statvis
Trust
Through
Data

Alberta Background Soil Quality System
Supporting Technical Document for Identification of Background Metals Data Records and Ranges
Phase 1: Prototype for Pilot Area

Report prepared for
InnoTech Alberta

February 5, 2023

Table of Contents

1. Introduction

Phase 1 of the Alberta Background Soil Quality System Project (ABSQS) is funded 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). The objective of the ABSQS is to develop a database of background metals and salinity parameters in the Province of Alberta for the purpose of decreasing the cost and time required to identify and remediate valid contaminants of potential concern (COPCs) on contaminated sites. The ABSQS is currently in Phase 1 which involves developing and testing the database for a pilot area. Acquisition of high-quality soils data is key to the ABSQS's overall success and to that end, owners of contaminated sites provided data from their sites under confidential data sharing agreements with InnoTech.

2. Background, Objectives, and Scope

2.1 Background

Salinity and metals parameters are some of the most common naturally elevated parameters in Alberta. 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 the Province of Alberta.

2.2 Objectives

Statvis Analytics Inc. (Statvis) was contracted by InnoTech to harmonize and clean the data, as well as develop a workflow to remove impacted soil data records leaving only background soil data records in the ABSQS database. The objective of the overall project is to work collaboratively with numerous actual and potential users of background soil data to develop the ABSQS. The ABSQS is intended to be used as a resource to assist industry and government in environmental management of contaminated sites. The objective of the current phase (Phase 1) of the project is to create a prototype version of the ABSQS for a pilot area. The pilot area was chosen based on the density of data provided in the various datasets. Starting with a pilot area vs. the full provincial scale is advantageous because:

- it allows the project team to test workflows,
- ensures stakeholder feedback can be incorporated before expanding to a provincial scale, and
- builds relationships with both data providers and users of the system.

In subsequent project phases, predictive mapping technologies will be applied to this dataset to create relevant spatial predictions of salinity and metals values across Alberta.

2.3 Scope of Work

For Phase 1 of the ABSQS, Statvis is responsible for compiling, harmonizing, and cleaning soil salinity and metals data into a geodatabase for subsequent analysis and use in predictive mapping platforms.

This document describes the workflow followed to complete these objectives for metals data records within the initial pilot area defined in Figure 1. A separate report was prepared describing the salinity workflow[a].

[a] Statvis Analytics Inc., 2023. Alberta Background Soil Quality System Supporting Technical Document for Identification of Background Salinity Data Records and Ranges Phase 1: Prototype for Pilot Area

Metals parameters of interest comprised antimony, arsenic, barium (non-barite), barite-barium, beryllium, boron (saturated paste), cadmium, chromium (hexavalent), chromium (total), cobalt, copper, lead, mercury, molybdenum, nickel, selenium, silver, thallium, tin, uranium, vanadium, and zinc. These 22 parameters were selected as they have regulatory numeric guidelines in the Alberta Tier 1 Soil and Groundwater Remediation Guidelines[b] (Tier 1) and as a result are the most commonly analyzed parameters by analytical laboratories. One reason these parameters are regulated and commonly analyzed is that they have potential to cause adverse effects in human and ecological receptors at higher concentrations.

