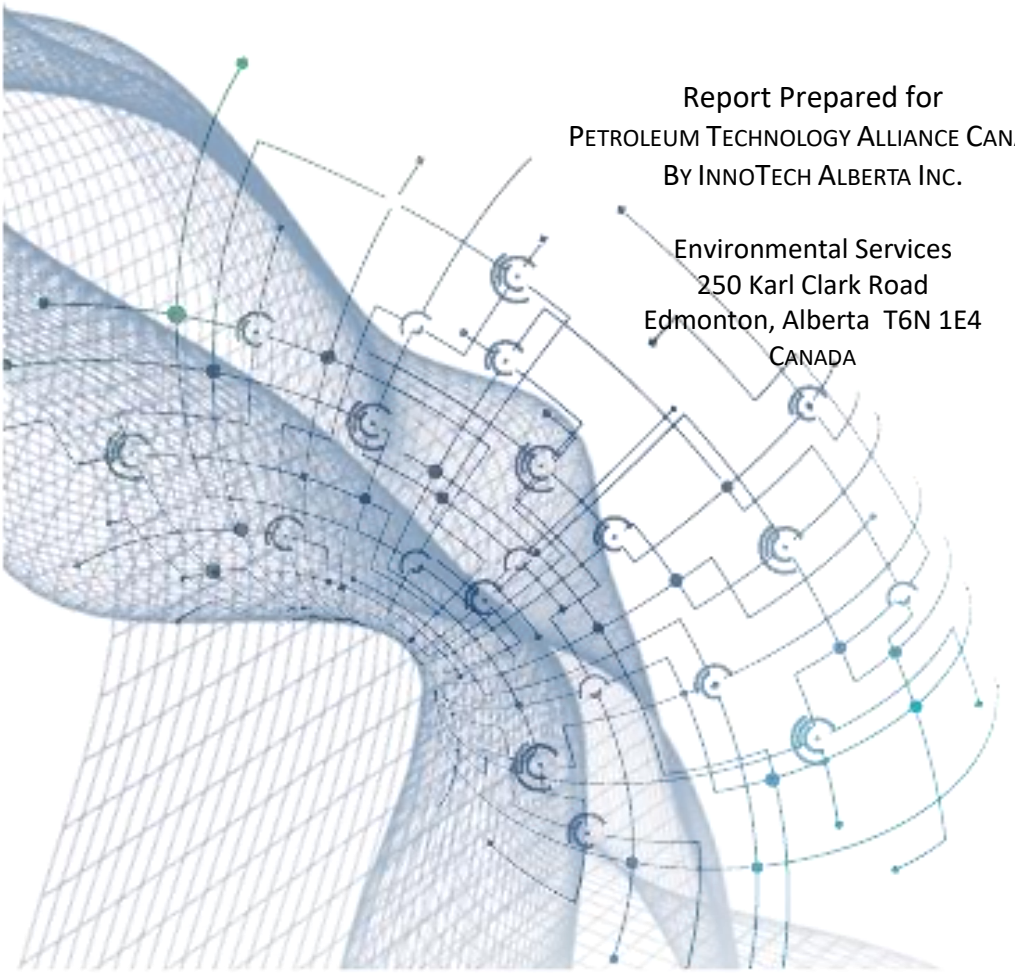
Annual Report on the Alberta Background Soil Quality System Project Phase 3

AUPRF 23-RRRC-02

Chibuike Chigbo, InnoTech Alberta Inc.

Paul Fuellbrandt, Statvis Analytics.

Preston Sorenson, University of Saskatchewan.



Report Prepared for
PETROLEUM TECHNOLOGY ALLIANCE CANADA
BY INNOTECH ALBERTA INC.

Environmental Services
250 Karl Clark Road
Edmonton, Alberta T6N 1E4
CANADA

CONFIDENTIAL

August 26, 2024

NOTICES OF REPORTS

1. This report was prepared as an account of work conducted by InnoTech Alberta Inc. (InnoTech Alberta), University of Saskatchewan, and Statvis Analytics Inc. (Statvis). All reasonable efforts were made to ensure that the work conforms to accepted scientific, engineering, and environmental practices, but InnoTech Alberta, University of Saskatchewan, and Statvis make no other representation and gives no other warranty with respect to the reliability, accuracy, validity, or fitness of the information, analysis, and conclusions contained in this report. Any and all implied or statutory warranties of merchantability or fitness for any purpose are expressly excluded. Reference herein to any specified commercial product, process, or service by tradename, trademark, manufacturer, or otherwise does not constitute or imply an endorsement or recommendation by InnoTech Alberta, University of Saskatchewan or Statvis.
2. Petroleum Technology Alliance Canada (PTAC) does not warrant or make any representations or claims as to the validity, accuracy, currency, timeliness, completeness, or otherwise of the information contained in this report, nor shall it be liable or responsible for any claim or damage, direct, indirect, special, consequential, or otherwise arising out of the interpretation, use, or reliance upon, authorized or unauthorized, of such information.

   The material and information in this report are being made available only under the conditions set out herein. PTAC reserves rights to the intellectual property presented in this report, which includes, but is not limited to, our copyrights, trademarks, and corporate logos. No material from this report may be copied, reproduced, republished, uploaded, posted, transmitted, or distributed in any way, unless otherwise indicated on this report, except for your own personal or internal company use.
3. The information contained in this report is confidential and proprietary to InnoTech Alberta, University of Saskatchewan and Statvis, and may not be distributed, referenced, or quoted without the prior written approval of InnoTech Alberta, University of Saskatchewan or Statvis.
4. Any authorized copy of this report distributed to a third party shall include an acknowledgement that the report was prepared by InnoTech Alberta and shall give appropriate credit to InnoTech Alberta and the author of the report.
5.

CITATION

This report may be cited as:

Chigbo, C., Sorenson, P., and Fuellbrandt, P. 2024. Annual Report on the Alberta Background Soil Quality System Project Phase 3. Prepared by InnoTech Alberta, Edmonton, Alberta for the Petroleum Technology Alliance Canada. Project AUPRF 23-RRRC-02. 29 pp.

REPORT PREPARATION

Prepared by:

(Signature on final)

Chibuike Chigbo, Ph.D., P.Biol., PMP
Experienced Researcher
InnoTech Alberta Inc.

EXECUTIVE SUMMARY

As part of the environmental regulatory framework to minimize risk to receptors, chemical parameter concentrations in soil or water exceeding regulatory guidelines, which can be attributed to industrial activities at the site, require remediation and/or monitoring. This is complicated by the fact that various parameters are naturally elevated in Alberta, with concentrations that exceed the generic (Tier 1) Soil and Groundwater Remediation guidelines. Where this occurs, environmental professionals must prove to the satisfaction of the applicable regulatory body that the elevated concentrations are of natural origin and not the result of industrial activities to avoid unnecessary remediation activities. This challenge is faced in environmental management of industrial sites while active and at end of life, and when responding to unintentional releases during product handling or transportation.

Salinity and certain metals are the most common naturally elevated parameters in Alberta. If salt and metal parameters are naturally elevated compared with Tier 1 guidelines, environmental professionals can mistake these naturally elevated concentrations for contamination, followed by unnecessary monitoring and remediation efforts. The challenge of proving, where applicable, that elevated parameter concentrations are of natural origin, has been identified by industry and practitioners as a root cause of cost uncertainty, multi-year timelines, reaching regulatory closure, and in some cases, requirements for unnecessary and unsustainable monitoring and remediation efforts.

Industry, government, and environmental consultants have identified a need for more effective identification of background salt and metals concentrations. There is currently no publicly available resource that maps or predicts background concentrations of these parameters for Alberta. The Alberta Background Soil Quality System (ABSQS) project was initiated by InnoTech Alberta (InnoTech), the Alberta Upstream Petroleum Research Fund (AUPRF) managed by the Petroleum Technology Alliance of Canada (PTAC), and the Clean Resources Innovation Network (CRIN) to address this gap. The overall objective of the ABSQS is to develop a database of background metals and salinity parameters in Alberta to decrease the cost and time required to identify and remediate contaminated sites. The ABSQS is intended to be used as a resource to assist industry and government in environmental management of sites that are actually contaminated.

The soil quality parameters (salinity and metals) analysis workflow derived to identify background data records provided stable and replicable results. Based on background metals fingerprints identified in the cleaned dataset, 18,513 data records of the 23,224 salinity data records in the master dataset were identified as being representative of background. The predictive soil mapping accuracy was greater for subsoil predictions compared to topsoil and all subsequent polygons were created using the subsoil prediction results due to the higher accuracy. Only a small percentage of the polygons had soil observation data present within the polygon, with approximately 5% of polygons having direct salinity data, and 2% having metals data. Therefore, most polygons had data extrapolated from the polygons with observations data based on the distribution of probabilities for the electrical conductivity (EC) and sodium adsorption ratio (SAR) class predictions. Prediction certainty was inversely related to salinity, with areas of higher salinity having more uncertainty.

The next phase in this project will be to execute a field sampling program to fill in the data gaps identified during this phase of the project, and to complete further testing and validation of the model.

TABLE OF CONTENTS

CITATION i
EXECUTIVE SUMMARY iii
LIST OF TABLES v
LIST OF FIGURES v
1.0 BACKGROUND 1
1.1 Preamble 1
1.2 Objective 2
1.3 Benefits 3
1.4 Scope 3
2.0 DATA 3
2.1 Methodology 4
2.1.1 Data Compilation 4
2.1.2 Data Harmonization 5
2.1.3 Data Cleaning 5
2.1.4 Additional Steps for Salinity Dataset 5
2.2 Data Exploration and Dimensionality Reduction 6
2.3 Identifying Background Patterns 8
2.4 Applying Background Patterns to the Full Dataset 9
2.5 Predictive Soil Mapping 9
2.6 Soil Property Summaries 13
2.7 Soil Property Extrapolation 14
2.8 Uncertainty Analysis 14
2.9 Field Sampling Plan 15
3.0 RESULTS 16
3.1 Salinity Results 16
3.2 Metals Results 17
3.3 Predictive Soil Mapping 18
3.4 Soil Property Summaries 19
3.5 Soil Property Extrapolation 21
3.6 Uncertainty Analysis 23
3.7 Field Sampling Plan 25
3.8 System distribution and communication plan 25
4.0 CONCLUSIONS 26
5.0 RECOMMENDATIONS 27
6.0 REFERENCES 28

LIST OF TABLES

Table 1. Stakeholder benefits from the Alberta Background Soil Quality System. 3

Table 2. Environmental covariates used for predictive soil mapping. 12

Table 3. Soil parameters summarized for each polygon. 14

Table 4. Accuracy, Kappa, and confusion matrix for topsoil electrical conductivity (EC) and sodium adsorption ratio (SAR). 18

Table 5. Accuracy, Kappa, and confusion matrix for subsoil electrical conductivity (EC) and sodium adsorption ratio (SAR). 18

LIST OF FIGURES

Figure 1. Project phases and tasks for the Alberta Background Soil Quality System project. 2

Figure 2. Full dataset and pilot area (labelled Prototype Area Boundary) data record sites. 4

Figure 3. Initial UMAP data model with distinct data node outlined in red. 6

Figure 4. UMAP data model after removal of sodium chloride node showing diffuse patterns consistent with background. 7

Figure 5. Example HCA dendrogram with bisecting lines to show various levels of granularity. 7

Figure 6. HCA dendrogram for background salinity dataset. 8

Figure 7. HCA dendrogram for background metals dataset. 9

Figure 8. Predictive Soil Mapping Process Flow Diagram 11

Figure 9. Statistical distributions of salinity parameters in the final background dataset.. 16

Figure 10. Statistical distributions of metals parameters in the final background dataset. 17

Figure 11. Mean electrical conductivity (EC) per polygon (dS m^{-1}). 19

Figure 12. Mean sodium adsorption ratio (SAR) values per polygon. 20

Figure 13. Mean selenium (mg kg^{-1}) per polygon. 21

Figure 14. Map outlining which polygons have extrapolated salinity data. 22

Figure 15. Map outlining which polygons have extrapolated metal data. 23

Figure 16. Map of polygon salinity prediction confidence. 24

Figure 17. Map illustrating target polygons in red and sample sites in white for field sampling to improve overall model performance for both salt and metal polygons. 25

1.0 BACKGROUND

The following background information on the research project, titled Background Metals and Salinity Database and Analysis Tool, was previously described by Shelby-James and Fuellbrandt (2022).

