Adding Value to Regional Geochemical Data through Exploratory Data Analysis

TREK Project Area, Central British Columbia (parts of NTS 093B, C, F, G)

Britt Bluemel MSc, REFLEX Geochemist

Geospatial intelligence the minute it comes out of the ground, anywhere in the world.

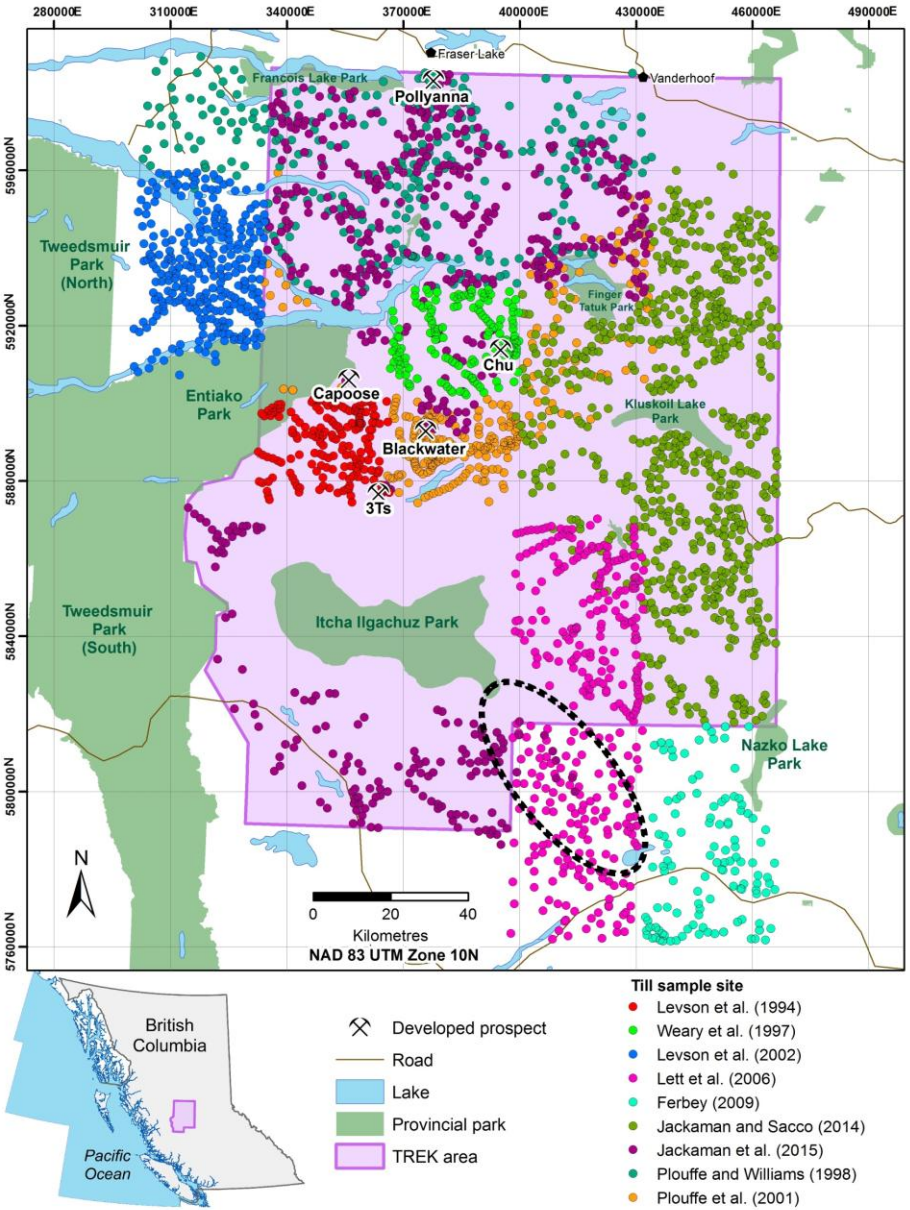
DATA ANALYSIS ACTION

Outline

- Introduction
- Project Challenges, Goals, and Objectives
- Work Completed
  - Preliminary Data Assessment
  - Exploratory Data Analysis
    1. Clustering
    2. Regression Analysis
    3. Multivariate Anomaly Assessment
- Targeting

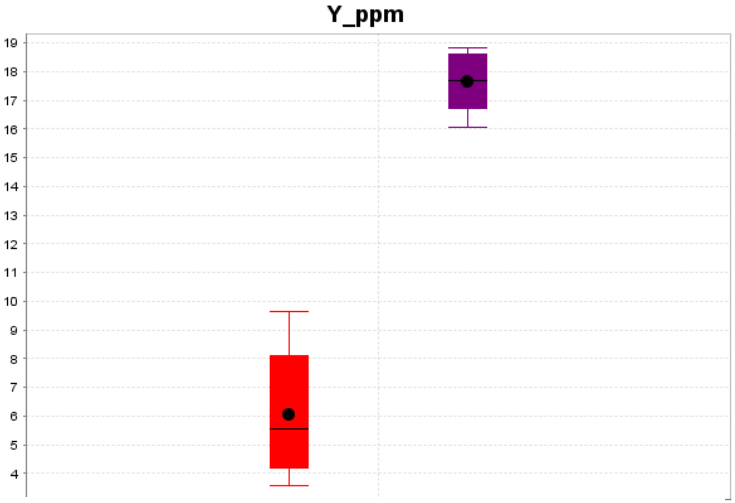
Introduction

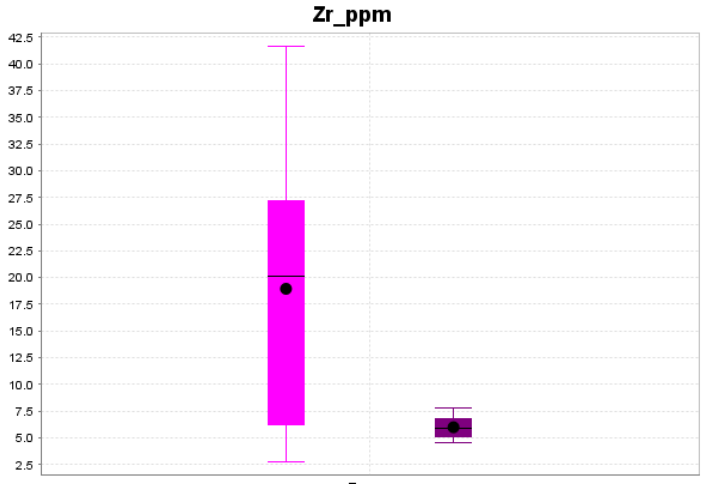
- Challenges
  - Exotic transported cover & multi-generational data
- Multiple protoliths contributing to tills... How to split?
  1. Spatial association with bedrock?
  2. Differentiate tills based on their geochemical signature?
  3. Assess mineral potential per group
- What was the point of this EDA exercise?
  - Create derivative geochemical products
  - Account for surficial processes
  - Delineate some new areas of mineralization potential

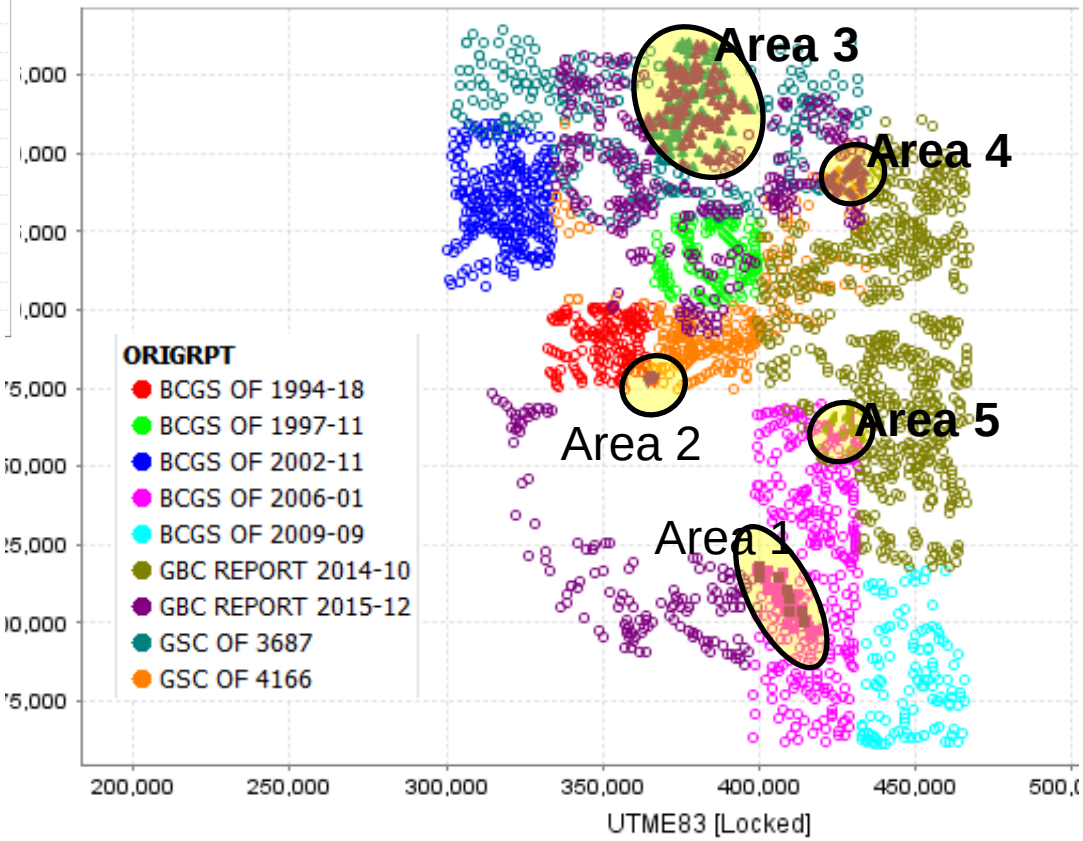


Step 1 – Preliminary Data Assessment

- Data cleaning (ie. Ignored 255 “no data” samples & a few outlier samples)
- Is data comparable between survey areas, without any levelling or prep?

