ALBERTA BACKGROUND SOIL QUALITY SYSTEM PROJECT: WORKSHOP SUMMARY

Natalie Shelby-James, InnoTech Alberta Inc.
Sarah Thacker, InnoTech Alberta Inc.
Chris Powter, Enviro Q&A Services
Paul Fuellbrandt, Statvis Analytics
Tomislav Hengl, EnvirometriX
Leandro Parente, EnvirometriX

REPORT PREPARED FOR
PETROLEUM TECHNOLOGY ALLIANCE CANADA
Reclamation Remediation Research Committee

20- RRRC-11

January 6, 2022

CITATION

This report may be cited as:

Shelby-James, N., S. Thacker, C.B. Powter, P. Fuellbrandt, T. Hengl and L. Parente, 2021. Alberta Background Soil Quality System Project: Workshop Summary. Prepared for the Petroleum Technology Alliance of Canada, Calgary, Alberta. Report 20-RRRC-11. 69 pp.

ACKNOWLEDGMENTS

The authors would like to acknowledge the financial contributions provided by the Alberta Upstream Research Fund (AUPRF) Program as well as the guidance and support provided by the Project Technical Steering Committee members (listed in Table 1) and Project Technical Champion (Rick Rohl, Arc Resources Ltd.). The contributions from all the working session participants are greatly appreciated.

EXECUTIVE SUMMARY

As part of the environmental regulatory framework to minimize risk to receptors, concentrations of chemicals in soil or water exceeding regulatory guidelines that can be attributed to industrial activities at a site require remediation and/or monitoring. This process is complicated by the fact that various chemical parameters are naturally elevated in Alberta, with concentrations that exceed the generic *Alberta Tier 1 Soil and Groundwater Remediation Guidelines*. Where this occurs, environmental professionals must prove to the satisfaction of the applicable regulatory body that elevated concentrations are natural and not the result of industrial activities.

The challenge of proving that elevated concentrations are of natural origin has been identified by industry and practitioners as a root cause of cost uncertainty, multi-year timelines, challenges in reaching regulatory closure, and in some cases, unnecessary and unsustainable monitoring, and remediation efforts.

The objective of this project is to work collaboratively with soil data users to develop the Alberta Background Soil Quality System (ABSQS) to be used as a tool to assist industry and government in environmental management. As the first step in the project, 64 people from government, industry, consulting, analytical labs and academia participated in a workshop designed to:

- Increase awareness and understanding of the Alberta Background Soil Quality System Project,
- Collect and incorporate feedback from workshop attendees related to barriers to use / technical acceptance and system performance expectations, and,
- Promote collaboration opportunities

There was considerable interest in the potential for the ABSQS to support regulatory compliance and participants did not identify any "show stoppers". Several questions were raised, and there was a desire for more information, much of which can be addressed by providing progress updates and specifically by providing more detail on the contents of the database, the geographic area of the pilot, examples of the background levels in polygons, and maps of the polygons in the pilot area.

Participants could help ensure project success by:

- Providing suggestions for, and/or access to, soil chemistry datasets (preferably georeferenced),
- Providing suggestions for, and/or access to, covariate data layers, and,
- Providing information on which soil chemical parameters are most often found to have naturally elevated background levels in various regions of the province

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

GLOSSARY

Artifact

A subset of outliers which are true errors in a data set that should be removed before further analysis.

Background

Concentrations of salinity and metals that are naturally occurring and unrelated to the discharge of pollutants or hazardous substances, or other anthropogenic activities.

Background Fingerprint

A visual depiction of the relationship between multiple chemicals representative of background conditions.

Bagging

A machine learning ensemble meta-algorithm designed to improve the stability and accuracy of machine learning algorithms used in statistical classification and regression.

Boosting

In machine learning, boosting is an ensemble meta-algorithm for primarily reducing bias, and variance.

Clustering

A method of statistical analysis that groups samples in such a way that they are more like other samples within the same group than they are to samples in other groups.

Data Cube

A complete, consistent, and current analysis-ready data stack consisting of multiple layers of relevant datasets adjusted to the same scale. Relevant data layers for the ABSQS data cube might include soil quality parameters, topography and landforms, potential industrial sources (e.g., well and pipeline location data), etc.

Data Sharing and Use Agreement

A signed agreement between InnoTech Alberta and a data provider (e.g., an energy company, a consultant, or a laboratory) specifying what data will be provided and how they will be used.

Dimensionality Reduction

A method of transforming data from a large set of variables (i.e., individual salinity and metal chemical parameters) into a smaller set that still contains most of the variability from the larger set.

Ensemble Machine Learning

Ensemble methods are meta-algorithms that combine several machine learning techniques into one predictive model to decrease variance (bagging), bias (boosting), or improve predictions (stacking).

Impacted

Samples with chemical signatures affected by anthropogenic activity plus those that are a mix of background and impacted signatures. Impacted samples are not background samples.

Machine Learning (ML)

Machine learning is a branch of artificial intelligence (AI) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving accuracy (IBM Cloud Education, 2020). Using statistical methods, algorithms are trained to make classifications or predictions, uncovering key insights within data mining projects.

Metadata

Data that provide information about other data; for example, in the ABSQS metadata may include sample location and sampling date.

Outliers

Measurements that are very large or small relative to the rest of the data and are suspected of misrepresenting the population from which they were collected. False outliers are measurements that are very large or small relative to the rest of the data but represent true extreme values of a distribution and indicate more variability in the population than was expected. True outliers (artifacts) are deleted from the dataset.

Parameter

An individual ion from a sample analysis.

Polygon

A defined area of the Province in which the background fingerprint has been defined and is different from the background fingerprints in adjoining polygons.

Predictive Soil Mapping

The development of a numerical or statistical model of the relationship among environmental variables and soil properties, which is then applied to a geographic database to create predictions and a resulting map (Scull et al., 2003).

Sample

A discrete soil sample from a georeferenced location and defined soil depth which has been analyzed for specific parameters.

Stacking

An ensemble machine learning method that combines the predictions from multiple machine learning models on the same dataset, like bagging and boosting. Stacking typically includes fitting of a meta-learner that relates base-learners with the target variable.

t-distributed Stochastic Neighbor Embedding (t-SNE)

A statistical method for visualizing high-dimensional data by giving each datapoint a location in a two or three-dimensional map (van der Maaten, 2021).

Uniform Manifold Approximation and Projection (UMAP)

A tool for presenting high-dimension data in a low-dimension graph (Coenen and Pearce, n.d.).

ACRONYMS

ABSQS	Alberta Background Soil Quality System
AGRASID	Agricultural Regions of Alberta Soil Inventory Database
ARD	Analysis-ready Data
AUPRF	Alberta Upstream Petroleum Research Fund

COG	Cloud Optimized GeoTIFF
DTM	Digital Terrain Model
EC	Electrical Conductivity
EML	Ensemble Machine Learning
ETM	Enhanced Thematic Mapper (Landsat)
EVI	Enhanced Vegetation Index
FIADB	Forest Inventory and Analysis Database
GIS	Geographic Information System
GLAD	Global Land Analysis & Discovery
HCA	Hierarchical Clustering Analysis
LCLU	Land-cover and Land-use
LiDAR	Light Detection and Ranging
LSD	Legal Subdivision
LST	Land Surface Temperature
ML	Machine Learning
MODIS	Moderate Resolution Imaging Spectroradiometer
NIR	Near Infrared
NSDB	National Soil Database
OWA	Orphan Well Association
PAH	Polycyclic Aromatic Hydrocarbons
PCA	Principal Components Analysis
PTAC	Petroleum Technology Alliance Canada
QGIS	Quantum Geographic Information Systems
RRRC	Remediation and Reclamation Research Committee
SAR	Sodium Adsorption Ratio
SOC	Soil Organic Carbon
STAC	SpatioTemporal Asset Catalog
SWIR	Shortwave Infrared
TSC	Technical Steering Committee
t-SNE	t-distributed Stochastic Neighbor Embedding
UMAP	Uniform Manifold Approximation and Projection
UWI	Universal Well Indicator

1.0 BACKGROUND

Where land has been used for industrial purposes, effective and sustainable ecological restoration and return to productive use are key to responsible land stewardship. As part of the environmental regulatory framework to minimize risk to receptors, concentrations of chemicals in soil or water exceeding regulatory guidelines that can be attributed to industrial activities at a site require remediation and/or monitoring. This process is complicated by the fact that various chemical parameters are naturally elevated in Alberta (Millennium EMS Solutions Ltd., 2016), with concentrations that exceed the generic *Alberta Tier 1 Soil and Groundwater Remediation Guidelines* (Tier 1; Alberta Environment and Parks, 2019). Where this occurs, environmental professionals must prove to the satisfaction of the applicable regulatory body that elevated concentrations are natural and not the result of industrial activities (see, for example, Luther (2020)).

The challenge of proving that elevated concentrations are of natural origin has been identified by industry and practitioners as a root cause of cost uncertainty, multi-year timelines, challenges in reaching regulatory closure, and in some cases, unnecessary and unsustainable monitoring and remediation efforts. Many of the backlogged legacy oil and gas wellsites and associated facilities in Alberta are stalled, or are being monitored for longer than necessary, for this reason.

Salinity and certain metals are the most common naturally elevated parameters in Alberta (for example, high naturally-occurring arsenic levels near Spirit River [Dudas, 1987], and high total chromium levels in the Peace River region [Soon and Abboud, 1990], have been documented). Challenges related to these naturally elevated parameters occur in environmental management of active and end-of-life industrial sites, and when responding to unintentional releases during product handling or transportation. If salt and metal parameters are naturally elevated compared with Tier 1 guidelines, environmental professionals can mistake these naturally elevated concentrations for contamination, triggering 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 the need for more effective identification of background salt and metals concentrations as one of their highest priorities (Utting and Heshka, 2019).

To prove that concentrations of one or more parameters are naturally elevated, background samples must be collected, often requiring additional mobilization of equipment and resources once site data have 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 years 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 (formerly pre-disturbance assessments); conservation, operation, and reclamation plans (formerly 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 utilizing standard methods prescribed by the government. If geo-referenced, the data have 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 metal concentrations.

1.1 PROJECT OBJECTIVE

The objective of this project is to work collaboratively with soil data users to develop the Alberta Background Soil Quality System (ABSQS) to be used as a tool to assist industry and government in environmental management. Key activities for the project (Figure 1) are listed below while the benefits of the project to stakeholders are shown in Figure 2.

- Compiling, cleaning, and integrating existing soil salinity and metals data into a geodatabase,
- Analyzing soil point data to fingerprint background and remove impacted samples,
- Developing predictive background soil maps – prototype in a select area of the Province and then scaling up to Provincial coverage, and,
- Developing a web application to house and allow users to interact with the System.

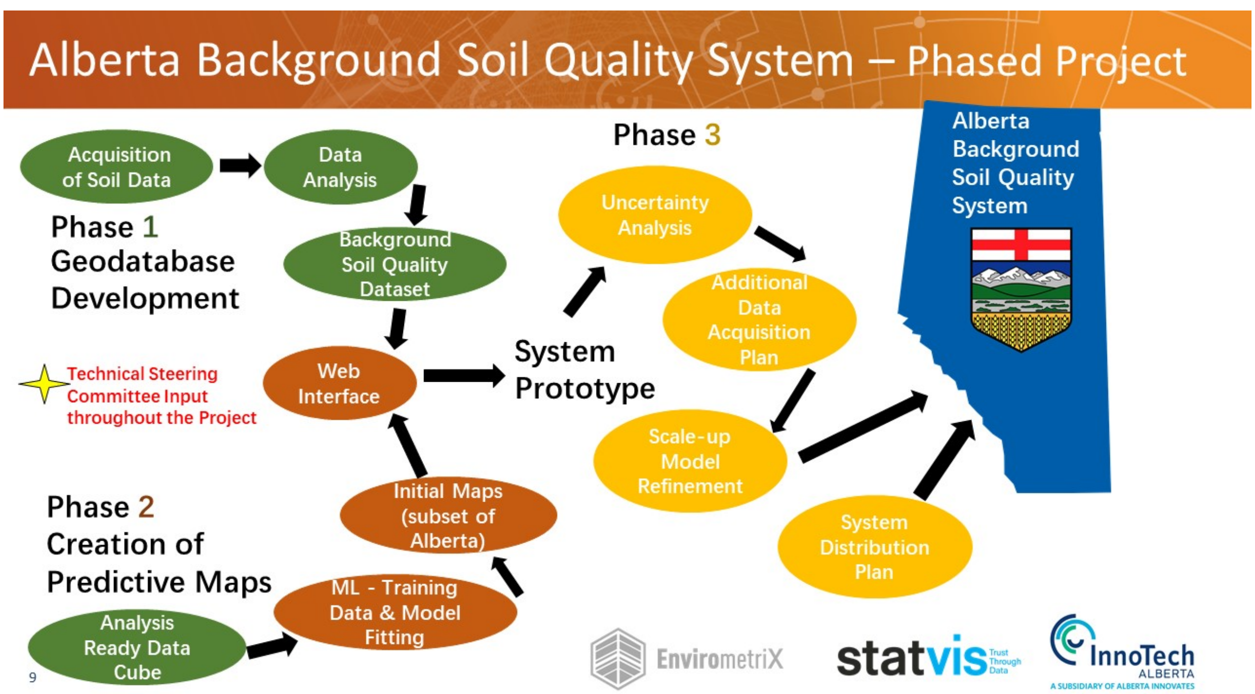


Figure 1. Alberta Background Soil Quality System project phases.

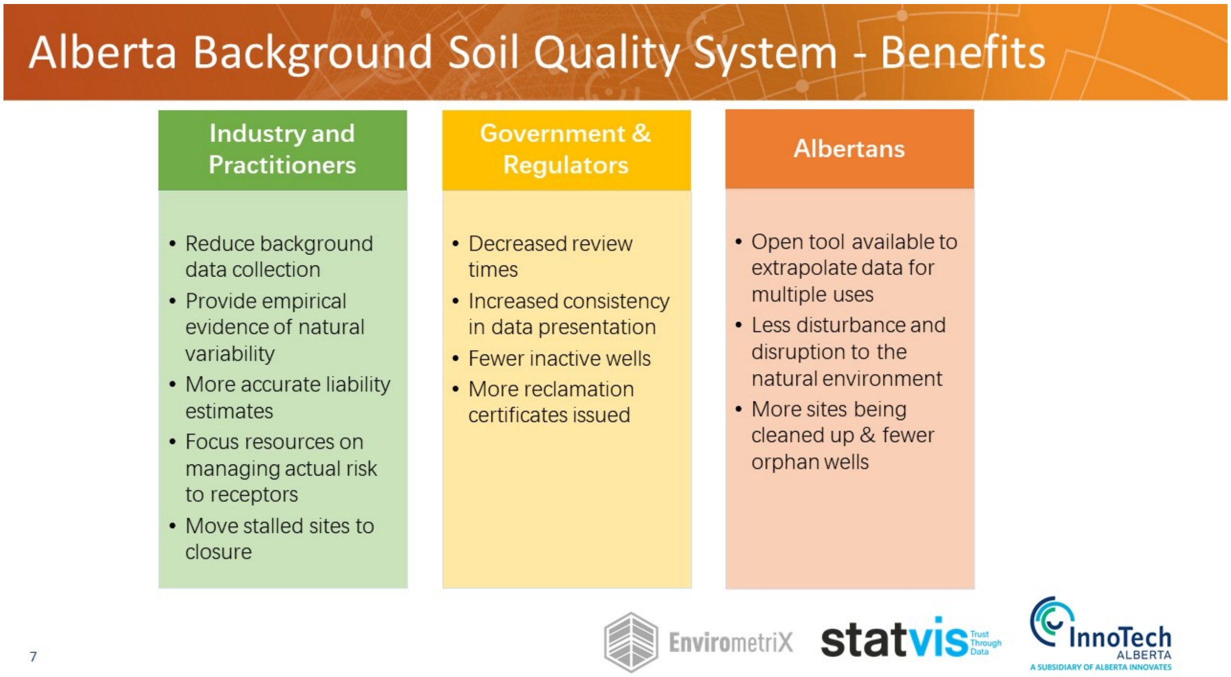


Figure 2. Benefits of the Alberta Background Soil Quality System to stakeholders.

1.2 PROJECT TEAM

The project team consists of Natalie Shelby-James and Sarah Thacker from InnoTech Alberta, Chris Powter from Enviro Q&A Services, Paul Fuellbrandt from Statvis Analytics, and Tom Hengl and Leandro Parente from EnvirometriX. In addition to the team, the project has a Technical Steering Committee (TSC) comprised of industry, regulators, soil data providers and technical advisors. The role of the TSC is to:

- Ensure the voice of the customers / end users is represented (i.e., outputs directly meet users' specific data needs),
- Ensure consistency with regulatory requirements and expectation,
- Provide support with obtaining data inputs, and,
- Provide support to achieving open and accessible data outputs

Membership of the Project TSC is provided in Table 1.

Table 1. Project Technical Steering Committee Membership

Member	Organization	TSC Role
Natalie Shelby-James	InnoTech Alberta	Co-Chair / Project Manager
Sarah Thacker	InnoTech Alberta	Core Project Team & TSC Support
Paul Fuellbrandt	Statvis	Co-Chair / Core Project Team / Technical Lead
Tom Hengl	Predictive Modelling	Core Project Team
Shawn Glessing	PTAC (Cenovus)	Industry Representative
Linda Eastcott	PTAC (Imperial)	Industry Representative
Rick Rohl	PTAC (Arc Resources)	Industry Representative
Sonia Glubish	PTAC (Canadian Natural)	Industry Representative
Kagen Newman	PTAC (Canadian Natural)	Industry Representative (alt)

Member	Organization	TSC Role
Rob Thompson	PTAC (Orphan Well Association)	Industry Representative
Sara Blacklaws	Alberta Energy Regulator	Alberta Government Representative
Shane Patterson	Alberta Environment and Parks	Alberta Government Representative
David Spiess	Alberta Agriculture and Forestry	Alberta Government Representative
Jonas Fen	Saskatchewan Energy and Resources	Saskatchewan Government Representative
Preston Sorenson	University of Saskatchewan	Academia
Tyler Prediger	Matrix Solutions Inc.	Environmental Consultant
Daniel Pollard	Wood	Environmental Consultant
Anthony Knafla	Equilibrium	Environmental Consultant

1.3 WORKSHOP

Invitations for the November 10, 2021, workshop were sent to 85 people in government, industry, industry associations, consulting companies, analytical labs and academic institutions. Sixty-four people, plus the seven team members, participated in the workshop (Appendix A).

The workshop goals were to:

- Increase awareness and understanding of the Alberta Background Soil Quality System Project,
- Collect and incorporate feedback from workshop attendees related to barriers to use / technical acceptance and system performance expectations, and,
- Promote collaboration opportunities

The workshop agenda is provided in Appendix B and the associated presentations are in Appendix C. As part of each discussion session, participants were asked to respond to a short survey to help set the stage for the conversation. Some participants were unable to access the survey tool, so a copy was sent by e-mail to gather additional feedback.

1.4 REPORT STRUCTURE

This report follows the workshop agenda:

- Section 2 – summary of the presentation on the soil database and the question-and-answer session that followed,
- Section 3 – summary of the presentation on the predictive mapping tool and the question-and-answer session that followed,
- Section 4 – key outcomes from the Workshop,
- Section 5 – references and some suggested reading,
- Appendix A – list of workshop attendees,
- Appendix B – workshop agenda, and,
- Appendix C – workshop presentations.

2.0 DEVELOPING THE SOIL DATABASE

Paul Fuellbrandt of Statvis Analytics presented information on the concept and development of the background soil database.

2.1 WORKSHOP PRESENTATION HIGHLIGHTS

Data will be acquired through data sharing agreements between data providers (e.g., energy company, consultant, analytical laboratory) and InnoTech Alberta. Three data tiers are envisioned: Tier 1 data – georeferenced datasets collected after 2015 to ensure standardized data collection and analytical methods; Tier 2 data – older georeferenced datasets that may require additional cleaning (quality assurance / quality control) and have less metadata; and Tier 3 data – non-georeferenced datasets with lower resolution and/or data accuracy (e.g., known general location such as a wellsite (UWI) or a legal subdivision (LSD)).

Data will be cleaned in three steps to develop a single comprehensive database:

- First, the extensive dataset will be reduced to the key components and/or clusters through a series of data dimensionality reduction steps using tools such as principal components analysis (PCA), hierarchical clustering analysis (HCA), t-distributed Stochastic Neighbor Embedding (t-SNE), and/or Uniform Manifold Approximation and Projection (UMAP),
- Second, background sample fingerprints (Figure 3) will be developed and assigned to a map polygon, and,
- All impacted (non-background) site data will be removed.

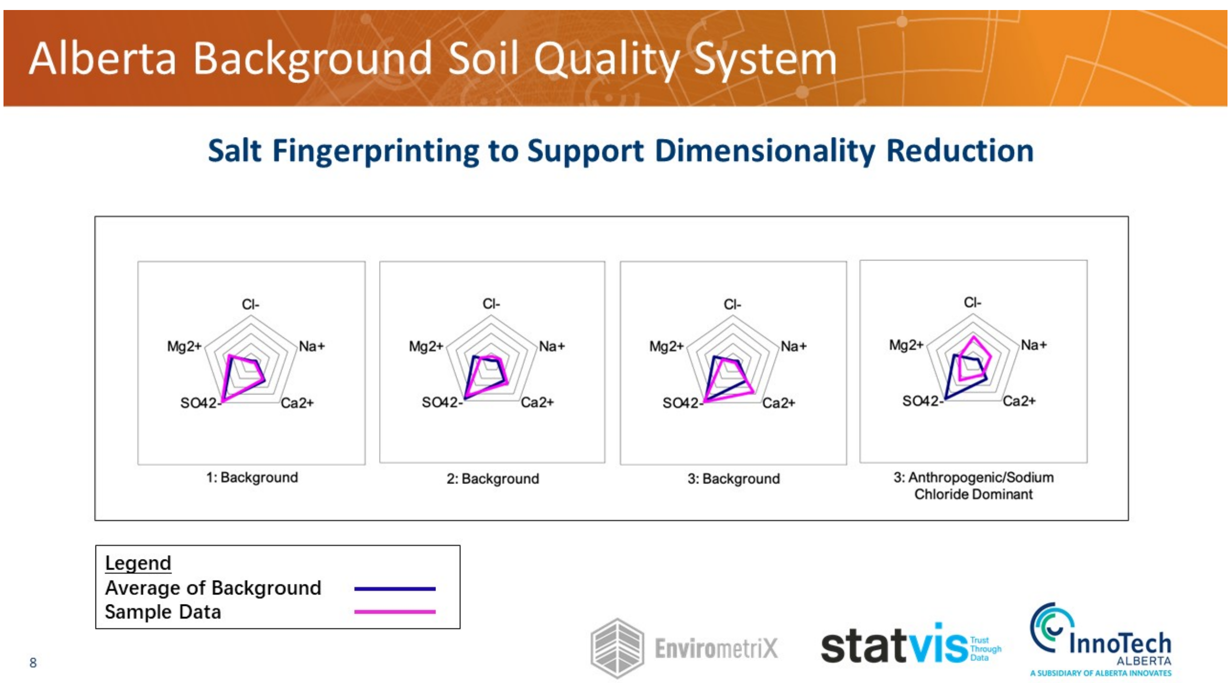


Figure 3. Examples of background and impacted salt fingerprints.

