EnvirometriX

Phase 1: Collation and Harmonization of Existing Soil and Covariate Data

Alberta Background Soil Quality System

Project leader: Natalie Shelby-James (natalie.shelby-james@InnoTechAlberta.ca)
Project reviewer: Ron J. Thiessen (ron.thiessen@innotechalberta.ca)

Calgary Research Park
3608 – 33 Street NW, Calgary
Alberta, Canada T9C 1T4
InnoTechAlberta.ca

Prepared by: **Tom Hengl** (EnvirometriX) and **Leandro Parente** (EnvirometriX / OpenGeoHub)
Wageningen, the Netherlands,
tom.hengl@envirometrix.net

Data access: https://diskstation.opengeohub.org:5001/fsdownload/A3tBiqwh6/salinity
Total size (folder): 14GB

Definitions

- **Canopy height**: estimated height of the tops of trees at different tree density levels. The global canopy height of the world at 30-m resolution was produced by Potapov et al. (2021). The 2019 global forest canopy height map is available at https://glad.umd.edu/dataset/gedi/.
- **Covariate layer**: gridded dataset or raster image with values assigned to pixels almost always imported and used as GeoTIFFs at different spatial resolutions,
- **DSM**: Digital Surface Model (SRTM DEM, AW3D, GLO-30) i.e. surface model that reports the elevations of a notional surface that reflects the tops of the trees and/or other objects, where present, including man-made structures, such as highways, buildings and similar. AW3D30 DSM can be downloaded from: https://www.eorc.jaxa.jp/ALOS/en/aw3d30/. GLO-30 can be obtained from: https://spacedata.copernicus.eu/web/cscda/dataset-details?articleId=394198.
- **DTM**: Digital Terrain Model or Digital Land Surface Model showing heights of a continuous bare ground (or terrain) surface (Hengl and Reuter, 2008). Currently the only globally consistent DTM is the MERIT DEM at 100-m resolution (Yamazaki et al., 2019). MERIT DEM can be downloaded from: http://hydro.iis.u-tokyo.ac.jp/~yamadai/MERIT_DEM/
- **Ensemble Machine Learning**: Typically a method where a meta-learner is used to combine multiple base learners (learner stacking) using cross-validation. In this work we used the mlr framework (https://mlr.mlr-org.com/) for Ensemble Machine Learning, which implements the Super Learner algorithm (Rhys, 2020). Other Ensemble Machine Learning frameworks include mlr3 (https://mlr3book.mlr-org.com/), and h2o (http://www.h2o.ai/).
- **Lights at night images**: usually estimate of the active light radiation directly quantifying industrialization level. Currently the highest quality open lights and night data is based on the Defense Meteorological Satellite Program (DMSP) satellites, i.e. the Joint Polar-orbiting Satellite System (JPSS) and the Visible and Infrared Imaging Suite (VIIRS) Day Night Band (DNB) on board of JPSS satellites. For more information refer to: https://eogdata.mines.edu/products/vnl/.
- **MODIS EVI**: Enhanced vegetation index is one of the key vegetation indices derived using EO data to represent density of vegetation. It provides consistent spatial and temporal comparisons of vegetation canopy greenness, a composite property of leaf area, chlorophyll and canopy structure and it can vary from month to month depending on an area.
- **MODIS LST**: Land Surface Temperature is estimated directly from MODIS terra products and usually comes at high precision but can be of limited accuracy in places with a lot of clouds, snow and similar. MODIS LST is available for nighttime and daytime periods, making it especially interesting to quantify urbanization, water content in soil and similar. MODIS LST is only available at coarse resolutions of 1 km.
- **Principal Component Analysis**: PCA is the procedure of converting original multi-response variables to the same number of orthogonal components (lower end components carry most of the correlated signal, last components often carry only pure noise). It is a statistical technique for reducing the dimensionality / information overlap of a dataset. It is a linear and computationally inexpensive technique where values can be easily back-transformed without information loss.
- **Predictive soil mapping**: PSM is a process of producing maps of soil variables using ground observations and laboratory measurements. It includes applying statistical and/or machine learning techniques to fit models for the purpose of producing spatial and/or spatiotemporal predictions of soil variables, i.e. maps of soil properties and classes at different resolutions. It is a multidisciplinary field combining statistics, data science, soil science, physical geography, remote

sensing, geoinformation science and a number of other sciences For more info see also: https://soilmapper.org.
- **Sentinel 5P**: ESA's Copernicus Sentinel missions are among the most ambitious Earth Observation programmes with over 20 planned satellite missions (can be best compared to NASA's Landsat programme). The Sentinel 5P mission is the first Copernicus mission dedicated to monitoring our atmosphere, especially to track CO2 and NOx emissions at high spatial resolution.