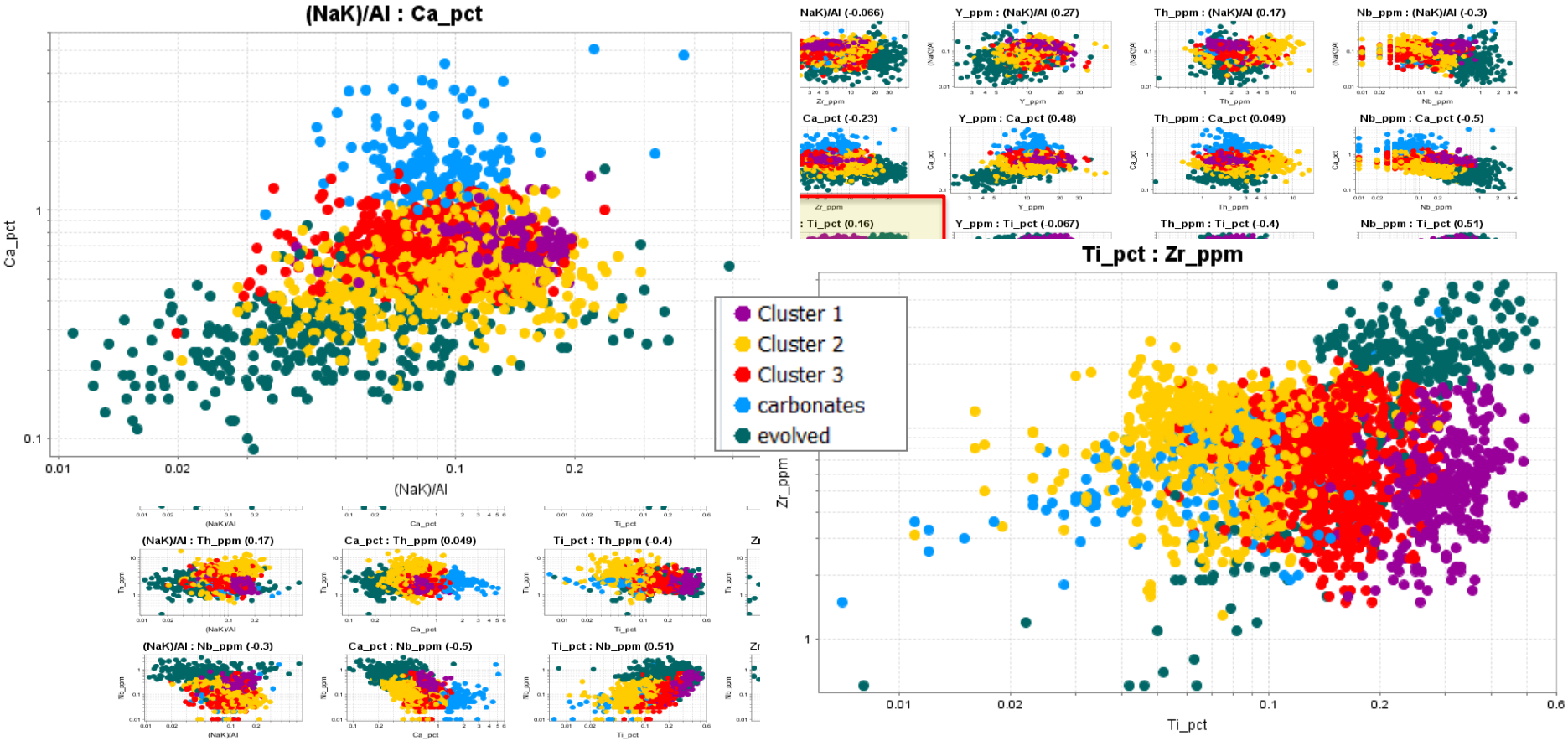






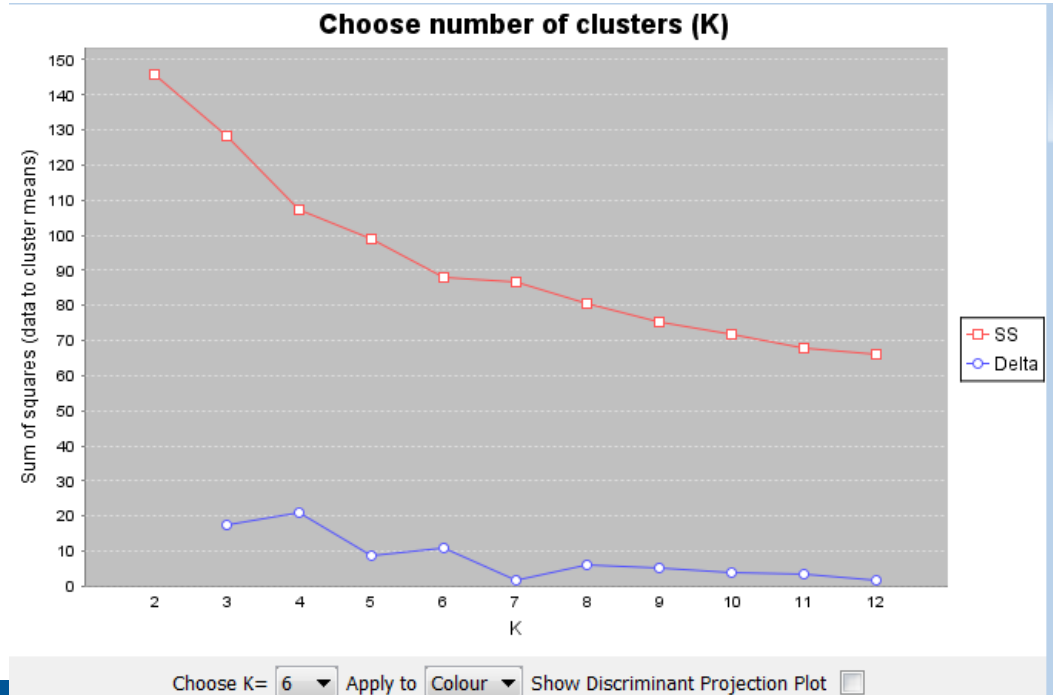
Step 2 – Exploratory Data Analysis

- EDA = detecting trends or structures within the data, which can “provide insight into the geochemical/geological processes”
- Clustering analysis... using immobiles (Ti, Zr, Y, Nb), major elements (Ca), useful trace elements (Th) and a “freshness” ratio (Na+K)/Al

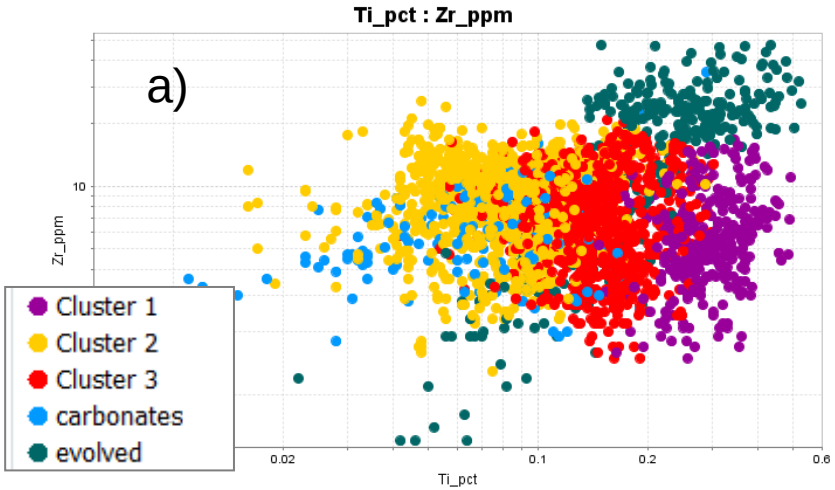


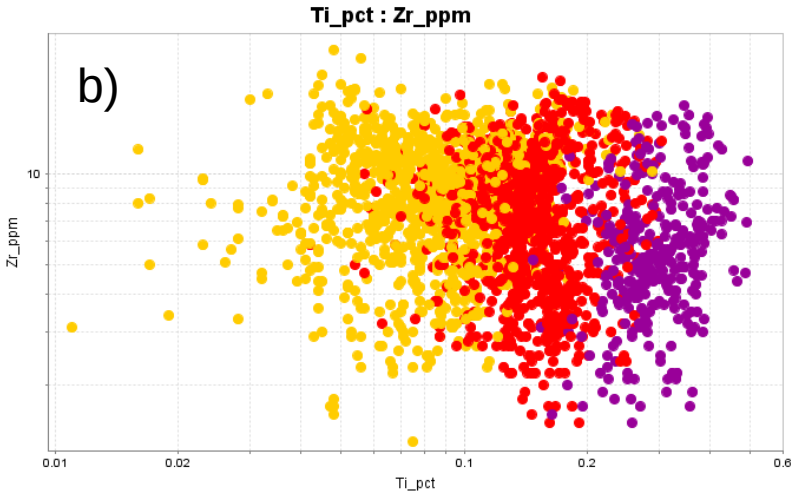
Part 1 - Clustering Analysis

- Cluster definition: a collection of data objects
  - Similar to one another within the same cluster.
  - Dissimilar to the objects in other clusters.
- Cluster analysis:
  - The main objective of clustering algorithms is to statistically represent distinct natural groupings of multidimensional data in as few groups as possible.
  - Based on a measure of distance between observations.
  - Used to discover structures in data without providing an explanation/ interpretation.
  - Useful for partitioning a very large population, *perhaps for data mining each sub-population separately.*
- Unsupervised!