1.1 PREAMBLE

Where land has been used for industrial purposes, effective and sustainable ecological restoration and return to productive use are key objectives for responsible land stewardship. As part of the environmental regulatory framework to minimize risk to receptors, chemical parameter concentrations in soil or water exceeding regulatory guidelines, which can be attributed to industrial activities at the site, require remediation or monitoring. This is complicated by the fact that various parameters are naturally elevated in Alberta, with concentrations that exceed the generic *Alberta Tier 1 Soil and Groundwater Remediation Guidelines* (Tier 1) (Alberta Environment and Parks, 2022). Where this occurs, environmental professionals must prove to the satisfaction of the applicable regulatory body that the elevated concentrations are natural and not the result of industrial activities.

The challenge of proving, where applicable, that elevated parameter concentrations are of natural origin, has been identified by industry and practitioners as a root cause of cost uncertainty, multi-year timelines, reaching regulatory closure, and in some cases, unnecessary and unsustainable monitoring and remediation efforts. The challenge is faced in environmental management of industrial sites while active and at end of life, and when responding to unintentional releases during product handling or transportation. Many of the backlog of legacy oil and gas well sites and associated facilities in Alberta are stalled or are being repeatedly monitored for this reason.

Salinity and certain metals are the most common naturally elevated parameters in Alberta. If salt and metal parameters are naturally elevated compared with Tier 1 guidelines, environmental professionals can mistake these naturally elevated concentrations for contamination, followed by unnecessary monitoring and remediation efforts. Key members of the Petroleum Technology Alliance Canada (PTAC)'s Alberta Upstream Petroleum Research Fund (AUPRF), environmental consultants, and regulators have identified a need for more effective identification of background salt and metals concentrations as one of their highest priorities.

To prove that concentrations of one or more parameters are naturally elevated, background samples must be collected, often requiring a second mobilization of equipment and resources once site data has been received from a laboratory. This has significant cost and timeline implications, not only for mobilization but also for obtaining permission for offsite sample collection and permitting. Liability estimates for some industrial sites are also inflated due to the inability to confirm elevated background concentrations.

Fortunately, soils have been analyzed, characterized, and mapped in Alberta for a variety of purposes including regulatory reporting and site evaluation, land use evaluation, local and regional land use planning, site-specific project planning, environmental impact assessments, global inventory modelling and soil classification (in agricultural regions). Considerable baseline soil information has been collected at point locations to support development of conservation and construction plans or pre-disturbance assessments, conservation and reclamation business plans, environmental impact assessments, detailed site assessments, and Phase 2 environmental site assessments. Although these data, which consist of both field observations and measurements and laboratory analyzed samples, were collected for specific purposes, in general, the information was collected using standard methods prescribed by the government. If georeferenced, the data has tremendous value and could be integrated into a

comprehensive database. This could then be leveraged with predictive mapping technologies to create relevant spatial predictions of soil variables, such as background soil salinity and elemental concentrations.

1.2 OBJECTIVE

InnoTech Alberta Inc (InnoTech), Statvis Analytics Inc. (Statvis), and the University of Saskatchewan are developing the Alberta Background Soil Quality System collaboratively with EnvirometriX B.V. (EnvirometriX) using a phased approach as shown in **Error! Reference source not found.**. Phasing the project allows stakeholders to provide valuable feedback that was incorporated into the design of subsequent phases. The overall objective of the Alberta Background Soil Quality System (ABSQS) is to develop a database of background metals and salinity parameters in Alberta to decrease the cost and time required to identify and remediate contaminated sites. The ABSQS is intended to be used as a resource tool to assist industry and government in environmental management of sites that are actually contaminated.

The project used a phased approach with the following key activities :

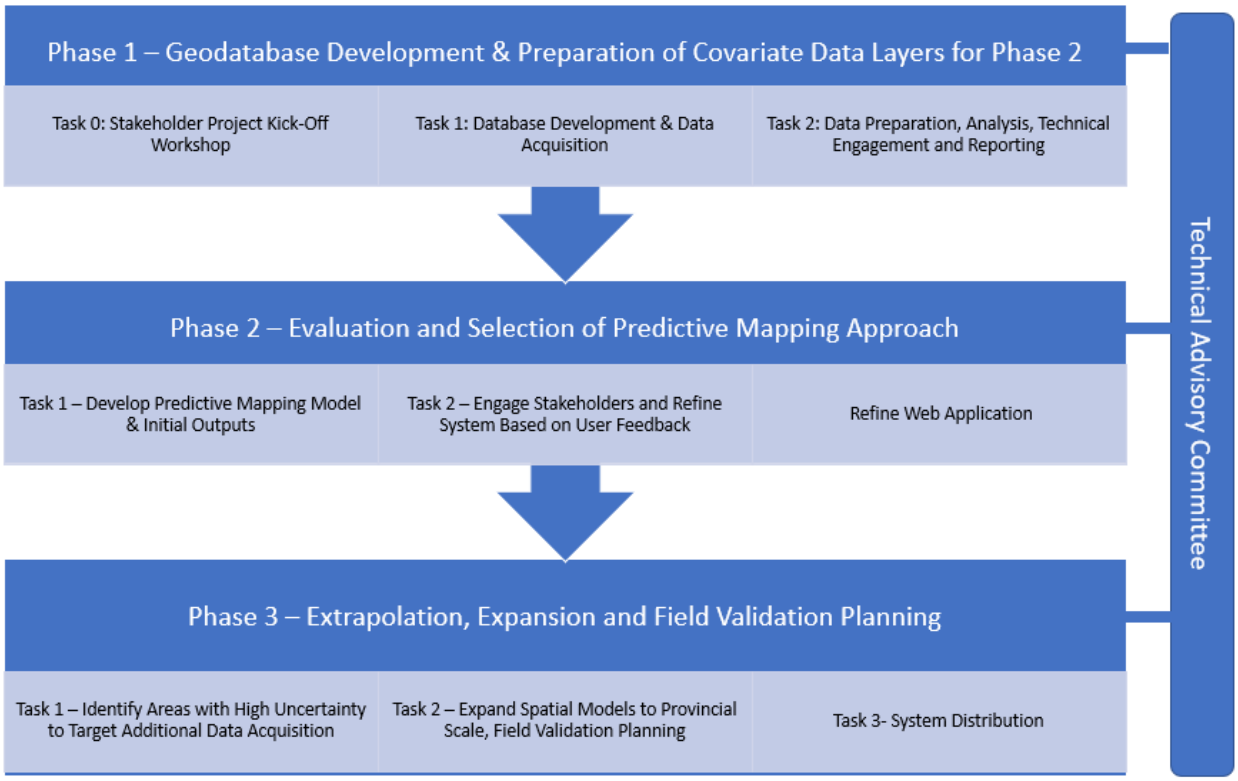


Figure 1. Project phases and tasks for the Alberta Background Soil Quality System project.

1.3 BENEFITS

It is anticipated that this project would lead to significant benefits for multiple stakeholders such as industry (e.g., conventional oil and gas, oil sands mining and in situ); municipal, provincial, and federal governments; Indigenous communities; and collectively for Albertans as described in **Error! Reference source not found.**.

Table 1. Stakeholder benefits from the Alberta Background Soil Quality System.

Stakeholder	Benefit
Industry and practitioners	Avoid replication of data collection, reducing costs and increasing certainty.
	Ability to provide empirical evidence of background concentrations and natural variability.
	More accurate (and lower) liability estimates by excluding naturally elevated parameters.
	Increased ability to focus resources on managing actual risk to receptors.
	Reduced liability by moving stalled sites to regulatory closure.
	Reduced time for reclamation certification.
Government and regulators	Reduced review times of environmental reports with elevated salinity and/or metals in background.
	Increased number of reclamation certificates being issued, decreasing the number of inactive wells in alignment with current policy goals
	Increased consistency in data presentation.
	Background maps for the entire province for use in land use planning.
Collective benefits for Albertans	Provision of open data available to extrapolate for multiple uses.
	Less disturbance and disruption of the natural environment due to assessment and remediation of uncontaminated soils with naturally elevated salinity or metals.
	Increased number of industrial sites being cleaned up, decreasing risk to the orphan well association and pressure on the economy due to bankruptcy induced by excess liability.

1.4 SCOPE

A Technical Steering Committee (TSC) was established and comprised of regulators, industry, environmental practitioners, academia, and the project team. TSC members served in a voluntary technical advisory capacity to ensure the project:

- is inclusive by including relevant stakeholders in workshops and consultations,
- aligns with regulatory requirements and expectations,
- deliverables directly meet users' specific data needs, and
- creates open and accessible data outputs.

This report summarizes the completed Phase 3 activities and results that were partially funded by PTAC.

2.0 DATA

Soil salinity and metals data was compiled, harmonized, and cleaned into a geodatabase for analysis and use in predictive mapping platforms. Locations of data records used in the project are shown in Figure 2. Salinity parameters of interest were calcium, chloride, magnesium, potassium, sodium, sulphate, electrical conductivity (EC), sodium adsorption ratio (SAR), and pH. These nine parameters were selected

as they are most commonly reported in laboratory salinity analytical packages. These parameters are commonly analyzed because they have potential to affect the ability of the soil to support plants and soil microbes as well as soil structure. They are also the most relevant parameters for complying with regulatory obligations and separating multiple salinity sources in Alberta.

Metals parameters of interest are antimony, arsenic, beryllium, cadmium, chromium, cobalt, copper, lead, mercury, molybdenum, nickel, selenium, thallium, uranium, vanadium, and zinc. These 16 parameters were chosen for their potential to cause an adverse effect on human health and the environment, and because data records provided had sufficient reporting and detection of these metals for valid statistical analysis.

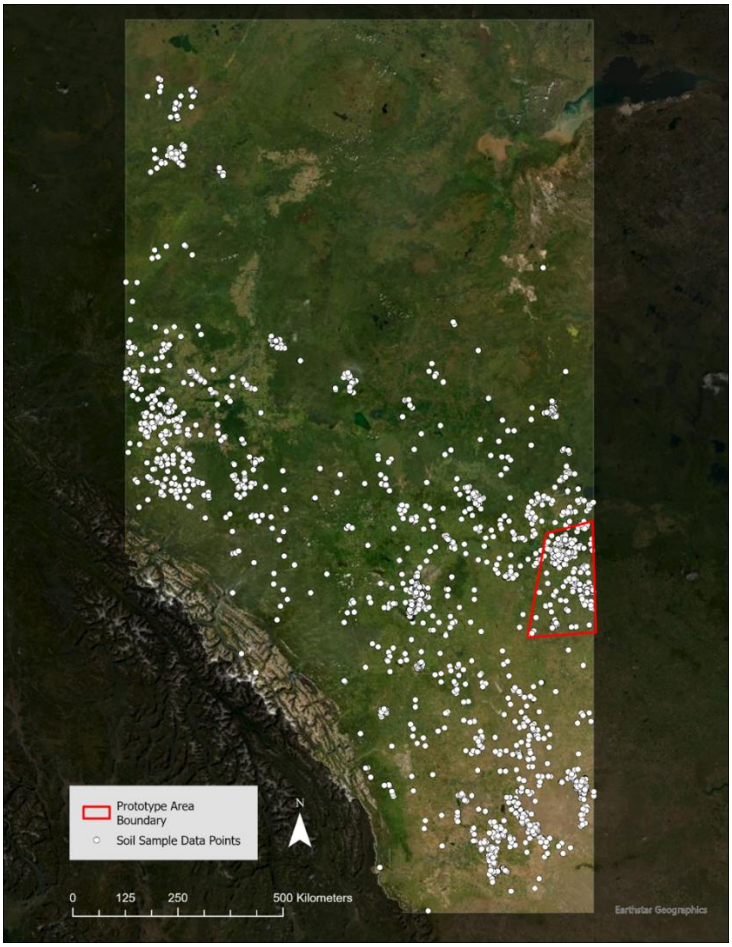


Figure 2. Full dataset and pilot area (labelled Prototype Area Boundary) data record sites.