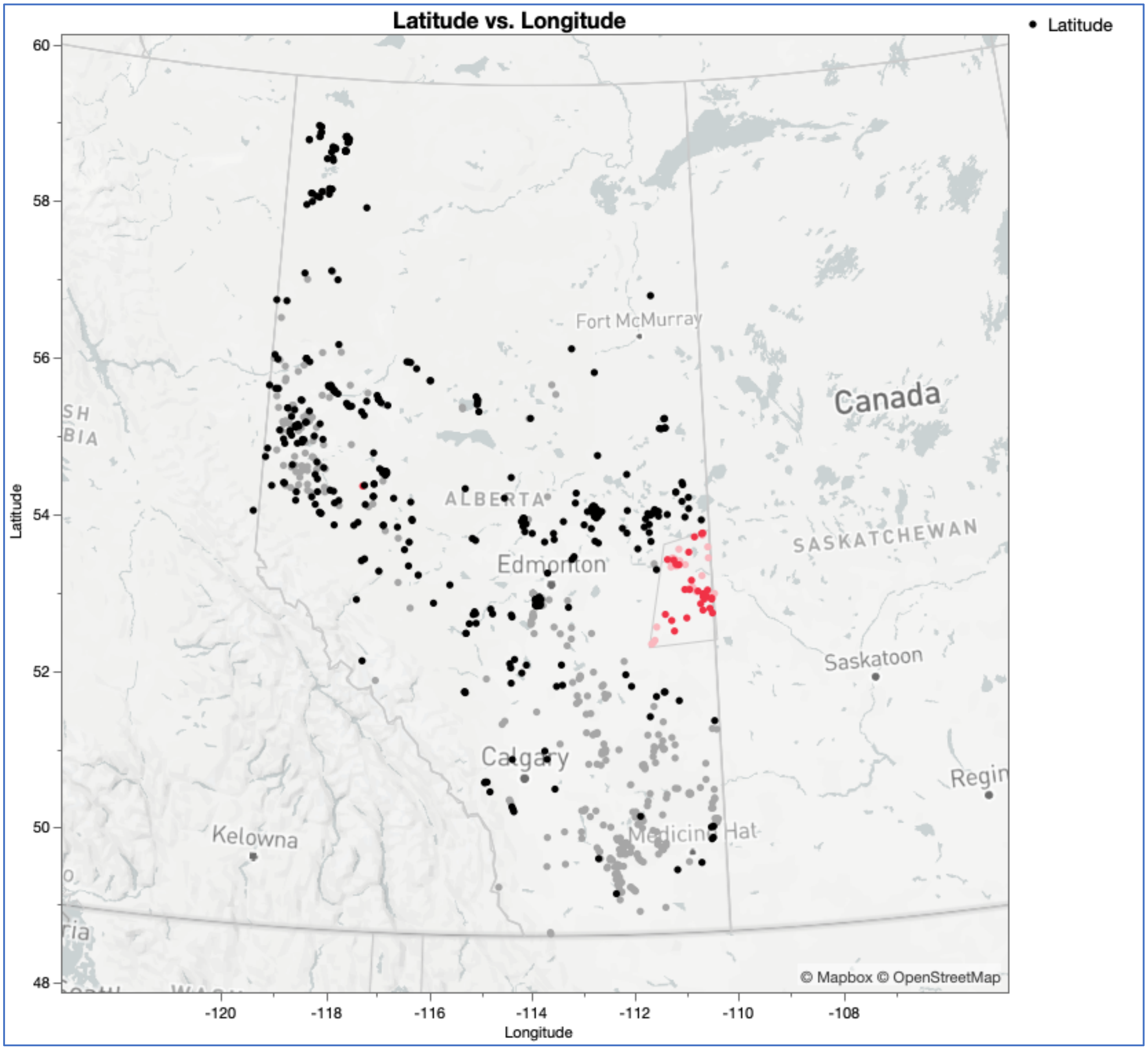

Figure 1. Full metals dataset in black and grey and pilot area dataset in red and pink.

[b] Alberta Environment and Protected Areas, 2023. Alberta Tier 1 Soil and Groundwater Remediation Guidelines.

3. 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. The chosen workflow proved the most effective for separating data records with anthropogenic influence from data records representative of background conditions.

3.1 Data Compilation

During the first eight months of the project (i.e., August 2021 to April 2022), Statvis engaged with data providers 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 protected as 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, Statvis engaged with the data provider, and the dataset was either re-exported to correct the issue or information was provided so that Statvis could manually correct the issue. Over 2,700 individual data files from eight different data owners were received for the pilot area.

3.2 Data Harmonization

Harmonization of datasets was then carried out. Harmonization is the process of 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 pilot area comprising 2,078 soil data records with one or more metal reported.

3.3 Data Cleaning

The master dataset was cleaned to remove incorrect, erroneous, or duplicated data as well as parameters that were reported inconsistently to prepare an analysis-ready dataset. Barite-barium, non-barite barium, boron, hexavalent chromium, and uranium were removed due to inconsistent reporting. Tin and silver had poor levels of detectability and were also removed. This left 15 metals parameters in the analysis ready dataset—antimony, arsenic, beryllium, cadmium, chromium (total), cobalt, copper, lead, mercury, molybdenum, nickel, selenium, thallium, vanadium, and zinc. Remaining non-detects where then imputed using multiplicative lognormal replacement – robust estimates. The analysis-ready dataset contained 1,405 samples with 15 metals available for statistical analysis.