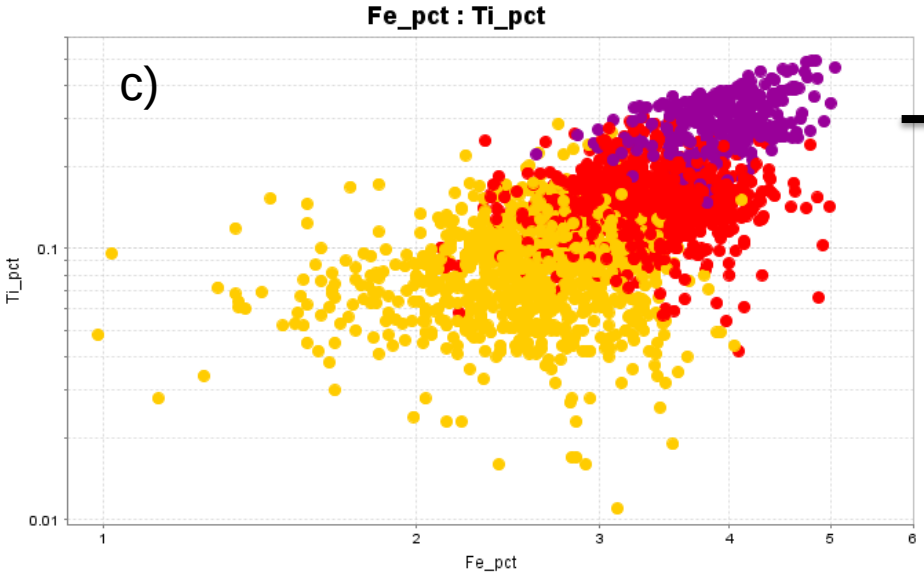


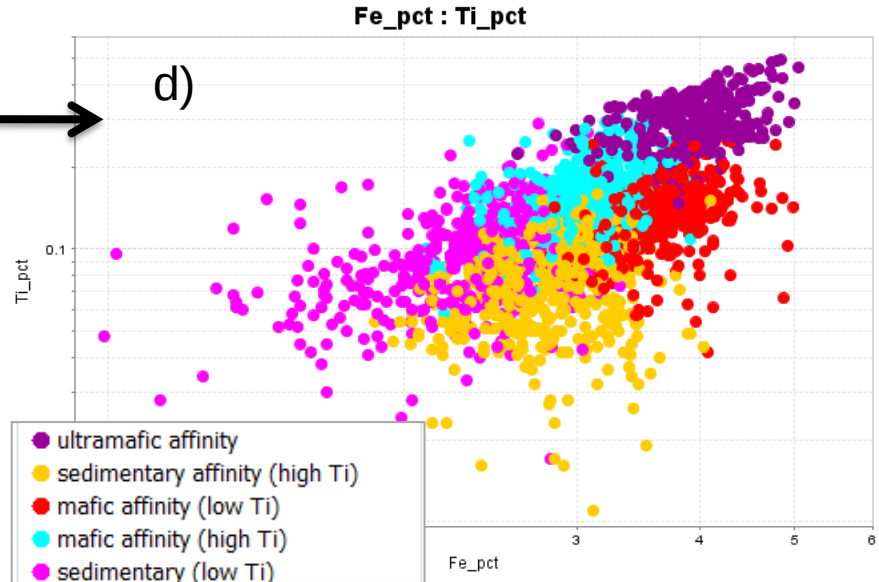
Refined clustering





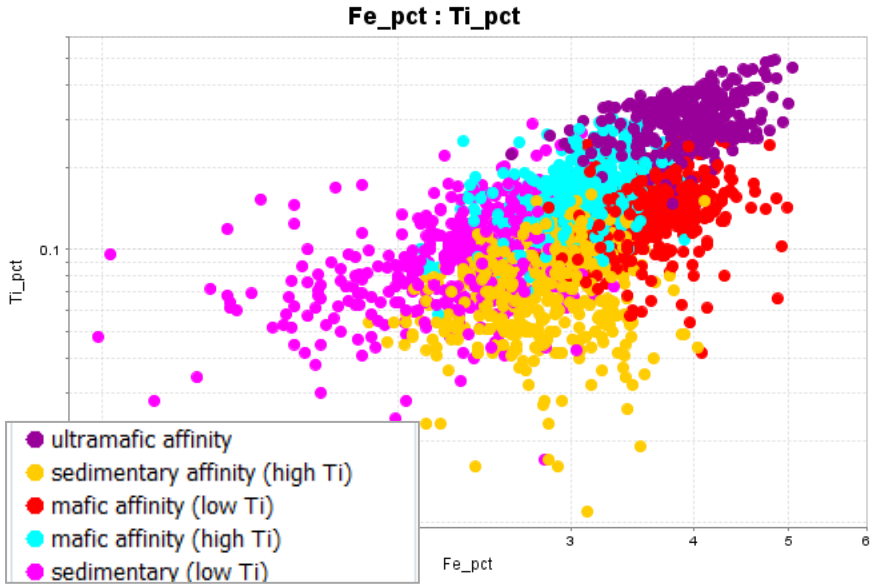
- Choose different elements that might help differentiate in the secondary environment (Ti, Al, Fe, Mg, Mn, Ca, Na, K, Cr, U, Ni)

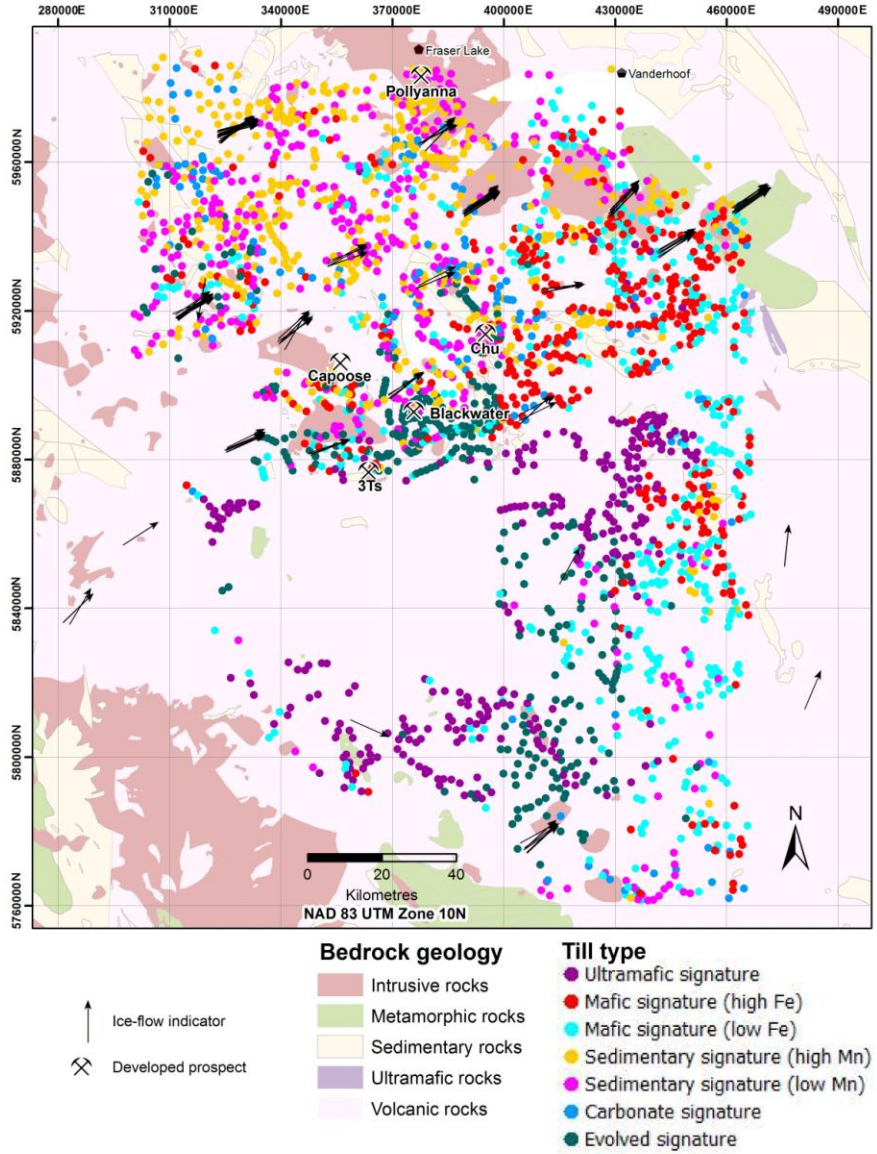




Final product = 7 till "types"

