG2: A 2-arc-minute resolution Earth Magnetic Anomaly Grid compiled from satellite, airborne and marine magnetic measurements.” Geochem. Geophys. Geosyst., doi:10.1029/2009GC002471. (2009)

16. Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." Proceedings of the 22nd ACM sigkdd international conference on knowledge discovery and data mining. (2016).

17. Chai, Tianfeng, and Roland R. Draxler. “Root mean square error (RMSE) or mean absolute error (MAE) Arguments against avoiding RMSE.” Literature Geoscientific model development 7.3 (2014): 1247-1250.

18. Reich, Yoram, and S. V. Barai. "Evaluating machine learning models for engineering problems." *Artificial Intelligence in Engineering* 13.3 (1999): 257-272.

19. Nielsen, Didrik. “Tree Boosting with XG Boost-Why Does XGBoost Win" Every" Machine Learning Competition?” MS thesis. NTNU, 2016.

20. Machine Learning Challenge Winning Solutions https://github.com/dmlc/xgboost/tree/master/demo learning-challenge-winning-solutions

21. LeCun, Yann, et al. "Gradient-based learning applied to document recognition." *Proceedings of the IEEE* 86.11 (1998): 2278-2324.

22. Zeiler, Matthew D., and Rob Fergus. "Visualizing and understanding convolutional networks." European conference on computer vision. Springer. (2014)