2.1 METHODOLOGY

To ensure creation of a high-quality dataset that reliable conclusions could be drawn from, datasets were collected and then prepared, explored, and analysed according to the methods described below. Multiple workflows were tested, and the workflow that proved the most effective for separating data records with anthropogenic influence from data records representative of background conditions was selected.

2.1.1 DATA COMPILATION

Data providers were engaged to request data and ensure data received was formatted correctly. The identity of data providers and details about the quantity and types of data provided are confidential under data sharing agreements. In several instances there were formatting issues or missing metadata (e.g., UTM zones, units, or analytical methods). To resolve these issues, we engaged with the data provider, and the dataset was either re-exported to correct the issue or information was provided so that the issue could be manually corrected.

2.1.2 DATA HARMONIZATION

Data were harmonized by combining multiple smaller datasets into one master dataset. Columns were matched based on parameters and metadata values and combined to create a master dataset for the project. The master dataset included 224,902 salinity data records from 5,887 unique locations and 23,224 metals data records from 1,911 unique locations. Many data records provided only had legal site description (LSD) for location information. When this was the case, multiple samples often had the same LSD and therefore have the same coordinates in the database. This is the reason for the comparatively lower number of sampling locations versus the number of data records.

2.1.3 DATA CLEANING

The master dataset was cleaned to remove records that would impede statistical analysis. In this case, cleaning refers to removing a record from the dataset used to define the ideal background pattern (referred to as a fingerprint). It is important to note that these data records were only removed from the dataset during the development of background fingerprints. Once these were defined, all data records were returned to the dataset to be classified as either representative of background or showing signs of anthropogenic influence.

Salinity records were cleaned (i.e., removed from the decision-making process) based on the following criteria:

- Samples with missing values.
- Samples with non-detect values of one or more cation/anion.
- Samples with cation/anion balance > 25%.
- Samples > 95th percentile of one or more cation/anion, SAR, or EC.

Metals records were cleaned (i.e., removed from the decision-making process) based on the following criteria:

- Samples with missing values.
- Samples with non-detect values of select metals parameters. Metals parameters were selected based on potential for causing adverse effects and sufficient detectability in the dataset for reliable statistical analysis.
- Samples with one or exceedance of an Alberta Tier 1 soil remediation guideline.
- High and low outliers.

After removal of these samples, 74,943 salinity and 13,333 metals data records remained.

2.1.4 *Additional Steps for Salinity Dataset*

The initial step in chemometric data analysis[1] was to apply the same workflow that was used for the pilot area. In the pilot area, a limit of 100 mg/kg chloride was used when identifying background fingerprints for diagnostic use in the master dataset. This limit came from a heuristic defined in the Subsoil Salinity Tool manual which states that background chloride concentrations "...vary between different regions and soil types but are generally below 100 mg/kg…" (Equilibrium Environmental, 2020).

Due to the size of the provincial-scale salinity dataset, more granularity was required to identify samples with patterns indicative of anthropogenic impact as large magnitude differences can hinder statistical analysis. Range normalization was applied to address this. Data was then analyzed using both principal component analysis (PCA) and uniform manifold approximation and projection (UMAP). UMAP was found to be superior in identifying unique fingerprints in the dataset that could then be further explored and determined to be background or representative of anthropogenic impact. Initial UMAP analysis identified a data node that was distinctly different from the rest of the data cluster as shown in Figure 3. After further analysis, the distinct data node was dominated by sodium chloride and determined to be indicative of anthropogenic influence. This data node was removed from the dataset for further analysis.

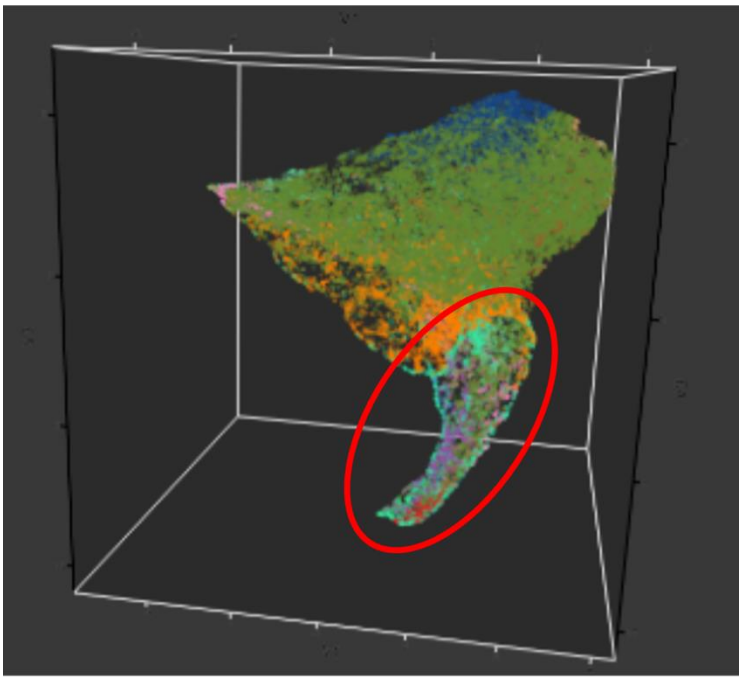


Figure 3. Initial UMAP data model with distinct data node outlined in red.

Re-analysis of the dataset after exclusion of the anthropogenic data node showed no distinct or separate nodes of data as shown in Figure 4. This is what would be expected from diffuse sources associated with natural background.

After these additional data analysis steps were completed for the salinity dataset, the workflow developed in Phase 1 of the ABSQS was applied.

[1] Chemometrics is the science of relating measurements made on a chemical system (including dynamic chemical processes) to the state of the system via application of mathematical or statistical algorithms.

2.2 DATA EXPLORATION AND DIMENSIONALITY REDUCTION

Hierarchical cluster analysis (HCA) was used alongside traditional statistical techniques (correlation plots and summary statistics for various subsets of the data) to identify clusters of data records representative of anthropogenic and non-anthropogenic (i.e., background) patterns. Clustering in general is a method of statistical analysis that clusters data records in such a way that they are more like other data records within the same cluster than they are to data records in other clusters. HCA is used to find discrete clusters with varying degrees of similarity (or dissimilarity) in a dataset. HCA builds a hierarchy of clusters and displays them on a dendrogram. A dendrogram is a tree-structured graph that shows the relationship between data records based on the length of the line connecting them. Shorter lines represent a closer relationship while longer lines indicate a larger difference between data records (Figure 5). As distance from individual data records increases, the dendrogram lines become longer

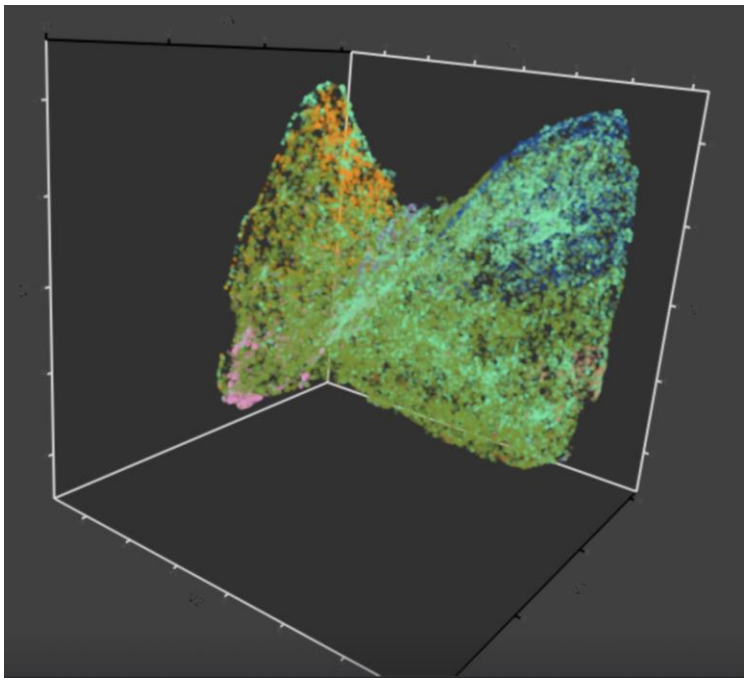


Figure 4. UMAP data model after removal of sodium chloride node showing diffuse patterns consistent with background.

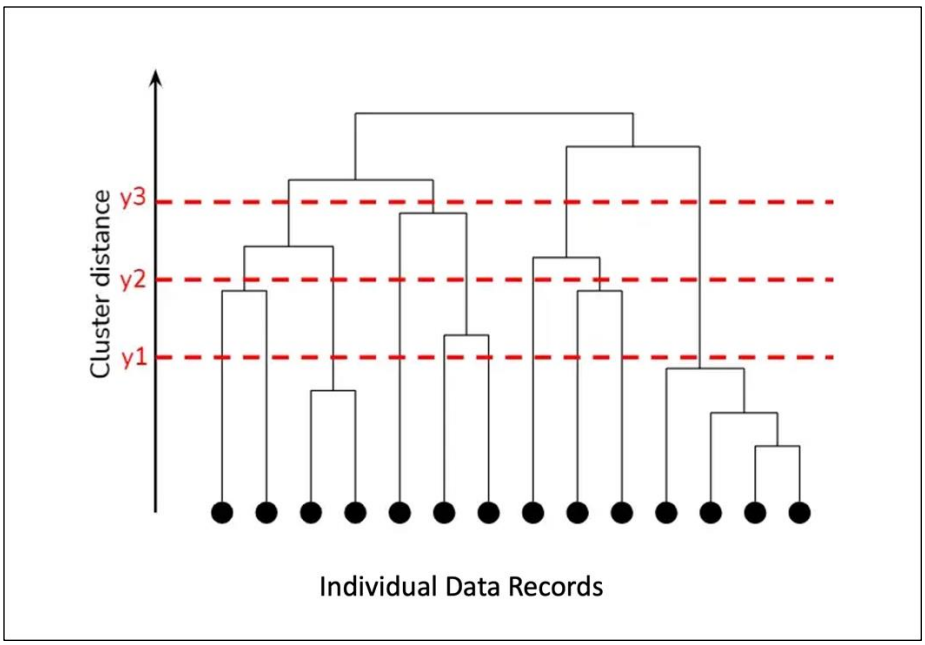


Figure 5. Example HCA dendrogram with bisecting lines to show various levels of granularity.

showing more dissimilarity between data records. In the example shown in Figure 5 three lines were drawn bisecting the dendrogram, labeled y1, y2, and y3 respectively, that show three options for clustering granularity. Line y1 splits the dataset into 10 clusters, y2 splits the dataset into seven clusters and y3 splits the dataset into four clusters. As the number of clusters increases, the relationships between individual data records in a cluster become more granular and specific. For example, line y1 provides so much granularity that several of the clusters have only one data record in them. The goal of exploring the ABSQS salinity dataset using HCA was to provide enough granularity that data records showing anthropogenic impacts could be separated from clusters representative of background conditions. To achieve this, boundary conditions for an ideal background dataset had to be defined.

2.3 IDENTIFYING BACKGROUND PATTERNS

An HCA dendrogram was completed for the ideal salinity and metals background datasets that also included a heat map of parameters. The dendrogram was explored at varying degrees of granularity to determine where relationships between parameters changed meaningfully. An example of a meaningful difference between clusters would be where data records in one cluster were dominated by a strong correlation between certain parameters while data records in an adjacent cluster were dominated by a strong correlation between different parameters. HCA dendrograms for the reduced background datasets for metals and salinity are provided in Figures 6 and 7, respectively.