Introduction

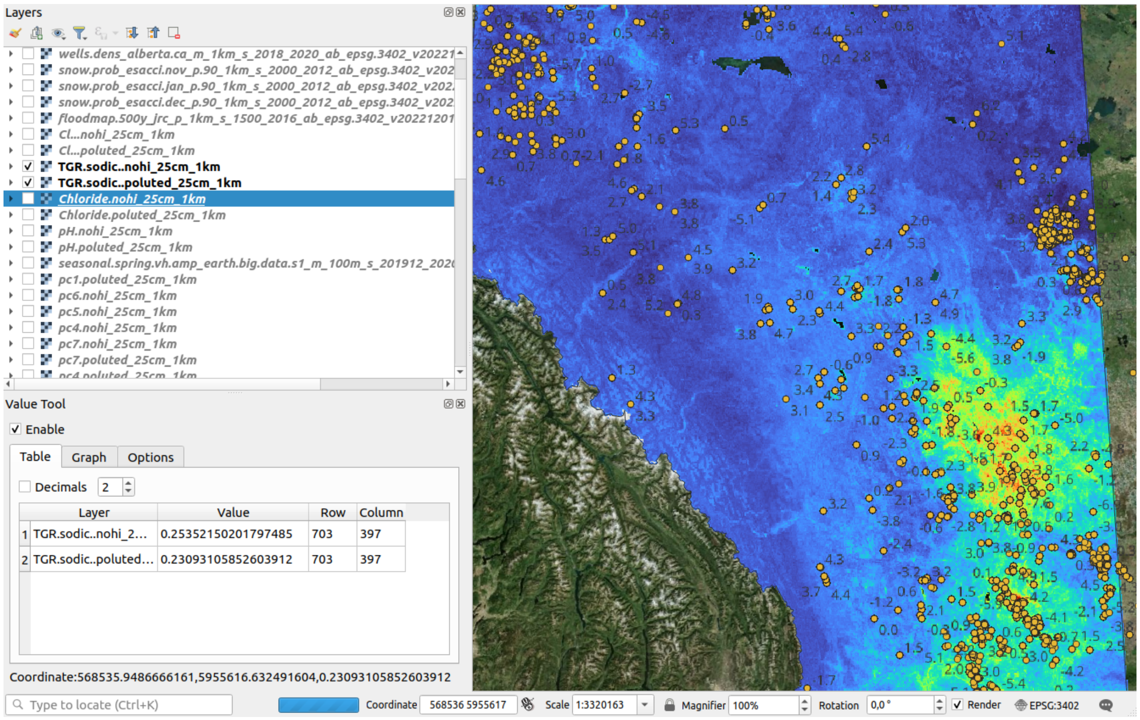
Phase 1 of the Alberta Background Soil Quality System Project (ABSQS) is funded by InnoTech Alberta (InnoTech), the Alberta Upstream Petroleum Research Fund (AUPRF) managed by the Petroleum Technology Alliance of Canada (PTAC) and the Clean Resources Innovation Network (CRIN). The objective of the ABSQS is to develop a database of background metals and salinity parameters in the Province of Alberta for the purpose of decreasing the cost and time required to identify and remediate valid contaminants of potential concern (COPCs) on contaminated sites. The ABSQS is currently in Phase 1 which involves choosing the model most suited to creating a replicable and defensible predictive map of salinity and metals parameters for a pilot area. This document describes production steps used to prepare covariate layers for the purpose of spatial modeling and predictive mapping of soil salinity and geochemicals (samples analyzed for macro-elements, soil pH, salinity, electrical conductivity and similar).

Background and Objectives

Background

The experimental methodology used to generate predictions is explained in the ensemble machine learning (EML) tutorial[1]. It implies modeling geochemical concentrations as a sum of components: industrial pollution + background (natural) concentrations. The goal of the predictive soil map (PSM) is to estimate background i.e. natural concentrations that can then be used as baselines for soil quality studies across the province. Covariates are based on global, regional and national datasets that have been downloaded, imported and stacked to perfectly match the bounding box of interest: Alberta province. The working resolution of covariates span from 1 km (climatic layers, density metrics) to 500m (lights at night), 250m (MODIS land products EVI etc, lithological and soil-type units), 100m (Sentinel-1 composites, MERIT DEM products) to 30m (Landsat images / composites; DTM derivatives) and beyond. Results of the initial model-fitting testing show that the most important covariates for mapping geochemicals (PC1, PC2) seem to be NO2 emission images, climatic variables including MODIS LST. Human impact covariates such as density of wells, nights at light and cropland percentage also come relatively high on the list of covariates, although are in general much less significant. De-coupling of the human-induced pollution (as explained in the EML tutorial) is possible but depends on how well are pollution processes represented. In summary, the first results indicate that extra effort needs to be put to represent sources of pollution with possibly buffer distances and classification of industry.

[1] https://opengeohub.github.io/spatial-prediction-eml/spatial-interpolation-in-3d-using-ensemble-ml.html



Layers

- □ wells.dens_alberta.ca_m_1km_s_2018_2020_ab_epsg.3402_v20221
- □ snow.prob_esacci.nov_p.90_1km_s_2000_2012_ab_epsg.3402_v202
- □ snow.prob_esacci.jan_p.90_1km_s_2000_2012_ab_epsg.3402_v202
- □ snow.prob_esacci.dec_p.90_1km_s_2000_2012_ab_epsg.3402_v202
- □ floodmap.500y_jrc_p_1km_s_1500_2016_ab_epsg.3402_v20221201
- □ Cl...nohi_25cm_1km
- □ Cl...poluted_25cm_1km
- ☑ TGR.sodic..nohi_25cm_1km
- ☑ TGR.sodic..poluted_25cm_1km
- □ Chloride.nohi_25cm_1km
- □ Chloride.poluted_25cm_1km
- □ pH.nohi_25cm_1km
- □ pH.poluted_25cm_1km
- □ seasonal.spring.vh.amp_earth.big.data.s1_m_100m_s_201912_202
- □ pc1.poluted_25cm_1km
- □ pc6.nohi_25cm_1km
- □ pc5.nohi_25cm_1km
- □ pc4.nohi_25cm_1km
- □ pc7.nohi_25cm_1km
- □ pc7.poluted_25cm_1km
- □ pc4.poluted_25cm_1km

Value Tool

☑ Enable

Table | Graph | Options

□ Decimals 2

	Layer	Value	Row	Column
1	TGR.sodic..nohi_2...	0.25352150201797485	703	397
2	TGR.sodic..poluted...	0.23093105852603912	703	397

Coordinate:568535.9486666161,5955616.632491604,0.23093105852603912

Type to locate (Ctrl+K) | Coordinate 568536 5955617 | Scale 1:3320163 | Magnifier 100% | Rotation 0,0 ° | ☑ Render | EPSG:3402

*Image: Example of predictions of soil geochemicals at coarse resolution (1km) based on the ca 65,000 training points / soil samples. Temporary access to this data is available **here**.*

Covariate layers are key to producing the most accurate possible maps of soil variables. In PSM many sources of covariates are used to increase accuracy of maps, most commonly these belong to groups such as (a) climate, (b) relief, (c) vegetation / land use, (d) geography, (e) parent material, (f) human influence. Processing steps used to prepare soil covariate layers are described in detail in: https://soilmapper.org/soil-covs-chapter.html.