3.4 Data Exploration and Dimensionality Reduction

Statvis used hierarchical cluster analysis (HCA) 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. During exploration of the analysis-ready dataset, selenium performed poorly in correlation tests and provided little predictive power. It was removed from decision-making processes, however, was included in the final concentration distributions.

Clustering in general is a method of statistical analysis that groups 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. See Figure 2 for an example HCA dendrogram. As distance from individual data records increases dendrogram lines become longer showing more dissimilarity between data records. In the example shown in Figure 2 three lines have been drawn bisecting the dendrogram, labeled y1, y2, and y3 respectively, to 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 the clusters have only one data record in them.

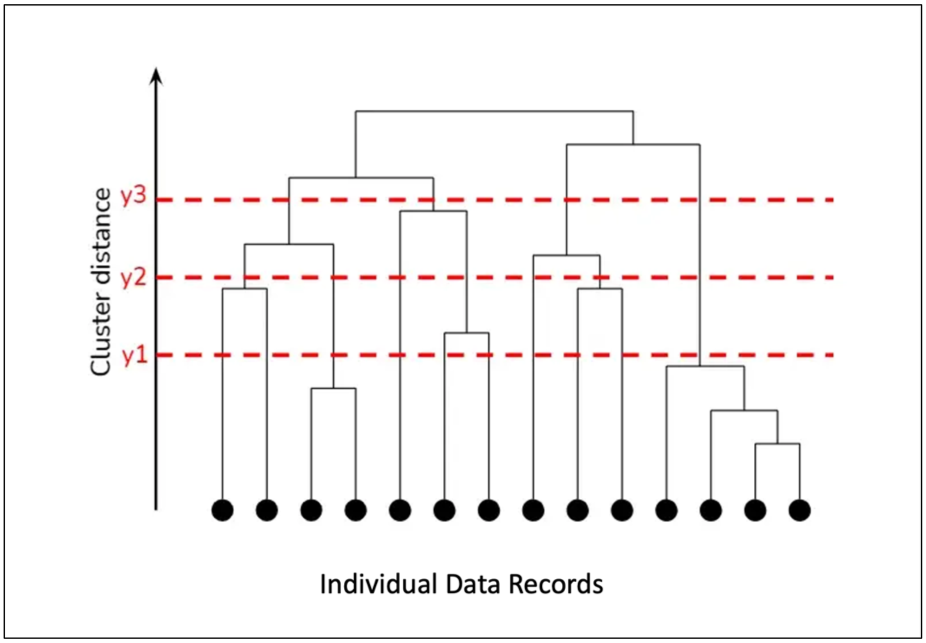

Figure 2. Example HCA dendrogram with bisecting lines to show various levels of granularity.

The goal of exploring the ABSQS metals 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.

3.5 Establishing an Ideal Background Dataset

To ensure a high degree of confidence in removal of impacted data records from the master background dataset for the pilot area, a conservative approach was used to select the ideal background dataset. In order for consistency with the provincial regulatory model, samples were compared to the applicable Tier 1 guidelines. Agricultural land use was selected as it is the most stringent and includes the most receptors and pathways of any land use. Of the 1,405 samples in the analysis-ready dataset for the pilot area, only 72 exceed one or more Tier 1 agricultural guideline. For the sake of conservatism these 72 samples were excluded from the ideal background dataset.

Box Cox normalization was then applied to the analysis-ready dataset and outliers were assessed. High and low outliers from the Box Cox normalized dataset were removed to create a dataset with a normal distribution. This removed 258 samples from the ideal background dataset.

After the above work was completed, the ideal background dataset contained 1,075 data records. The ideal background dataset was then explored to identify definitive patterns present.

3.6 Identifying Background Metals Patterns

An HCA dendrogram was completed for the ideal background dataset using percent normalized data. The HCA also included a heat map of the 15 metals parameters in the analysis-ready dataset (antimony, arsenic, beryllium, cadmium, chromium (total), cobalt, copper, lead, mercury, molybdenum, nickel, selenium, thallium, vanadium, and zinc). The dendrogram was explored at varying degrees of granularity to determine where relationships between metals parameters changed meaningfully. An example of a meaningful difference between clusters would be where data records in one cluster were dominated by different metals than those in an adjacent cluster.