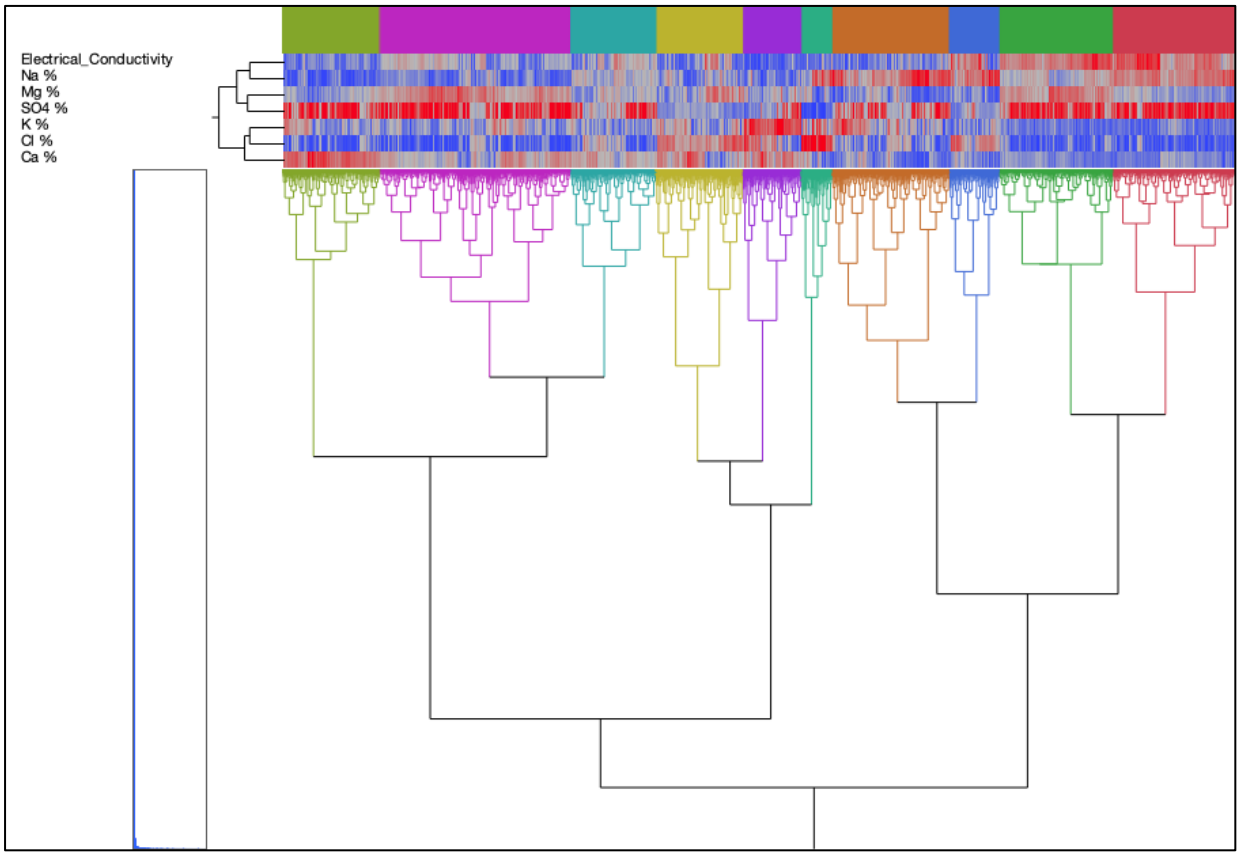


Figure 6. HCA dendrogram for background salinity dataset. The box to the left of the dendrogram is a histogram of the dataset showing the majority of samples were low concentration.

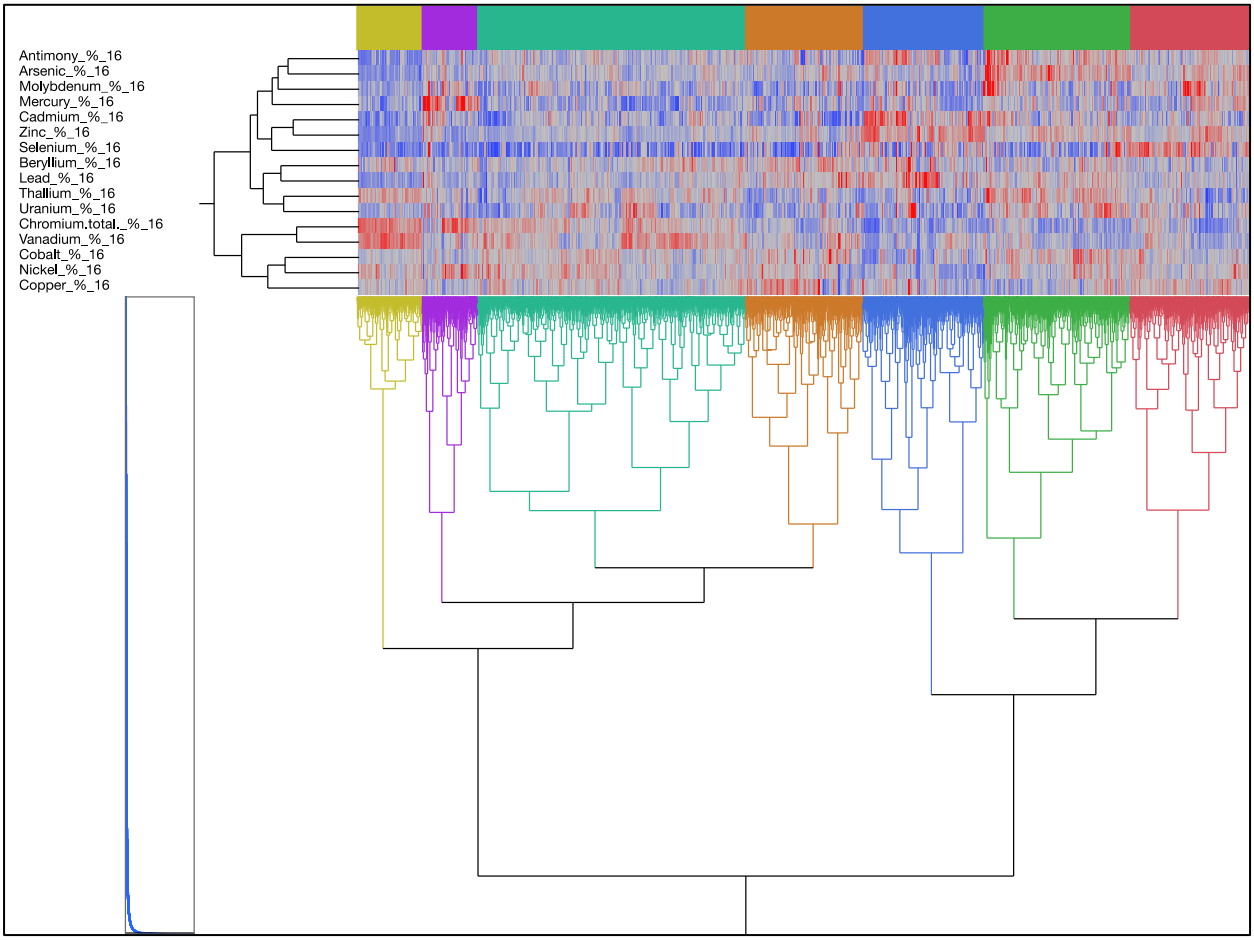


Figure 7. HCA dendrogram for background metals dataset.

2.4 APPLYING BACKGROUND PATTERNS TO THE FULL DATASET

Boundary conditions were defined for each background cluster in the ideal background dataset. Boundary conditions were defined using the minimum and maximum percent contribution of each parameter in a background cluster. To be designated as belonging to a background cluster, a data record had to have measured values of all salinity or metals parameters within the minimum and maximum boundary conditions set for a particular cluster. The master dataset was compared to these boundary conditions and records that did not fit into one or more background clusters were removed from the final dataset as they were considered to show anthropogenic influence. Boundary conditions for each parameter are shown in the results section.

2.5 PREDICTIVE SOIL MAPPING

The predictive soil mapping (PSM) objectives were assessed (Figure 8). However, due to limitations of the data the regression model results were not sufficient for decision making as the R^2 values were less than 0.4. Two main reasons are likely responsible for the low performance:

- Coordinates: The absence of coordinates for borehole locations led to a significant reduction in data. This caused data aggregation to have a coarser spatial resolution.
- Spatial Resolution: The transition to a coarser spatial resolution decreased the ability to detect salinity. Saline features often manifest at finer scales than the training data resolution after aggregating.

Additionally, the end members included too many multigroup classifications making it unsuitable for mapping. Therefore, an alternative classification approach was undertaken as described in this section. Metals were not mapped separately but were instead summarized for each salinity polygon.

Precise location data was not available for the soil point data, and only location data to the legal land description was available. For that reason, data was aggregated based on a 250 m resolution, as given the center of a legal subdivision (LSD) would be 200 m from the boundaries, with the additional 50 m to account for potential overlap for points located on the edge of LSDs. The maximum value for electrical conductivity (EC) and sodium adsorption ratio (SAR) within each 250 m resolution pixel was used for model training and validation. Additionally, EC and SAR values were calculated for the topsoil layer, defined as 0 to 0.15 metres below ground surface (m bgs), and subsoil which was defined as 0.15 to 1.5 m bgs. This data was then classed into Alberta's salinity and sodicity rating categories of good, fair, poor, and unsuitable (Alberta Environment, 2001). The topsoil ratings were used for the 0 to 0.15 m bgs layer, and the subsoil ratings were used for 0.15 to 1.5 m bgs. Environmental covariate data was retrieved using Google Earth Engine for the entirety of Alberta for predictive soil mapping. All covariates are listed in Table 2.

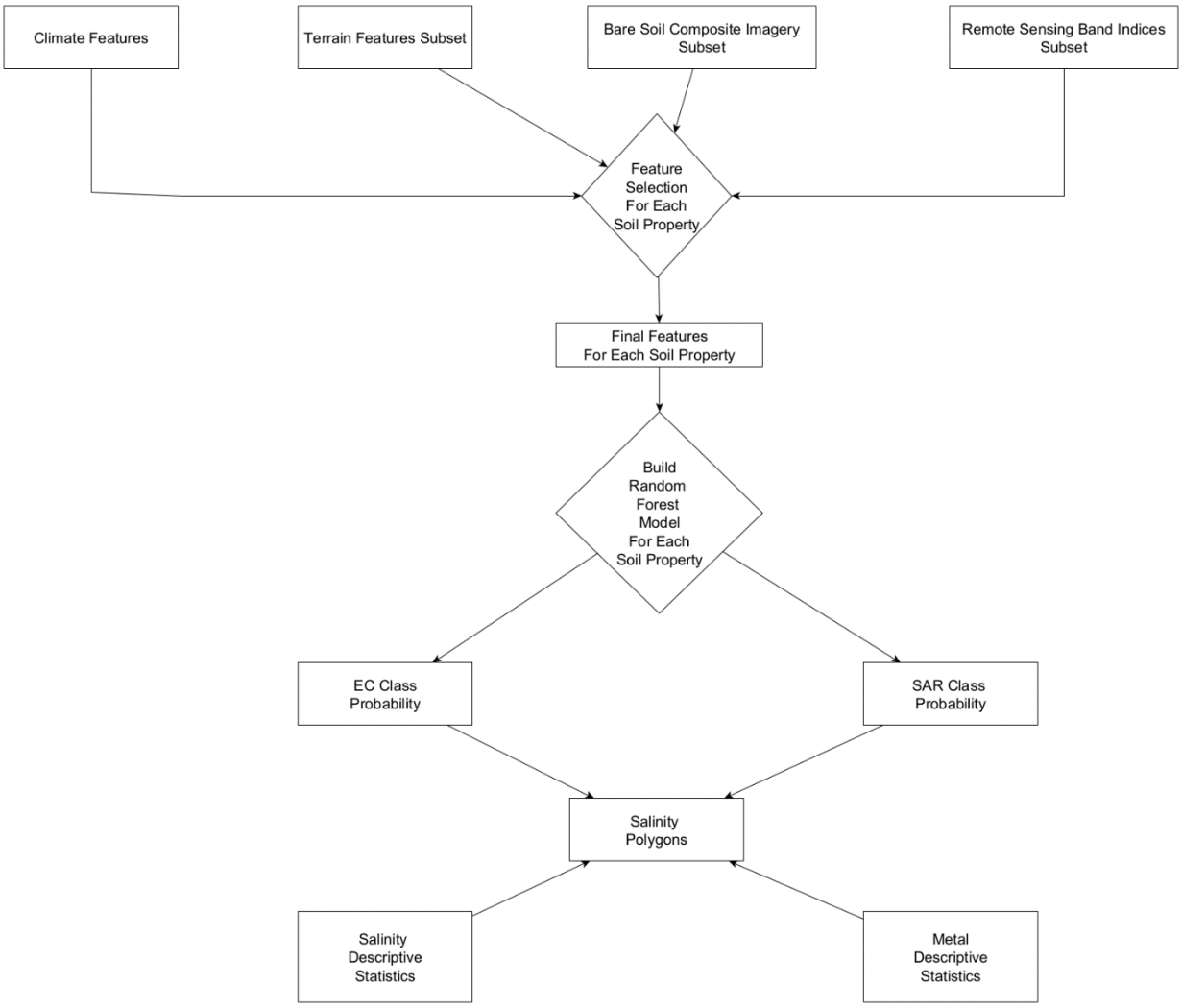


Figure 8. Predictive Soil Mapping Process Flow Diagram

Table 2. Environmental covariates used for predictive soil mapping.