Objective

The objective of Phase 1 of the ABSQS is to choose the covariate datasets and machine learning model that provide the most replicable and reliable predictions of the chosen salinity and metals parameters chosen for inclusion in the ABSQS.

Methodology

For the purpose of the Alberta Background Soil Quality System we specifically focus on representing two main groups of covariates: (1) natural soil forming factors and processes (parent material / lithology, climate, relief, vegetation etc), and (2) covariates quantifying human-based pollution sources e.g. density and intensity of urban, industrial and agricultural sources of geochemicals in the area. For the first group of covariates (natural soil forming factors) we use the following specific layers:

1. Parent material map / lithology map of **Alberta i.e. Surficial Geology of Alberta**, Generalized Digital Mosaic (DIG 2013-0002);

2. Soil type indicator maps based on **Agricultural Regions of Alberta Soil Information Database** Version 4.1;
3. Digital Terrain Model of Alberta based on GLO30 and AW3D30 global products;

For the second group we use the following four key covariate layers:

1. **Lights at night images** based on VIIRS Nighttime Light (global 500m product) with minimum, average and maximum values of radiation;
2. **Density of wells** and **facilities**, derived using the kernel density filter at 1km;
3. **Cropland percent map** (Global cropland expansion in the 21st century | GLAD) quantifying intensity of agricultural production and hence also fertilization and similar.
4. **Distance to cities / accessibility** (Nelson et al., 2019);
5. **Sentinel 5P L3 NO2** emission data (Copernicus Sentinel-5P Mapping Portal);
6. **Buffer distances** to key sources of emission.

To produce predictions of geochemicals assuming no human impact, one can simply fit models using all data, then during prediction synthetically set values of lights at nights, density of industrial objects and similar to 0. Likewise, setting buffer distance in prediction space to some high number e.g. maximum distance should result in the same effect of producing predictor space equivalent to natural vegetation i.e. no human influence.

MODIS LST nighttime images could also be considered to quantify the human-induced pollution sources as LST of urban areas and industry is often few degrees higher than under natural vegetation. This type of human impact is more difficult to de-couple as MODIS LST would need to be reprocessed using complex rules i.e. pixel per pixel.

All layers described here are prepared using spatiotemporal referencing included in the filename and referent year also included in the directly. Based on the distribution of soil samples, we prepared all covariate layers following the span of seven (7) years i.e. 2015–2021. The folder “static” contains covariate layers that represent a longer span of years e.g. long-term climatic variables (1980–2020), DTM variables, parent material and similar. These are assumed to be constant or already aggregated over a span of years, hence they do not vary through time.

For all layers we consistently use the bounding box of Alberta and the EPSG::

```
tmerc.prj = "+proj=tmerc +lat_0=0 +lon_0=-115 +k=0.9992 +x_0=500000 +y_0=0 +ellps=GRS80 +towgs84=0,0,0,0,0,0,0 +units=m +no_defs"
te = raster::extent(raster::raster("./grids1km/static/elev_aw3d30_m_1km_s_2018_2020_ab_epsg.3402_v20221130.tif"))
te
xmin = 170000
ymin= 5426000
xmax = 866000
ymax = 6660000
```

In this initial phase we are only testing modeling distribution of geochemicals using the initial set of covariates. This has three practical purposes:

1. We can highlight and subset covariate layers that are most important i.e. that come highest at the variable importance analysis.

2. We can develop a tailored-based modeling framework and test accuracy of the predictions i.e. establish a baseline accuracy assuming using standard modeling frameworks.
3. We can potentially detect points / samples that have a high impact on modeling performance but are potentially of questionable quality / outliers or blunders. Removing even 0.5% of the training points that are potentially in error (blunders, typos or similar) can significantly benefit map accuracy, especially where ML methods are used.

During this project we will be testing two frameworks for mapping concentrations of geochemicals / salinity:

(a) **Composition-decomposition method** based on the Random Forest with a multi-response system.
(b) **Individual models (per variable)** based on point data separated as background / polluted values.

The composition-decomposition method consist of four main steps:

1. Normalize, scale (0 mean, 1 stdev) and convert the target variables to principal component matrix (PC1, PC2, … PCp);
2. Overlay points and covariates, then fit a (1) multiresponse Random Forest model[2] (PC1, PC2, PCp ~ X1, X2, Xk).
3. Predict values of PC1, PC2, … PCp at all prediction locations.
4. Back-transform the values from PCs to original variables, then also re-scale to original scale.