Table 2 summarizes the key differences between the ABSQS background soil database and traditional methods of developing background databases.

Table 2. Summary of the differences between Alberta Background Soil Quality System database and traditional background databases.

Alberta Background Soil Quality System	Traditional Background Database
Background fingerprinting based on pattern recognition not concentration	Pre-determined concentration limits for background
Modern data science, machine learning data dimensionality reduction now widely available	Simple statistics (correlations, etc.) were the only tools widely available until recently
Impacted samples are initially included to learn differences between impacted and background and then are removed	Impacted samples removed based on assumptions, location, experience, etc.
RESULT: Data-driven process removes potential for invisible bias by continuous cross-validation / statistical learning using robust testing	**RESULT:** Potential for missing naturally elevated parameters

During the presentation, participants were asked through an online questionnaire to identify the top three barriers (out of 11 choices) to successful implementation of the soil background quality database. Figure 4 summarizes the responses. Concerns with inconsistent sample depths, ownership of the database and availability of data were most frequently cited.

2.2 DATABASE QUESTION AND ANSWER SESSION

Q: A lot of the data sets will come from disturbed soil profiles. Is this a concern?

A: The patterns of true background will still show, so though it is a consideration, it will be mitigated during the data cleaning process.

Q: In your examples you mentioned several times you had a single sample or soil profile with three subsamples from different depths. Do you expect that there would be separate backgrounds by depth, or is it a location background?

A: There will be different depth backgrounds, but we are not pre-determining what those depths will be. They will be determined using the data driven process that we discussed. It's going to have to be a discussion with the Technical Steering Committee to decide if we take a soil science approach to splitting the depths (i.e., based on soil horizons/pedons) or a regulatory approach (i.e., use the depths specified in existing guidance). We expect topsoil is going to be quite a bit different than subsoil, but we are unsure on how many subsoil layers there will be. We will have to decide where the meaningful breaks are in terms of the depth profiles based on analysis of the dataset. *Post-workshop Note: We can also fit 3D models of target variables where depth would be one of covariates. Tom's examples from the USA data demonstrated how to do it and found that for several geochemicals depth is in the top 5 most important variables.*

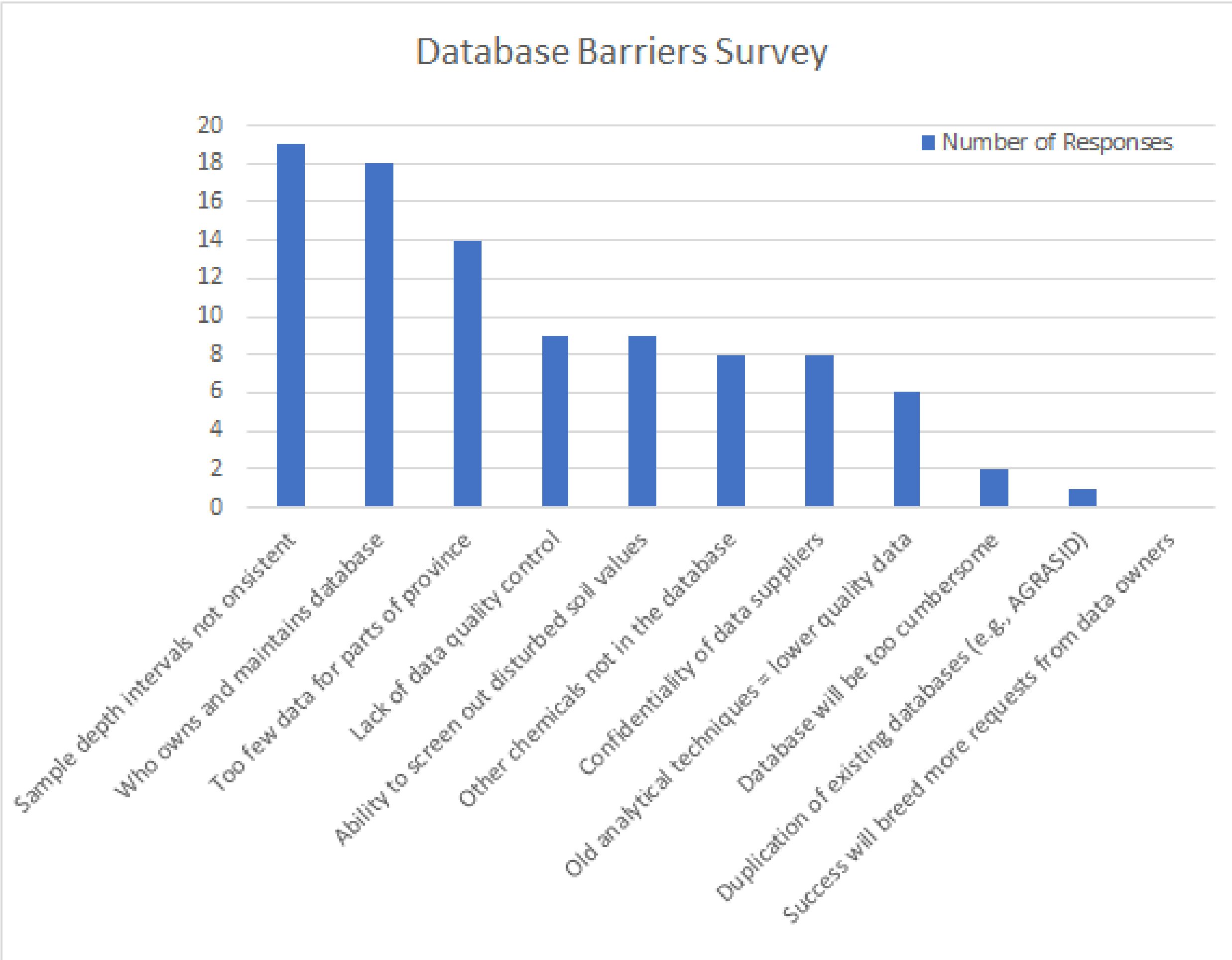


Figure 4. Participant responses to database barriers survey.

Q: How are the polygons in the map going to be determined?

A: It's going to be a data-driven process based on the relationships of soil data (background fingerprints) to the covariate data. We are going to map the fingerprints, and the boundaries will be where the fingerprint starts to change. Information such as the slope, the parent material, and the soil formation process, are all going to influence where the patterns change. Some of these factors can be represented in covariate layers – for example, a digital elevation model can provide slope information.

Q: Does that mean that different polygons throughout the Province could have the same fingerprint?

A: Yes, that's a great observation. We would expect different polygons from different areas to have similar fingerprints if they've got similar soil formation processes and chemistry.

Q: Will the dataset include sites within urban areas?

A: There's a lot more impacts in urban areas that make it harder to find true background, but some background samples may be provided so, they're not being excluded. I think acquisition of urban area samples may be considered lower-level tier data. They are still a very valid set of data. We will consider urban area samples and do have access to the information.

Q: What areas of the province are currently included in the tool or planned? I guess for the smaller area that we're starting testing on.

A: We have an idea of where we may start but we don't have all our data sharing agreements in place yet, so it's premature to say exactly what area we'll start with. I think the idea is to start with an area that is covered by the Agricultural Regions of Alberta Soil Inventory Database (AGRASID) so that we can use AGRASID as a partial validation for the final data set.

Q: Can you run over again the definition of background for us please?

A: In the context of this project, it's salinity and metals at concentrations that are naturally occurring, which means they're unrelated to the discharge of pollutants or hazardous substances or other anthropogenic activities.

Q: The context for my question was, how do you define background spatially if you have an impact and how do you defend your selection?

A: Generally, at an impacted site, we would start with the known background (i.e., off-site, upgradient, undisturbed) samples and then look at the remaining samples using the tools we described earlier (e.g., hierarchical clustering, principal components analysis) to identify samples that are impacted, those that have some level of impact, and those that appear to be unimpacted even though they were not initially identified as definitely background. It is this last set of samples that are the most valuable to the project as they provide evidence that samples that might otherwise be considered impacted can in fact be designated as background. They also help strengthen the confidence in the background fingerprint by adding to the n= value for the number of samples defining a polygon. This same process will be used for metals and then the whole dataset will be used to define the background polygons. The real strength of the process is the way we use multiple lines of evidence, and once they all start to agree with each other, then we know we're on a good track.

Q: Can you talk about the current or planned data sources? You mentioned AGRASID and you talked about data sharing agreements, but can you tell us anything more?

A: The primary data sources for us are going to be large oil and gas operators in this province. That's where most of the data exists. It's going to come via those operators, the consultants who manage the sites, or directly from the labs. Current or past research projects may also have useful data.

Q: Any interest in expanding into the great province of Saskatchewan?

A1: Absolutely we can help with that. In terms of the project, we are expanding a step at a time. Pilot area, the Province of Alberta, then potentially other areas.

A2: PTAC is very interested in expanding into Saskatchewan. Obviously, we're interested in expanding it to other areas where the outcomes would provide clarity for the regulators. Certainly, central and the southwest of Saskatchewan in the southern part of the province is paramount, so hopefully we can look forward to roll this out into Saskatchewan as well.

Q: You used several examples where you said that you know if concentration is X then it obviously can't be background. I assume the system works the other way around and counters biases that say this must be background?

A: Yes, there will be both positive and negative surprises that come out of this. For example, I paint the province with this broad brush based on my experience whereby I know that chloride goes up to this level or selenium goes up to this level when in fact it is variable based on the area you're looking at. I think Tom's work will show we can easily map salinity at a provincial level using vegetation cover and maybe depth to bedrock.

Q: Is there a maximum depth you will look at where you include bedrock data and I guess I'll add, does it matter if the sample is in groundwater (saturated) or not?

A: I'm thinking that the maximum depth will probably be around 6 m. One reason that I say that is because I think most samples will be from depths above that. So yes, there will be a depth limit but we're not going to set it right now. Once we explore the data and see what level of confidence we have in it, we can set a cut off where the confidence in the predictions starts to go down.

Q: Is one of the final goals of this good work to update Tier1 Guidelines?

A1: I'll start with the InnoTech perspective as a neutral convener, our intention is to provide this information to the regulators and the industry. It's not part of our scope to advocate for policy changes.

A2: From the PTAC RRRC perspective we are here to inform and provide recommendations. Obviously, there are regulators that really support the initiatives, but we're just bringing forward recommendations. We also do not set policy, so it will be up to the regulators to tell us what they need to support changes to Tier 1 should they want to go that way. There's a range of potential outcomes for this project and updating Tier 1 is at the far end of that spectrum. The primary goal of the project is to develop a system that allows operators to meet the current Tier 1 requirements for demonstrating that background is greater than the Tier 1 guideline value.

Q: You use salinity as an example, but can you confirm that we're doing metals in this part of the project as well?

A: Yes, it will be metals and salinity. I used salinity as a case study to demonstrate some of the basic principles.

Q: When you talked about salinity and showed us the examples you had the key ions and you had electrical conductivity (EC). Will you also be including SAR and pH?

A: Yes, certainly those parameters will be in the fingerprints.

Q: Bedrock samples have many false positives, especially for trace metals. How will you deal with that?

A: Hopefully the core logs can help us identify bedrock and when we see a change in the fingerprint we can decide if and where the cut off should be made between soil and bedrock. And then each layer would have its own fingerprint. We can use data layers from the Alberta Geological Survey (i.e., layers that identify) shallow bedrock as covariate data in the mapping tool. I expect that parent material and bedrock are going to be one of the strong definers of the shape of the polygons.

Q: What other class type variables do you intend to bring into the analysis? You mentioned soil horizons as one.

A: I'm really glad this question was asked because it's one of the things that we're always looking for input on. In Tom's presentation we're going to hear about some of the covariate data that we are planning on using. Datasets like digital elevation model (LiDAR), satellite imagery, precipitation maps, etc. The beautiful thing about using machine learning is we don't just look at the relationship between say, depth to bedrock and ion concentration as a single correlation between two variables. Machine learning allows us to find the relationship between all the covariate data sets and all the chemicals to define the polygon. We would really like your input into the core covariate layers we should be including to make this the most useful and accurate tool for your needs.

Q: Who will own and maintain the database?

A1: One of the project components is to develop a plan for system distribution and that includes engagement about housing and maintaining the database and mapping tools.

A2: I agree, the long-term sustainability of these tools is essential. But I have found that as you go through the development process of products such as this, there are changes that can influence how and where the tool will be managed from a long-term perspective. So, I think that it's important to demonstrate efficacy and effectiveness first and then determine what the specifics are going to be for long term storage and maintenance. I think we have to concentrate on the proof of concept, and the efficacy and effectiveness, before we want to start determining what the business model for long-term sustainability will look like.

Q: One of the other concerns raised in the survey was that there are too few data for some places in their Province, or in other words, do you have a minimum number of data points necessary to identify and populate a polygon?

A: This is probably a question that's better put off until the predictive mapping component because the predictive mapping will produce a map of the predicted concentrations and a map of the predicted error of those predictions so we can see where confidence is lower. And then in phase three of this project, there is some opportunity to go out and fill some of those data gaps. But to be honest, I don't think we're going to have anywhere where we don't have enough information just from looking at a map of wellsites in the province. However, if we do identify low-data regions we can go to producers in the area and get their data.

Q: What is the size of your test data set?

A: That is to be determined, but I'm guessing it'll be on the order of 20,000 to 50,000 samples.

3.0 PREDICTIVE SOIL MAPPING

Tom Hengl and Leandro Parente of EnvirometriX presented information on the concept and development of the predictive soil mapping tool.

3.1 WORKSHOP PRESENTATION HIGHLIGHTS

Predictive soil mapping involves development of a numerical or statistical model of the relationship among environmental variables and soil properties, which is then applied to a geographic data base to create a predictive map (Scull et al., 2003). For more information on the use of predictive soil mapping see Drozdowski et al. (2019) and Hengl and McMillan (2019).

There are six steps in preparing a predictive soil map (Figure 5):

- Point and gridded data preparation,
- Preparation of regression/classification matrix,
- Fine-tuning, feature selection, optimization / model training,
- Model validation / accuracy assessment,
- Prediction and visualization of results, and,
- Re-analysis and model improvements

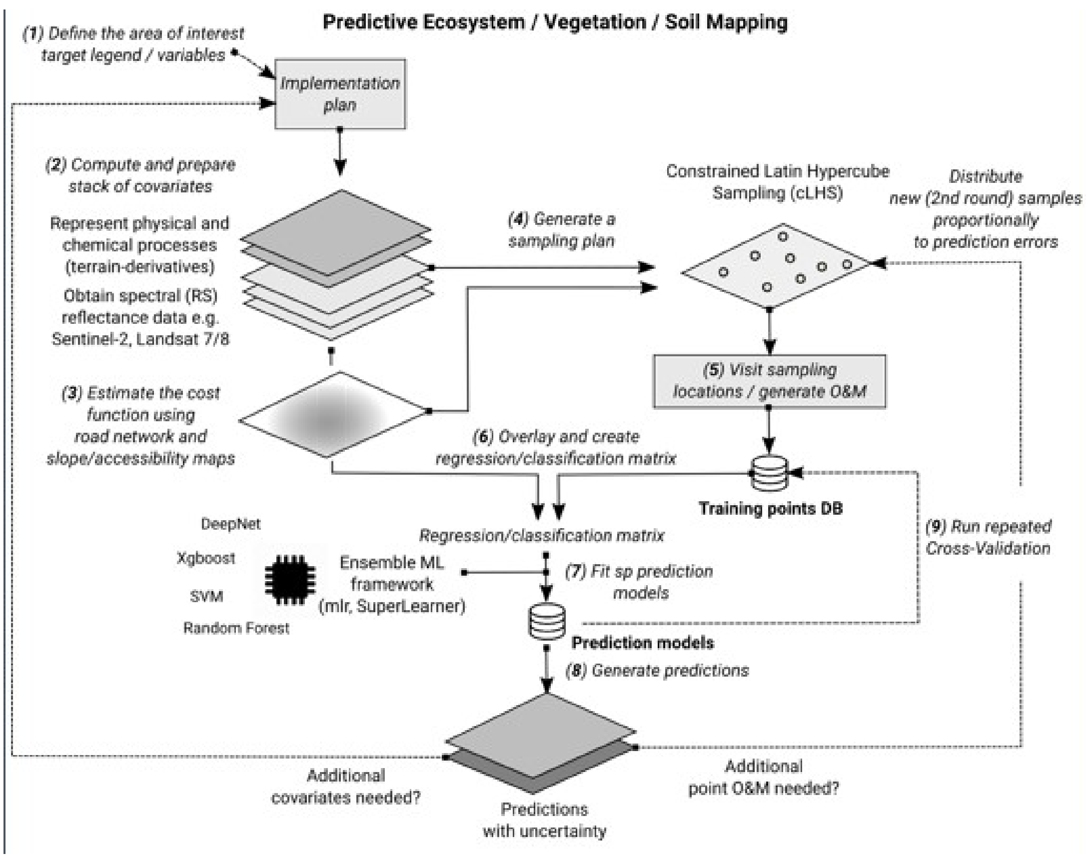


Figure 5. Steps in creating a predictive soil map.
From Hengl and McMillan (2019).

The ABSQS project requires development of an Alberta-specific data cube (i.e., a complete, consistent, and current analysis-ready data stack consisting of multiple layers of relevant datasets adjusted to the same scale) that will form the basis of the predictive soil map. The greater the number of high-quality, high relevance layers, the more accurate the predictive capabilities of the model.

EnvirometriX has shown how larger geochemical databases (e.g., USGS National Geochemical Database: Soil[1]) can be used to demonstrate predictive capabilities of standard PSM methods. In the example shown in Figure 6, almost 200 input covariates representing distance to urban areas, vegetation / biomass, and land surface temperature regime (MODIS LST) were used to predict soil Pb levels in the US. Ensemble Machine Learning (EML) showed that soil depth, MODIS LST daytime images for various months, and travel time to cities were within the top 10 covariates predicting Pb soil concentrations, suggesting that a large part of Pb concentration is explained by distance from industrial sources.

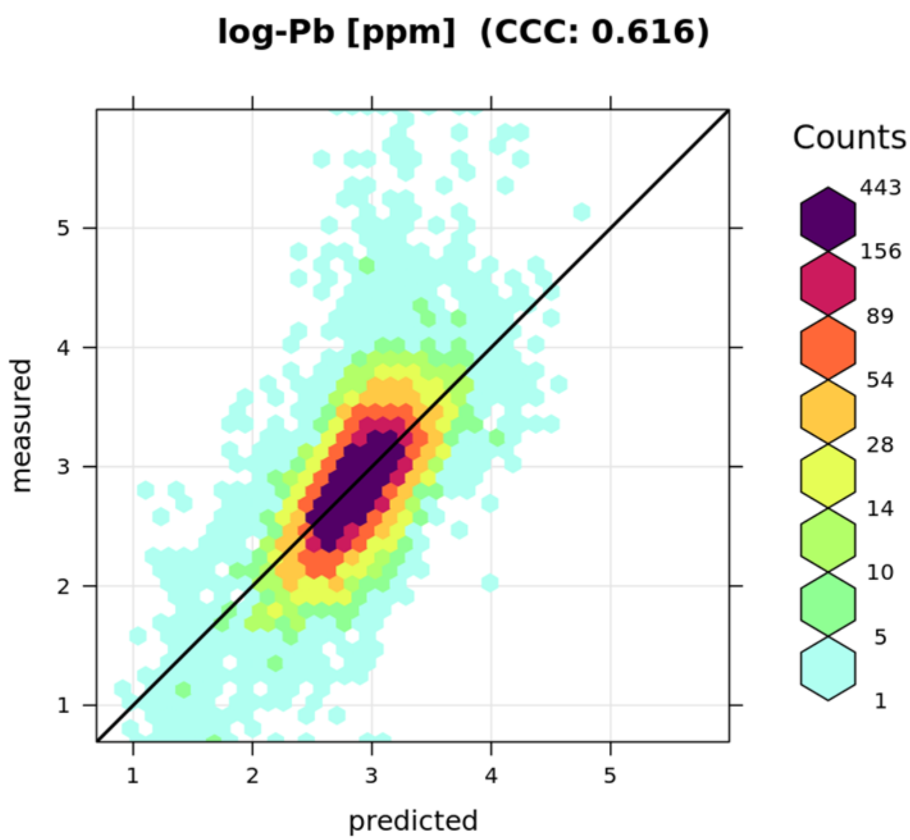


Figure 6. Predicted Pb soil concentrations in the continental US using the USGS National Geochemical Database: Soil and Ensemble Machine Learning.

[1] See https://mrdata.usgs.gov/ngdb/soil/

3.2 Mapping Question and Answer Session

Q: Would these be new tools you have to develop, or would they leverage those that might already be accessible?

A: If you are referring to the Earth observation data, we will make a unique data cube for Alberta, so nothing you can find off the shelf. It will require a lot of processing and training of the model for the geochemical mapping.

Q: Is there a minimum number of layers that you need, and do you need specific layers to make this work?

A: We tend to want to use all available layers. We will look for any relationship we can find. If there is a relationship with anything that is already available, we will find it. So, there is no minimum, but the maximum will be constrained by time and cost of available data. It is going to be a wide search and then we will optimize models by running the Feature Selection function.

Q: Can you talk a little bit more about how you're going to deal with the different data scales? Especially since the soils data are point data versus, say, satellite data that are at 3 to 10 km resolution.

A: The scale is a very important issue and when you are looking at a multivariate problem like this you will have data at multiple scales. We have tools available to manage data at different scales (e.g., Hengl et al (2021)).

Q: Some organic soils have drastically different density from other soils. This means that mg/kg data can be distorted compared to mineral soils. Any concern addressing these issues?

A1: It is a concern, but we may be able to address it if we have bulk density data associated with the geochemical data.