Covariate	Data Source	Time Period	Resolution
Standard deviation of elevation with a 21 x 21 focal window	ALOS1 3x3 focal smoothed digital elevation model	2006 to 2011	250 m
Standard deviation of elevation with a 21 x 21 focal window	ALOS1 9x9 focal smoothed digital elevation model	2006 to 2011	250 m
Bare soil composite imagery	Landsat 8 with 10x10 focal smoothing	May to October 2013 to 2023	250 m
Raw Bands 25th Percentile	Landsat 8 with 10x10 focal smoothing	May to October 2013 to 2023	250 m
Raw Bands 50th Percentile	Landsat 8 with 10x10 focal smoothing	May to October 2013 to 2023	250 m
Raw Bands 75th Percentile	Landsat 8 with 10x10 focal smoothing	May to October 2013 to 2023	250 m
Raw Bands 25th Percentile	Sentinel-2 with 10x10 focal smoothing	May to October 2017 to 2023	250 m
Raw Bands 50th Percentile	Sentinel-2 with 10x10 focal smoothing	May to October 2017 to 2023	250 m
Raw Bands 75th Percentile	Sentinel-2 with 10x10 focal smoothing	May to October 2017 to 2023	250 m
Vertical Horizontal Polarization	Sentinel-2 with 10x10 focal smoothing	May to October 2015 to 2023	250 m
Median NDVI[2]	Landsat 8 with 10x10 focal smoothing	May to October 2013 to 2023	250 m
Standard Deviation of NDVI	Landsat 8 with 10x10 focal smoothing	May to October 2013 to 2023	250 m
CRSI[3]	Landsat 8 with 10x10 focal smoothing	May to October 2013 to 2023	250 m
REIP[4]	Sentinel-2 with 10x10 focal smoothing	May to October 2017 to 2023	250 m
Rasterized soil classification polygons	AGRASID	NA	250 m
Rasterized surficial geology polygons	Alberta Geological Survey Surficial Geology Polygons	NA	250 m
Mean Annual Air Temperature	Copernicus Climate Change Service ERA5 Climate Reanalysis	1979 to 2020	250 m
Annual Total Precipitation	Copernicus Climate Change Service ERA5 Climate Reanalysis	1979 to 2020	250 m

[1] Advanced Land Observation Satellite

[2] Normalized Difference Vegetation Index

[3] Canopy Response Salinity Index

[4] Red Edge Inflection Point

For the topsoil and subsoil for EC and SAR, random forest classification models were built using the ranger package in R (Wright & Ziegler, 2017). Each model had the following model development process:

- Covariates that were correlated with another covariate by more than 0.95 were removed, and only one of the covariates was kept.
- Data was split into training and validation data, with 75 percent of the data retained for training and 25 percent for validation.
- An initial model was built using all the covariates after removal of the highly correlated covariates.
- A backwards feature selection process was then used whereby incremental models were built with the least important variable according to the model removed each time.
- The set of features that minimized out-of-bag prediction error were then selected.
- A final model was then built.

For each random forest model, case weights were set to equalize the probability that data from each salinity and sodicity class were selected during bootstrapping of the random forest models. Class weights were also set to internally equalize the importance of each class during model optimization. The importance value was set to impurity and the split rule was set to extra trees. The random forest models were also built as probability random forests to be able to predict the probability that each point belonged to each class based on random forest tree agreements. Model performance was then evaluated based on accuracy and kappa scores using the validation data set only.

Once the predictive models were generated, the probability random forest models were used to predict the probability of each rating category for EC and SAR in the topsoil and subsoil. As accuracy for subsoil EC and SAR were the highest, as there were generally lower values for EC and SAR for topsoil, the subsoil prediction results were used to create polygons. The probability that a given pixel was classified as having 'good' EC and SAR was determined. Both resulting rasters were then resampled to 20 m using bicubic sampling to create smoother lines during vectorization.

The probabilities for each parameter was binned to intervals of 0-0.2, 0.2-0.4, 0.4-0.6, 0.6-0.8, and 0.8-1. The EC and SAR subsoil probabilities were then combined to create a label for each pixel and the results were vectorized to create polygons.

2.6 SOIL PROPERTY SUMMARIES

After generation of polygons, values for each salinity and metal parameter were determined by calculating summary statistics based on the soil sample data within each polygon. Summary statistics were calculated for the following depths:

- 0 to 0.15 m bgs
- 0.15 to 1.5 m bgs
- 1.5 to 3.0 m bgs
- 3.0 to 6.0 m bgs

The following summary statistics were calculated for each soil parameter (Table 3) for each depth:

- Min
- 5^{th} Percentile
- 25^{th} Percentile
- 50^{th} Percentile

- 75th Percentile
- 95th Percentile
- Max
- Mean
- Standard Deviation
- Standard Error of the Mean
- Number of Observations

Table 3. Soil parameters summarized for each polygon.

Type	Parameters	Polygons with Observations
Salinity	pH, Electrical Conductivity, Sodium Adsorption Ratio, Calcium (mg kg^{-1}), Magnesium (mg kg^{-1}), Sodium (mg kg^{-1}), Potassium (mg kg^{-1}), Chloride (mg kg^{-1}), Sulphate (mg kg^{-1}), Potassium (meq), Magnesium (meq), Calcium (meq), Sodium (meq), Sulfate (meq), Chloride (meq), Total Anion, Total Cation, Potassium (%), Magnesium (%), Calcium (%), Sodium (%), Sulfate (%), Chloride (%)	845
Metals	Antimony (mg kg^{-1}), Arsenic (mg kg^{-1}), Beryllium (mg kg^{-1}), Cadmium (mg kg^{-1}), Total Chromium (mg kg^{-1}), Cobalt (mg kg^{-1}), Copper (mg kg^{-1}), Lead (mg kg^{-1}), Mercury (mg kg^{-1}), Molybdenum (mg kg^{-1}), Nickel (mg kg^{-1}), Selenium (mg kg^{-1}), Thallium (mg kg^{-1}), Uranium (mg kg^{-1}), Vanadium (mg kg^{-1}), Zinc (mg kg^{-1})	316

2.7 SOIL PROPERTY EXTRAPOLATION

Most polygons did not have any data observations directly in the polygons; hence values were extrapolated for most polygons. For polygons with one or more data observations the average values for the class probabilities for EC and SAR subsoil were determined for each polygon. For each polygon without any observations, the nearest neighbor polygon was determined using the FFN package in R (Beygelzimer et al., 2024). Nearest neighbor was not calculated spatially, but rather the nearest neighbor in terms of the probabilities for good, fair, poor, and unsuitable EC and SAR classes.

The values from the nearest neighbor polygon with observations were then assigned to each polygon without observation data. An additional field was added to the database (extrap) to note if the data was extrapolated or not. A value of N indicates the data is derived from observational data, whereas a value of Y indicates the data is extrapolated.

2.8 UNCERTAINTY ANALYSIS

Following the soil property extrapolation, the prediction uncertainty was estimated for each polygon. Using the average class probabilities for each polygon, the maximum probability across all classes was determined for both EC and SAR. The maximum EC class probability and maximum SAR probability were then averaged. Fundamentally, the probabilities are calculated based on the proportion of trees within the random forest model that agree on the final prediction. Higher maximum probability values indicate greater agreement amongst the decision trees within a random forest model, and therefore higher certainty in the predictions.

Once the maximum class probability was calculated for each polygon, the resulting values were scaled between 0 and 1 based on the minimum and maximum probabilities. This approach is also referred to as min-max normalization. This process resulted in each polygon having a certainty score between 0 and 1, with 0 being complete uncertainty and 1 being complete certainty.

2.9 FIELD SAMPLING PLAN

The field sampling plan was developed with the following methodology:

1. Polygons were subset to only those with site data for both salts and metals and only polygons that had confidence levels less than 0.5 were kept, prioritizing sampling from lower confidence locations.
2. Polygons were then subset to only include those with an average subsoil EC and SAR value that would be considered fair or poorer.
3. The Alberta Energy Regulator shapefile of wellsites in Alberta (Alberta Energy Regulator, 2024) was then intersected with these polygons.
4. Thirty random Canadian Natural Resources Limited (CNRL) wellsites were then selected.
5. Buffers with a radius of 200 m and 100 m were generated around the wellsite. The difference between the two buffers was calculated to create a ring 100 m wide and 100 m away from the wellsite.
6. Ten samples were randomly placed within this ring for each of the 30 wellsites for a total of 300 sample locations.

3.0 RESULTS

Results for analysis of salinity and metals datasets are provided below.

3.1 SALINITY RESULTS

Statistical distributions of the nine salinity parameters analyzed (EC, SAR, pH, chloride, sulphate, sodium, calcium, magnesium, and potassium) across all depths represented are shown in Figure 9.

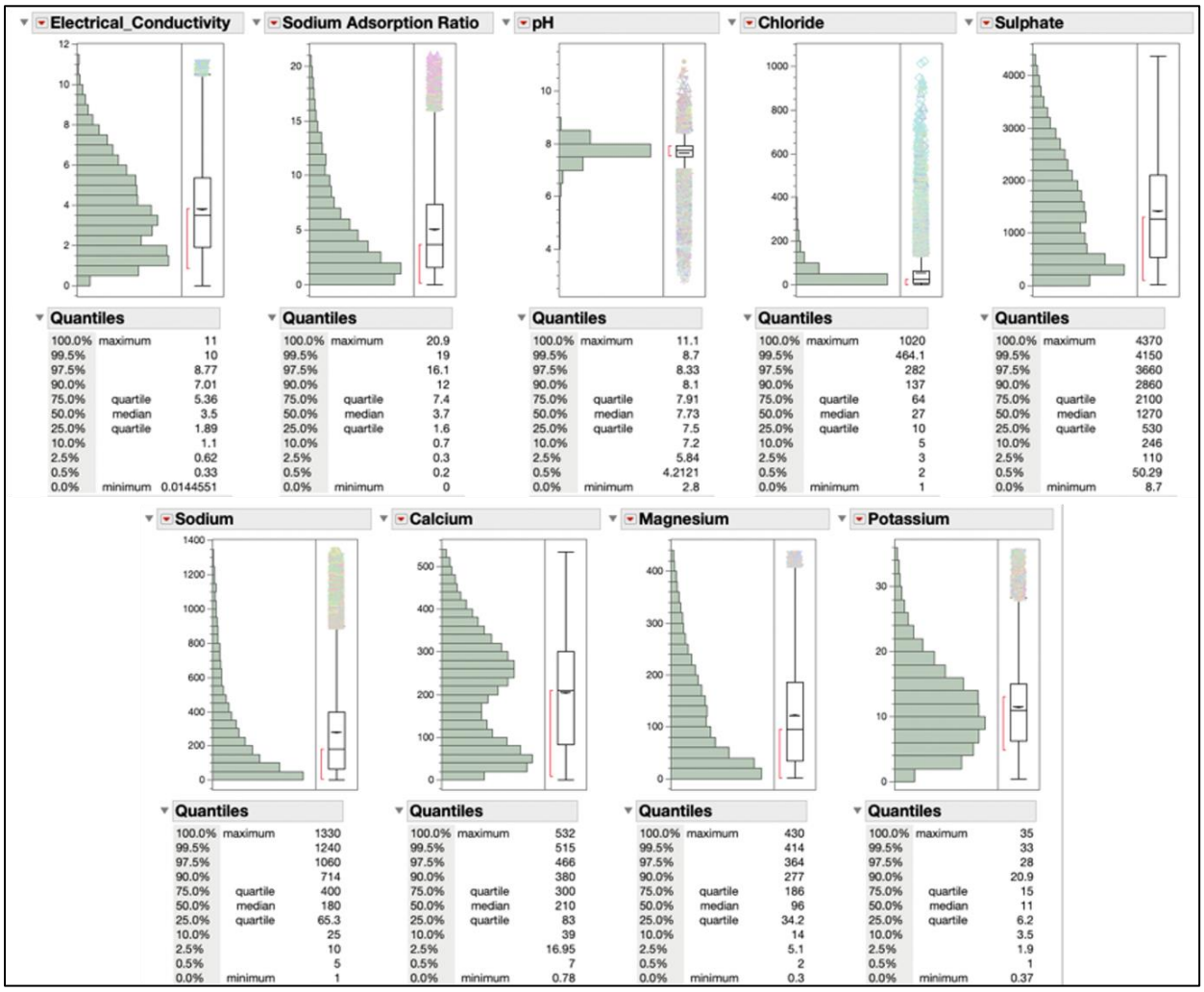


Figure 9. Statistical distributions of salinity parameters in the final background dataset. Histograms show the number of records at a particular concentration or value. Box plots provide a visual representation of quantiles.