The advantages of the composition-decomposition method are:

- It fairly compact as one needs to fit a single model only to map large number of target variables; this model than deals efficiently with multicollinearity of the target variables;
- Since only portion of the Principal Components need to be mapped (e.g. PCs leading to 99% of compression of signal), it is also computationally very efficient as it reduces prediction process;
- It is relatively generic process and can be automated + it is very efficient for gap-filling missing values for target variables (assuming that all target variables follow close-to-normal distribution);
- Although the interpretation of components is complex, if the target variables are correlated and especially if groupings exist, then these can be mapped and interpreted by soil experts;
- During the decomposition, one can easily set all human-impact measures to none (e.g. 0) so that a consistent estimate of background concentrations can be produced;

If the number of target variables is large (e.g. >>10), if the target variables can be easily normalized with e.g. log-transformation or similar, and if the target variables are significantly cross-correlated (hence Principal Component transformation efficiently reduces the space), and if there are not too many missing values (e.g. <30%), then the composition-decomposition method is probably the most logical choice to use to produce predictions of geochemicals.

Some disadvantages of the composition-decomposition method are:

- In the case correlation between target variables is weak or nonlinear, use of PCs does not significantly helps reduce variable space anyway, so there is less incentive to use it at the first place;

[2] https://www.randomforestsrc.org/articles/mvsplit.html

- Some target variables can not be easily transformed to normal distribution hence gap-filling of missing values in target variables could potentially lead to bias;
- Derivation of prediction errors is somewhat more complex as one needs to derive a compositional error by summing up errors of all components (e.g. prediction error for PC1 + prediction error for PC2 etc);
- The method can not be extended to Ensemble ML systems i.e. it can currently only be used in the randomForestSRC package;

In practice, the composition-decomposition method could lead to lower accuracy than if individual models are used, hence testing and comparing the two is recommended. We have now only tested using composition-decomposition method and only predicting at coarser spatial resolutions (1 km) hence these preliminary results should be considered as a test-run only.

Results

As expected, the target geochemicals and salinity measures seem to be highly correlated with PC1 and PC2 explaining already 45% + 18% of variation in data. This justifies use of PCA analysis, especially to reduce the number here.of variables for spatial modeling from 32 to 14 or less. Plot below shows the PC1-PC2 biplot with general grouping of variables. Fitted models and predictions of geochemicals are available ***here***.

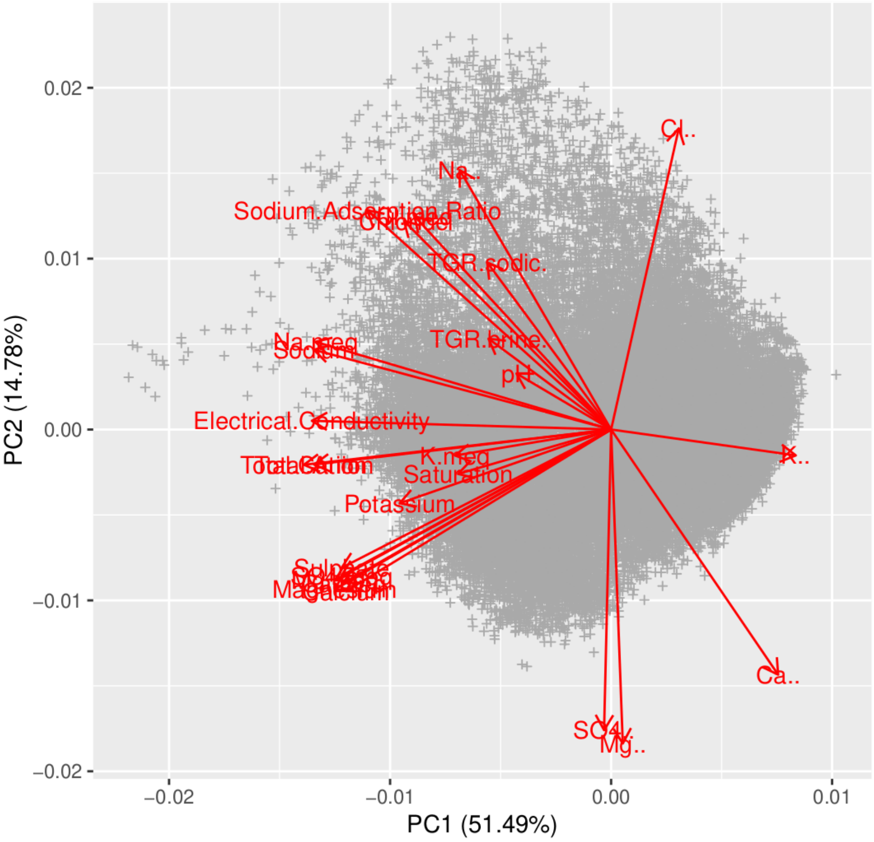
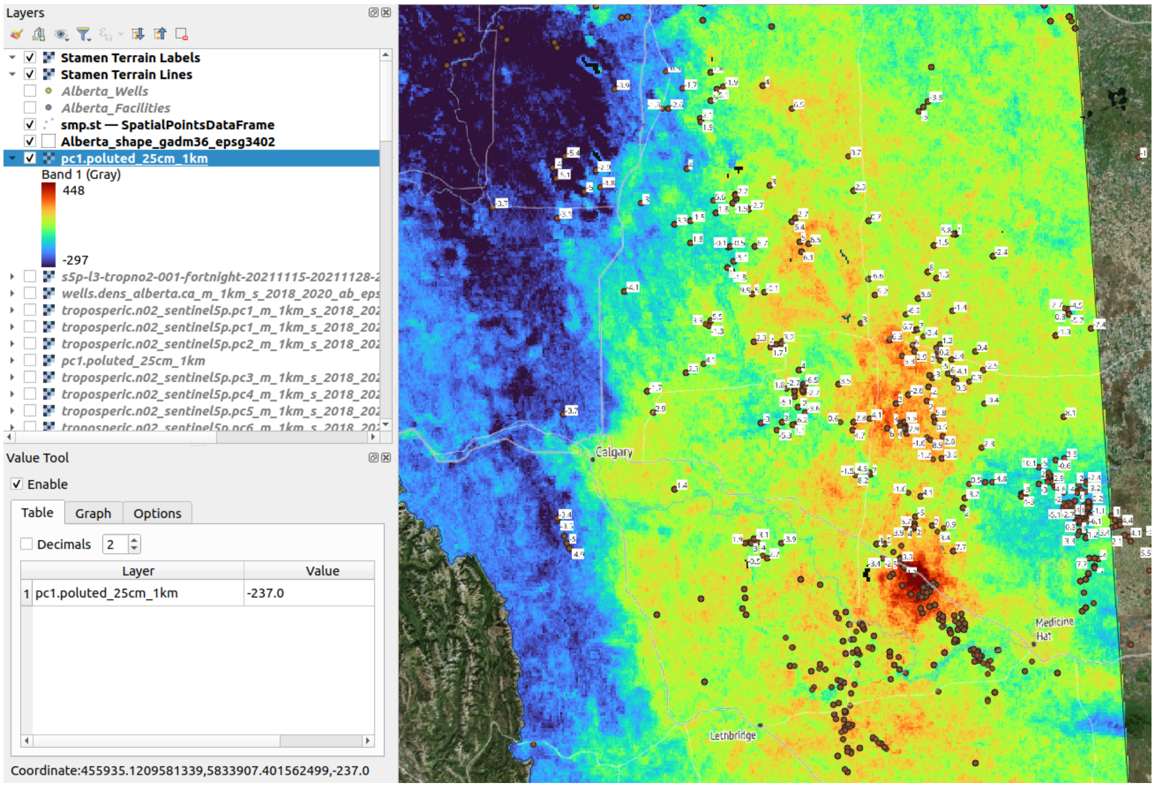