A2: (Paul) We will probably have bulk density for some samples, but as I don't have the data sets in hand yet, I can't say what percentage of samples will have the necessary data. It may help us to work through this issue with regulators to determine the best way to approach it, because we will have obvious areas of organic soil that are different. How would you like us to move forward with these? Would you like us to try to map them in mg/kg and hope for the best? Do the best that we can with the data we have, or do you want to say that for organic soils we provide additional guidance in terms of the sampling requirements? Either of those would be a positive outcome for the industry.

Q: How much of Alberta is organic soils?

A1: We are still exploring what would be an applicable spatial domain for producing predictions (and how much can we extrapolate). These are issues we need to decide as the project progresses. (*Post-workshop Clarification*: *Our intention is to start with a smaller area of the province first, and then expand to the rest of the province. We may need to look at additional covariate data layers for areas with organic soils, compared to upland sites*).

A2: Some of the covariate layers that we're hoping to use are shapefiles for different land uses and different land covers of various areas in the province. I don't have a number now in terms of the percentage of peatlands in the province.

A3: The initial testing is going to be in areas where we have AGRASID data in the White Area, so not addressing organic soils to start. However, it's certainly going to be an unfolding issue for us and I'm excited to see how it gets addressed.

A4: About 16% of the province is peatlands (see Turchenek and Pigot [1988]).

Q: The province only owns a small subset of LIDAR. For the rest of the province there is a license which restricts the ability to share the data and the products derived from the data. This is something that we run across quite frequently, so again, it gets back to is there a minimum data set that we are willing to use if we can't access layers that are either too expensive or unavailable to us?

A: We are aware of the limited LiDAR datasets owned by the Province, and we are going to do some tests on different LiDAR scales and see what they do to the predictive model accuracy. We know there are cost implications to consider.

Q: Earlier you had mentioned that the users won't get access to the underlying soils data. What about access to the data cube?

A1: From what I understand, no. The users will access the products through a web app mapping application; there will be some layers that users can select, but the files will be embedded into the database.

A2: Yes, that's right. The web app will interact with the outputs of the predictions and not the mechanism of making the predictions themselves.

4.0 WORKSHOP OUTCOMES

There was considerable interest in the potential for the ABSQS to support regulatory compliance and no "show stoppers" were identified by participants. Several questions were raised, and there was a desire for more information, much of which can be addressed by providing progress updates and specifically by providing more detail on the contents of the database, the geographic area of the pilot, examples of the background levels in polygons, and maps of the polygons in the pilot area.

Participants could help ensure project success by:

- Providing suggestions for, and/or access to, soil chemistry datasets (preferably georeferenced),
- Providing suggestions for, and/or access to, covariate data layers, and,
- Providing information on which soil chemical parameters are most often found to have naturally elevated background levels in various regions of the province

4.1 POST-WORKSHOP COMMENTS

Participants were provided a post-workshop opportunity to provide additional comments on the workshop and project. A summary of key comments is provided below:

- I had hoped that there would be an opportunity to view the tool
- I wanted to see more about what the tool would look like
- It would be interesting to know the planned schedule for this project, particularly the release date for the system
- It would be great to have future updates and to know when this will be live for beta testing or final release. This is relevant to many ongoing projects
- I think it is a great project / tool and would aid with some sites where in a quick snapshot seem to have salt issues, but incorporating the data from the tool, it may bring to light natural salts in the area and help to explain some anomalous data
- I would like to hear from industry how they view this project and what they are hoping to get from it
- At the appropriate time we need to consider use of predicted versus actual data in regulatory decisions
- I would like to see how the salinity and metals will look once it is all integrated
- I would like to see this expanded to naturally occurring hydrocarbons and PAHs when applicable
- Would like to discuss the 6 m as maximum depth
- Maybe more focus on Alberta ecoregions that we are going to focus on. We mentioned this at the TSC – work in areas where there is higher confidence to prove the statistical approach, then move to other areas where less data
- I have access to a large database but it is all Type 3 non-georeferenced data – how can we best use that?

- It is difficult to understand the link between the data and the remote sensing layers but this may become more evident in the future
- The modelling presentation was very technical. However, it may be worthwhile having a specific session for those few who do want to dive deeper behind the curtain

5.0 REFERENCES

Alberta Environment and Parks, 2019. Alberta Tier 1 Soil and Groundwater Remediation Guidelines. Alberta Environment and Parks, Land Policy Branch, Policy and Planning Division, Edmonton, Alberta. Land Policy, 2019, No.1. 198 pp. https://open.alberta.ca/dataset/842becf6-dc0c-4cc7-8b29-e3f383133ddc/resource/a5cd84a6-5675-4e5b-94b8-0a36887c588b/download/albertatier1guidelines-jan10-2019.pdf

Coenen and Pearce, n.d. Understanding UMAP. GitHub Inc. https://pair-code.github.io/understanding-umap/

Drozdowski, B., C. Aumann and C.B. Powter, 2019. Predictive Soil Mapping Seminar: Summary Report. InnoTech Alberta, Edmonton, Alberta. 188 pp.

Dudas, M.J., 1987. Accumulation of Native Arsenic in Acid Sulphate Soils in Alberta. Canadian Journal of Soil Science 67: 317-331. https://cdnsciencepub.com/doi/pdf/10.4141/cjss87-028

Hengl, T. and R.A. MacMillan, 2019. Predictive Soil Mapping with R. OpenGeoHub Foundation, Wageningen, the Netherlands. 370 pp. https://soilmapper.org/

Hengl, T., M.A.E. Miller, J. Križan, K.D. Shepherd, A. Sila, M. Kilibarda, O. Antonijević, L. Glušica, A. Dobermann, S.M. Haefele, S.P. McGrath, G.E. Acquah, J. Collinson, L. Parente, M. Sheykhmousa, K. Saito, J.-M. Johnson, J. Chamberlin, F.B.T. Silatsa, M. Yemefack, J. Wendt, R.A. MacMillan, I. Wheeler and J. Crouch, 2021. African Soil Properties and Nutrients Mapped at 30 M Spatial Resolution Using Two-Scale Ensemble Machine Learning. Scientific Reports 11: 6130. https://www.nature.com/articles/s41598-021-85639-y#article-info

IBM Cloud Education, 2020. Machine Learning. IBM, Armonk, New York, USA. https://www.ibm.com/cloud/learn/machine-learning

Luther, S., 2020. Assessing Source of Trace Metals at a Wellsite in Alberta to Support Remediation and reclamation planning. IN: RemTech Conference 2020. Environmental Services Associate=ion of Alberta. 28 slides. https://esaa.org/wp-content/uploads/2021/04/RT2020-Luther12.pdf

Millennium EMS Solutions Ltd., 2016. Evaluation of Background Metal Concentrations in Alberta Soils. Prepared for Petroleum Technology Alliance Canada, Calgary, Alberta by Millennium EMS Solutions Ltd., Calgary, Alberta. 23 pp. plus appendices. https://auprf.ptac.org/wp-content/uploads/2016/09/PTAC-15-SGRC-02-Evaluation-of-Background-Metal-Concentrations-in-Alberta-Soils-combined.pdf

Scull, P., J. Franklin, O.A. Chadwick and D. McArthur, 2003. Predictive soil mapping: A review. Progress in Physical Geography: Earth and Environment 27(2):171-197. doi:10.1191/0309133303pp366ra

Soon, Y.K. and S.A. Abboud, 1990. Trace Elements in Agricultural Soils of Northwestern Alberta. Canadian Journal of Soil Science 70: 277-288. https://cdnsciencepub.com/doi/pdf/10.4141/cjss90-029

Turchenek, L. and M. Pigot, 1988. Peatland Distribution in Alberta. Alberta Research Council, Edmonton, Alberta. Map 212. https://static.ags.aer.ca/files/document/MAP/Map_212.pdf

Utting, N. and N. Heshka, 2019. Orphan and Abandoned Well Research Needs Assessment. Natural Resources Canada, CanmetENERGY, Devon, Alberta. Report CDEV-2019-0096-RT. 23 pp. https://www.nrcan.gc.ca/sites/www.nrcan.gc.ca/files/canmetenergy/files/pubs/CDEV-2019-0096-RT%20-%20Final%20-

%20Orphan%20and%20Abandoned%20Well%20Research%20Needs%20Assessment%20accessible.pdf

van der Maaten, L., 2021. t-SNE. Facebook AI Research, New York, New York, USA. https://lvdmaaten.github.io/tsne/

5.1 Additional Reading

The following sections provide additional resources that help set context for the ABSQS project. Brief summaries of relevant information from each resource are provided.

5.1.1 *Using Background Soil Data in Regulatory Work*

Alberta Environment and Parks, 2019. Alberta Tier 2 Soil and Groundwater Remediation Guidelines. Land Policy Branch, Policy and Planning Division. 150 pp. https://open.alberta.ca/dataset/aa212afe-2916-4be9-8094-42708c950313/resource/157bf66c-370e-4e19-854a-3206991cc3d2/download/albertatier2guidelines-jan10-2019.pdf

> For the purpose of applying Alberta Tier 1 or Alberta Tier 2 Soil and Groundwater Remediation Guidelines, the background concentration of a substance in soil or groundwater is defined as:
>
> > 1. The natural concentration of that substance in the absence of any input from anthropogenic activities or sources, or;
> > 2. The background concentration in the surrounding area as a result of generalized non-point anthropogenic sources.
>
> In some situations, the background concentration of a substance can be a significant proportion of, or even exceed, the Tier 1 guidelines. In cases when the background concentration is demonstrated to be greater than Alberta Tier 1 guidelines, the remediation level shall be set to background or to guidelines developed using Tier 2 procedures.
>
> Background concentrations will vary with soil parent material, soil depth, and hydrologic regime. These factors lead to spatial variations in background concentrations that may or may not be predictable. To gain a good understanding of background conditions at a site, it is necessary to take sufficient representative samples from soils with similar characteristics to the affected site, but which are taken from outside the area affected by contamination. Sample depth and landscape position, soil profile characteristics and parent material should be recorded for all samples.

Alberta Environment, 2001. Salt Contamination Assessment and Remediation Guidelines. Alberta Environment, Environmental Sciences Division, Edmonton, Alberta. Publication No. T/606. 88 pp. https://open.alberta.ca/dataset/d53c62c1-7dec-4396-aa8a-2a01703d2060/resource/b7bee18b-c7cf-4f85-957d-bcd2dc68a13a/download/2001-saltcontaminationremediationguidelines.pdf

> Soils in some areas in Alberta have naturally occurring levels of salt or sodium that are detrimental to plant growth. These naturally occurring concentrations can be higher than those resulting in some salt spill contaminated soils. The most common naturally occurring salts found in Alberta soils consist of sodium (Na^+), magnesium (Mg^{++}), and calcium (Ca^{++}) ions in combination with sulphate ($S0_4^-$), bicarbonate ($HC0_3^-$), and to a lesser extent chloride (Cl-) ions. Contaminant concentrations must meet remediation objectives (background levels, generic guidelines or site-specific risk-based objectives) or better.

Government of Alberta, 2009. Soil Monitoring Directive. Alberta Environment, Edmonton, Alberta. 22 pp. plus appendix. https://open.alberta.ca/dataset/237d66fc-2347-41f4-820b-

e950c7e75542/resource/df94722f-7333-429b-9344-0be8f5ebd063/download/2009-soilmonitoringdirective-may2009a.pdf

> For an undeveloped site where the baseline soil condition for a soil parameter is suspected to be affected by any regional or local geochemical abnormality, that parameter should be included for analysis. In particular, the Baseline Soil Monitoring Program should include any parameter where the background condition may exceed the Alberta Tier 1 Soil and Groundwater Remediation Guidelines as amended. For example, analyses of soil soluble sodium or sulphate should be included at a site where geochemical enrichment of those elements is expected as the result of natural groundwater discharge.
>
> A proposed site-specific soil quality standard for a given substance shall be determined in accordance with the following: a) the natural background concentration of the substance ...

Government of British Columbia, n.d. Background Concentration Databases. Government of British Columbia, Victoria, British Columbia. https://www2.gov.bc.ca/gov/content/environment/air-land-water/site-remediation/contaminated-sites/the-remediation-process/background-concentrations/background-databases

> Background concentrations in soil for select metals has been compiled for eight administrative regions (Cariboo, Kootenay, Lower Mainland, Omineca-Peace, Skeena, Southern Interior, Vancouver Island, and Vancouver). Background sampling has also been in the Metro Vancouver area, due to the higher redevelopment rate of commercial and industrial properties in that region. These data have been used to adjust contaminated sites' soil standards in the Contaminated Sites Regulation (CSR) and to calculate regional estimates of background concentrations in the soil for use under the CSR. An Excel database is available.

British Columbia Ministry of Environment and Climate Strategy, 2021. Protocol 4 for Contaminated Sites: Establishing Local Background Concentrations in Soil. Version 11. British Columbia Ministry of Environment and Climate Strategy, Victoria, British Columbia. 10 pp. https://www2.gov.bc.ca/assets/gov/environment/air-land-water/site-remediation/docs/protocols/p4__jan2021_revisions_final_signed.pdf

> This protocol describes options for establishing a local background concentration in soil for use in the investigation and remediation of a contaminated site and to carry out soil relocation, where naturally occurring substance concentrations exceed the applicable numerical soil standards of the *Contaminated Sites Regulation*. A local background concentration can be established by either directly applying regional background concentration estimates provided by the ministry for specified inorganic substances or by using the procedures outlined in this protocol for determining site-specific background concentrations in soil.

Geiselbrecht, A., S. Rouhani, K. Thorbjornsen, D. Blue, S. Nadeau, T. Gardner-Brown and S. Brown, 2019. Important Considerations in the Derivation of Background at Sediment Sites. Integrated Environmental Assessment and Management 15(3): 448–457. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6850622/

> A thorough understanding of a site is critical to selecting the background reference areas from which representative background concentrations can be derived. Representative background concentrations should account for contributions from those background chemical inputs (natural and anthropogenic sources) that will continue affecting the site even after remediation. Perceived outliers should not be eliminated from the background data set just because they are

the highest or lowest values. Geochemical evaluation of trace metals is a useful tool for deriving representative background concentrations.

5.1.2 *Alberta Information and Data*

Government of Alberta, n.d. Agricultural Regions of Alberta Soil Inventory Database (AGRASID). Government of Alberta, Edmonton, Alberta. https://www.alberta.ca/agricultural-regions-of-alberta-soil-inventory-database.aspx

The Agricultural Regions of Alberta Soil Inventory Database (AGRASID) is a spatial database of soils for Alberta's Agricultural area. This database is ideal for regional, and field scale, land use assessment and decision-making.

Government of Alberta, 2020. Government of Alberta, Agricultural Land Resource Atlas. Government of Alberta, Edmonton, Alberta. https://maps.alberta.ca/genesis/rest/services/Agricultural_Land_Resource_Atlas/Latest/MapServer

This site lists 22 layers of data available for the agricultural region of Alberta. Layers relevant to the ABSQS are listed below.

Government of Alberta, 2020. Organic Matter Content of Cultivated Soils. Agricultural Land Resource Atlas. Government of Alberta, Edmonton, Alberta. https://maps.alberta.ca/genesis/rest/services/Agricultural_Land_Resource_Atlas/Latest/MapServer/13

This map displays the percentage of organic matter in the surface layer of cultivated soils in the agricultural region of Alberta. Soil organic matter (SOM) is derived primarily from the decomposition of plant biomass. SOM improves both the physical and chemical properties of soil and has beneficial effects on agricultural soil quality. SOM is reported on the map as a percentage using the following classes: less than 2 (very low), 2 to 4 (low), 4 to 6 (medium), 6 to 8 (high) and greater than 8 (very high).

Government of Alberta, 2020. Saline Soils. Agricultural Land Resource Atlas. Government of Alberta, Edmonton, Alberta. https://maps.alberta.ca/genesis/rest/services/Agricultural_Land_Resource_Atlas/Latest/MapServer/15

The data represents the occurrence of saline soils in the agricultural area of Alberta. A Saline Soil is a non-alkali (pH less than 8.5 and exchangeable-sodium less than 15%) soil containing soluble salts in great enough quantities that they interfere with the growth of most crop plants. This resource was created in 2002 using ArcGIS.

Government of Alberta, n.d. Salinity Maps of Selected Counties. Government of Alberta, Edmonton, Alberta. https://www.alberta.ca/salinity-maps-of-selected-counties.aspx

In the 1990s, the Alberta government partnered with interested counties to map visible salinity from aerial photographs and other data sources. Maps for 22 Counties are available.

Government of Alberta, 2020. Solonetzic Soils. Agricultural Land Resource Atlas. Government of Alberta, Edmonton, Alberta. https://maps.alberta.ca/genesis/rest/services/Agricultural_Land_Resource_Atlas/Latest/MapServer/19

This map displays the distribution of Solonetzic soils in the agricultural region of Alberta. Solonetzic soils have developed on saline parent material that is high in sodium and have a characteristic hardpan layer that has formed in the subsoil. This hardpan is very hard when dry

and has low permeability when wet. This results in restricted root and water penetration that may limit the productivity of these soils. Solonetzic soils occur in association with Chernozemic soils and, to a lesser extent, with Luvisolic soils.

Government of Alberta, 2020. Soil Texture. Agricultural Land Resource Atlas. Government of Alberta, Edmonton, Alberta. https://maps.alberta.ca/genesis/rest/services/Agricultural_Land_Resource_Atlas/Latest/MapServer/18

This map illustrates the distribution of soil parent material textures in the agricultural region of Alberta. Soil texture is defined by the relative proportions of the sand, silt and clay particles present. Soil textures are identified by classes using the Soil Texture Triangle. The Soil Texture Triangle identifies the textural class of a soil at the intersection of the percent sand (x-axis) and the percent clay (y-axis). The percent silt of the soil is the remainder to add up to 100 percent. For presentation on this map, the texture classes of soil parent materials identified with each Agricultural Region of Alberta Soil Inventory Database (AGRASID) soil landscape polygon were combined into four more general groups - fine, medium, moderately coarse and very coarse.

Government of Alberta, 2020. Derived Ecosite Phase v2.0. Government of Alberta, Edmonton, Alberta. https://open.alberta.ca/opendata/gda-ae37f83c-c994-47a9-b2f0-39ba1da0e64c

Derived Ecosite Phase (DEP) v2.0 is a digital and spatial representation of ecological sites and phases in those areas of Alberta (mostly the Green Area) where both Alberta Vegetation Inventory (AVI) and LiDAR are available. The AVI is an imagery-based digital inventory developed to identify the type, extent, and conditions of vegetation, where it exists and what changes are occurring. The most up-to-date ecological site phases can be found in the Ecological Site Guides. Guides are broken into individual Natural Subregions.

Government of Alberta, 2020. Alberta Merged Wetland Inventory. Government of Alberta, Edmonton, Alberta. https://geodiscover.alberta.ca/geoportal/rest/metadata/item/bfa8b3fdf0df4ec19f7f648689237969/html

The Alberta Merged Wetland Inventory depicts wetlands within the province for the period 1998 to 2017 classified to the five major classes in the Canadian Wetland Classification System. These five major classes include bog, fen, marsh, swamp and shallow open water. For the purposes of this inventory, shallow open water includes all open water. The Alberta Merged Wetland Inventory is a generalized, merged product of 35 component wetland inventories that utilized different types of source data from different years, different data capture specifications and different classifications. Considerable variation in the level of detail and accuracy is present in this dataset. Accuracy assessments have been included where available, but it should be noted that the geoprocessing applied to the data may have introduced additional error.

Canon, K. and L. Leskiw, 1999. Soil Quality benchmarks in Alberta. IN: Proceedings of 36th Annual Alberta Soil Science Workshop, February 16-18, 1999, Calgary, Alberta. 4 pp. https://open.alberta.ca/dataset/a3627017-105a-468f-853f-11f9ea26f507/resource/c952ec8f-153c-4ce8-94fc-7ab5ba653e0e/download/afrd-aesa-soil-quality-benchmarks-in-alberta.pdf

Long-term benchmark soil sampling started in the fall of 1998 to monitor soil quality across Alberta landscapes and is part of the AESA (Alberta Environmentally Sustainable Agriculture) Soil Quality Monitoring Initiative. These sites were chosen to be representative of the soil-landscape patterns and agronomic practices within a given ecodistrict. There are two goals for this program. The first is to determine the state of soil quality across Alberta and the second is to determine the risk of change in soil quality with various management practices.

Penney, D., K. Cannon and D. Keyes, 2003. Preliminary Analyses of Five Years of Soil Data from the AESA Soil Quality Benchmark Sites. IN: Proceedings of 40th Annual Alberta Soil Science Workshop, February 18-20, 2003, Edmonton, Alberta. 6 pp. https://open.alberta.ca/dataset/ede1fca2-58cb-4b48-913b-3b264a51846c/resource/7f4fc4d9-b7ac-4c60-b650-2585c62bd960/download/afrd-aesa-preliminary-analyses-five-years-soil-data.pdf

The AESA (Alberta Environmentally Sustainable Agriculture) Soil Quality Benchmark Program was established in 1998 to provide a monitoring network across Alberta. From each of 43 Ecodistricts within seven Ecoregions, one site was chosen to represent the soil-landscape patterns and agronomic practices of that Ecodistrict. Soil properties that tend to change slowly over time (pH, EC, P and K) were fairly consistent from year to year at many of the sites, but at least 3 of the above properties were quite variable at 13 of the 43 sites. The variation from year to year in light fraction (LF) organic matter and NO_3 was high. Significant differences in soil properties occurred across the three slope positions (upper, mid and lower) at many of the sites. These important differences would be masked if average values had been obtained from composite samples taken across slope positions (field composite samples).

Dudas, M.J. and S. Pawluk, 1977. Heavy Metals in Cultivated Soils and in Cereal Crops in Alberta. Canadian Journal of Soil Science 57: 329-339. https://cdnsciencepub.com/doi/pdf/10.4141/cjss77-037

The content of cadmium, cobalt, copper, mercury, manganese, nickel, strontium, lead and zinc in several agricultural soils in Alberta was investigated. The abundances of these heavy metals were found to be low and represent levels naturally present in uncontaminated soils.

5.1.3 *Predictive Mapping Resources*

Nadeau, L.B., C. Li and H. Hans, 2004. Ecosystem Mapping in the Lower Foothills Subregion of Alberta: Application of Fuzzy Logic. The Forestry Chronicle 80(3): 359-365. http://pubs.cif-ifc.org/doi/pdf/10.5558/tfc80359-3

Fuzzy logic technology can be used to computerize essential elements of ecosystem identification, and the outputs can be linked to a Geographic Information System for map production. A pilot study was undertaken on the application of this technology to the Alberta Ecological Land Classification database and the resulting ecosite map for a township located in central Alberta (Tp42R9W5).