Based on background fingerprints identified in the cleaned salinity dataset, 151,542 data records of the 224,902 salinity data records in the master dataset were identified as being representative of background.

3.2 METALS RESULTS

In the metals database, 16 metals had sufficient data records and detection rates in the dataset provided for analysis including antimony, arsenic, beryllium, cadmium, chromium, cobalt, copper, lead, mercury, molybdenum, nickel, selenium, thallium, uranium, vanadium, and zinc. Statistical distributions of the 16 metals parameters across all depths represented are shown in Figure 10.

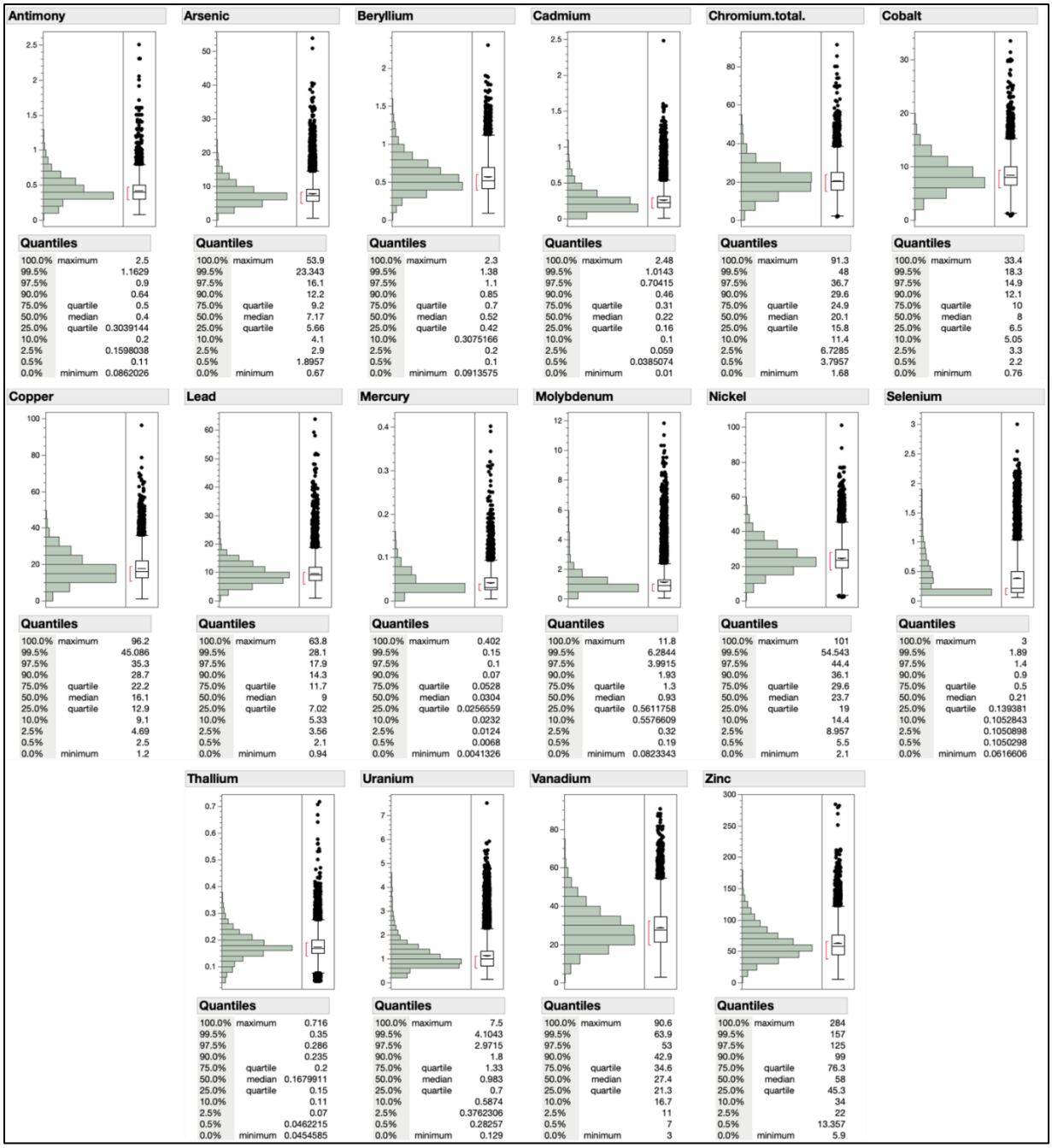


Figure 10. Statistical distributions of metals parameters in the final background dataset.

Based on background metals fingerprints identified in the cleaned dataset, 18,513 data records of the 23,224 metals data records in the master dataset were identified as being representative of background.

3.3 PREDICTIVE SOIL MAPPING

Accuracy results for the predictive soil mapping of topsoil are provided in Table 4, and results for subsoil are provided in Table 5. Results are on a point by point basis based on a 75-25 train-test split. Predictive soil mapping accuracy and Kappa values were greater for the subsoil predictions as compared to the topsoil. All subsequent polygons were created using the subsoil prediction results due to the higher accuracy.

Table 4. Accuracy, Kappa, and confusion matrix for topsoil electrical conductivity (EC) and sodium adsorption ratio (SAR).

EC				
	Good	Fair	Poor	Unsuitable
Good	226	55	41	4
Fair	55	49	37	5
Poor	22	37	26	8
Unsuitable	1	5	6	1
Accuracy	0.53			
Kappa	0.22			
SAR				
	Good	Fair	Poor	Unsuitable
Good	289	75	25	13
Fair	35	32	6	13
Poor	16	11	9	12
Unsuitable	7	6	5	12
Accuracy	0.60			
Kappa	0.24			

Table 5. Accuracy, Kappa, and confusion matrix for subsoil electrical conductivity (EC) and sodium adsorption ratio (SAR).

EC				
	Good	Fair	Poor	Unsuitable
Good	533	70	74	2
Fair	48	45	39	0
Poor	25	25	79	3
Unsuitable	1	4	6	2
Accuracy	0.69			
Kappa	0.38			
SAR				
	Good	Fair	Poor	Unsuitable
Good	493	85	39	19
Fair	78	44	22	19
Poor	23	21	16	18
Unsuitable	13	13	19	34
Accuracy	0.61			
Kappa	0.28			

3.4 SOIL PROPERTY SUMMARIES

Each of the background soil properties were summarized based on the statistics discussed in the methodology. However, it is important to note that only a small percentage of polygons had observation data within the polygons. In total there were 16,473 polygons that resulted from the polygon creation process. The results for EC are illustrated in Figure 11, and the results for SAR are in Figure 12.

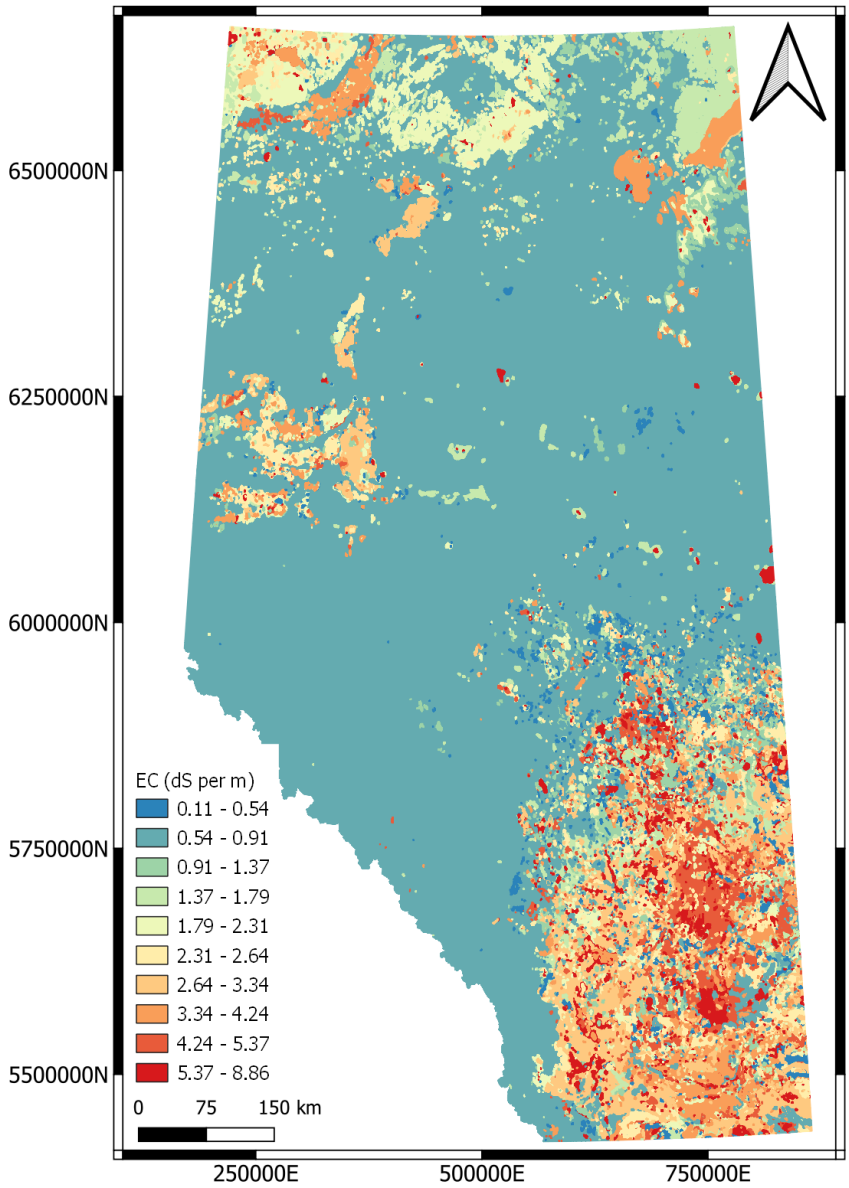


Figure 11. Mean electrical conductivity (EC) per polygon (dS m^{-1}).