Image: Results of PCA using ALL geochemicals showing some clear groupings in data e.g. K, Mg and Saturation.

Results of Random Forest modeling show that significant models can be fitted to explain distribution of the target variable. The results of model fitting using strict cross-validation (spatial blocking using 250m blocks to remove overlapping points) indicates:

```
> print(m.s1, outcome.target = "PC1")
                         Sample size: 38070
                     Number of trees: 85
           Forest terminal node size: 2
       Average no. of terminal nodes: 917.2941
No. of variables tried at each split: 165
              Total no. of variables: 169
              Total no. of responses: 15
         User has requested response: PC1
       Resampling used to grow trees: by.user
    Resample size used to grow trees: 16829
                            Analysis: mRF-R
                              Family: regr+
                      Splitting rule: mv.mse *random*
       Number of random split points: 10
                     (OOB) R squared: 0.42251309
   (OOB) Requested performance error: 10.29135303
```

The R-square drops significantly for PC2+, which can be expected as also the variance in the signal drops significantly towards higher level PC. Note however that the out-of-bag prediction errors are still smaller for higher order PCs, so we can conclude that, on average, the accuracy of predictions does not drop significantly.



☑ Stamen Terrain Labels
☑ Stamen Terrain Lines
☐ Alberta_Wells
☐ Alberta_Facilities
☑ smp.st — SpatialPointsDataFrame
☑ Alberta_shape_gadm36_epsg3402
☑ pc1.poluted_25cm_1km
☐ s5p-l3-tropno2-001-fortnight-20211115-20211128-2
☐ wells.dens_alberta.ca_m_1km_s_2018_2020_ab_eps
☐ troposperic.n02_sentinel5p.pc1_m_1km_s_2018_202
☐ troposperic.n02_sentinel5p.pc1_m_1km_s_2018_202
☐ troposperic.n02_sentinel5p.pc2_m_1km_s_2018_202
☐ pc1.poluted_25cm_1km
☐ troposperic.n02_sentinel5p.pc3_m_1km_s_2018_202
☐ troposperic.n02_sentinel5p.pc4_m_1km_s_2018_202
☐ troposperic.n02_sentinel5p.pc5_m_1km_s_2018_202
☐ troposperic.n02_sentinel5p.pc6_m_1km_s_2018_202

☑ Enable

☐ Decimals 2

	Layer	Value
1	pc1.poluted_25cm_1km	-237.0

Coordinate:455935.1209581339,5833907.401562499,-237.0

Image: Predicted patterns of PC1 for top-soil (25 cm depth).

Variable importance analysis for PC1 and PC2 show that, in principle, Sentinel-5P NO2 emission images and climatic variables seem to be the main drivers of the distribution of geochemicals. Surprisingly parent material maps, DTM derivatives, night light images do not come up high in variable importance. This could be due to the following potential reasons:

- Quality of the parent material map is limited; more detailed actual parent material map needs to be prepared that reflects better soil mineralogy, not only geomorphology;
- Urban areas (lights at night) are systematically underrepresented with almost no points inside big urban areas and hence modeling not catching the relationship correctly;
- Sources of pollution might not be correctly presented with the kernel density maps of wells and facilities for Alberta;

The results of testing also further show that back-transformation of values to the original scale is computationally efficient and does not seem to introduce any bias (systematic over- or under-estimation or over-smoothing). The values of the observed and predicted values seem to be on a 1:1 line indicating unbiased estimation.