3.7 Applying Background Patterns to the Full Pilot Area Dataset

Larger and smaller clusters were examined and explored to identify boundaries and ratios for metals found in background soils in the pilot area. Boundary conditions for the 15 metals parameters (antimony, arsenic, beryllium, cadmium, chromium (total), cobalt, copper, lead, mercury, molybdenum, nickel, selenium, thallium, vanadium, and zinc) were defined for each background cluster in the ideal background dataset. Boundary conditions were defined using the minimum and maximum percent contribution of each of the 15 metals parameters in each background cluster from the ideal background dataset. To be designated as belonging to a background cluster, a data record had to have measured values of all 15 metals parameters within the minimum and maximum boundary conditions set for a particular cluster. The analysis-ready dataset of 1,405 records was compared to these boundary conditions and records that did not fit into one or more background cluster were removed from the final dataset as they were considered to show anthropogenic influence.

4. Results

The HCA dendrogram and heatmap for antimony, arsenic, beryllium, cadmium, chromium (total), cobalt, copper, lead, mercury, molybdenum, nickel, selenium, thallium, vanadium, and zinc generated from the ideal background dataset are shown in Figure 3. There are two main HCA clusters, with cluster one being separated into three smaller clusters—coloured red, green, and blue respectively. Cluster two is coloured brown. These four clusters represent different background patterns present in the pilot area.

Based on the four background patterns identified in the 1,075 data records included in the ideal background dataset, 1,183 data records of the 1,405 data records in the entire pilot area dataset were identified as representative of background.