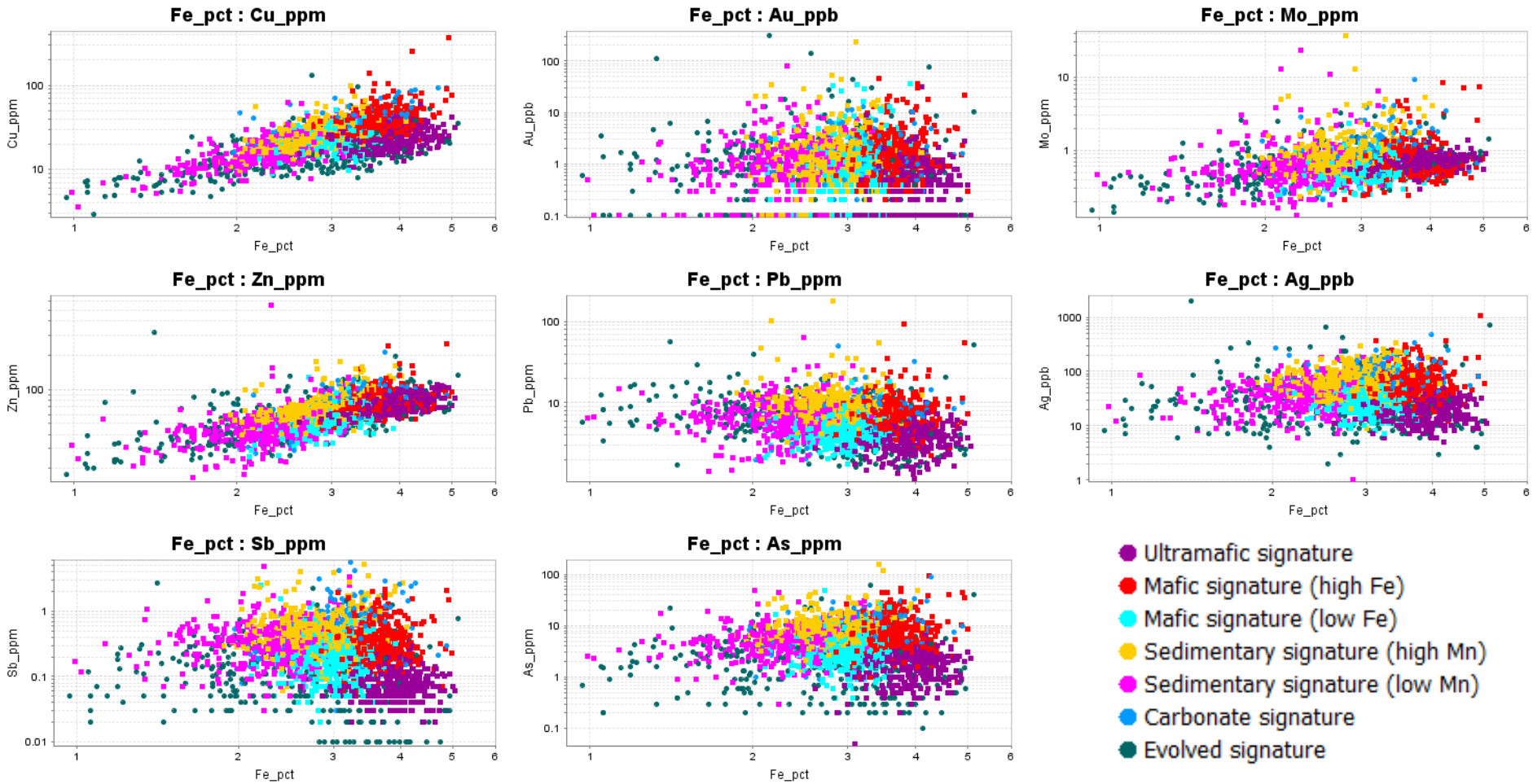
- Ti, Al, Fe, Mg, Mn, Ca, Na, K, Cr, U, Ni
- Results "hang" together spatially





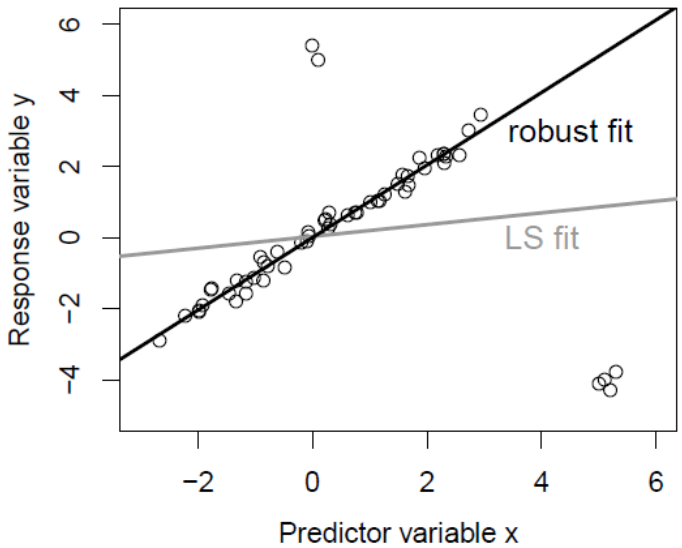
The fun part! Find some anomalies!

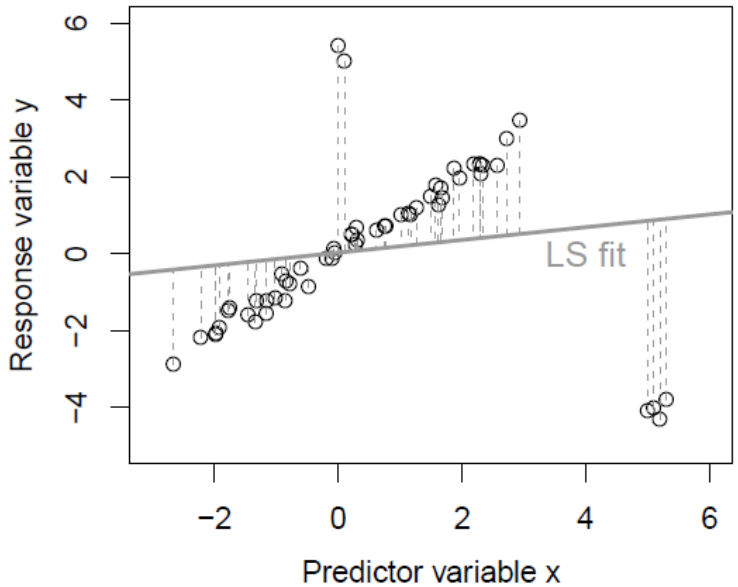
- Regression analysis... Cu Mo Zn Pb Ag Sb As Au Ni Hg Tl U
  - Regress against Fe and Al