Cordeiro, M.R.C., G. Lelyk, R. Kröbel, G. Legesse, M. Faramarzi, M. Badrul Masud and T. McAllister, 2018. Deriving a dataset for agriculturally relevant soils from the Soil Landscapes of Canada (SLC) database for use in Soil and Water Assessment Tool (SWAT) simulations. Earth System Science Data 10: 1673–1686. https://cms.eas.ualberta.ca/faramarzilab/wp-content/uploads/sites/14/2018/09/essd-10-1673-2018_published.pdf

The objective of this work was to preprocess the Soil Landscapes of Canada (SLC) database to offer a country-level soils dataset in a format ready to be used in Soil and Water Assessment Tool (SWAT) simulations.

Heung, B., 2017. Regional-Scale Digital Soil Mapping in British Columbia using Legacy Soil Survey Data and Machine-Learning Techniques. Ph.D. Thesis. Department of Geography, Simon Fraser University,

Burnaby, British Columbia. 164 pp.
http://summit.sfu.ca/system/files/iritems1/17377/etd10172_BHeung.pdf

Describes a framework for developing training data from British Columbia conventional soil survey maps and compares various machine-learning techniques for predicting the spatial patterns of qualitative soil data such as soil parent material and soil classes.

Sorenson, P.T., S.J. Shirtliffe and A. Bedard-Haughn, 2021. Predictive Soil Mapping Using Historic Bare Soil Composite Imagery and Legacy Soil Survey Data. Geoderma 401: 115316.
https://www.sciencedirect.com/science/article/pii/S0016706121003967

This study focused on using bare soil composite imagery for Saskatchewan along with legacy soil data (1958–1998) with high location uncertainty to predict soil organic carbon, clay, and cation exchange capacity. The bare soil composite images were created from Landsat 5 imagery (1985 to 1995) using Google Earth Engine.

Sothe, C., A. Gonsamo, J. Arabian and J. Snider, 2022. Large Scale Mapping of Soil Organic Carbon Concentration with 3D Machine Learning and Satellite Observations. Geoderma 405: 115402.
https://www.sciencedirect.com/science/article/pii/S0016706121004821

The authors tested a three-dimensional (3D) machine learning approach and 40 spatial predictors derived from 20 years of optical and microwave satellite observations to estimate the spatial and vertical distributions of SOC concentration in Canada in six depth intervals up to 1 m.

Smith, D.B., W.F. Cannon, L.G. Woodruff, F. Solano, J.E. Kilburn and D.L. Fey, 2013. Geochemical and Mineralogical Data for Soils of the Conterminous United States. U.S. Geological Survey, Central Mineral and Environmental Resources Science Center, Denver, Colorado, USA. U.S. Geological Survey Data Series 801. 19 pp. https://pubs.usgs.gov/ds/801/pdf/ds801.pdf

In 2007, the U.S. Geological Survey initiated a low density (1 site per 1,600 square kilometers, 4,857 sites) geochemical and mineralogical survey of soils of the conterminous United States. The resulting dataset provides an estimate of the abundance and spatial distribution of chemical elements and minerals in soils of the conterminous United States and represents a baseline for soil geochemistry and mineralogy against which future changes may be recognized and quantified.

Smith, D.B., W.F. Cannon, L.G. Woodruff, F. Solano and K.J. Ellefsen, 2014. Geochemical and Mineralogical Maps for Soils of the Conterminous United States. U.S. Geological Survey, Central Mineral and Environmental Resources Science Center, Denver, Colorado, USA. Open-File Report 2014–1082. https://pubs.usgs.gov/of/2014/1082/pdf/ofr2014-1082.pdf

This report releases geochemical and mineralogical maps derived from the data noted in the Smith et al. (2013) report along with a histogram, boxplot, and empirical cumulative distribution function plot for each element or mineral.

Ramcharan, A., T. Hengl, T. Nauman, C. Brungard, S. Waltman, S. Wills and J. Thompson, 2018. Soil Property and Class Maps of the Conterminous United States at 100-Meter Spatial Resolution. Soil Science Society of America Journal 82(1): 186-201.
https://acsess.onlinelibrary.wiley.com/doi/full/10.2136/sssaj2017.04.0122

Three US soil point datasets – the National Cooperative Soil Survey Characterization Database, the National Soil Information System, and the Rapid Carbon Assessment dataset – were combined with a stack of over 200 environmental datasets and gSSURGO polygon maps to generate complete coverage gridded predictions at 100-m spatial resolution of six soil properties (percentage of organic C, total N, bulk density, pH, and percentage of sand and clay) and two US soil taxonomic classes (291 great groups [GGs] and 78 modified particle size classes [mPSCs]) for the conterminous United States.

Hengl, T., M. Nussbaum, M.N. Wright, G.B.M. Heuvelink and B. Gräler, 2018. Random Forest as a Generic Framework for Predictive Modeling of Spatial and Spatio-temporal Variables. PeerJ 6:e5518. https://peerj.com/articles/5518/

This paper presents a random forest for spatial predictions framework where buffer distances from observation points are used as explanatory variables, thus incorporating geographical proximity effects into the prediction process.

Hengl, T., M. Nikolić and R.A. MacMillan, 2013. Mapping Efficiency and Information Content. International Journal of Applied Earth Observation and Geoinformation 22: 127-138. https://www.sciencedirect.com/science/article/abs/pii/S0303243412000402

This paper proposes two compound measures of mapping quality to support objective comparison of spatial prediction techniques for geostatistical mapping: (1) mapping efficiency – defined as the costs per area per amount of variation explained by the model, and (2) information production efficiency – defined as the cost per byte of effective information produced.

Molnar, C., 2021. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. https://christophm.github.io/interpretable-ml-book/

Machine learning has great potential for improving products, processes and research. But computers usually do not explain their predictions which is a barrier to the adoption of machine learning. This book is about making machine learning models and their decisions interpretable.

5.1.4 *Websites*

Alberta Agriculture and Forestry, n.d. Alberta Soil Information Viewer. https://soil.agric.gov.ab.ca/agrasidviewer/

A searchable mapped version of the AGRASID database.

Alberta Environment and Parks, n.d. Environmental Site Assessment Repository (ESAR) – http://www.esar.alberta.ca/esarmain.aspx

An online, searchable database that contains scientific and technical information about the province's assessed and reclaimed sites.

Government of Alberta, n.d. GeoDiscover Alberta. https://geodiscover.alberta.ca/geoportal/#searchPanel

GeoDiscover Alberta provides enhanced details on Alberta's geospatial data.

Government of Canada, 2021. The National Soil Database. https://sis.agr.gc.ca/cansis/nsdb/index.html

The NSDB is a collection of geospatial datasets which contain soil, landscape, and climatic data for all of Canada. It serves as the national archive for land resources information that was collected by federal and provincial field surveys, or created by land data analysis projects. The NSDB includes GIS coverages at a variety of scales, and the characteristics of each named soil series.

APPENDIX A – WORKSHOP ATTENDEES

Sixty-four people attended the workshop, five of whom could not be identified – the rest of the participants are listed below. The seven members of the project team are also listed in the blue boxes.

Name	Organization
Natalie Shelby-James	InnoTech Alberta
Sarah Thacker	InnoTech Alberta
Paul Fuellbrandt	Statvis
Tomislav Hengl	EnvirometriX
Leandro Parente	EnvirometriX
Preston Sorenson	University of Saskatchewan
Chris Powter	Enviro Q&A Services
Arne Larson	Earthmaster Environmental Strategies Inc.
Vaishalie Anand	Alberta Environment and Parks
Jennifer Arnold	Wood PLC
David Bergstrom	Alberta Energy Regulator
Sara Blacklaws	Alberta Energy Regulator
Trevor Burgers	Millennium Environmental Management Solutions
Julie Burghardt	Alberta Environment and Parks
Jessie Chao	Wood PLC
Geordie Clyde	Alberta Energy Regulator
Tami Dolen	City of Edmonton
Bonnie Drozdowski	InnoTech Alberta
Catrina Duffy	Alberta Energy Regulator
Linda Eastcott	Imperial Oil
Brian Eaton	InnoTech Alberta
Leanne Erickson	Alberta Energy Regulator
Jonas Fenn	Saskatchewan Ministry of Energy and Resources
Shawn Glessing	Cenovus
Sonia Glubish	Canadian Natural Resources Limited
Mark Grant	Torxen
Emily Herdman	InnoTech Alberta
David Jones	Cardinal Energy
Tim Kulka	ATCO

Name	Organization
Jess Leatherdale	Enbridge
Simone Levy	InnoTech Alberta
Kaylan Lundquist	Cenovus
Sheila Luther	Matrix Solutions
Steven Lyster	Alberta Energy Regulator
Erik Martin	Vertex
Claire McFee	Golder
Hollis McGrath	Alberta Energy Regulator
Trippett McKnight	Cenovus
Symon Mezbahuddin	Alberta Agriculture and Forestry
Collen Middleton	Waterline Resources
Kari Moody	City of Edmonton
Anthony Neumann	Element Materials Technology
Kagen Newman	Canadian Natural Resources Limited
Terry Obal	Bureau Veritas
Shane Patterson	Alberta Environment and Parks
Hartzheim, Paul	Canadian Association of Petroleum Producers
Colin Peters	Alberta Energy Regulator
Liana Phoenix	AGAT Labs
Daniel Pollard	Wood PLC
Kathryn Pooley	Alberta Energy Regulator
Tyler Prediger	Matrix Solutions
Jim Purves	Northshore Environmental
Laura Rathgeber	AGAT Labs
Rick Rohl	ARC Resources
Wanda Sakura	Orphan Well Association
Tammy Sargeant	Alberta Environment and Parks
Bachitter Singh	Suncor Energy
Beth Strukoff	Wood PLC
Neal Tanna	InnoTech Alberta
Andy Taylor	Cenovus
Angela Taylor	Stantec

Name	Organization
Megan Valvasori	Cardinal Energy
Tim Wyatt	TerraLogix Solutions
Shannon Yacyshyn	Alberta Environment and Parks
Christine Yang	Alberta Energy Regulator
Elena Zimmerman	Alberta Energy Regulator

Alberta Background Soil Quality System Project Workshop

InnoTech Alberta, Virtual - Microsoft Teams
November 10, 2021 (9:00 am to 12 pm)

AGENDA

The **Workshop Objectives** are to**:**

- Increase awareness and understanding of the **Alberta Background Soil Quality System Project,**
- Solicit and incorporate feedback from workshop attendees related to barriers to use/technical acceptance and System performance expectations, and
- Promote collaboration opportunities.

Time	Topic	Speaker
9:00 – 9:10 am	Welcome and Land Acknowledgement	Chris Powter
9:10 – 9:15 am	Project Background and Benefits	Natalie Shelby-James
9:15 – 9:20 am	Workshop Objectives and Project Team Introductions	Natalie Shelby-James / Paul Fuellbrandt / Tom Hengl
9:20 – 9:50 am	Soil Quality Database Development: Soil Data Acquisition, Compilation, Fingerprinting 'Background'	Paul Fuellbrandt
9:50 – 10:10 am	Interactive Discussion - Soil Chemistry Database & Approach for Separating Background from Impacted Soil Data - Project Challenges: Barriers to Execution & Acceptance	Chris Powter
10:10 – 10:30 am	*Break*	
10:30 – 11:10 am	Predictive Soil Quality Maps: Overview of Approach	Tom Hengl / Leandro Parente
11:10 – 11:20	Look Ahead: Alberta Background Soil Quality System Features	Paul Fuellbrandt
11:20 – 11:55	Interactive Discussion - Predictive Soil Quality Mapping Approach - Project Challenges: Barriers to use/ acceptance and performance expectations	Chris Powter
11:55 -12:00	Closing Remarks	Natalie

APPENDIX C – WORKSHOP PRESENTATIONS

Four presentations were made during the workshop:

- Introductory Remarks by Natalie Shelby-James
- Soil Chemistry Data Acquisition Approach by Paul Fuellbrandt
- Soil Quality Predictive Mapping Approach by Tomislav Hengl and Leandro Parente
- Look Ahead: System Feature and Future Expansions by Pau Fuellbrandt and Natalie Shelby-James

Introductory Remarks by Natalie Shelby-James

Alberta Background Soil Quality System

Outline

- Welcome and Land Acknowledgement
- Workshop Objectives, Problem, Project Overview and Benefits
- Project Team Introductions
- Soil Quality Database Development
- **Interactive Survey & Discussion**
- ***Break***
- Advancements in Predictive Soil Mapping & Overview of Mapping Approach
- Look Ahead: System Features
- **Technical Q&A / Discussion**
- Closing Remarks

2

EnvirometriX statvis InnoTech ALBERTA

Alberta Background Soil Quality System

Workshop Objectives

- Increase awareness and understanding of the **Alberta Background Soil Quality System Project (The Project)**
- Collect and incorporate feedback from workshop attendees related to
  - barriers to use / technical acceptance and
  - system performance expectations
- Promote collaboration opportunities

4

EnvirometriX statvis Trust Through Data InnoTech ALBERTA A SUBSIDIARY OF ALBERTA INNOVATES

Alberta Background Soil Quality System

Problem

- It is well known that certain areas of Alberta have naturally elevated chemical parameter concentrations in both soil and water.
- There is currently no publicly available, scientifically vetted resource that accurately maps or predicts background soil chemistry (e.g., salts and metals) for Alberta
- Industry and practitioners incur high cost, schedule, and regulatory barriers to prove that elevated parameters are of natural origin
- Regulators also face efficiency challenges with the number of applications requiring detailed reviews
- Alberta has >150,000 inactive wells that do not have reclamation certificates and there is political will and support to accelerate the review process.

5

EnvirometriX statvis Trust Through Data InnoTech ALBERTA A SUBSIDIARY OF ALBERTA INNOVATES

Alberta Background Soil Quality System

Project Objective & Key Activities

Work collaboratively with prospective users of background soil quality data to develop the **Alberta Background Soil Quality System** to be used as a resource tool to assist industry and government in environmental management.

Key activities for the Project Include:

- Compiling, cleaning and integrating existing soil salinity and metals data into a geodatabase
- Analysis of soil point data to fingerprint background and remove impacted samples
- Developing predictive background soil maps – prototype in a select area of the Province and then scaling up to Provincial coverage.
- Developing a web application to house, and allow users to interact with the System

6

Alberta Background Soil Quality System - Benefits

Industry and Practitioners	Government & Regulators	Albertans
• Reduce background data collection • Provide empirical evidence of natural variability • More accurate liability estimates • Focus resources on managing actual risk to receptors • Move stalled sites to closure	• Decreased review times • Increased consistency in data presentation • Fewer inactive wells • More reclamation certificates issued	• Open tool available to extrapolate data for multiple uses • Less disturbance and disruption to the natural environment • More sites being cleaned up & fewer orphan wells

7

Alberta Background Soil Quality System

Core Project Team

InnoTech ALBERTA — A SUBSIDIARY OF ALBERTA INNOVATES

Natalie Shelby-James
(Project Lead)
- Project management
- Engagement & consultation (industry, regulators and data providers)
- Data sharing agreements
- Co-chair of Technical Steering Committee

statvis — Trust Through Data

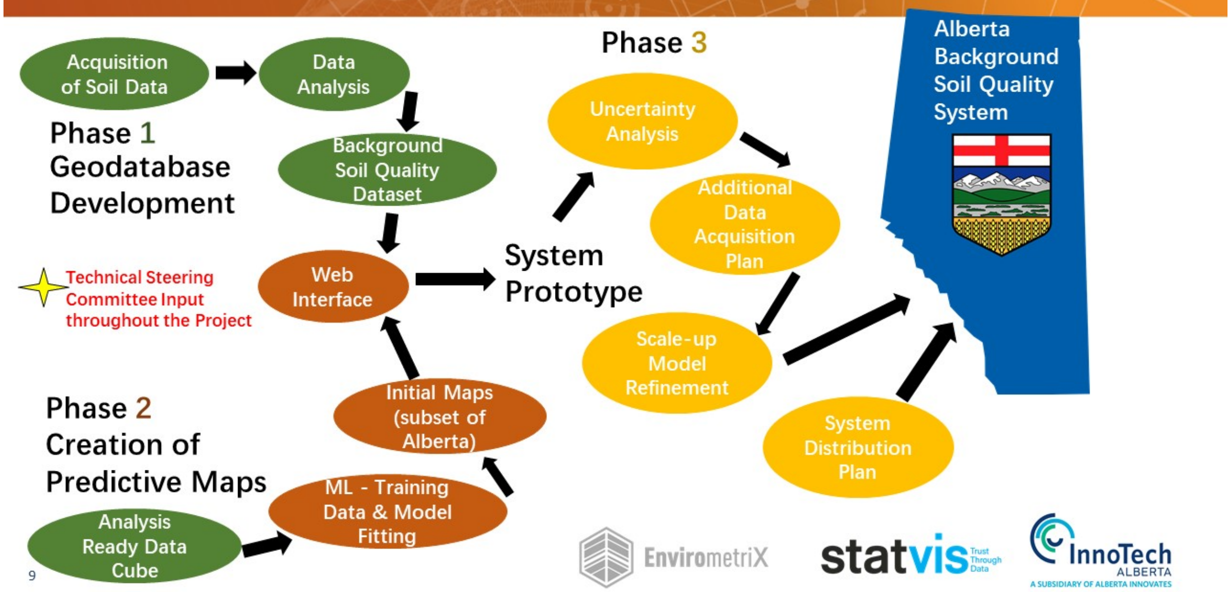
Paul Fuellbrandt
- Soils dataset & database development
- Development of web application to interact with database
- Field verification sampling proposal
- Co-chair of Technical Steering Committee

EnvirometriX

Tom Hengl
- Develop predictive soil model (PSM)
- Auxiliary data collation, cleaning and integration with the PSM
- Develop uncertainty map
- Model expansion to Provincial scale

8

Alberta Background Soil Quality System – Phased Project



9

Alberta Background Soil Quality System

Technical Steering Committee

- Guidance to the project team to ensure:
  - Voice of the customer / end users are represented (i.e., outputs that directly meet users' specific data needs)
  - Consistency with regulatory requirements and expectation
  - Support with obtaining data inputs
  - Support to achieving open and accessible data outputs
- Membership: Core Project Team, Industry, Regulators, Soil Data Providers and Technical Advisors

10

Alberta Background Soil Quality System

Phase 1: Soil Quality Database Development – Definitions

- **Background:** concentrations of salinity and metals that are naturally occurring and unrelated to the discharge of pollutants or hazardous substances, or related to other anthropogenic activities
- **Impacted:** 'not background' – samples with chemical signatures affected by anthropogenic activity + those that are a mix of background and impacted signatures
- **Background fingerprints:** a visual depiction of the relationship between multiple chemicals representative of background conditions.
- **Dimensionality Reduction:** transforming data from a large set of variables (i.e., individual salinity and metal chemical parameters) into a smaller set that still contains most of the information from the larger set.
- **Clustering:** a method of statistical analysis that groups samples in such a way that they are more similar to other samples within the same group than they are to samples in other groups.

11

Soil Chemistry Data Acquisition Approach by Paul Fuellbrandt

Soil Chemistry Data Acquisition Approach

Data Tiers

1. **Georeferenced datasets (point data) collected >2015**
   - Standardized data collection & analytical methods
   - Less compilation efforts, comprehensive datasets (higher quality), additional metadata
2. **Older georeferenced datasets**
   - Still of value, but may require additional cleaning (QA/QC), less metadata
3. **Non-georeferenced datasets (e.g., UWI info – 110 x 110 m)**
   - Still of value, but lower resolution / data accuracy. Will be used to supplement primary dataset.

2

EnvirometriX statvis Trust Through Data InnoTech ALBERTA A SUBSIDIARY OF ALBERTA INNOVATES

Alberta Background Soil Quality System

Data Sharing and Use Agreements Prioritization

1. **PTAC RRRC Producer members who have expressed interest in participating + OWA**
2. **Other non RRRC PTAC participants with known large datasets**
3. **As needed – data gaps**
   - Sand & gravel (aggregate), pipeline companies, transmission lines

3

Alberta Background Soil Quality System

Fingerprinting Background Soil Quality

- **Data compilation:** creation of a single comprehensive database.
- **Dimensionality reduction:** PCA, HCA, t-SNE, and/or UMAP
- **Background fingerprinting:** known background samples used to create fingerprints
- **Removal of impacted samples:** background fingerprints train dimensionality reduction models
- **Model testing:** verified data used to confirm that known distributions match distributions of background data

4

Alberta Background Soil Quality System

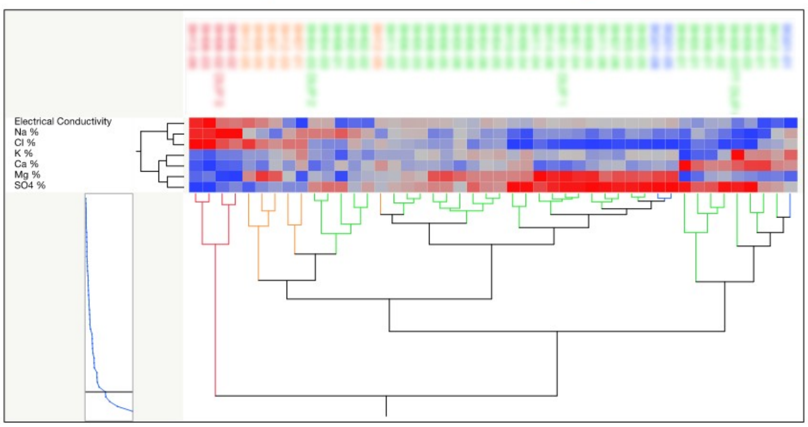
Dimensionality Reduction For a Single Wellsite

EnvirometriX statvis Trust Through Data InnoTech ALBERTA A SUBSIDIARY OF ALBERTA INNOVATES

5

Alberta Background Soil Quality System

Dimensionality Reduction For a Single Wellsite

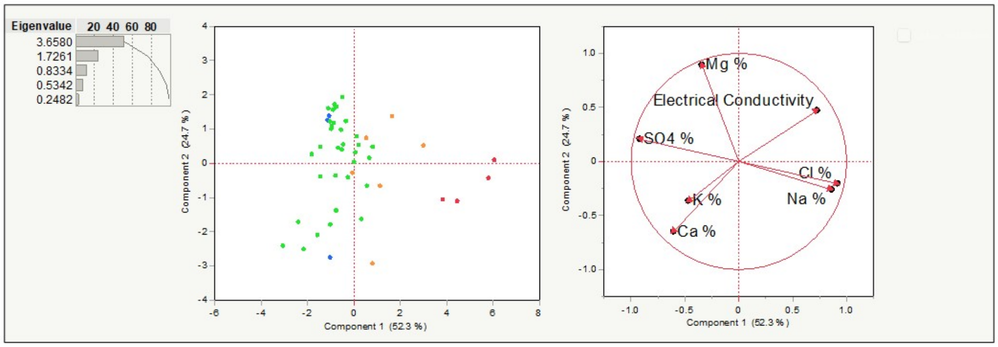


EnvirometriX statvis Trust Through Data InnoTech ALBERTA A SUBSIDIARY OF ALBERTA INNOVATES

6

Alberta Background Soil Quality System

Dimensionality Reduction for a Single Wellsite



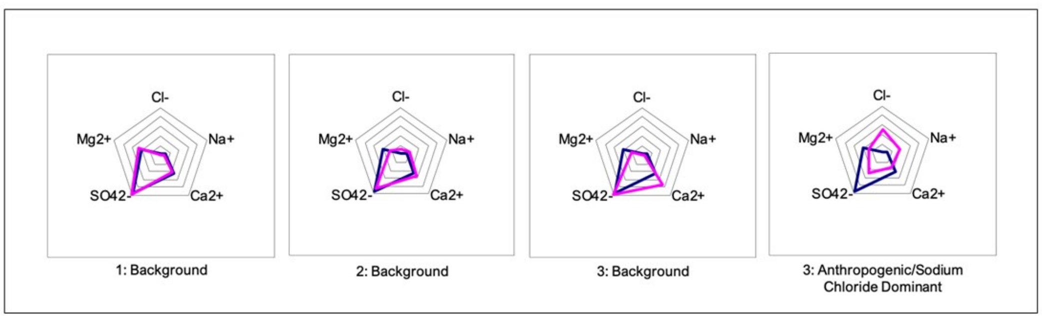
Principal Component Analysis plot with background samples in blue and green, sodium chloride dominated samples in red and mixed samples in orange.