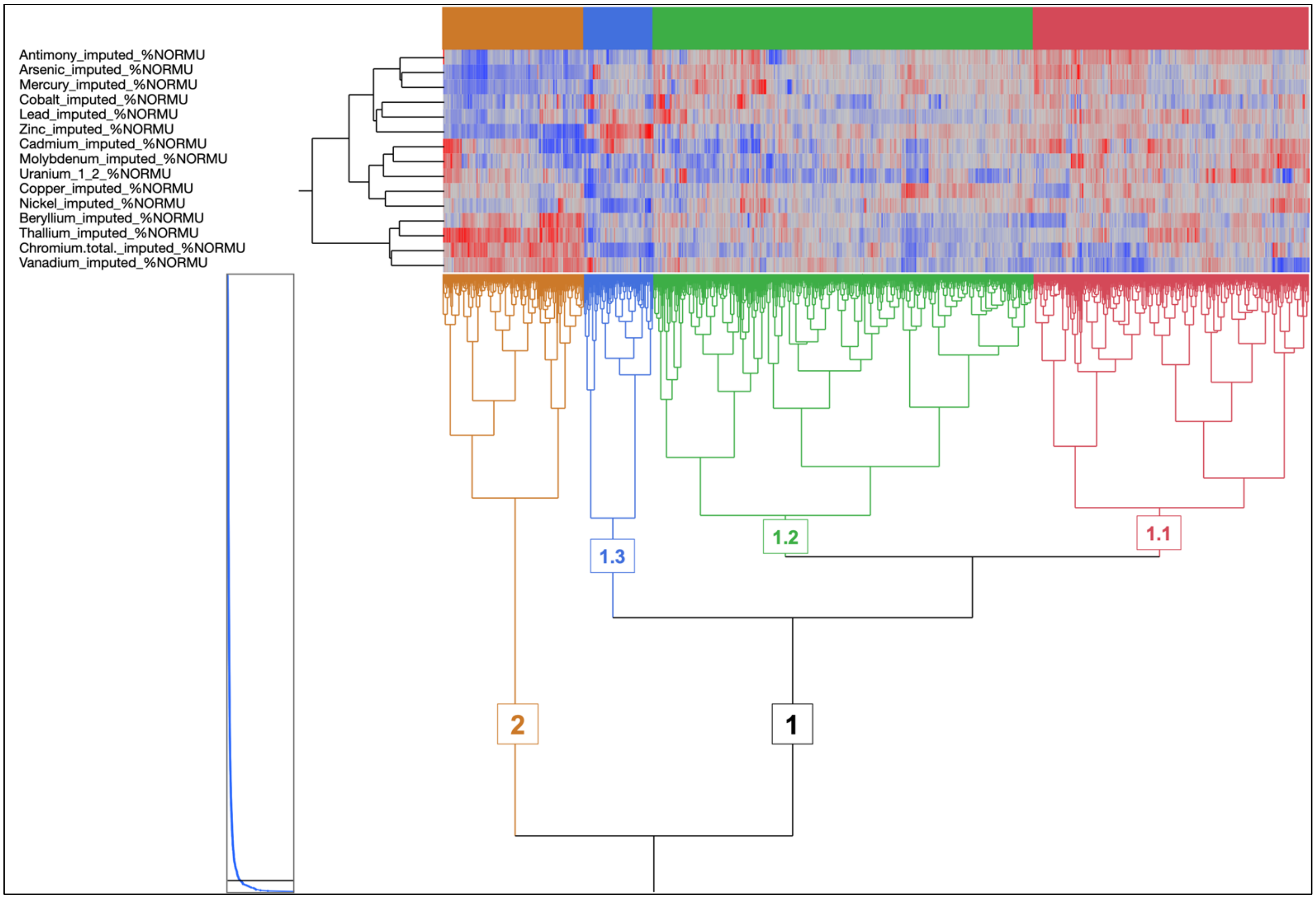

Figure 3. HCA dendrogram with four main coloured clusters representing background fingerprints in the pilot area.

5. Conclusions

Measured values of metals parameters in background may exceed regulatory guidelines. The objectives of the current scope of work defined for Statvis in Phase 1 of the ABSQS were to harmonize and clean the data, as well as develop a workflow to remove impacted data records leaving only background data records in the ABSQS database. These objectives were achieved within the defined scope of work.

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 15 metals parameters chosen for the pilot area of the ABSQS (antimony, arsenic, beryllium, cadmium, chromium (total), cobalt, copper, lead, mercury, molybdenum, nickel, selenium, thallium, vanadium, and zinc) are the most regularly reported across the metals analytical packages found in the provided datasets.

The metals analysis workflow derived 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 1,405 data records in the pilot area dataset, 1,183 data records were identified as background.

6. Recommendations

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

- The creation of a prototype of the ABSQS for a pilot area (Phase 1) was successful and the prototype should be shared with stakeholders for testing and feedback.
- The ABSQS should be expanded to the full provincial scale (i.e., carry on to Phases 2 and 3).
- The 22 parameters of interest should be carried forward to future phases of the ABSQS, however, the final parameters retained in the ABSQS database should be determined based on metals parameters remaining after compilation, harmonization, cleaning, and initial exploration of the full provincial-scale dataset.
- 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.

7. Statement of Limitations

This report was prepared for the exclusive use of the client identified herein. The report, which specifically includes all tables, figures, and appendices, is based on data and information collected or provided during the work conducted by Statvis Analytics Inc. and is based solely on the conditions of the site and data obtained by Statvis Analytics Inc. as described in this report. Information and data provided to Statvis Analytics Inc. has not been independently verified.

The services performed as described in this report were conducted in a manner consistent with the level of care and skill normally exercised by other environmental professionals currently practicing under similar conditions.

Any use a third party makes of this report, or any reliance on or decisions to be made based on it, are the responsibilities of such third parties. Statvis Analytics Inc. accepts no responsibility for damages, if any, suffered by any third party as a result of decisions made or actions based on this report.

The content of this report is based on data and information collected or provided during our assessment, our present understanding of site conditions and our professional judgement in light of such information at the time of this report. This report provides a professional opinion and therefore no warranty is expressed, implied, or made as to the conclusions and recommendations offered in this report. This report does not provide a legal opinion regarding compliance with applicable laws. It should be noted that regulatory statutes and the interpretation of regulatory statutes are subject to change. The findings and conclusions of this report are valid only as of the date of this report. If new information is discovered in future work, Statvis Analytics Inc. should be engaged to re-evaluate the conclusions of this report and provide amendments as required.

8. Closure

We trust that the information presented in this report meets your current requirements. Should you have any questions or require additional information, please do not hesitate to contact the undersigned.

Sincerely,

Statvis Analytics Inc.

Paul

Court Sandau

Paul Fuellbrandt, P.Ag., PMP
Principal Environmental Scientist

Court Sandau, PhD
Principal Chemist