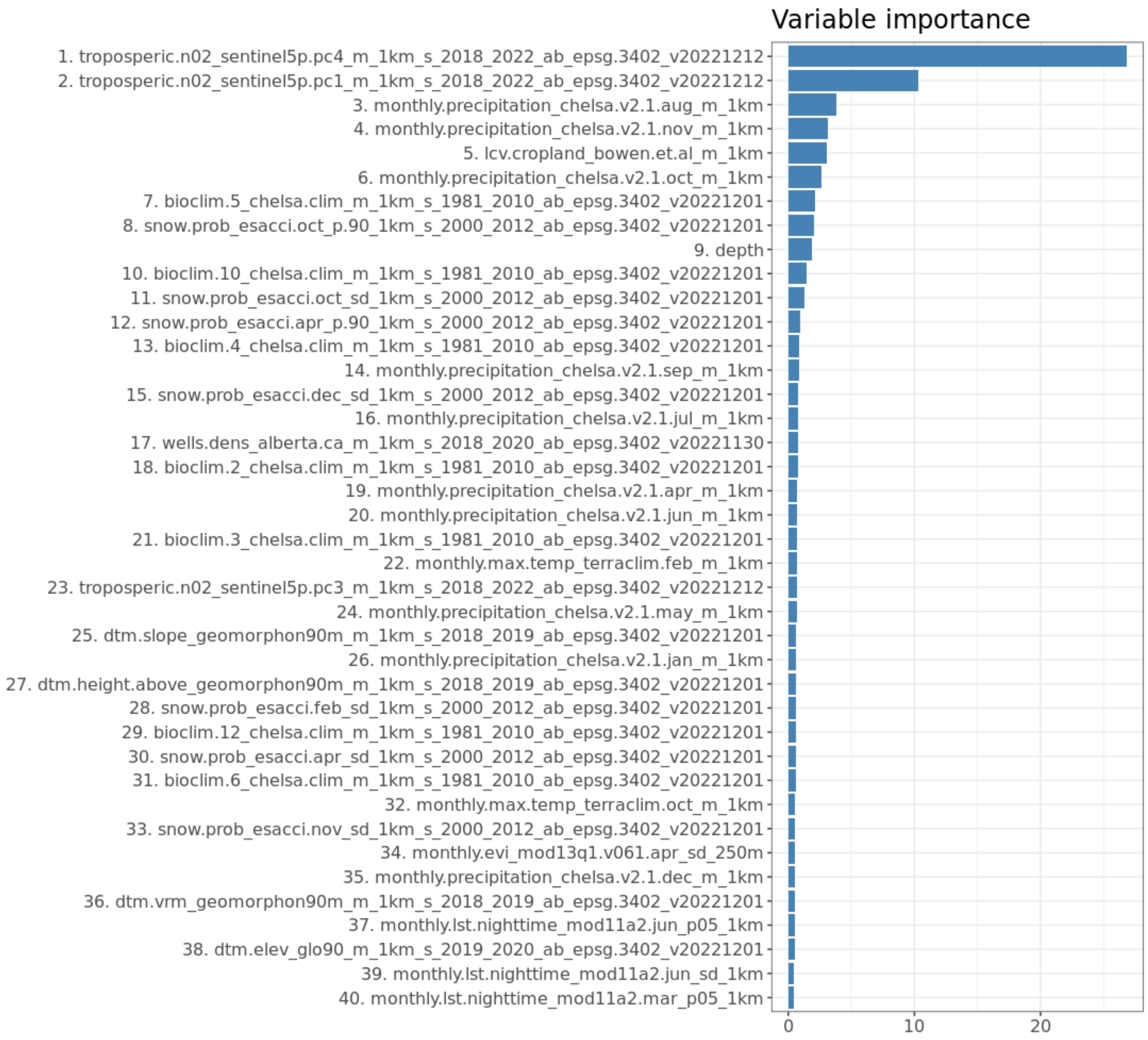
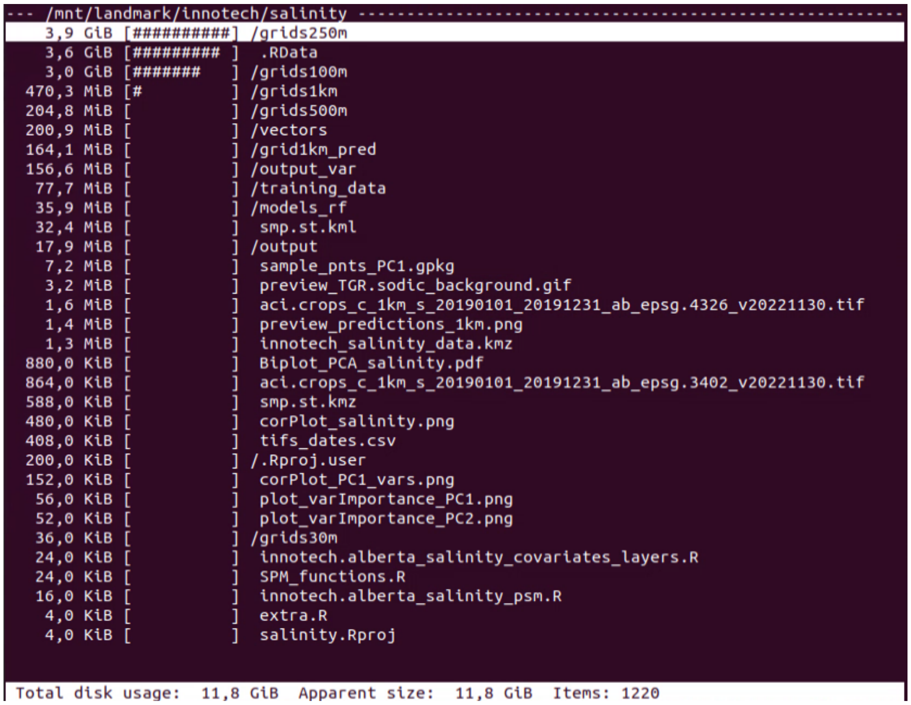


Image: Variable importance for modeling PC1.