EnvirometriX statvis Trust Through Data InnoTech ALBERTA A SUBSIDIARY OF ALBERTA INNOVATES

7

Alberta Background Soil Quality System

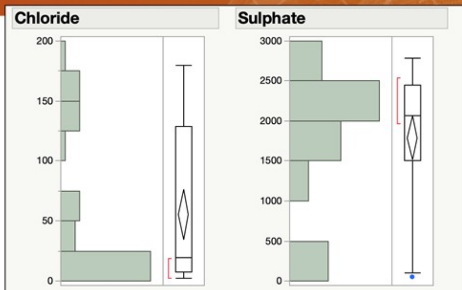
Salt Fingerprinting to Support Dimensionality Reduction



Legend
Average of Background
Sample Data

EnvirometriX statvis Trust Through Data InnoTech ALBERTA A SUBSIDIARY OF ALBERTA INNOVATES

8



Chloride – Quantiles

100.0%	maximum	180
99.5%		180
97.5%		180
90.0%		153.8
75.0%	quartile	129
50.0%	median	19.5
25.0%	quartile	7.5
10.0%		2.5
2.5%		2.5
0.5%		2.5
0.0%	minimum	2.5

Chloride – Summary Statistics

Mean	55.472222
Std Dev	61.261494
Std Err Mean	10.210249
Upper 95% Mean	76.20013
Lower 95% Mean	34.744315
N	36

Sulphate – Quantiles

100.0%	maximum	2790
99.5%		2790
97.5%		2790
90.0%		2552
75.0%	quartile	2450
50.0%	median	2070
25.0%	quartile	1510
10.0%		273.6
2.5%		49
0.5%		49
0.0%	minimum	49

Sulphate – Summary Statistics

Mean	1788.9722
Std Dev	820.49596
Std Err Mean	136.74933
Upper 95% Mean	2066.5881
Lower 95% Mean	1511.3563
N	36

Salt Fingerprinting to Support Dimensionality Reduction

Salinity Parameter	***Maximum Value in Background Groupings (mg/kg unless otherwise specified)***
Chloride	180
Sulphate	2,790
Sodium	352
Calcium	539
Magnesium	380
Potassium	35

EnvirometriX statvis Trust Through Data InnoTech ALBERTA A SUBSIDIARY OF ALBERTA INNOVATES

9

Alberta Background Soil Quality System

Fingerprinting Background Soil Quality - Continued

Database: Only data designated and verified to be 'background' will go into the final database, predictive maps, and web application.

Data owner & TSC verification: validation (i.e., viewing and testing the database and web interface are being built.

EnvirometriX statvis Trust Through Data InnoTech ALBERTA A SUBSIDIARY OF ALBERTA INNOVATES

10

Alberta Background Soil Quality System

AB Background Soil Quality System

- Background fingerprinting based on pattern recognition not concentration
- Modern data science, ML data dimensionality reduction now widely available
- Impacted samples included originally to learn differences between impacted and background

RESULT: Data-driven process removes potential for invisible bias

VS.

Other Background Databases

- Pre-determined concentration limits for background
- Simple statistics (correlations etc.) were the only tools widely available until recently
- Impacted samples removed based on assumptions, location, experience, etc.

RESULT: Potential for missing naturally elevated parameters

11

EnvirometriX statvis Trust Through Data InnoTech ALBERTA A SUBSIDIARY OF ALBERTA INNOVATES

Soil Quality Predictive Mapping Approach by Tomislav Hengl and Leandro Parente

EnvirometriX role in the project:

- Predictive Soil Mapping (PSM): overlay, data mining, fine-tuning, Ensemble Machine Learning, prediction, validation, uncertainty assessment;
- Auxiliary data collation (Analysis-Ready Data Cube for Alberta), cleaning and integration with the PSM;
- Developing uncertainty map;
- Model expansion to Provincial scale (upscaling to: (A) general multivariate case, (B) further automation, (C) multiscaling / finer resolutions);

2

EnvirometriX statvis Trust Through Data InnoTech ALBERTA A SUBSIDIARY OF ALBERTA INNOVATES

In this presentation

- Basics of **PSM based on Machine Learning**;
- Recent **trends and developments in PSM**:
    - Ensemble ML;
    - 2D, 3D, 2D+T / 3D+T Machine Learning;
    - Example with USGS geochemical data;
- Building an **Environmental Data Cube of Alberta**:
    - Compiling publicly available datasets (250-m, 100-m, 30-m, 10-m);
    - Why is it important? Which are the best practices?
- Conclusions and next steps.

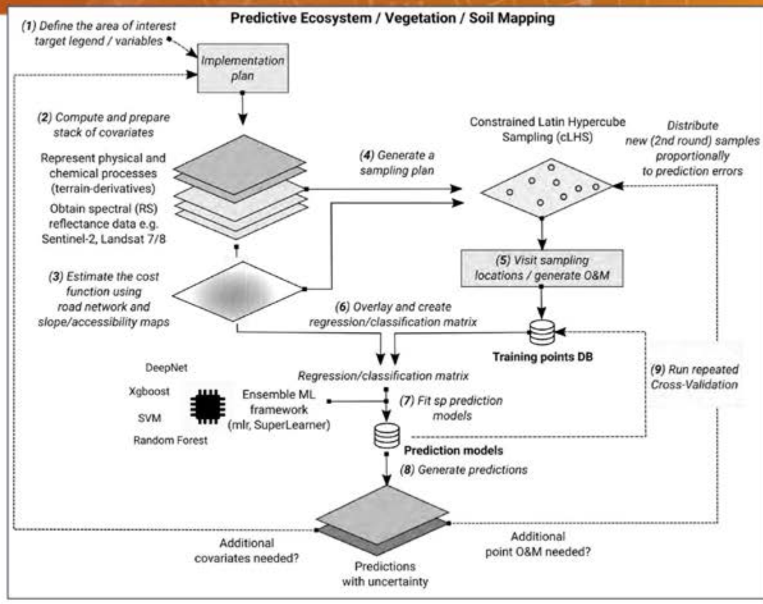
EnvirometriX statvis Trust Through Data InnoTech ALBERTA A SUBSIDIARY OF ALBERTA INNOVATES

3

What is Predictive Soil Mapping and how does Machine Learning helps make better maps?

EnvirometriX statvis InnoTech ALBERTA A SUBSIDIARY OF ALBERTA INNOVATES

Predictive soil mapping in a nutshell



11

Predictive soil mapping in a nutshell (6 steps)

Standard processes:

1. **Point and gridded data preparation**:
   a. Preparation of point data (harmonization, standardization, removal of artifacts, transformation);
   b. Preparation of covariate layers (stacking, gap-filling, artifact reduction);
2. **Preparation of regression/classification matrix**:
   a. Spatial overlay, transformation, removal of zero-variance covariates,
3. **Fine-tuning, feature selection, optimization / model training**:
   a. Pre-selection of best-learners;
   b. Fine-tuning / benchmarking / feature selection;
   c. Model stacking;

12

Predictive soil mapping in a nutshell (6 steps)

4. **Model validation / accuracy assessment**:
    a. Cross-validation with model refitting;
    b. Estimation of prediction errors (unbiased / realistic estimate of uncertainty);
5. **Prediction and visualization of results**:
    a. Production of predictions and uncertainty (per pixel);
    b. Aggregation and visualization of results / expert-based assessment;
    c. Post-modeling diagnostic (**Interpretable Machine Learning**);
6. **Re-analysis and model improvements**:
    a. Production of new sampling plans (uncertainty-based sampling);
    b. Total reanalysis and re-production of predictions;

EnvirometriX statvis Trust Through Data InnoTech ALBERTA A SUBSIDIARY OF ALBERTA INNOVATES

13

Recent trends in PSM: Ensemble Machine Learning (mlr)

EnvirometriX statvis InnoTech ALBERTA A SUBSIDIARY OF ALBERTA INNOVATES

Going further: Ensemble approach to ML

When it comes to ensemble prediction using ML, four packages seem to be highly fitted for the purpose:

★ h2o

★ caretEnsemble

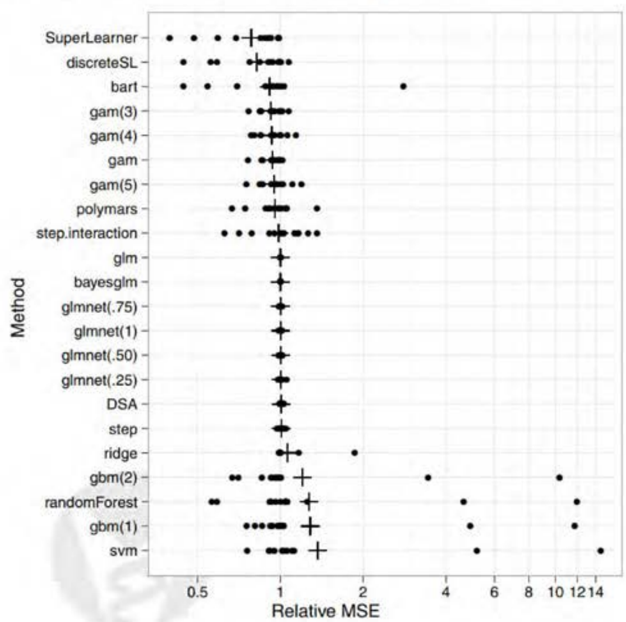
★ mlr

★ SuperLearner / subsemble

EnvirometriX statvis InnoTech ALBERTA

18

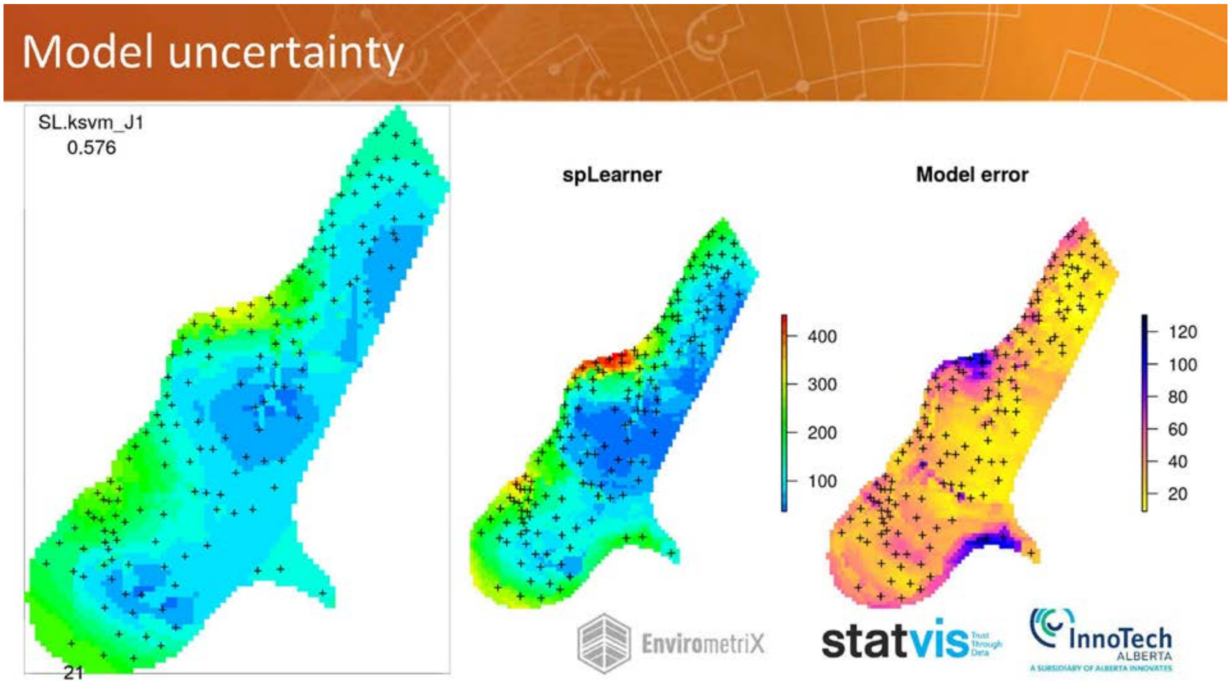
Figure 3: 10-fold cross-validated relative mean squared error compared to glm across 13 real datasets. Sorted by the geometric mean, denoted with the plus (+) sign.

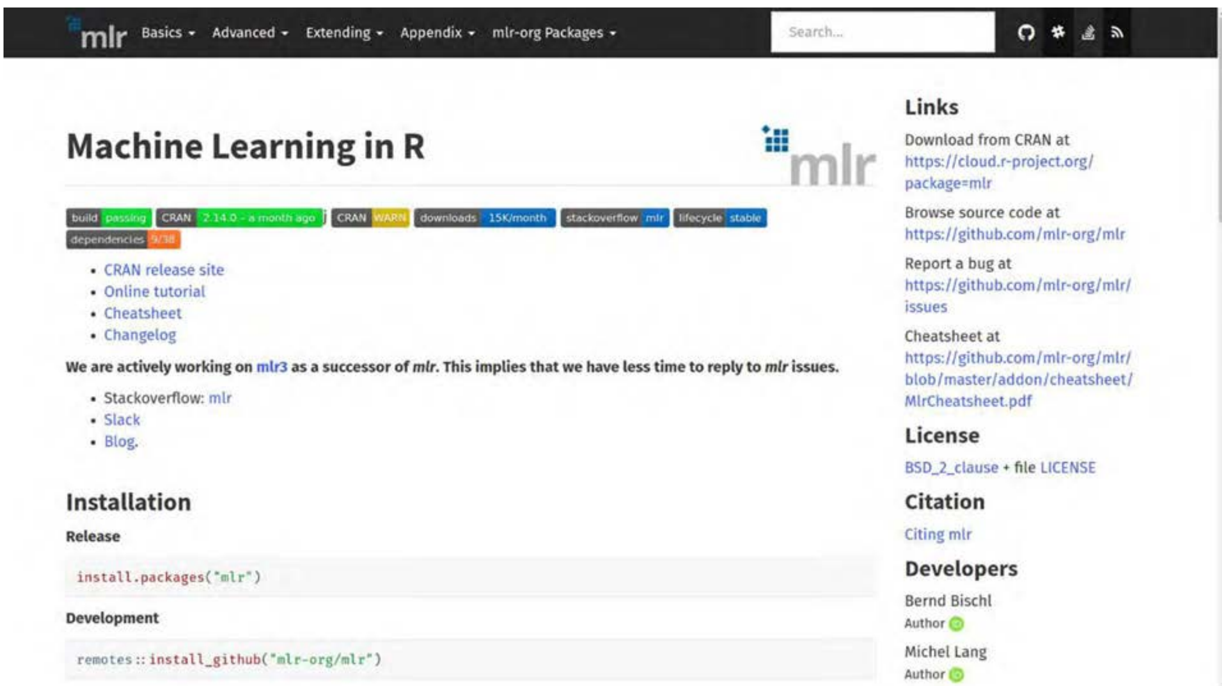


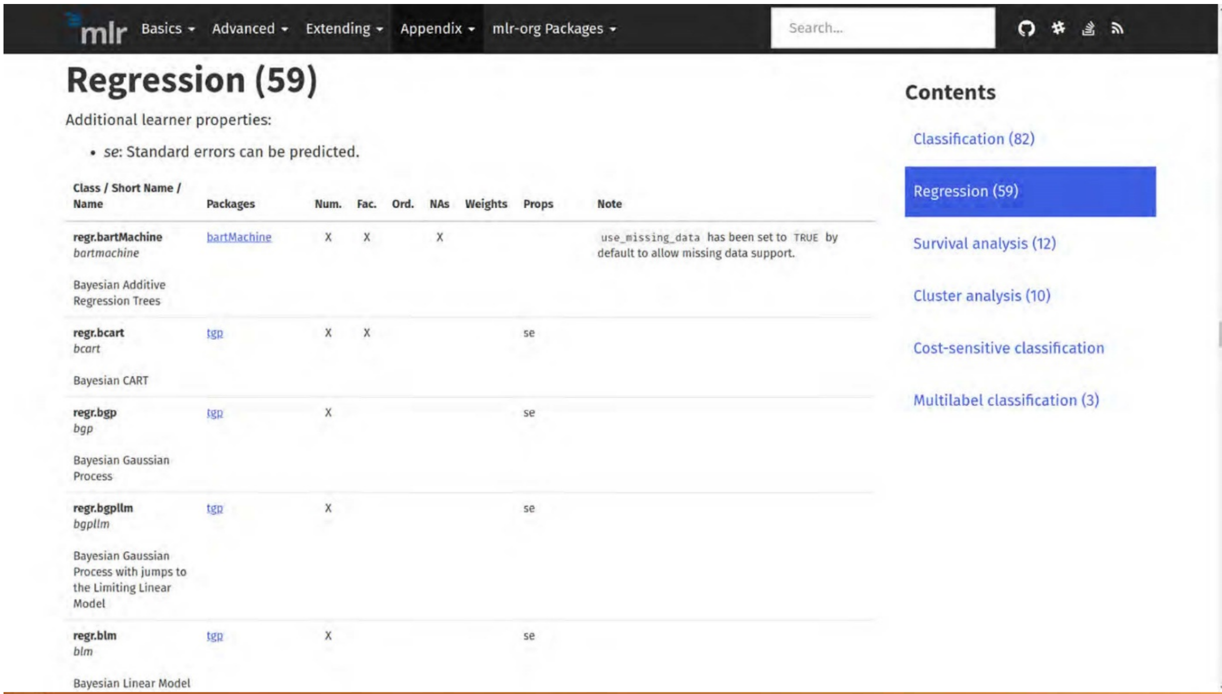
"Ensemble methods are meta-algorithms that combine several machine learning techniques into one predictive model in order to **decrease variance** (bagging), **bias** (boosting), or **improve predictions** (stacking)."
https://blog.statsbot.co/ensemble-learning-d1dcd548e936

This however comes at costs:
- higher computational load,
- higher RAM requirements,

statvis InnoTech ALBERTA


Model uncertainty
SL.ksvm_J1
0.576
spLearner
Model error
400
300
200
100
120
100
80
60
40
20
EnvirometriX
statvis
InnoTech
ALBERTA
21


mlr
Basics
Advanced
Extending
Appendix
mlr-org Packages
Search...
Machine Learning in R
mlr
build passing
CRAN 2.14.0 - a month ago
CRAN WARN
downloads 15K/month
stackoverflow mlr
lifecycle stable
dependencies 9/38
• CRAN release site
• Online tutorial
• Cheatsheet
• Changelog
We are actively working on mlr3 as a successor of mlr. This implies that we have less time to reply to mlr issues.
• Stackoverflow: mlr
• Slack
• Blog.
Installation
Release
install.packages("mlr")
Development
remotes::install_github("mlr-org/mlr")
Links
Download from CRAN at https://cloud.r-project.org/package=mlr
Browse source code at https://github.com/mlr-org/mlr
Report a bug at https://github.com/mlr-org/mlr/issues
Cheatsheet at https://github.com/mlr-org/mlr/blob/master/addon/cheatsheet/MlrCheatsheet.pdf
License
BSD_2_clause + file LICENSE
Citation
Citing mlr
Developers
Bernd Bischl
Author
Michel Lang
Author



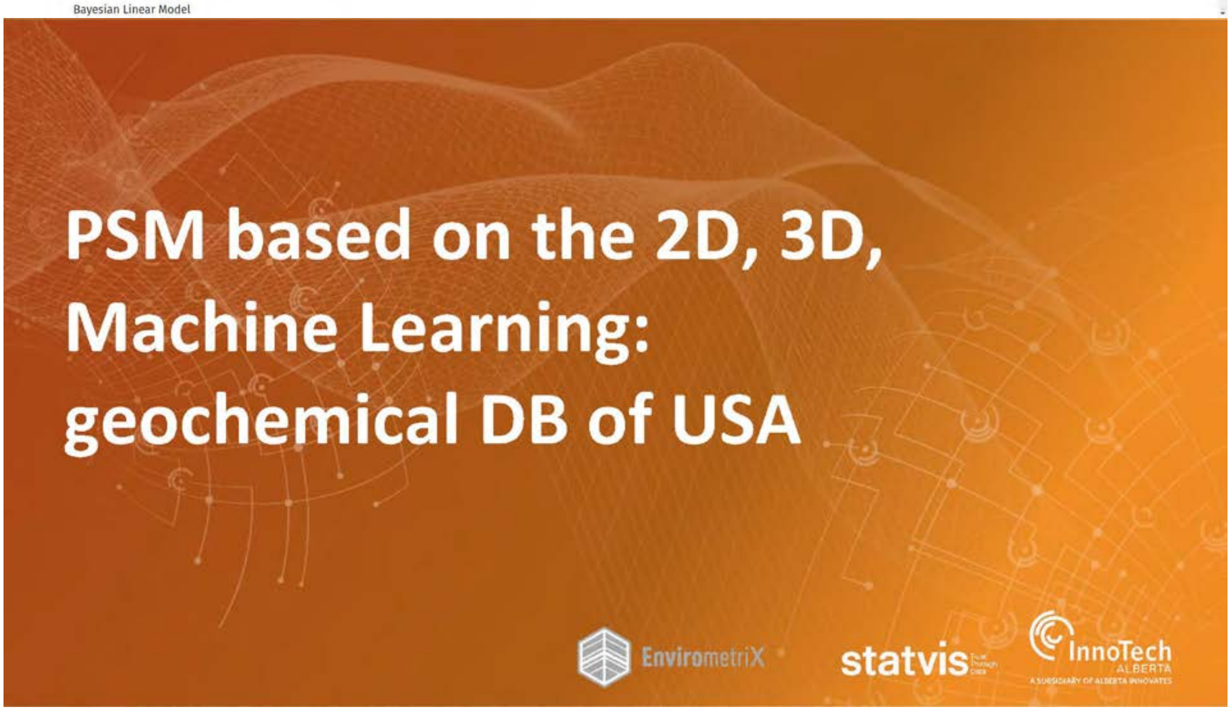
Regression (59)

Additional learner properties:

- *se*: Standard errors can be predicted.