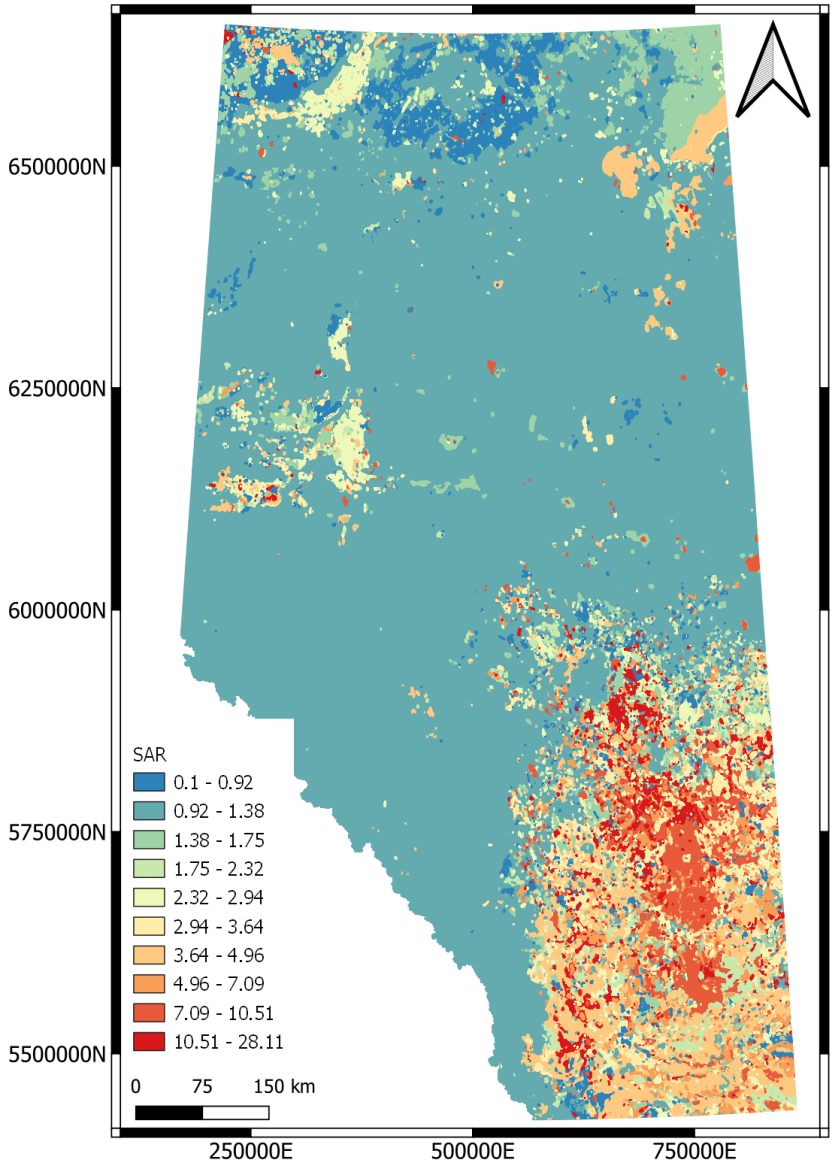


Figure 12. Mean sodium adsorption ratio (SAR) values per polygon.

Along with salinity data, metal data was also summarized successfully based on the observational data. Summary data was provided for each metal parameter. For illustrative purposes, the results for selenium are arbitrarily provided in Figure 13.

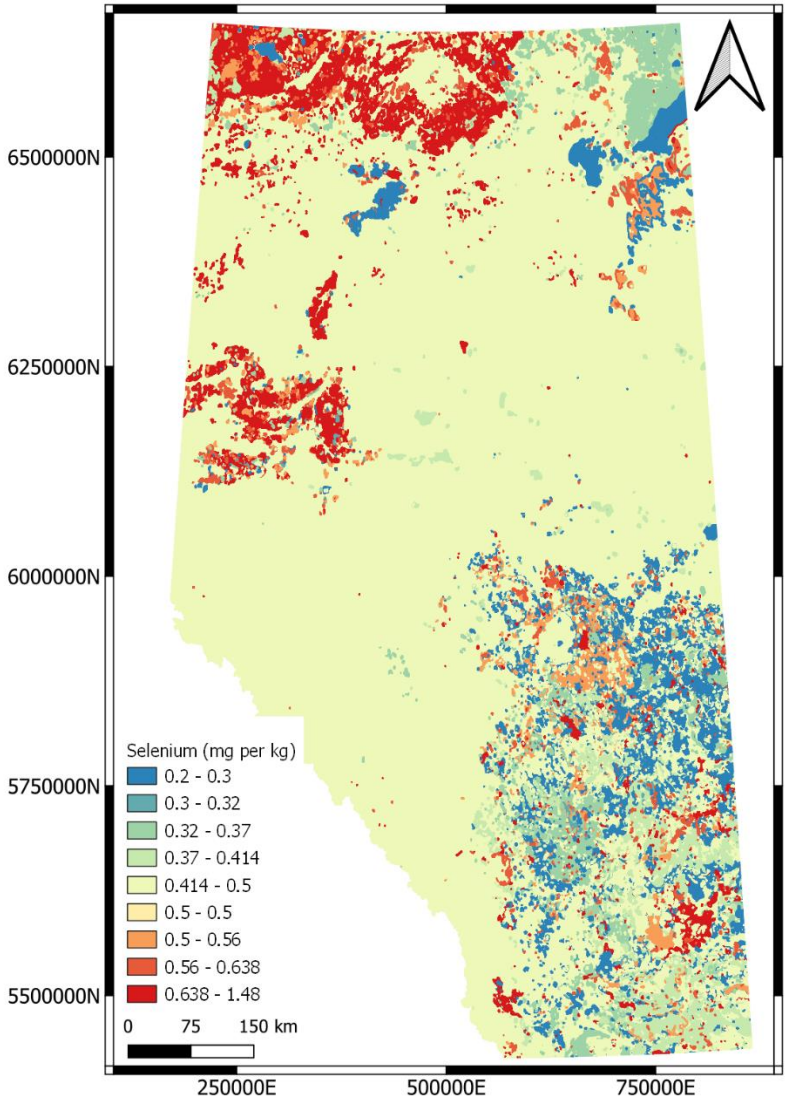


Figure 13. Mean selenium (mg kg $^{-1}$) per polygon.

3.5 SOIL PROPERTY EXTRAPOLATION

Only a small percentage of the polygons had soil observation data present within the polygon. For the salinity polygons 845 out of 16,473 total polygons have direct soil observation data. For the metals data, 316 polygons out of 16,473 total polygons have direct soil observation data (Table 3). That equals approximately 5% of polygons having direct salinity data, and 2% having metals data. Most polygons have data extrapolated from the polygons with observations data based on the distribution of probabilities for the EC and SAR class predictions. Figure 14 indicates which polygons have extrapolated salinity data, and Figure 15 indicates which polygons have extrapolated metals data.

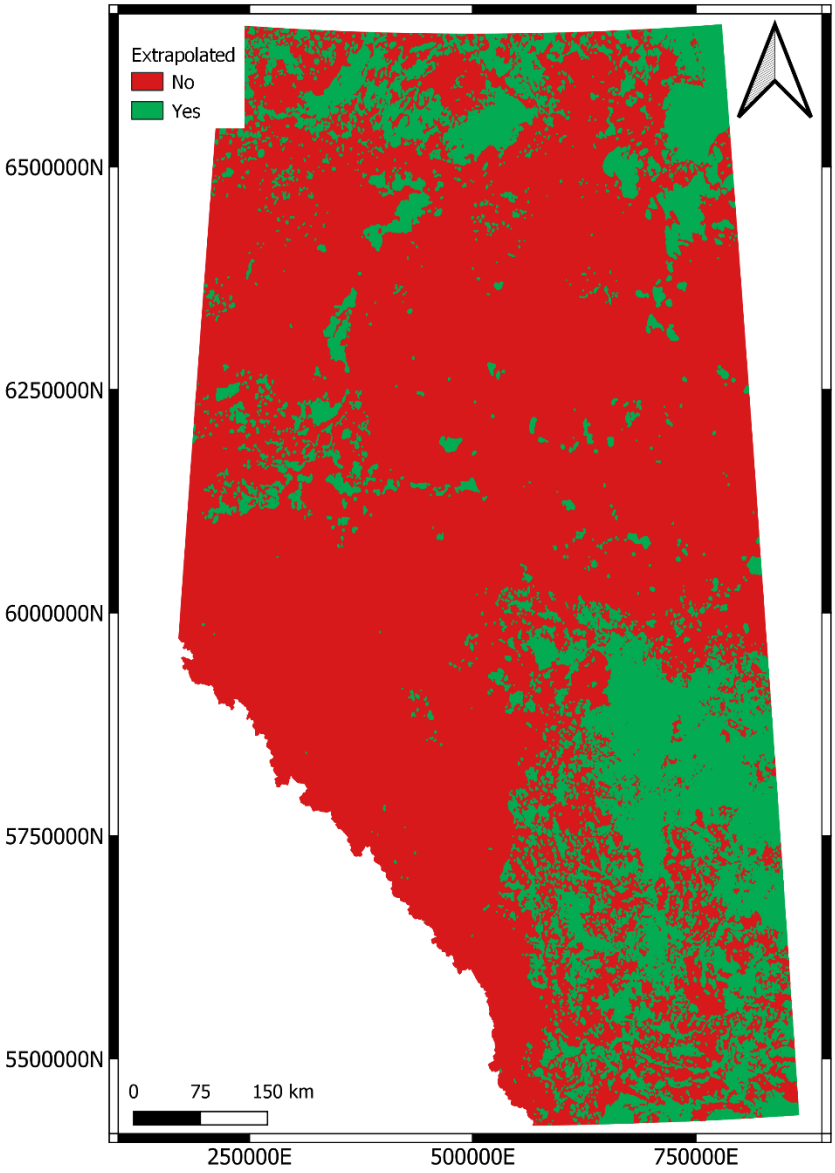


Figure 14. Map outlining which polygons have extrapolated salinity data. Red areas have soil observation data within the polygon, and green areas have extrapolated data.

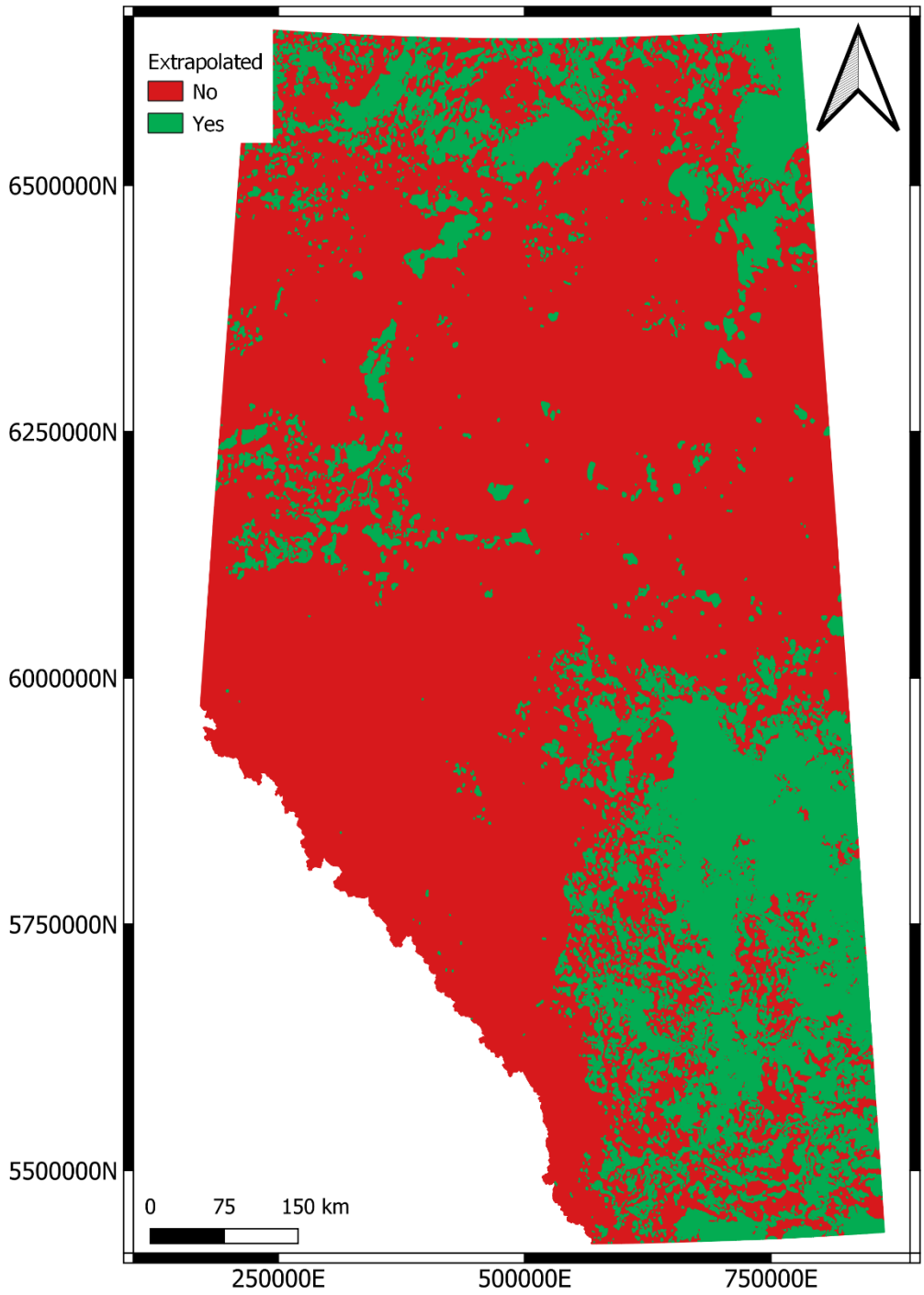


Figure 15. Map outlining which polygons have extrapolated metal data. Red areas have soil observation data within the polygon, and green areas have extrapolated data.