Data Access

All layers used in this report can be accessed via the OpenGeoHub S3 service or downloaded individually from **here**. The available layers are distributed in multiple directories based on the native spatial resolution of the covariates used. The current structure of the folder with all inputs and outputs is shown below:



```
--- /mnt/landmark/innotech/salinity -----------------------------------------------------
    3,9 GiB [##########] /grids250m
    3,6 GiB [######### ]  .RData
    3,0 GiB [#######   ] /grids100m
  470,3 MiB [#         ] /grids1km
  204,8 MiB [          ] /grids500m
  200,9 MiB [          ] /vectors
  164,1 MiB [          ] /grid1km_pred
  156,6 MiB [          ] /output_var
   77,7 MiB [          ] /training_data
   35,9 MiB [          ] /models_rf
   32,4 MiB [          ]  smp.st.kml
   17,9 MiB [          ] /output
    7,2 MiB [          ]  sample_pnts_PC1.gpkg
    3,2 MiB [          ]  preview_TGR.sodic_background.gif
    1,6 MiB [          ]  aci.crops_c_1km_s_20190101_20191231_ab_epsg.4326_v20221130.tif
    1,4 MiB [          ]  preview_predictions_1km.png
    1,3 MiB [          ]  innotech_salinity_data.kmz
  880,0 KiB [          ]  Biplot_PCA_salinity.pdf
  864,0 KiB [          ]  aci.crops_c_1km_s_20190101_20191231_ab_epsg.3402_v20221130.tif
  588,0 KiB [          ]  smp.st.kmz
  480,0 KiB [          ]  corPlot_salinity.png
  408,0 KiB [          ]  tifs_dates.csv
  200,0 KiB [          ] /.Rproj.user
  152,0 KiB [          ]  corPlot_PC1_vars.png
   56,0 KiB [          ]  plot_varImportance_PC1.png
   52,0 KiB [          ]  plot_varImportance_PC2.png
   36,0 KiB [          ] /grids30m
   24,0 KiB [          ]  innotech.alberta_salinity_covariates_layers.R
   24,0 KiB [          ]  SPM_functions.R
   16,0 KiB [          ]  innotech.alberta_salinity_psm.R
    4,0 KiB [          ]  extra.R
    4,0 KiB [          ]  salinity.Rproj


 Total disk usage:  11,8 GiB  Apparent size:  11,8 GiB  Items: 1220
```

Image: Preview of the folder content.

To open the data in QGIS or similar simply follow an import procedure as illustrated below. For example, the filename to use for parent material is:

```
> in.tif =
"/vsicurl/https://s3.us-west-1.wasabisys.com/canada/ab/grids250m/static/surf.lithology_alberta
.ca_c_250m_s_2000_2018_ab_epsg.3402_v20221130.tif"
> library(terra)
terra 1.6.17
> rast(in.tif)
class       : SpatRaster
dimensions  : 4936, 2784, 1  (nrow, ncol, nlyr)
resolution  : 250, 250  (x, y)
extent      : 170000, 866000, 5426000, 6660000  (xmin, xmax, ymin, ymax)
coord. ref. : NAD83(CSRS) / Alberta 10-TM (Forest) (EPSG:3402)
source      : surf.lithology_alberta.ca_c_250m_s_2000_2018_ab_epsg.3402_v20221130.tif
name        : surf.lithology_alberta.ca_c_25~00_2018_ab_epsg.3402_v20221130
```

You can also access this data as a geospatial DB i.e. by using the terra package in R and/or GDAL supported packages. For example to overlay points using Cloud-Optimized GeoTIFFs best use the terra::extract function with max 10 parallel threads e.g.:

```
ov = parallel::mclapply(in.tif.lst, function(i){terra::extract(rast(i), xy.lst)}, mc.cores =
10)
```

For more info about how to use COG files see also this video: https://av.tib.eu/media/55228

Conclusions

The first results of modeling spatial distribution of geochemicals / salinity across the province of Alberta are promising. The composition-decomposition method can potentially be used to separate industrial pollution from natural background concentrations, however this would need to be tested in the Phase 2. Variable importance analysis for PC1 and PC2 show that, in principle, Sentinel-5P NO2 emission images and climatic variables seem to be the main drivers of the distribution of geochemicals. Surprisingly parent material maps, DTM derivatives, night light images do not come up high in variable importance. Prepared layers are available via a web-folder and can also be accessed as Cloud-Optimized GeoTIFFs (see instructions above). These will be gradually extended in phase 2, depending on the results of modeling / feature selection results.

Recommendation

Although the first results of testing PSM show that the geochemicals can be mapped using the initial point data, predictions will likely be limited and hence extrapolation problems need to be carefully considered. The following training point problems need to be especially tackled:

1. Points over-represent industrial areas; urban areas, especially with >50,000 citizens are almost non-represented and hence it is difficult to detect the relationship between lights and night and geochemical concentrations.

2. Many points (>30%) have relatively poor location accuracy and are incomplete, hence using finest resolution covariates (e.g. <250m spatial resolution) could become cumbersome as it is impossible to match the values of covariates with values of target variables. One solution could be to use a weighted Machine Learning where quality of soil samples is used as weight into training (high quality data gets exponentially higher weight).

3. Parent-material maps have not shown to correlate significantly to training points, however the soil polygon maps e.g. salinity fraction are likely to correlate much better as they reflect better parent material. Additional effort is needed to prepare this data.