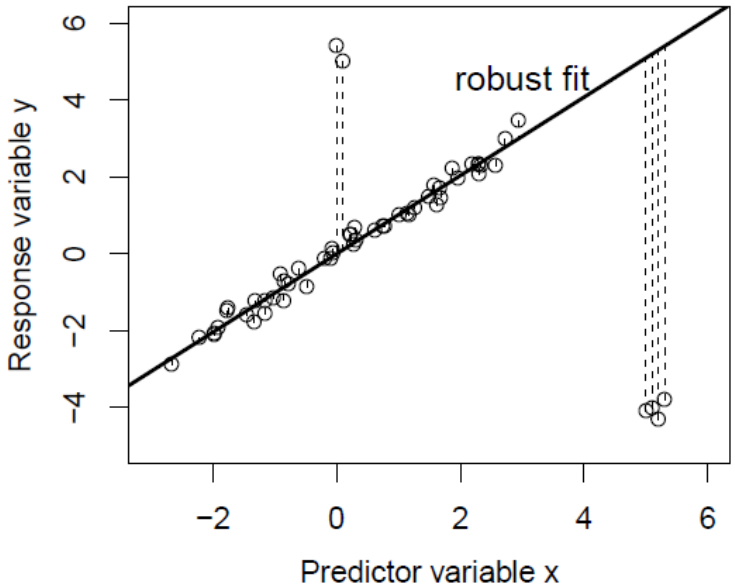


Part 2 - Regression Analysis

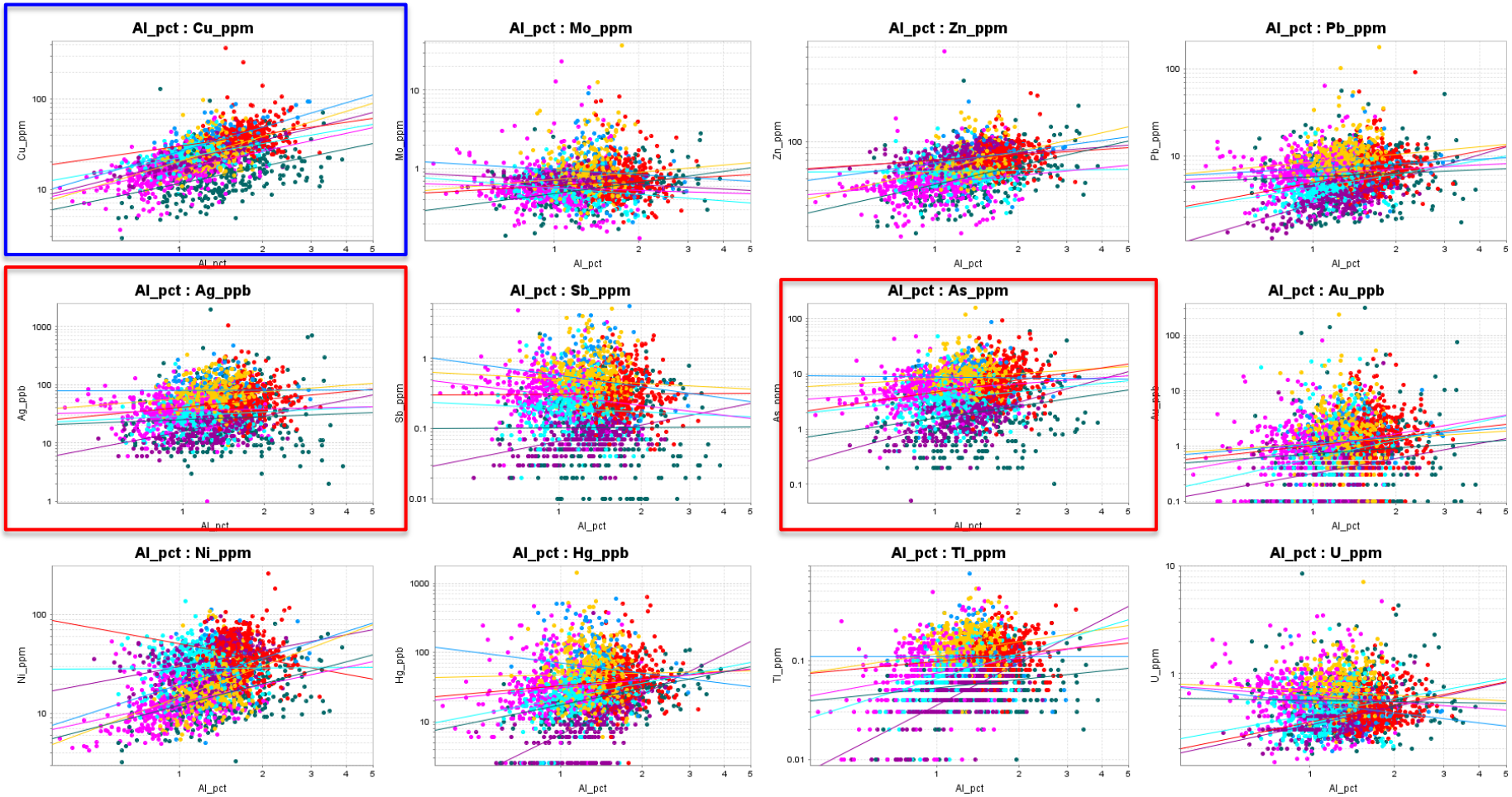
- Robust
- Identify and Quantify outliers of a response variable *with respect to* a predictor variable
- Is there a controlling element in the surficial environment? Fe? Al?
- Is there correlation between a predictor and my pathfinder (response variable)?



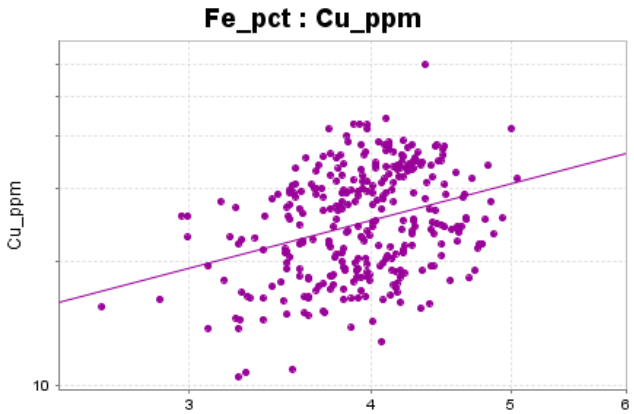


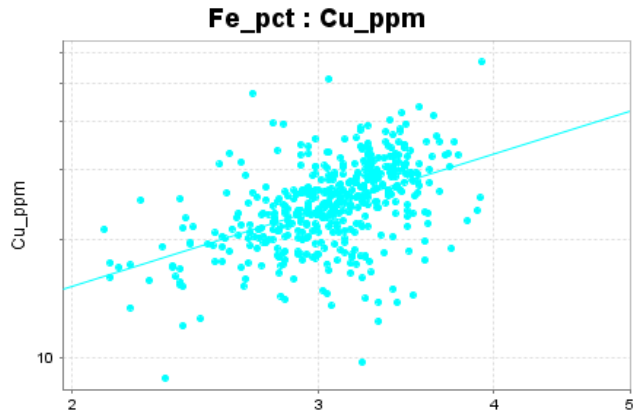


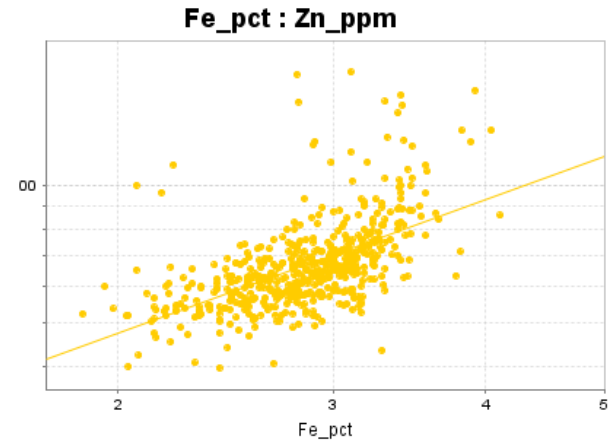
Pathfinder regressions against Al

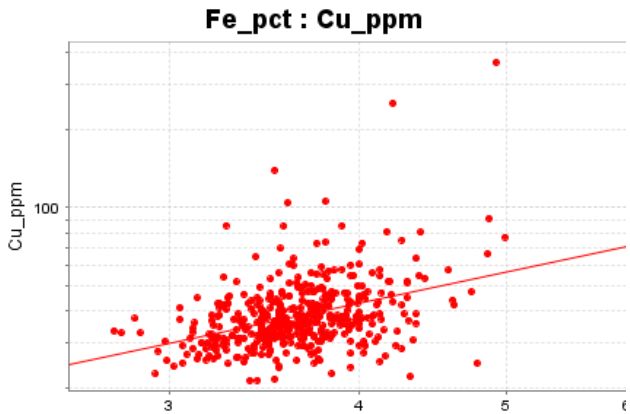


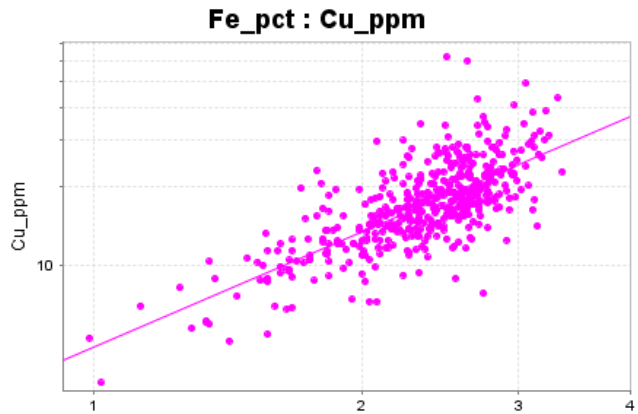
Example – pathfinder regressions "by colour group"



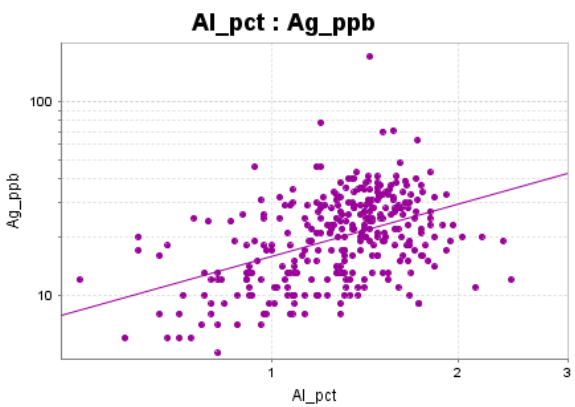


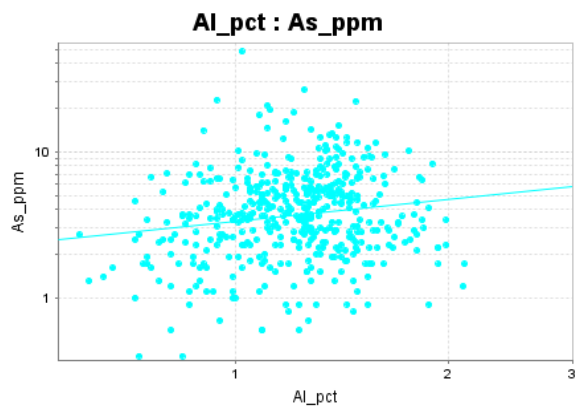


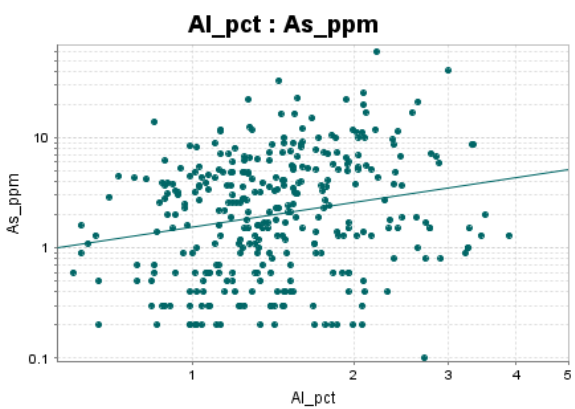




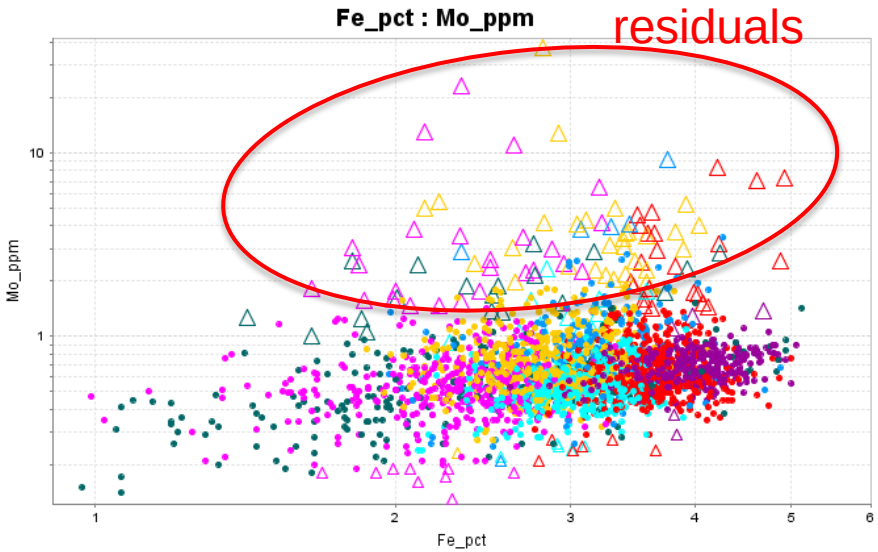
- Ultramafic signature
- Mafic signature (high Fe)
- Mafic signature (low Fe)
- Sedimentary signature (high Mn)
- Sedimentary signature (low Mn)
- Carbonate signature
- Evolved signature