3.6 UNCERTAINTY ANALYSIS

Prediction certainty was inversely related to salinity, with areas of higher salinity having more uncertainty. This is because as salinity increased the model would have more decision trees spreading predictions across fair, poor, and unsuitable classes. Polygons in the good category tended to have a high proportion of decision trees in the random forest consistently predicting good. Prediction certainty

results are illustrated in Figure 16. As the values were scaled between zero and one, the values indicate the relative certainty. Values of one are the polygons with the highest certainty and values of zero have the lowest certainty.

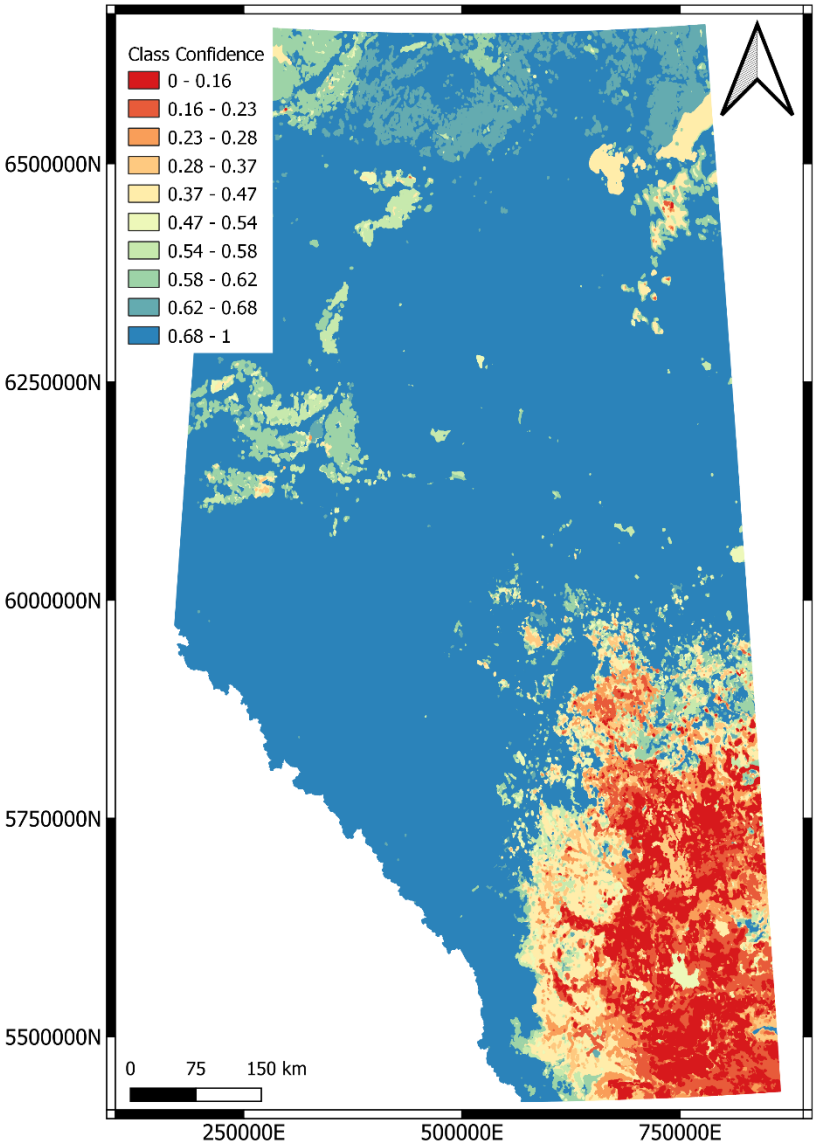


Figure 16. Map of polygon salinity prediction confidence. Values of one are the polygons with the highest certainty and values of zero have the lowest certainty.

3.7 FIELD SAMPLING PLAN

The target polygons for salt and metal sampling in SE Alberta are illustrated in Figure 17. Wellsites with sampling locations are illustrated by the white dots, with 10 sampling locations at each of the 30 selected wellsites. Overall, this stratified random sampling plan has 300 sample locations in total.

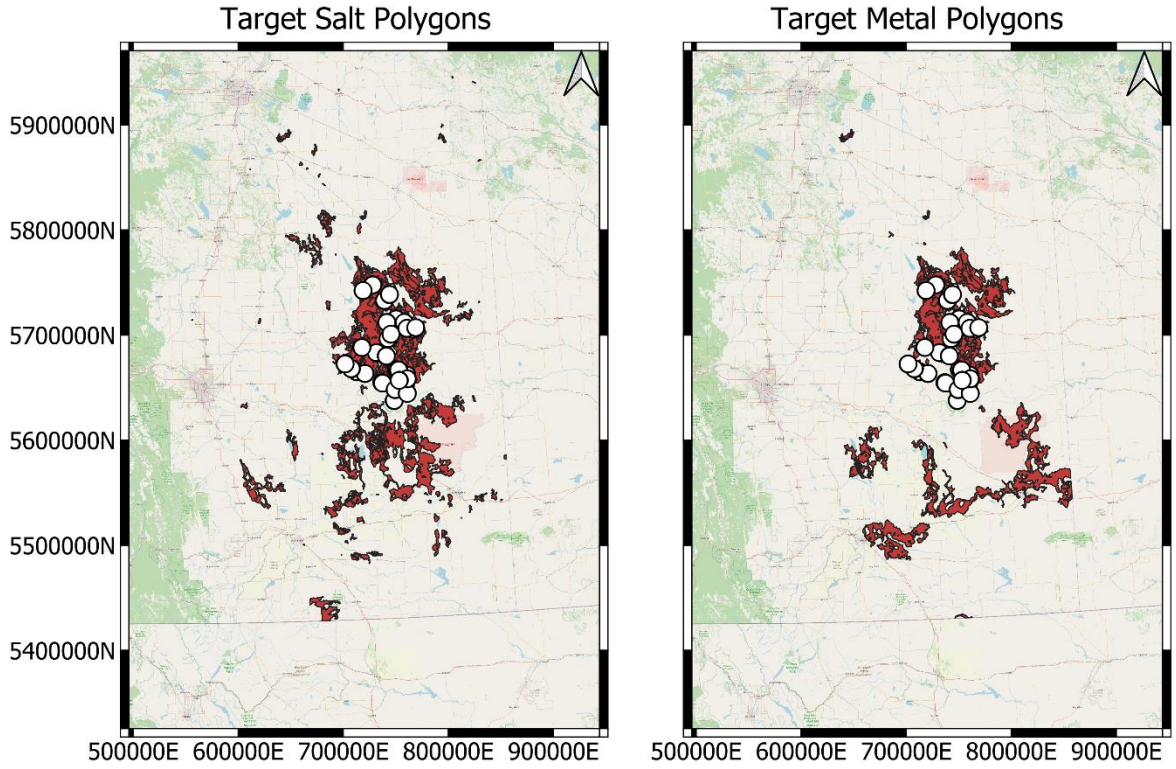


Figure 17. Map illustrating target polygons in red and sample sites in white for field sampling to improve overall model performance for both salt and metal polygons.

3.8 SYSTEM DISTRIBUTION AND COMMUNICATION PLAN

The communication and distribution will happen over time based on the results of this project. The objective of the communication and distribution plan is to create or increase awareness about the outcome of this project, educate the public, engage stakeholders and encourage participation and feedback, and practitioners or their clients to be providing data on an ongoing basis to update the database and models. The key message will be that a tool has been developed and is available to assist industry and government in environmental management of contaminated sites. Communication channels will be through the following:

1. Webinars and Lunch and Learn: We would discuss with the Environmental Services Association of Alberta (ESAA) or Canadian Conservation and Land Management (CCLM) to host a webinar or Canadian Land Reclamation Association (CLRA) for a lunch and learn. We would ask some of the organizations to promote the event and send it out on their email list. The plan is to end up with as many people as possible participating in the webinar or lunch and learn, and for those that are unable to attend to have access to the video presentation and the slide deck to check out the tool.
2. Presentations at workshops/conferences: We plan to make presentations at CLRA and the Remediation Technologies (REMTECH) symposium to promote the availability of the tool.

3. Webpage: A dedicated project webpage hosted at InnoTech's website will be created with updates and access to the tool, including reports and resources.

After the execution of the field sampling program (Phase 4) to fill in data gaps identified during the current phase of this project, and to complete further testing and validation of the model, all of the data will be integrated into a final predictive soil mapping model to predict updated soil property values. The resulting spatial data will be integrated into the web platform and will be publicly available.

4.0 CONCLUSIONS

Throughout the process of collecting data from data providers several challenges were identified. Important metadata items – including units of measure and analytical methods used – were sometimes not provided. This lengthened the time required for the data collection phase. Collecting coordinates for each individual data point leads to increased predictive power in analysis of patterns and trends in soil chemistry datasets. The nine salinity parameters (pH, EC, SAR, chloride, sulphate, calcium, magnesium, potassium, and sodium) and 16 metals parameters (antimony, arsenic, beryllium, cadmium, chromium, cobalt, copper, lead, mercury, molybdenum, nickel, selenium, thallium, uranium, vanadium, and zinc) chosen for the ABSQS are the most regularly reported across salinity and metals analytical packages found in the provided datasets. The data analysis workflow developed to identify background data records—so that impacted data records could be removed from the final ABSQS database—provided stable and replicable results. Of the 224,902 salinity and 23,224 metals data records in the master dataset, 151,542 salinity and 18,513 metals data records were identified as being representative of background.

Regression modelling was not successful for predictive soil mapping due to the lack of accurate coordinates requiring coarser spatial resolution mapping. Mapping soil classes was more successful with subsoil salinity class accuracy being 0.69. To ensure that eventual decisions with the mapping are based on direct field data, the classes were used just for creating polygons. Direct field observation data was used to summarize salinity and metal data for each polygon, with those polygons not having points having extrapolated data. Only a small number of the total polygons had soil observation data present.

5.0 RECOMMENDATIONS

Based on the results of the salinity data analysis, the following are recommended:

- The ABSQS should be completed by continuing with Phase 4 which includes developing and executing a field sampling program to fill in data gaps.
- As a significant amount of the data provided for the ABSQS came directly from analytical laboratories, an attempt should be made to add geospatial coordinates to the list of metadata items included in lab databases going forward.
- Future phases or similar projects should allot additional time for the data collection phase.
- Although a clear template for data formatting was provided, future phases or similar projects should provide additional guidance on data and metadata requirements (i.e., must-haves vs. nice-to-haves).
- Datasets with geospatial coordinates for individual data records should continue to be solicited opportunistically.
- The system will benefit from future incorporation of more site data over time.
- Regression modelling may give useful results if coordinate data is collected with laboratory data. All polygon data was provided accompanying this report as .gpkg files.

6.0 REFERENCES

Alberta Energy Regulator. (2024). *ST37: List of Wells in Alberta.* Retrieved from https://www1.aer.ca/ProductCatalogue/10.html

Alberta Environment and Parks. (2022). *Alberta Tier 1 soil and groundwater remediation guidelines.* Government of Alberta. Retrieved from https://open.alberta.ca/publications/1926-6243

Alberta Environmental Sciences Division. (2001). *Salt contamination assessment & remediation guidelines.* Alberta Environment, Environmental Service, Environmental Sciences Division.

Beygelzimer, A. K. (2024). *_FNN: Fast Nearest Neighbor Search Alogrithms and Applications_(R package version 1.1.4).*

Equilibrium Environmental. (2020). *Subsoil Salinity Tool Version 3.0 Manual.* https://www.alberta.ca/system/files/custom_downloaded_images/aep-subsoil-soil-salinity-user-manual.pdf.

Shelby-James, N., & Fuellbrandt, P. (2022). *The Alberta background soil quality system—Phase 1 (20-RRRC-11) [Revised scope of work].* InnoTech Alberta Inc. and Statvis Analytics Inc.

Wright, M. N., & Ziegler, A. (2017). ranger : A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software, 77(1).* Retrieved from https://doi.org/10.18637/jss.v077.i01