Class / Short Name / Name	Packages	Num.	Fac.	Ord.	NAs	Weights	Props	Note
regr.bartMachine *bartmachine* Bayesian Additive Regression Trees	bartMachine	X	X		X			`use_missing_data` has been set to `TRUE` by default to allow missing data support.
regr.bcart *bcart* Bayesian CART	tgp	X	X				se	
regr.bgp *bgp* Bayesian Gaussian Process	tgp	X					se	
regr.bgpllm *bgpllm* Bayesian Gaussian Process with jumps to the Limiting Linear Model	tgp	X					se	
regr.blm *blm* Bayesian Linear Model	tgp	X					se	

Contents



USGS National Geochemical Database (Soil)


USGS
science for a changing world
USGS Home Contact USGS Search USGS
Data Series 801

Geochemical and Mineralogical Data for Soils of the Conterminous United States

By David B. Smith, William F. Cannon, Laurel G. Woodruff, Federico Solano, James E. Kilburn, and David L. Fey

Abstract

In 2007, the U.S. Geological Survey initiated a low-density (1 site per 1,600 square kilometers, 4,857 sites) geochemical and mineralogical survey of soils of the conterminous United States as part of the North American Soil Geochemical Landscapes Project. Sampling and analytical protocols were developed at a workshop in 2003, and pilot studies were conducted from 2004 to 2007 to test and refine these recommended protocols. The final sampling protocol for the national-scale survey included, at each site, a sample from a depth of 0 to 5 centimeters, a composite of the soil A horizon, and a deeper sample from the soil C horizon or, if the top of the C horizon was at a depth greater than 1 meter, from a depth of approximately 80–100 centimeters. The <2-millimeter fraction of each sample was analyzed for a suite of 45 major and trace elements by methods that yield the total or near-total elemental content. The major mineralogical components in the samples from the soil A and C horizons were determined by a quantitative X-ray diffraction method using Rietveld refinement. Sampling in the conterminous United States was completed in 2010, with chemical and mineralogical analyses completed in May 2013. The resulting dataset provides an estimate of the abundance and spatial distribution of chemical elements and minerals in soils of the conterminous United States and represents a baseline for soil geochemistry and mineralogy against which future changes may be recognized and quantified. This report (1) describes the sampling, sample preparation, and analytical methods used; (2) gives details of the quality control protocols used to monitor the quality of chemical and mineralogical analyses over approximately six years; and (3) makes available the soil geochemical and mineralogical data in downloadable tables.

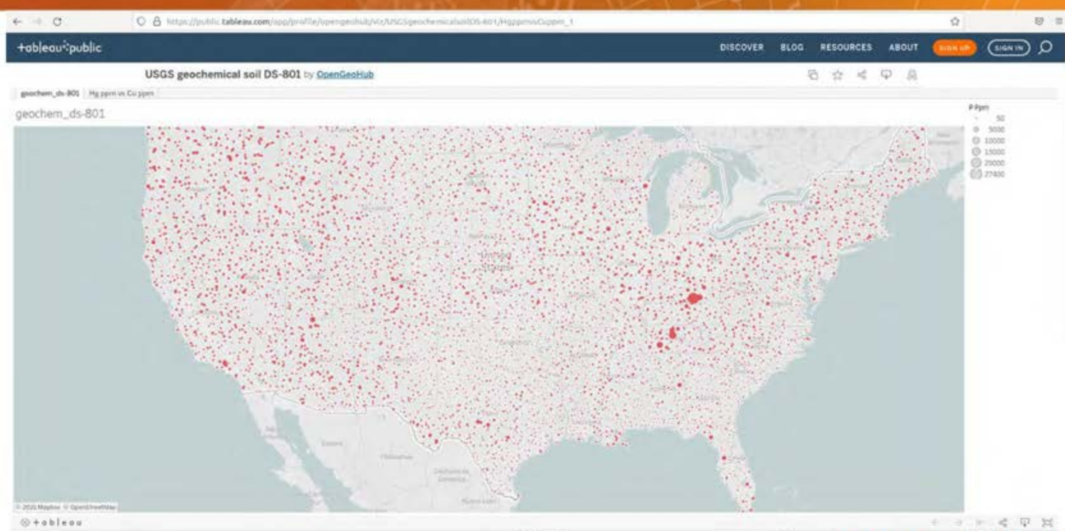
First posted October 25, 2013

- Report PDF (11.3 MB)
- Downloads Directory
  Contains: Appendixes 1–5

For additional information contact:
Director, Central Mineral and Environmental Resources Science Center
U.S. Geological Survey
Box 25046, MS-973
Denver, CO 80225
http://minerals.cr.usgs.gov/

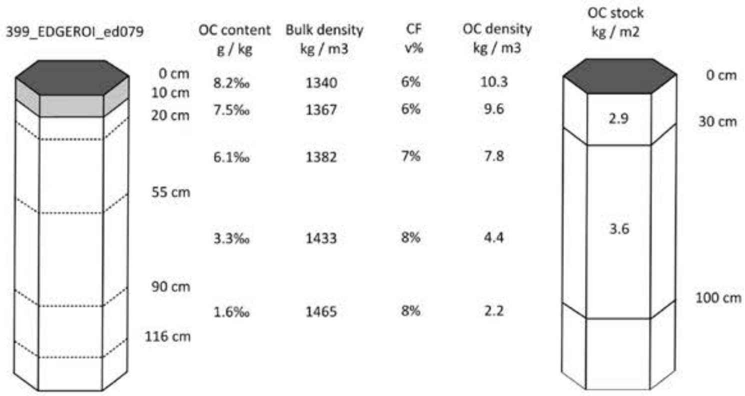
Part or all of this report is presented in Portable Document Format (PDF); the latest version of Adobe Reader or similar

To interactively explore data use:



28 https://mrdata.usgs.gov/ngdb/soil/

3D = vertical depth at which samples are taken



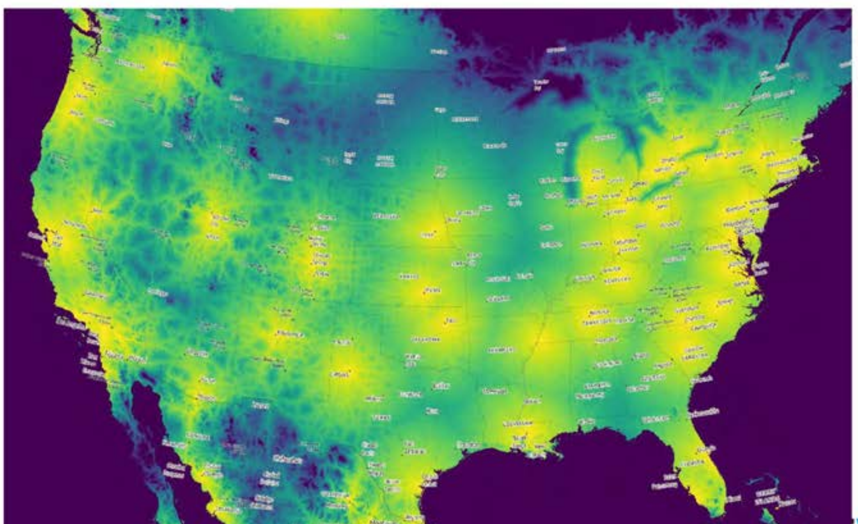
Example of an actual soil profile from Australia

https://soilmapper.org/SOC-chapter.html

EnvirometriX statvis Trust Through Data InnoTech ALBERTA A SUBSIDIARY OF ALBERTA INNOVATES

29

Covariate layers at 1-km resolution (about 230 used)



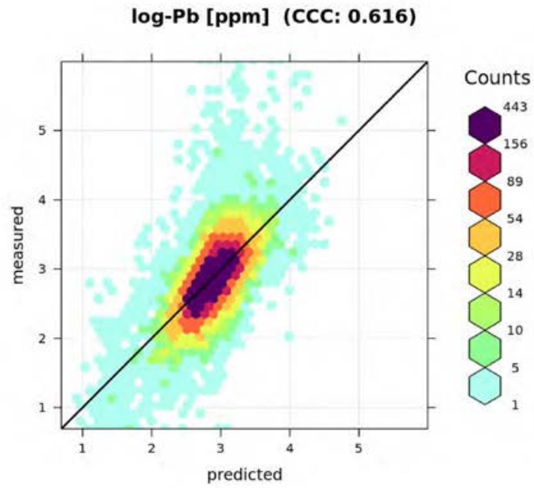
- Distance to cities;
- MODIS LST (monthly daytime and nighttime);
- Lights at night images;
- MODIS EVI (monthly);
- Terrain / hydrological indices;
- ...
- Depth

statvis Trust Through Data InnoTech ALBERTA A SUBSIDIARY OF ALBERTA INNOVATES

31 https://www.nature.com/articles/s41597-019-0265-5

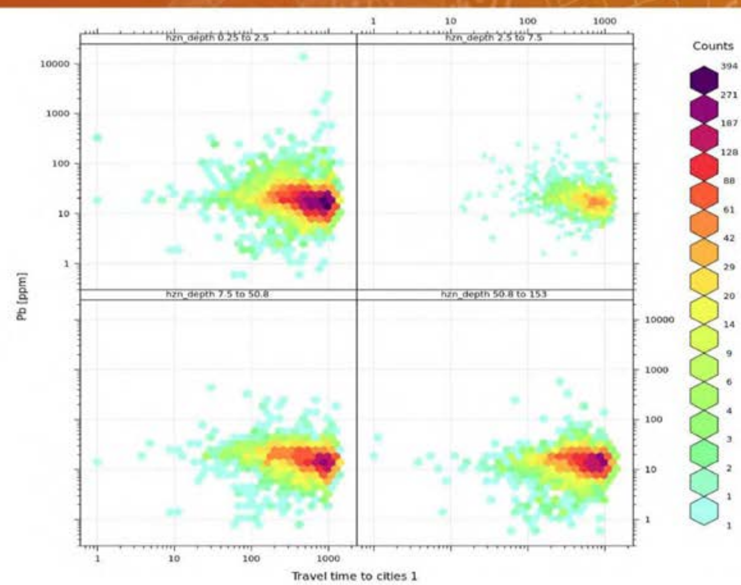
Results: Pb [ppm]

```
Results of ensemble model fitting 'ranger', 'xgboost', ...:

Variable: log.pb
R-square: 0.446
Fitted values sd: 0.361
RMSE: 0.403

EML model summary:
Call:
stats::lm(formula = f, data = d)

Residuals:
    Min      1Q  Median      3Q     Max
-2.4230 -0.1973 -0.0100  0.1683  6.3319

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -0.28123    0.03232  -8.703  < 2e-16 ***
regr.ranger    0.66183    0.02898  22.837  < 2e-16 ***
regr.xgboost   0.27315    0.02710  10.081  < 2e-16 ***
regr.cvglmnet  0.16133    0.01994   8.089 6.51e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4028 on 14260 degrees of freedom
Multiple R-squared:  0.4456,    Adjusted R-squared:  0.4455
F-statistic:  3821 on 3 and 14260 DF,  p-value: < 2.2e-16
```

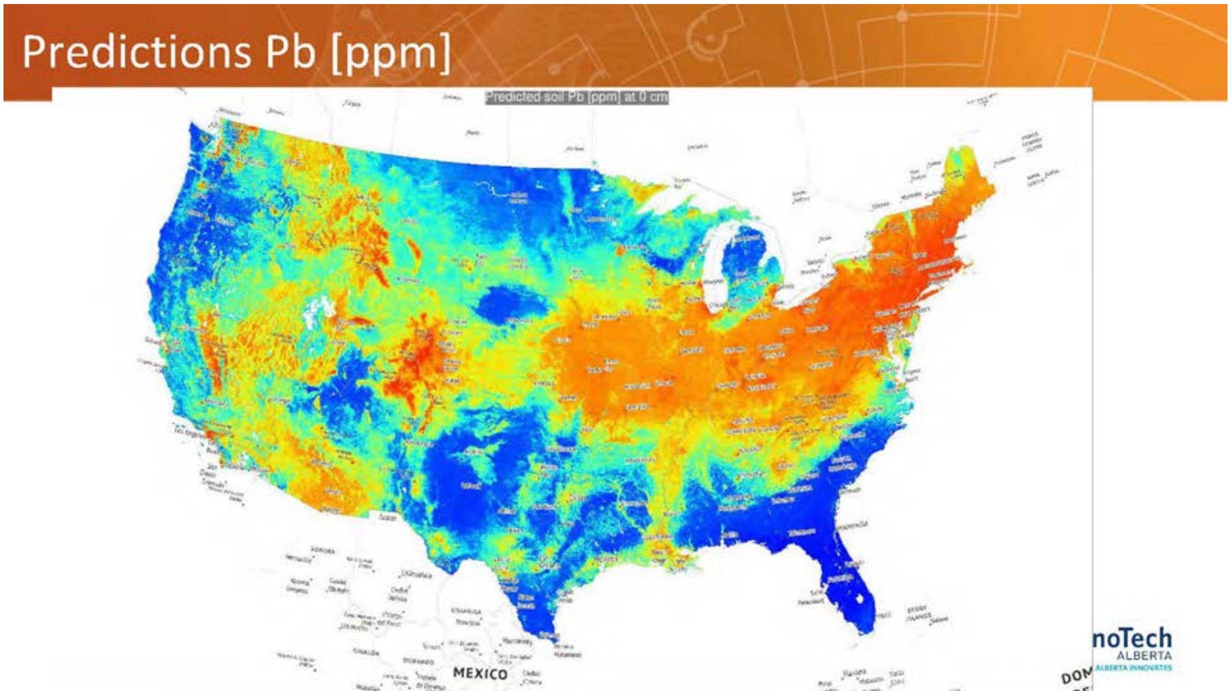


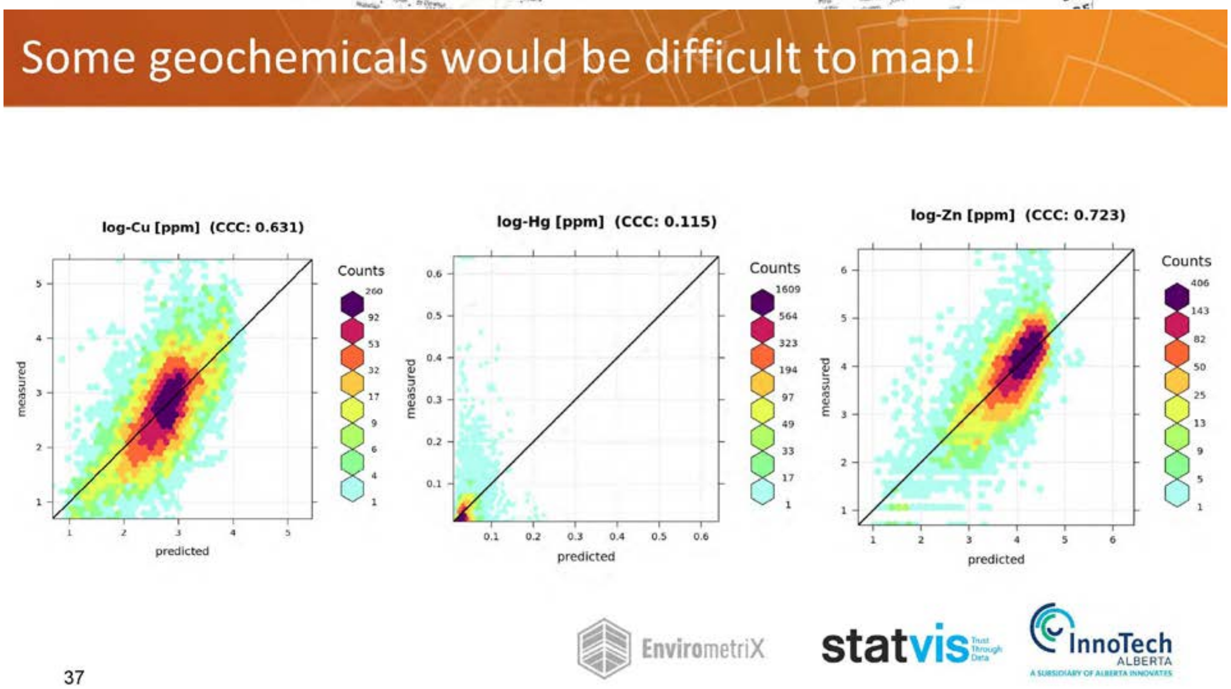
32

Results: cities increase Pb [ppm]



34

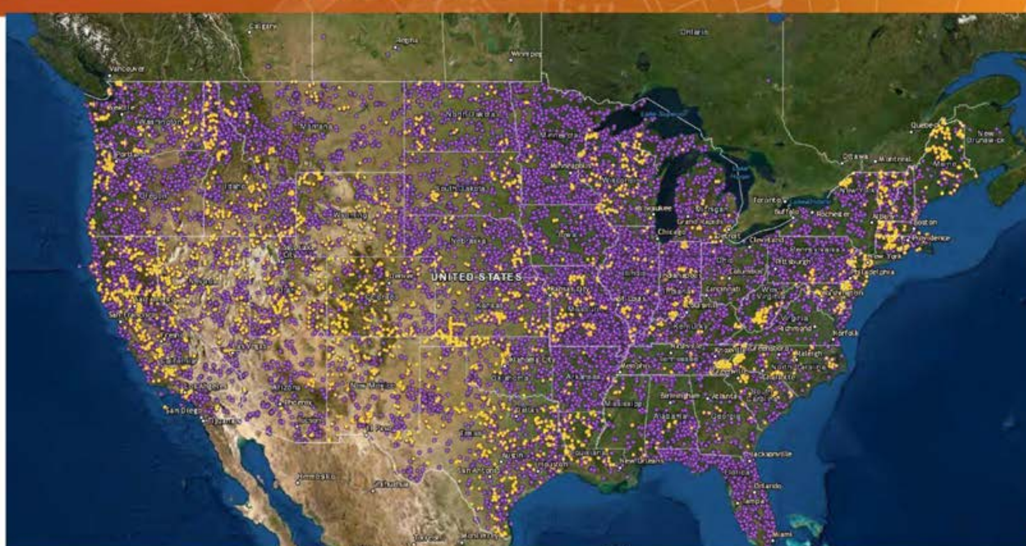

Predictions Pb [ppm]
Predicted soil Pb [ppm] at 0 cm
MEXICO
InnoTech
ALBERTA


Some geochemicals would be difficult to map!
log-Cu [ppm] (CCC: 0.631)
log-Hg [ppm] (CCC: 0.115)
log-Zn [ppm] (CCC: 0.723)
measured
predicted
Counts
EnvirometriX
statvis
InnoTech
ALBERTA
A SUBSIDIARY OF ALBERTA INNOVATES
37

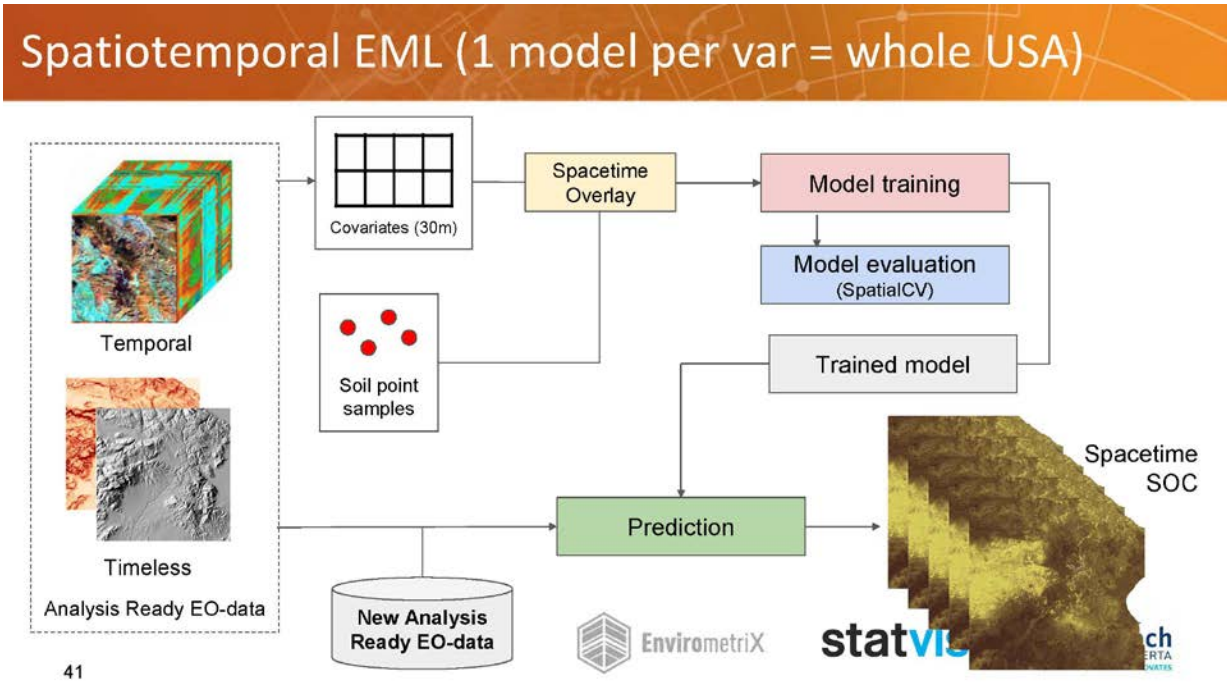
PSM based on the 2D+T / 3D+T Machine Learning