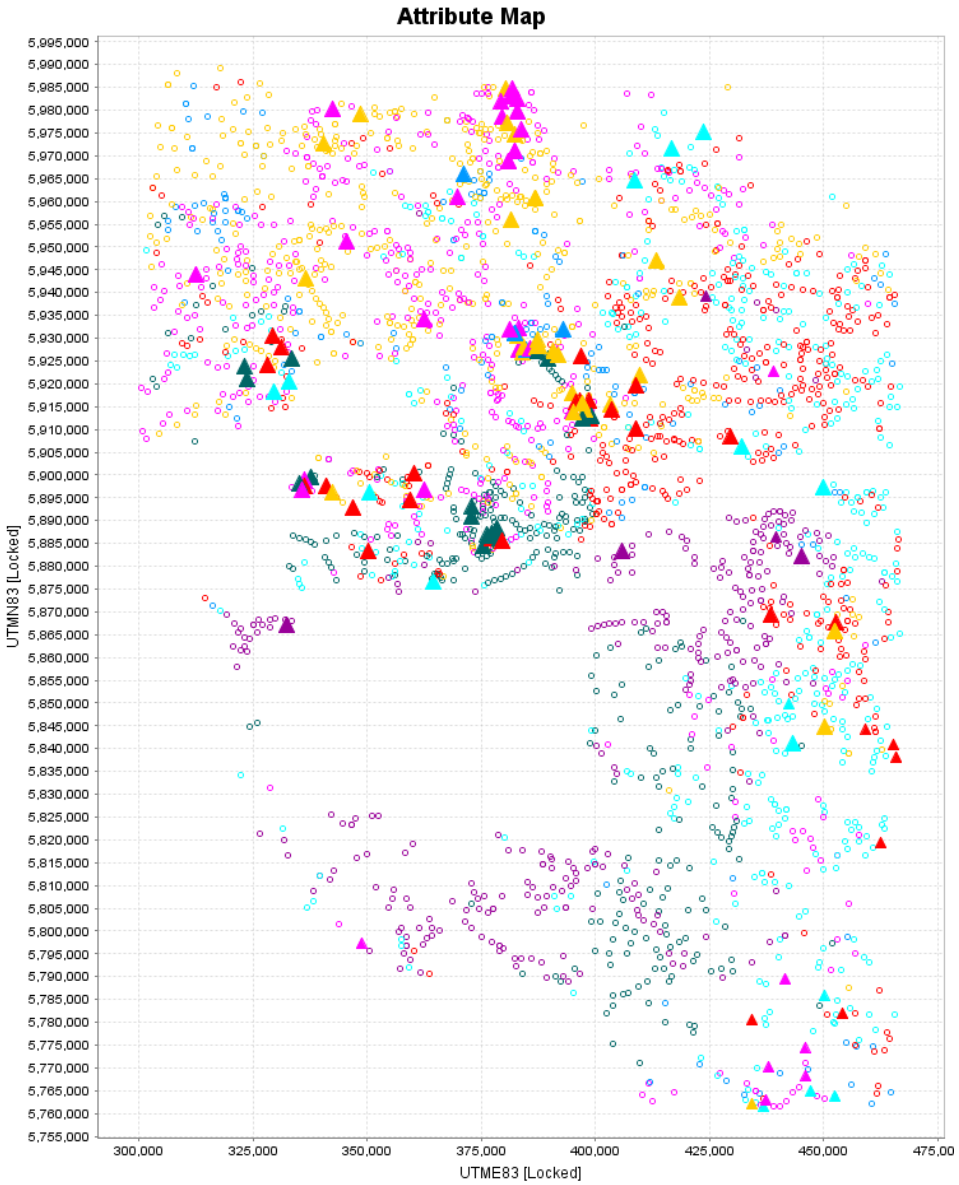




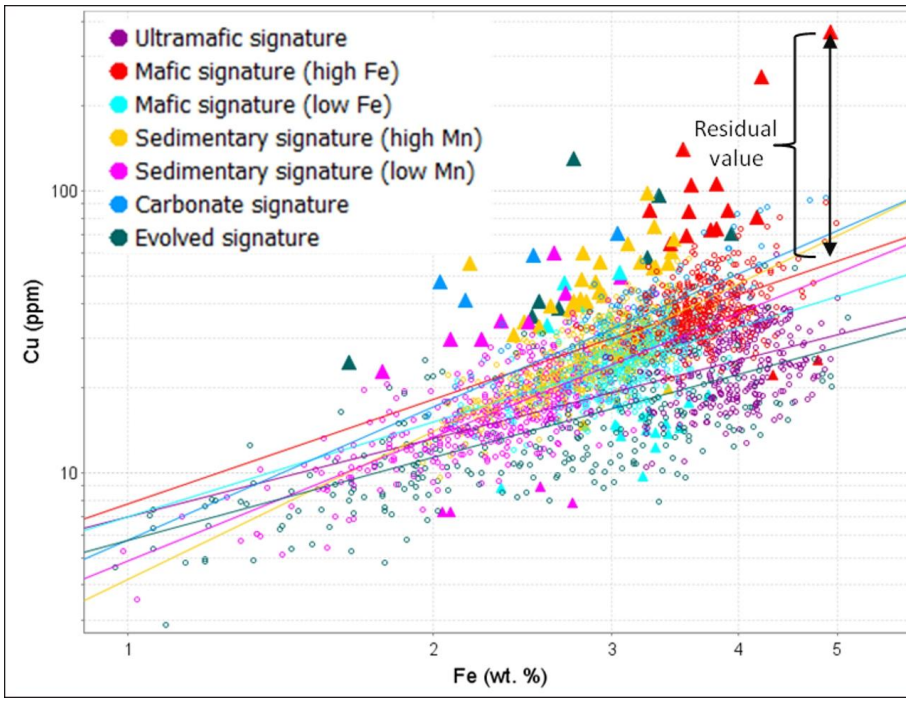
Regression – Mo as a function of Fe

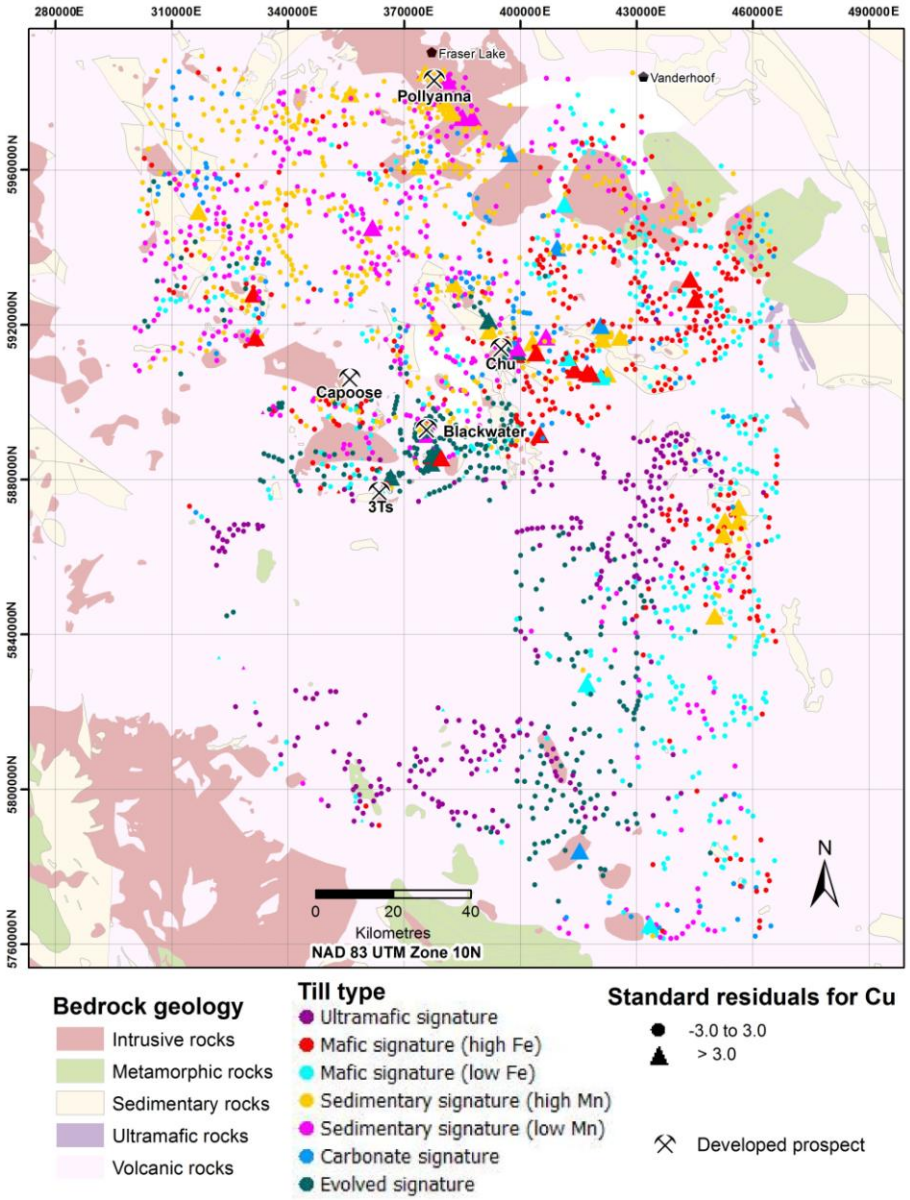


- Ultramafic signature
- Mafic signature (high Fe)
- Mafic signature (low Fe)
- Sedimentary signature (high Mn)
- Sedimentary signature (low Mn)
- Carbonate signature
- Evolved signature