Several new covariate layers can be added to help improve mapping geochemicals. Most importantly we need to put effort into preparing soil salinity (based on soil type) fraction maps based on Agricultural Regions of Alberta Soil Information Database Version 4.1. These seem to be available but need to be gap-filled and prepared for analysis.

Another important source of information will be Landsat 8/9 and Sentinel 2 MS products, which we have also prepared but have not used at this phase for modeling as these also need to be converted to bare-earth spectra (Demattê et al., 2020). Adding finer resolution, tailor-based covariates that help explain soil forming processes and migration of chemicals, and adding additional point samples in areas of extrapolation are probably the best strategy to help increase accuracy of these maps.

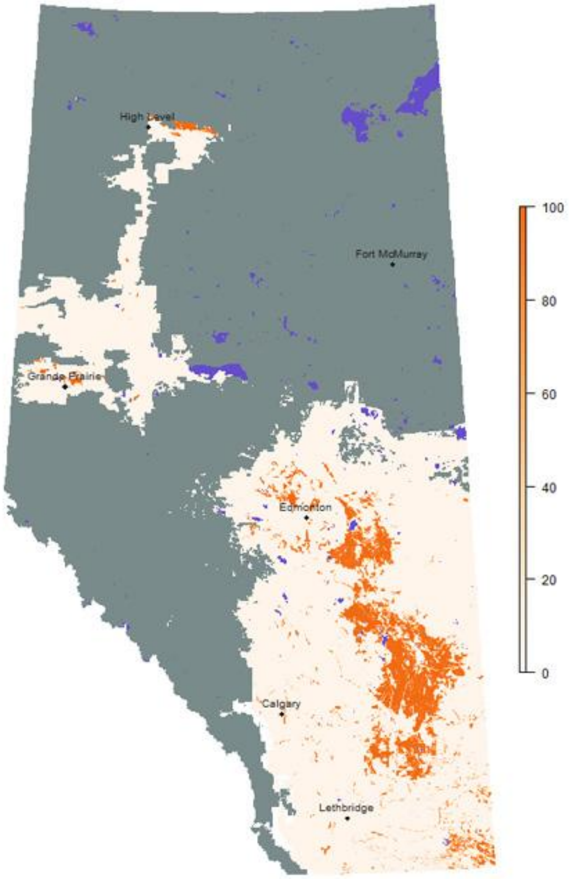


Image: Example of saline soil fraction based on the [ABMI Soil Layers](). These maps are currently not used in modeling because only PNG is available publicly and also the map is incomplete (gap-filling is certainly possible but needs to be implemented carefully).

Important references:

1. Brun, P., Zimmermann, N. E., Hari, C., Pellissier, L., & Karger, D. N. (2022). Global climate-related predictors at kilometre resolution for the past and future. Earth System Science Data Discussions, 1-44. https://doi.org/10.5194/essd-2022-212
2. Demattê, J. A., Safanelli, J. L., Poppiel, R. R., Rizzo, R., Silvero, N. E. Q., Mendes, W. D. S., ... & Lisboa, C. J. D. S. (2020). Bare earth's surface spectra as a proxy for soil resource monitoring. Scientific reports, 10(1), 1-11. https://doi.org/10.1038/s41598-020-61408-1
3. Elvidge, C. D., Baugh, K., Zhizhin, M., Hsu, F. C., & Ghosh, T. (2017). VIIRS night-time lights. International Journal of Remote Sensing, 38(21), 5860-5879. https://doi.org/10.1080/01431161.2017.1342050
4. Hengl, T., MacMillan, R.A., (2019). Predictive Soil Mapping with R. OpenGeoHub foundation, Wageningen, the Netherlands, 370 pages, www.soilmapper.org, ISBN: 978-0-359-30635-0.
5. Kellndorfer, J., Cartus, O., Lavalle, M., Magnard, C., Milillo, P., Oveisgharan, S., ... & Wegmüller, U. (2022). Global seasonal Sentinel-1 interferometric coherence and backscatter data set. Scientific Data, 9(1), 1-16. https://doi.org/10.1038/s41597-022-01189-6

6. Nelson, A., Weiss, D. J., Etten, J. van, Cattaneo, A., McMenomy, T. S., & Koo, J. (2019). A suite of global accessibility indicators. Scientific Data, 6(1), 1–9. doi:10.1038/s41597-019-0265-5
7. Potapov, P., Li, X., Hernandez-Serna, A., Tyukavina, A., Hansen, M. C., Kommareddy, A., ... & Hofton, M. (2021). Mapping global forest canopy height through integration of GEDI and Landsat data. Remote Sensing of Environment, 253, 112165. https://doi.org/10.1016/j.rse.2020.112165
8. Potapov, P., Turubanova, S., Hansen, M. C., Tyukavina, A., Zalles, V., Khan, A., ... & Cortez, J. (2022). Global maps of cropland extent and change show accelerated cropland expansion in the twenty-first century. Nature Food, 3(1), 19-28. https://doi.org/10.1038/s43016-021-00429-z
9. Rhys, H. I. (2020). Machine Learning with R, the tidyverse, and mlr. Manning Publications.
10. Song, X. P., Hansen, M. C., Stehman, S. V., Potapov, P. V., Tyukavina, A., Vermote, E. F., & Townshend, J. R. (2018). Global land change from 1982 to 2016. Nature, 560(7720), 639-643. https://doi.org/10.1038/s41586-018-0411-9
11. Yamazaki, D., Ikeshima, D., Sosa, J., Bates, P. D., Allen, G. H., & Pavelsky, T. M. (2019). MERIT Hydro: a high-resolution global hydrography map based on latest topography dataset. Water Resources Research, 55(6), 5053-5073. https://doi.org/10.1029/2019WR024873