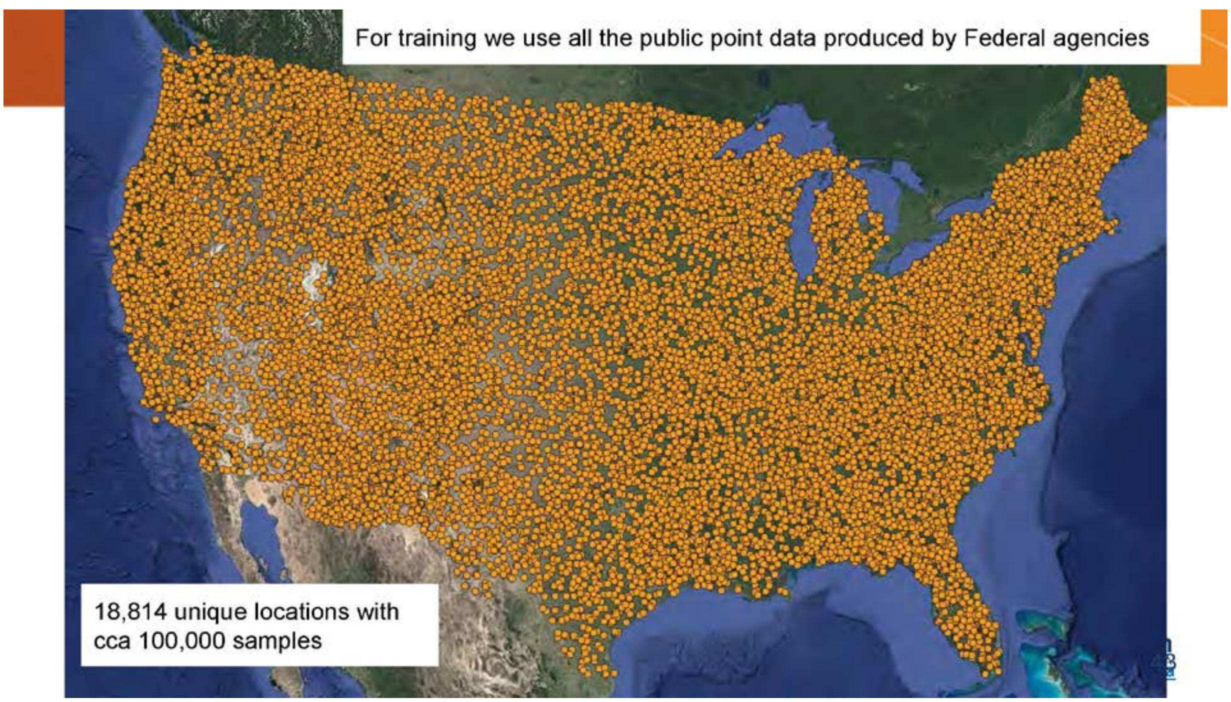
EnvirometriX statvis InnoTech ALBERTA A SUBSIDIARY OF ALBERTA INNOVATES

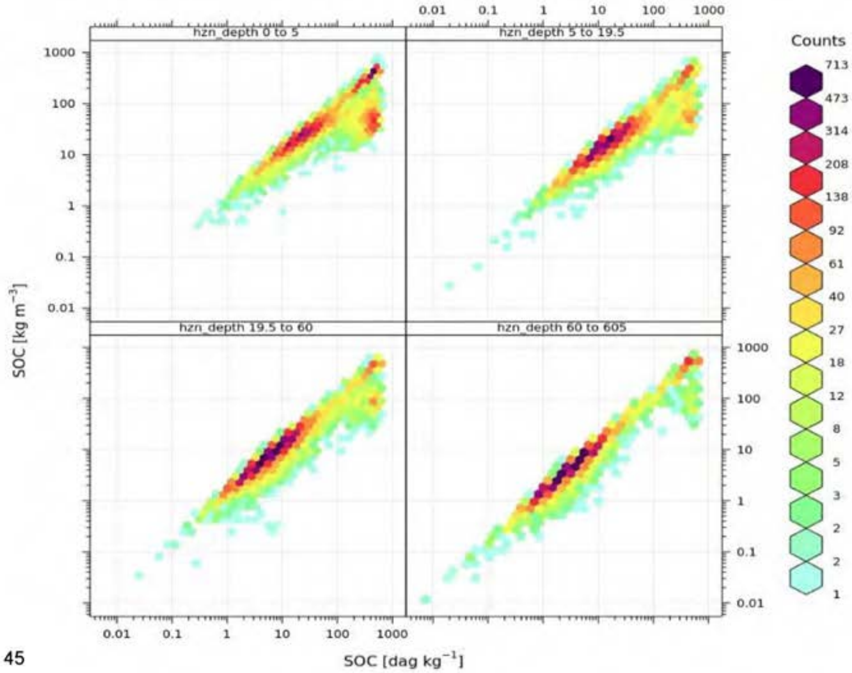
Soil Carbon mapping spacetime (USA48)



40 https://ncsslabdatamart.sc.egov.usda.gov/


Spatiotemporal EML (1 model per var = whole USA)
Covariates (30m)
Spacetime Overlay
Model training
Model evaluation (SpatialCV)
Temporal
Soil point samples
Trained model
Spacetime SOC
Timeless
Prediction
Analysis Ready EO-data
New Analysis Ready EO-data
EnvirometriX
41


For training we use all the public point data produced by Federal agencies
18,814 unique locations with cca 100,000 samples



Without FIADB relationship between SOC % and SOC kg/m3 is **more common for mineral soils**; consequently, we have decided to take the FIADB outside of modeling;

45

Key to getting better: better covariates

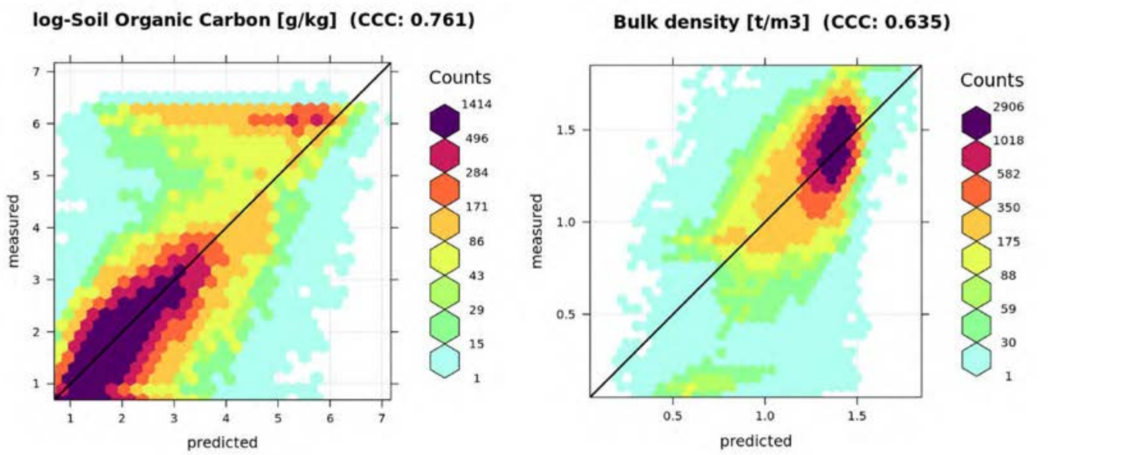
Over 400 covariate layers including:

- Landsat P25, P50, P75 bands spring, summer, autumn for 2000-2020 (8TB data);
- MODIS LST monthly P5, P50 and P95 day-time, night-time;
- Crop frequencies soybeans, wheat, corn, cotton (2010-2020),
- Climate normals 1990-2020 downscaled;
- Snow probability long-term;
- DTM derivatives;
- Soil taxonomy and PSCS 100-m resolution layers;
- ...

https://nassgeodata.gmu.edu/CropScape/

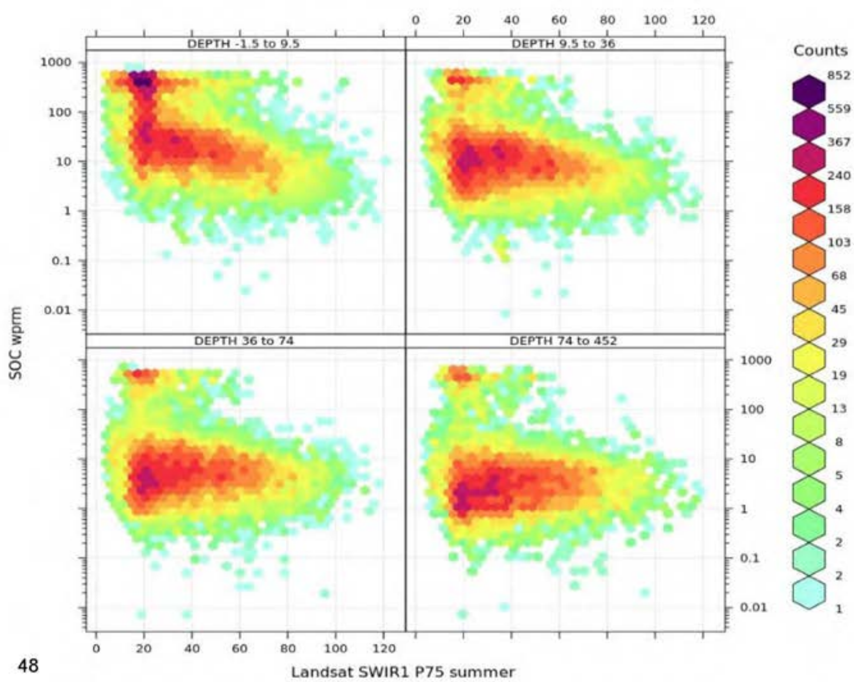
46

Results (spatial 5-fold Cross-Validation)



https://gitlab.com/geoharmonizer_inea/odse-workshop-2021/-/tree/main/R-training

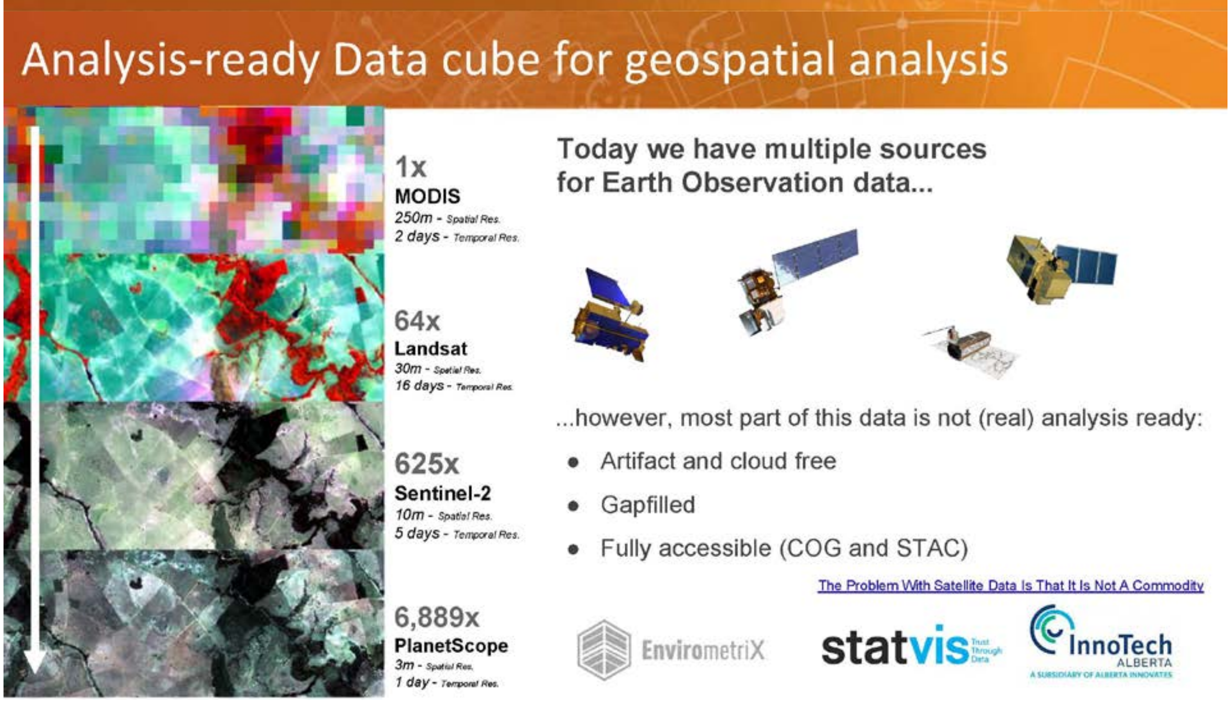
47



Great news! correlation with Landsat bands P75 summer is a good sign indicating that SOC can be mapped at high spatial resolution. With soil-depth relationship is less distinct.

48


Building an Environmental Data Cube for Alberta: Why is it important? Which are the best practices?
EnvirometriX
statvis
InnoTech
ALBERTA
A SUBSIDIARY OF ALBERTA INNOVATES


Analysis-ready Data cube for geospatial analysis
1x
MODIS
250m - Spatial Res.
2 days - Temporal Res.
64x
Landsat
30m - Spatial Res.
16 days - Temporal Res.
625x
Sentinel-2
10m - Spatial Res.
5 days - Temporal Res.
6,889x
PlanetScope
3m - Spatial Res.
1 day - Temporal Res.
Today we have multiple sources for Earth Observation data...
...however, most part of this data is not (real) analysis ready:
Artifact and cloud free
Gapfilled
Fully accessible (COG and STAC)
The Problem With Satellite Data Is That It Is Not A Commodity
EnvirometriX
statvis
InnoTech
ALBERTA
A SUBSIDIARY OF ALBERTA INNOVATES

Analysis-ready Data cube for geospatial analysis

Basic definitions and target requirements:

- **"A data cube"**: complete consistent current Analysis-Ready data stack;
- Ideal case: perfectly stacked layers to the same grid (multivariate matrices);
- Cloud-ready data format: Cloud Optimized GeoTIFF (COG - raster database);
- Ideally: fully documented / fully automated i.e. can be reproduced, updated and extended;

52

EnvirometriX statvis Trust Through Data InnoTech ALBERTA A SUBSIDIARY OF ALBERTA INNOVATES

Why invest time into building a Data Cube?

Basic definitions and target requirements:

- The larger the Data Cube, the higher the chance we can find correlation with target variables;
- Most importantly we can improve mapping accuracy / **reduce risks of making decisions**;
- We can also better understand what are the key drivers / soil forming factors;

53

EnvirometriX statvis Trust Through Data InnoTech ALBERTA A SUBSIDIARY OF ALBERTA INNOVATES

Spectral Reflectance ARD - From 2000 to 2020

GLAD Landsat ARD

- Globally consistent analysis ready data (ARD) for multi-decadal LCLU monitoring
- 16-day time-series composites from Landsat 5, 7 and 8 (TM, ETM+ and OLI)
- Per-pixel observation quality flag
- MODIS (MOD44C) surface reflectance calibrated
- Product organized by 1 × 1 degree tiles
- Automatically download through HTTP API
- Product under Creative Commons Attribution License

FORCE

- An all-in-one remote sensing processing framework for Sentinel-2 A/B MSI and Landsat 5, 7 and 8 (TM, ETM+ and OLI)
- Advanced cloud and cloud shadow detection
- Integrated atmospheric, topographic and BRDF correction
- Reprojection and grinding capabilities
- Different strategies to generate composites (e.g. best available pixel, spectral temporal metrics)
- Free software under GNU License v.3

EnvirometriX statvis Trust Through Data InnoTech ALBERTA A SUBSIDIARY OF ALBERTA INNOVATES

54

Spectral Reflectance ARD

FORCE
(All levels)

GLAD Landsat ARD
(Level 3 product)

esa USGS science for a changing world
(Level 1 and on-demand level 2 products)

Level 4
(Model output)

Level 3
(Temporal composites and gridded data)

Level 2
(Atmospheric correction)

Level 1
(Radiometrically calibrated and georectified data)

Proper input for machine learning approaches (Level 3 and Level 4)

Asrar, G.; Greenstone, R. (Eds.) MTPE EOS Reference Handbook; NASA/Goddard Space Flight Center: Greenbelt, MD, USA, 1995; p. 281.

EnvirometriX statvis Trust Through Data InnoTech ALBERTA A SUBSIDIARY OF ALBERTA INNOVATES

55

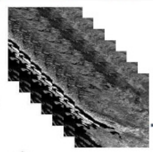
GLAD Landsat ARD

QA codes		
1	land	clear-sky
2	water	clear-sky
3	cloud	
4	cloud shadow	
5	hillshade	clear-sky
6	snow	clear-sky
7	haze	
8	cloud buffer	
9	shadow buffer	
10	shadow high likelihood	
11	additional cloud buffer over land	clear-sky
12	additional cloud buffer over water	clear-sky
14	additional shadow buffer over land	clear-sky
15	land, water detected but not used	clear-sky
16	additional cloud buffer over land, water detected but not used	clear-sky
17	additional shadow buffer over land, water detected but not used	clear-sky

Image layer	Spectral band	Variable (units)	Scaling	Data format
1	Blue	reflectance (unitless)	x 40,000	UInt16
2	Green	reflectance (unitless)	x 40,000	UInt16
3	Red	reflectance (unitless)	x 40,000	UInt16
4	NIR	reflectance (unitless)	x 40,000	UInt16
5	SWIR 1	reflectance (unitless)	x 40,000	UInt16
6	SWIR 2	reflectance (unitless)	x 40,000	UInt16
7	Thermal	brightness temperature (degree C)	x 100	UInt16
8	QA	code, see table below		UInt16

Interval ID	start	end
1	1-Jan	16-Jan
2	17-Jan	1-Feb
3	2-Feb	17-Feb
4	18-Feb	4-Mar
5	5-Mar	20-Mar
6	21-Mar	5-Apr
7	6-Apr	21-Apr
8	22-Apr	7-May
9	8-May	23-May
10	24-May	8-Jun
11	9-Jun	24-Jun
12	25-Jun	10-Jul
13	11-Jul	26-Jul
14	27-Jul	11-Aug
15	12-Aug	27-Aug
16	28-Aug	12-Sep
17	13-Sep	28-Sep
18	29-Sep	14-Oct
19	15-Oct	30-Oct
20	31-Oct	15-Nov
21	16-Nov	1-Dec
22	2-Dec	17-Dec
23	18-Dec	31-Dec

Composite intervals

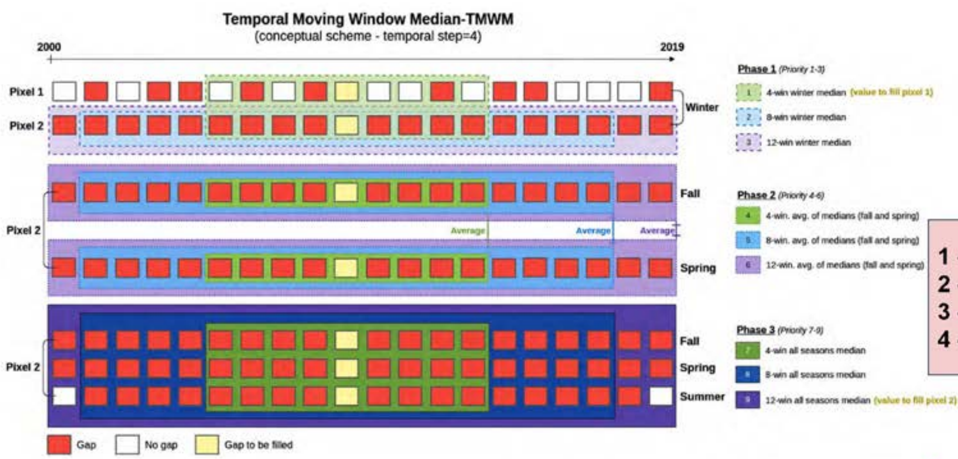


1 - Cloud removal,
2 - temporal aggreg.,
3 - gapfilling,
4 - mosaicking.

EnvirometriX statvis InnoTech ALBERTA

56

GLAD Landsat ARD

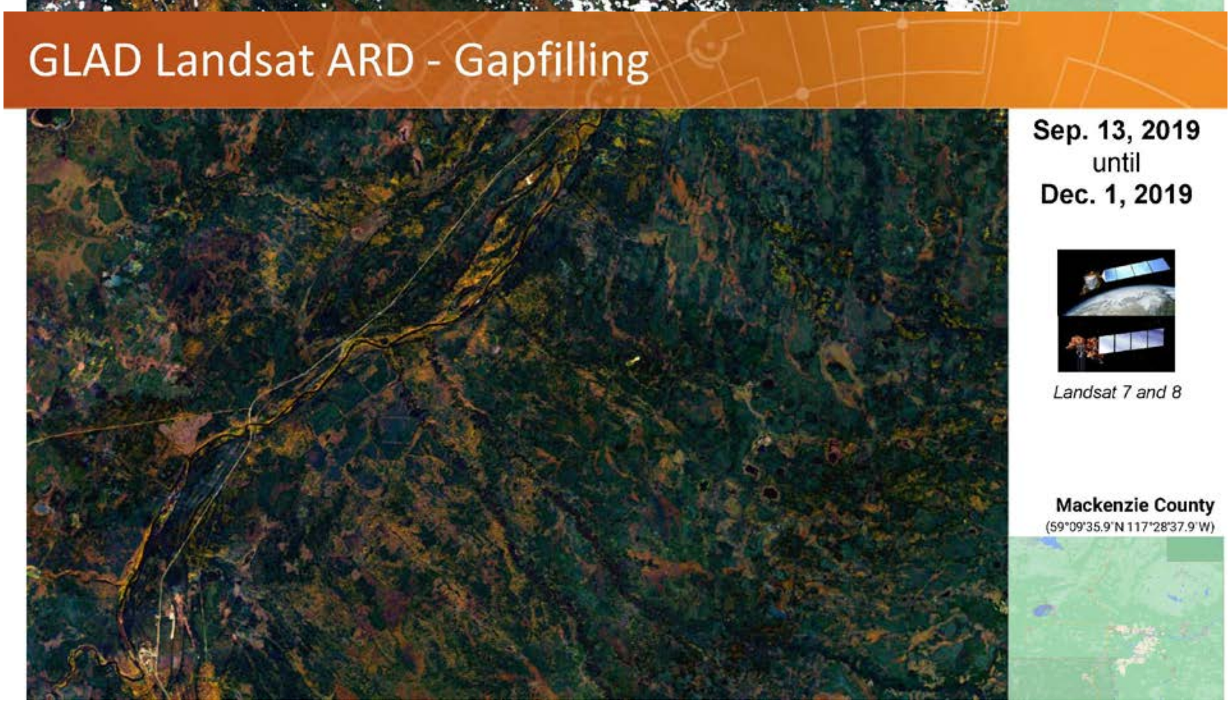


1 - Cloud removal,
2 - temporal aggreg.,
3 - gapfilling,
4 - mosaicking.