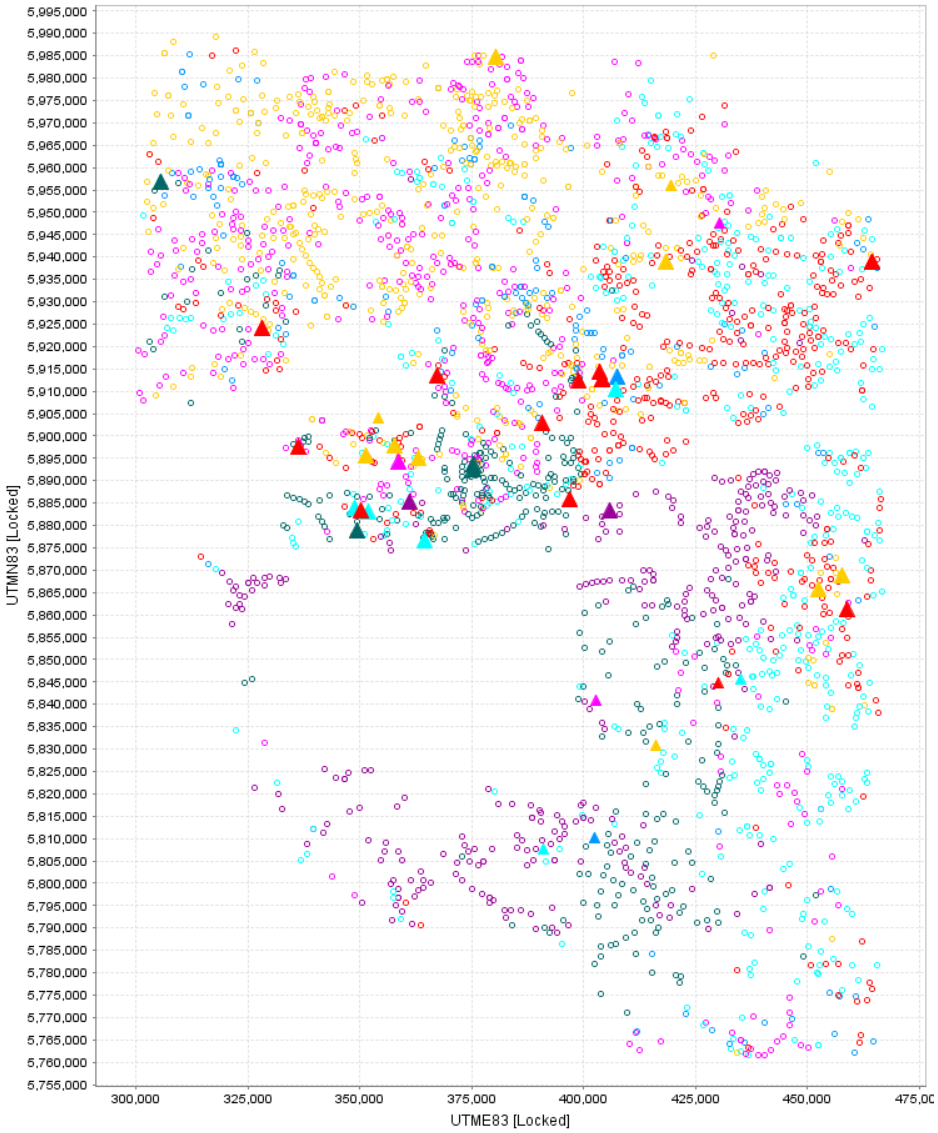


Regression – Cu as a function of Fe

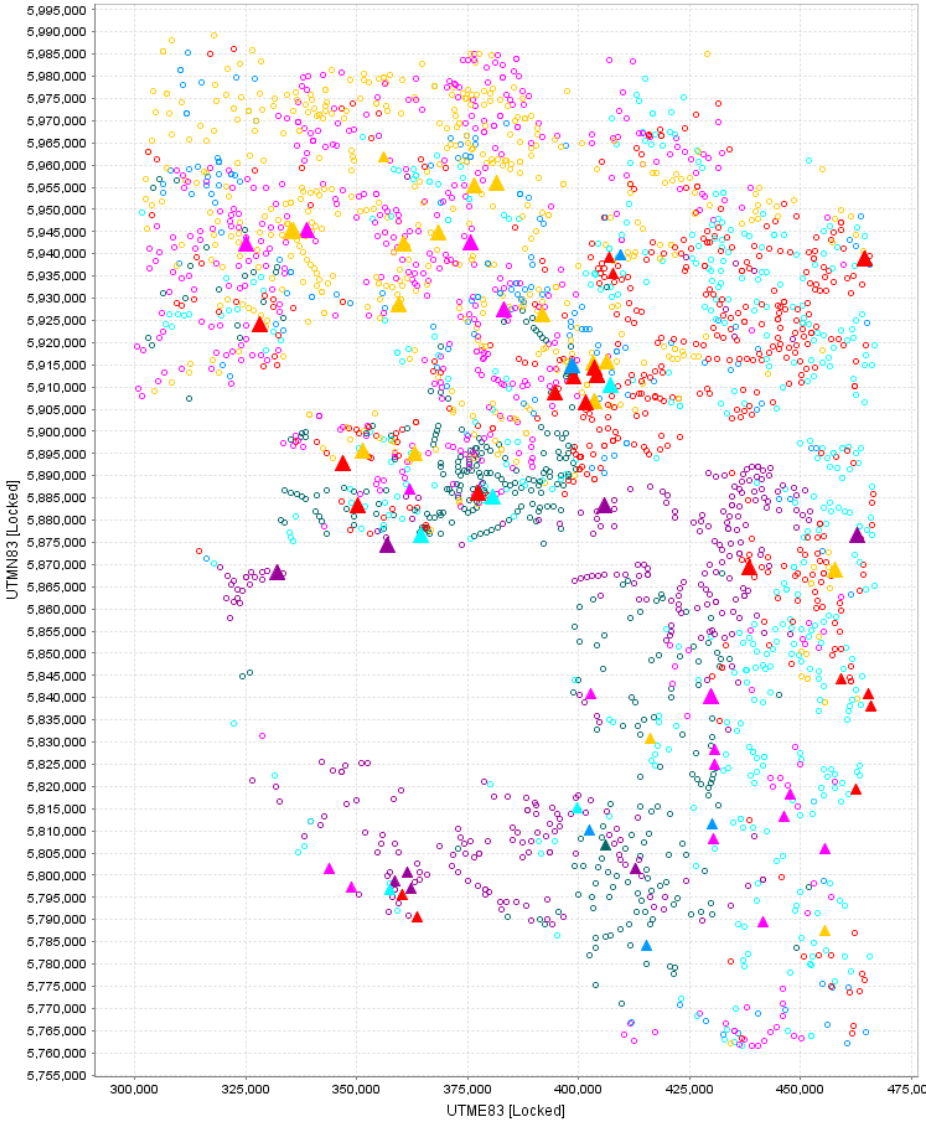




Silver (over Al)