EnvirometriX statvis InnoTech ALBERTA

57


GLAD Landsat ARD - With gaps
Sep. 13, 2019
until
Dec. 1, 2019
Landsat 7 and 8
Mackenzie County
(59°09'35.9"N 117°28'37.9"W)


GLAD Landsat ARD - Gapfilling
Sep. 13, 2019
until
Dec. 1, 2019
Landsat 7 and 8
Mackenzie County
(59°09'35.9"N 117°28'37.9"W)

Preparation of Landsat layers

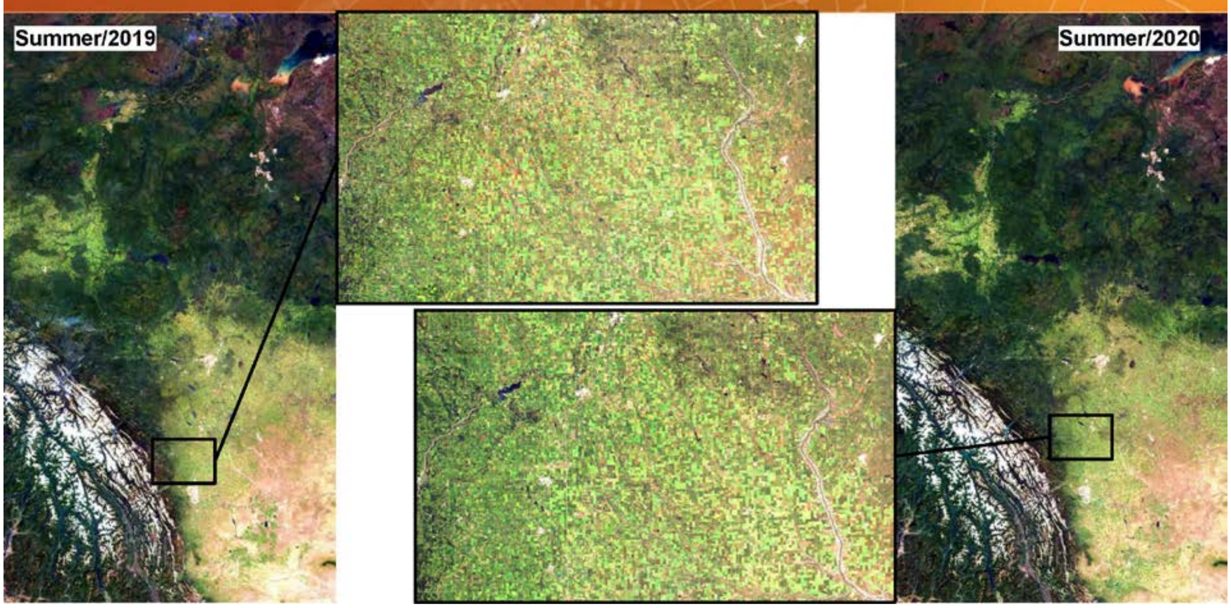
List of covariate layers currently under preparation for Alberta

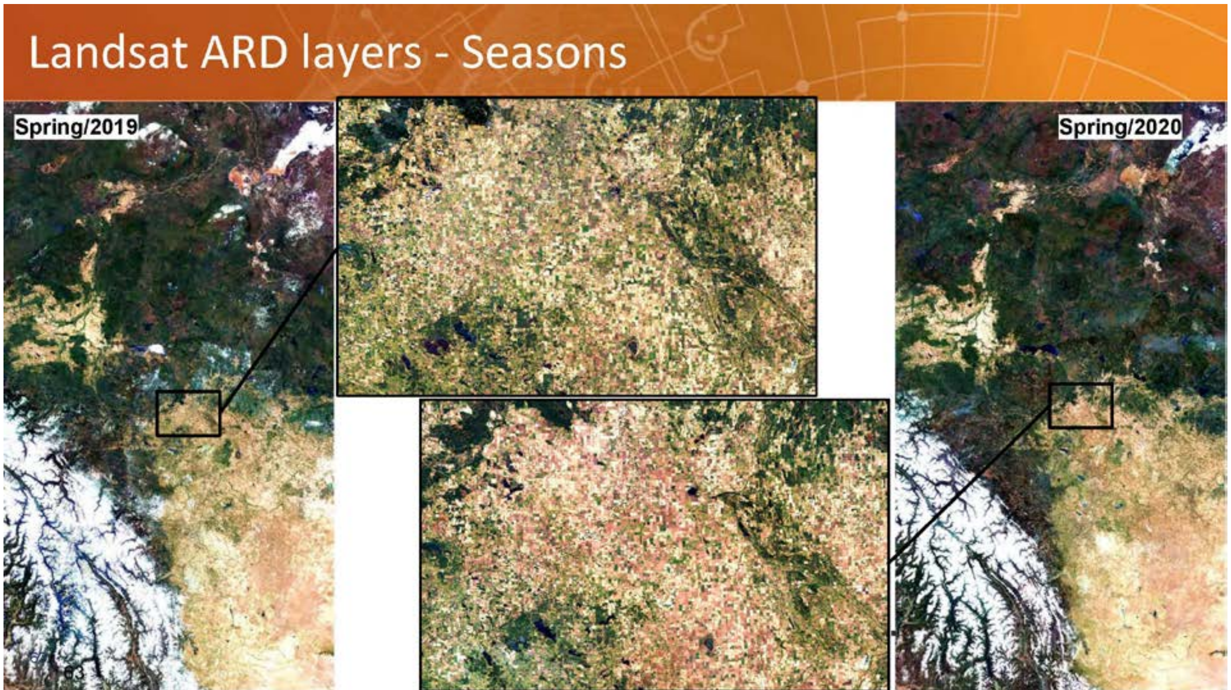
- **Bands:** blue, green, red, NIR, SWIR1, SWIR2, Thermal
- **Time window:**
  - Winter: December 2 of previous year until March 20 of current year
  - Spring: March 21 until June 24 of current year
  - Summer: June 25 until September 12 of current year
  - Fall: September 13 until December 1 of current year

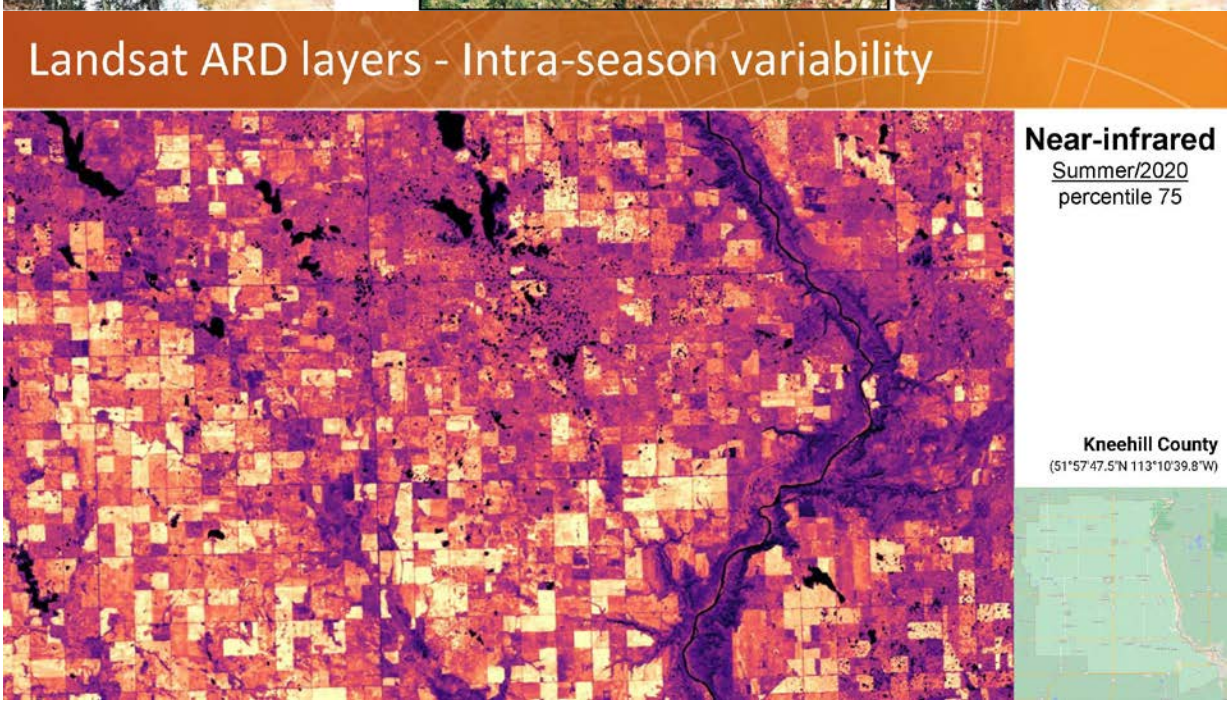
- **Season temporal reduction:** 3 percentiles (p25, p50, p75)

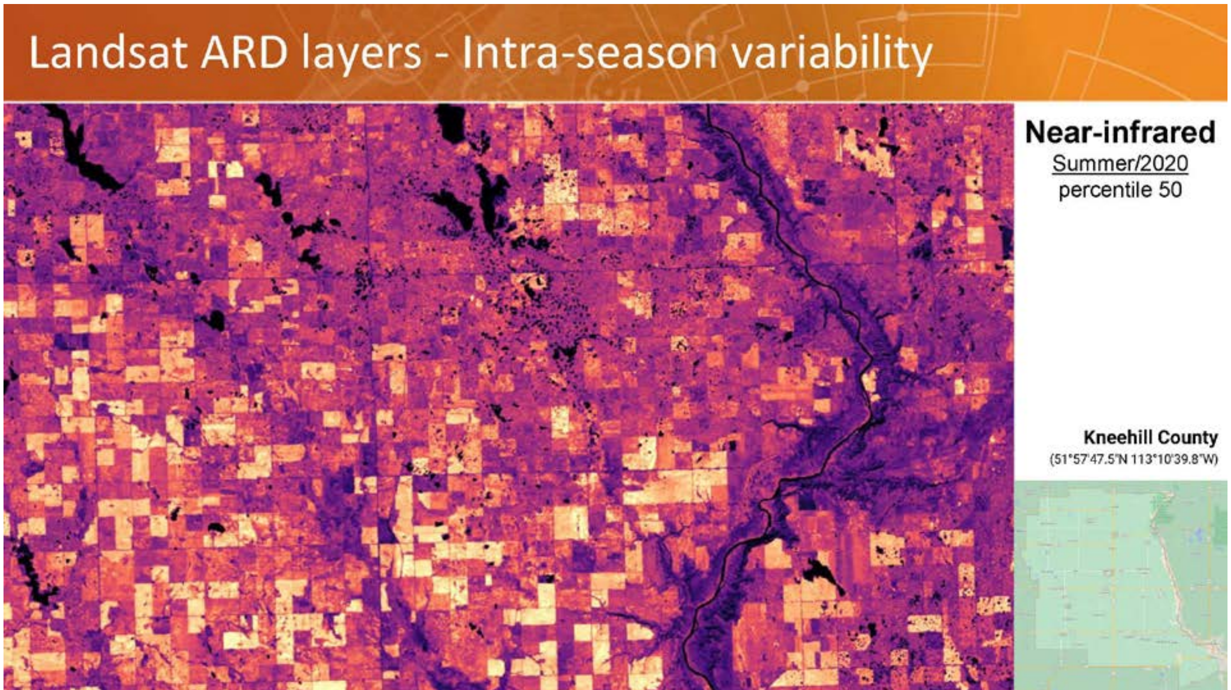
61

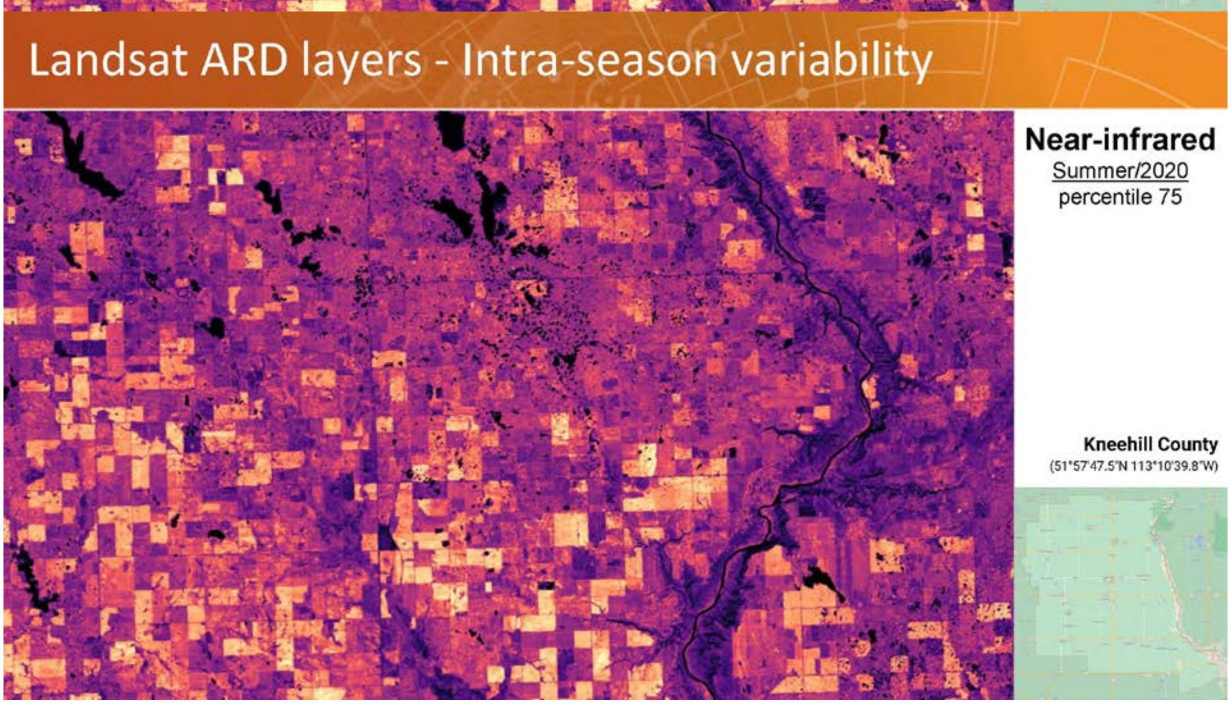
Landsat ARD layers - Seasons




Landsat ARD layers - Seasons
Spring/2019
Spring/2020


Landsat ARD layers - Intra-season variability
Near-infrared
Summer/2020
percentile 75
Kneehill County
(51°57'47.5"N 113°10'39.8"W)

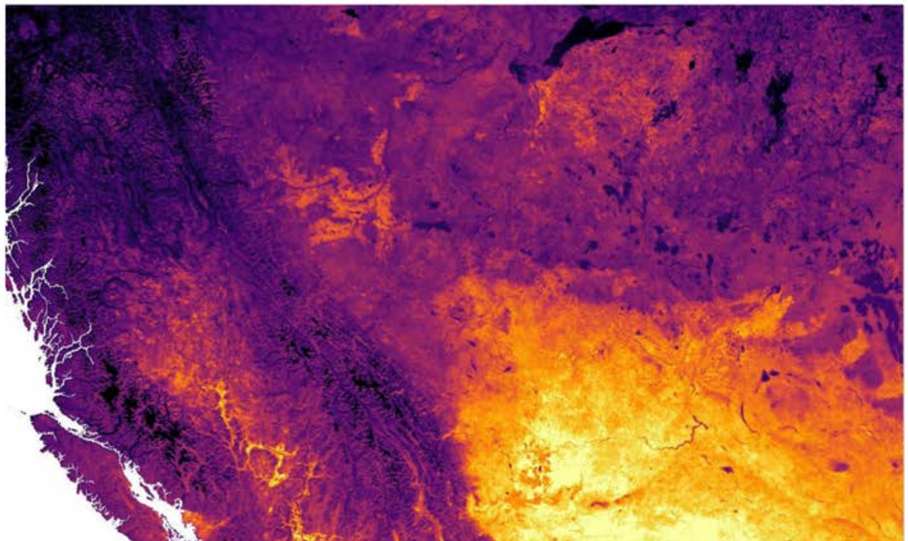

Landsat ARD layers - Intra-season variability
Near-infrared
Summer/2020
percentile 50
Kneehill County
(51°57'47.5"N 113°10'39.8"W)


Landsat ARD layers - Intra-season variability
Near-infrared
Summer/2020
percentile 75
Kneehill County
(51°57'47.5"N 113°10'39.8"W)

Preparation of Climate data for Alberta

Interpolated climatic layers and MODIS Land Surface Temperature

- Day and night time
- Monthly
- 20 years (2000-2020)



Long-term day time temperature September (2000-2020)

68

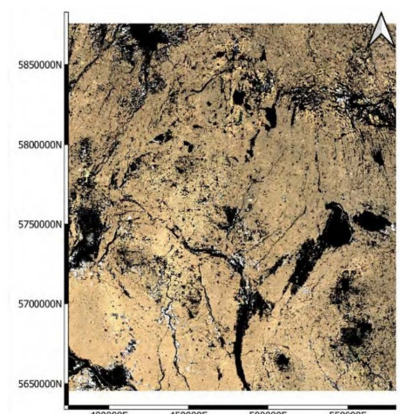
Bare Soil Composite imagery for Alberta

Bare Soil Composite Imagery - Landsat 5 for Pre-conservation Tillage Era

Example of area east of Saskatoon

- Visible light bands (B1, B2, B3) and NIR Band (B4) good for predicting SOC
- SWIR Bands (B5 and B7) are good for predicting clay content

Sorenson, P.T., Shirtliffe, S.J., Bedard-Haughn, A., 2021. Predictive Soil Mapping Using Historic Bare Soil Composite Imagery and Legacy Soil Survey Data. Geoderma 401, 115316. https://doi.org/10.1016/j.geoderma.2021.115316.



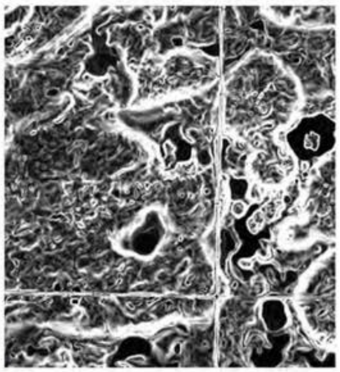
69

EnvirometriX statvis Trust Through Data InnoTech ALBERTA A SUBSIDIARY OF ALBERTA INNOVATES

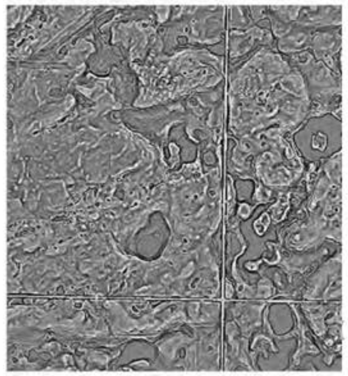
Terrain Variables

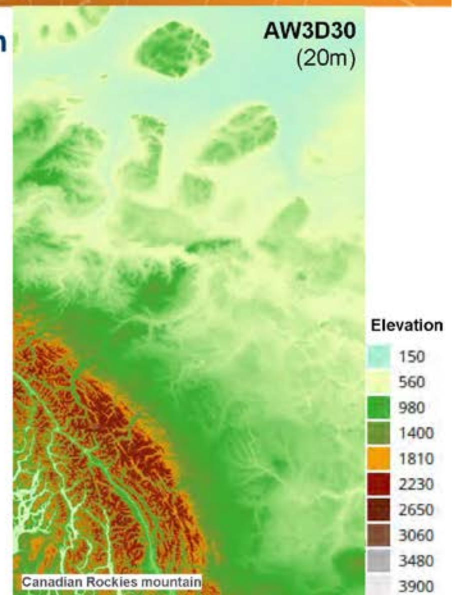
Terrain variables focused on hydrological / deposition and accumulation processes, geomorphology, morphometric features

Slope



Profile Curvature





70

Data delivery

Fully operational data cube / geospatial database

- All data analysis-ready perfectly stacked as COG via our S3 service;
- All metadata: SpatioTemporal Asset Catalog (STAC);
- Minimum and extended tutorial explaining how to access and use data in R, python, QGIS;

71

EnvirometriX statvis Trust Through Data InnoTech ALBERTA A SUBSIDIARY OF ALBERTA INNOVATES

EML problems still to solve

1. **Extrapolation problems** (quality of spatial sampling)
2. Computation intensity **very high**
3. **Over-fitting problems** (spatial de-clustering)
4. How to generate prediction intervals / geostatistical simulations?

80

The most important things you should take away from this presentation

Closing remarks

Summary points

- Modern Predictive Soil Mapping is **a fusion of stats, machine learning, soil science, proximal and remote sensing and process based modeling**; it remains however driven by accuracy;
- Most of our modeling now focuses on building **2D+T, 3D+T (spatiotemporal) models using Ensemble Machine Learning**,
- Best results (lowest costs!) you get by doing well-planed, multi-stage PSM,
- For data distribution we recommend **Cloud-Optimized GeoTIFFs + STAC**,

73

73

Closing remarks

Summary points

- 3D+T (dynamic soil information) is the currently the preferred modeling framework. It has many advantages:
  - It can be used as a generic framework for PSM;
  - It helps increase accuracy and understand key drivers / key processes;
- And also some disadvantages:
  - It needs much more training data / it gets quite computational;
  - It suffers from extrapolation problems;

74

74

Look Ahead: System Feature and Future Expansions by Pau Fuellbrandt and Natalie Shelby-James

Look Ahead: System Features – Web Application

- 'Open' geodatabase of soil quality data (salinity and metals) with a predictive map (initially for a smaller region and then scaled up to the Province)
  - Polygons, with range and mean concentrations & confidence interval for each parameter
  - Users will not have access to underlying data
  - System is at minimum accepted and ideally endorsed by system end users and regulators

2

EnvirometriX statvis Trust Through Data InnoTech ALBERTA A SUBSIDIARY OF ALBERTA INNOVATES

Alberta Background Soil Quality System

Look Ahead: System Features – Future Expansions

- Large database of background for future research projects
- Can be expanded to other parameters (natural F3, toluene, PAHs)
- Can be expanded to other physical characteristics (hydraulic conductivity would be valuable)

3

EnvirometriX statvis Trust Through Data InnoTech ALBERTA A SUBSIDIARY OF ALBERTA INNOVATES

Alberta Background Soil Quality System

Collaboration Opportunities

- We're extremely open and encouraging additional collaboration to enable more sources and quality of point data
- If you'd like to learn more / discuss collaboration – email

Natalie.Shelby-James@innotechalberta.ca

4

EnvirometriX statvis Trust Through Data InnoTech ALBERTA A SUBSIDIARY OF ALBERTA INNOVATES