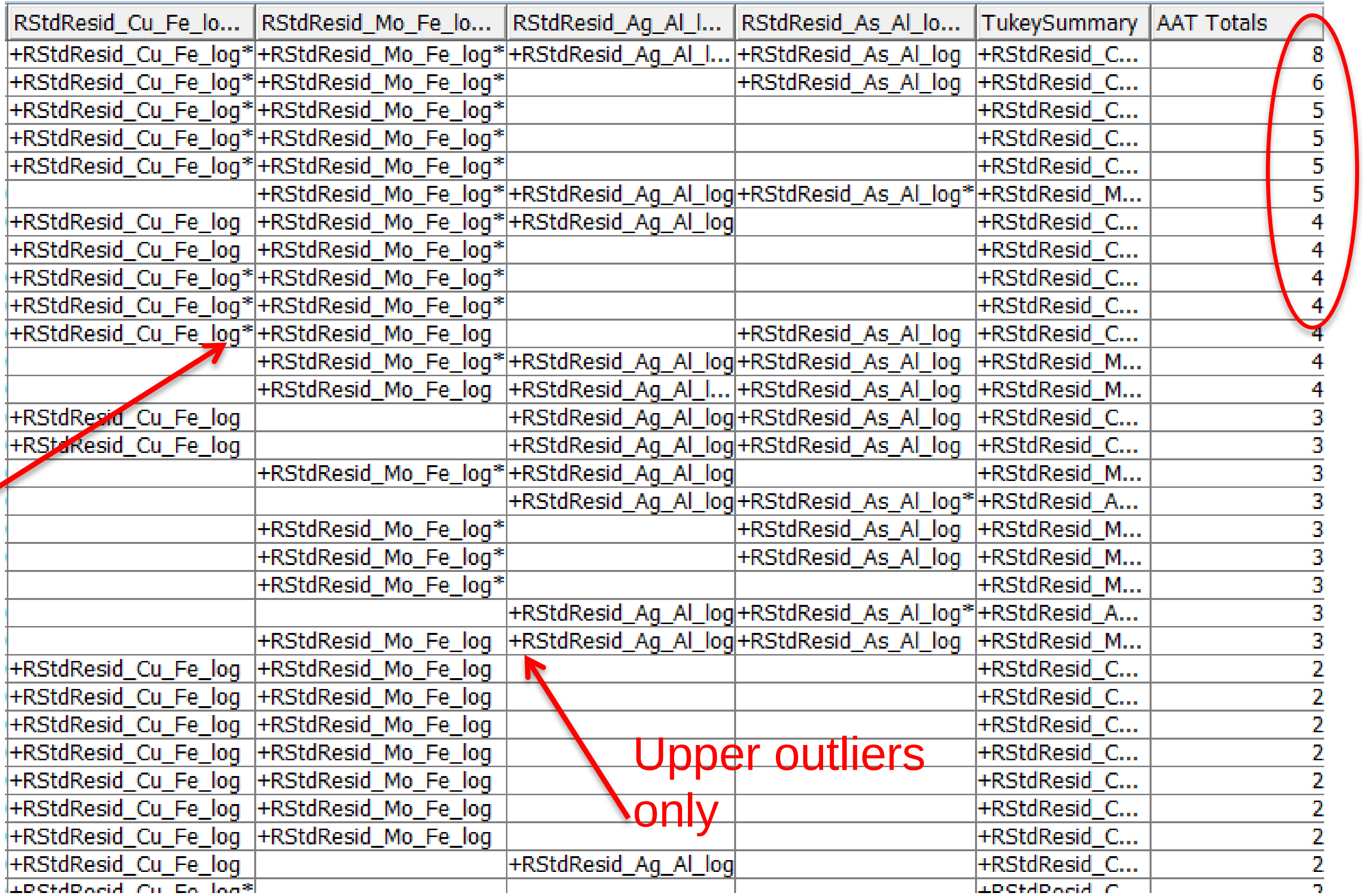


Arsenic (over Al)



Part 3 - Multivariate Pathfinder Assessment

- Cu, Mo (residuals over Fe) and Ag, As (residuals over Al)
- Combination of Tukey outliers and Mahalanobis distances

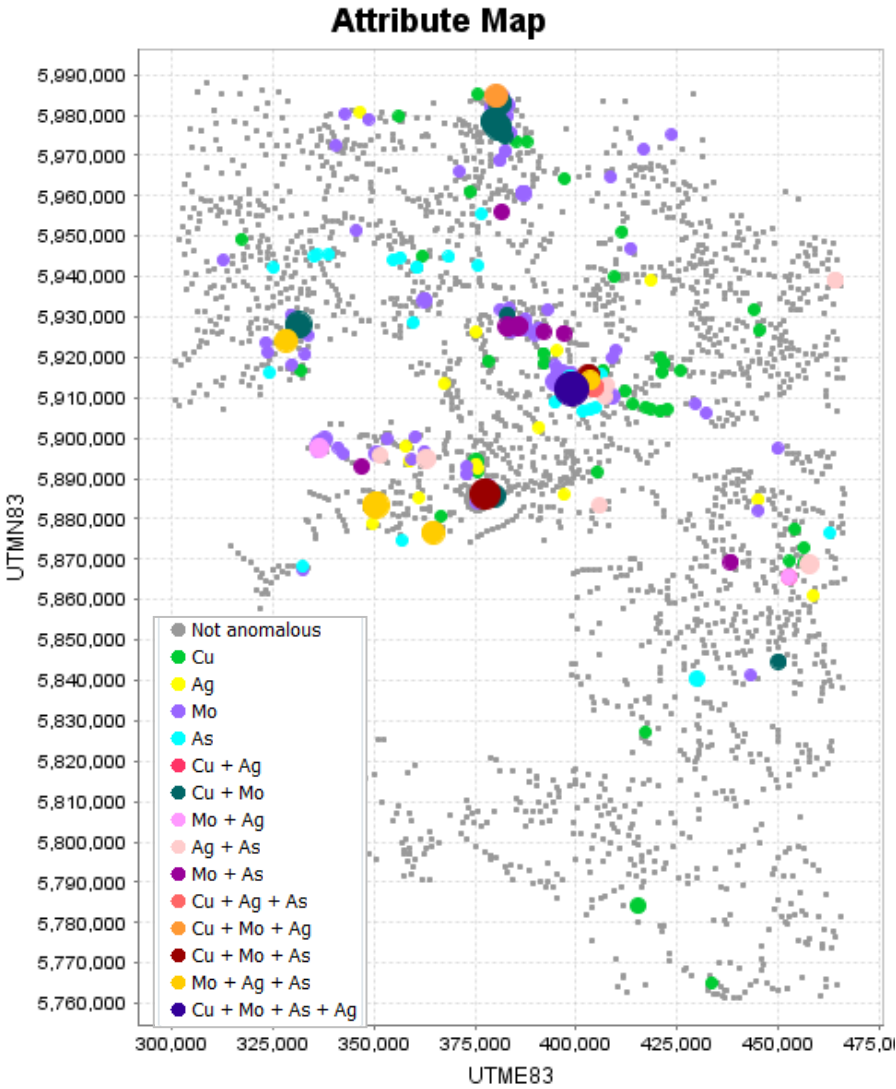


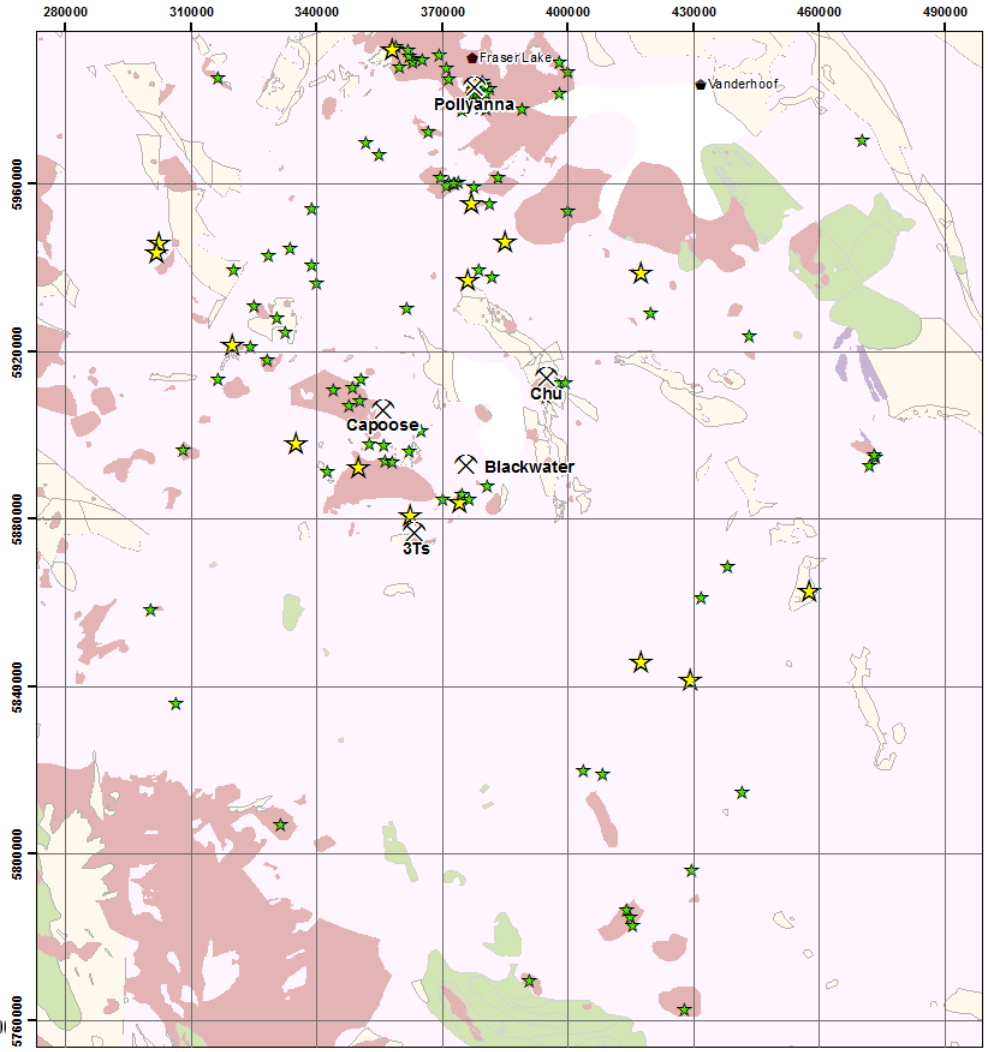
RStdResid_Cu_Fe_lo...	RStdResid_Mo_Fe_lo...	RStdResid_Ag_Al_l...	RStdResid_As_Al_lo...	TukeySummary	AAT Totals
+RStdResid_Cu_Fe_log*	+RStdResid_Mo_Fe_log*	+RStdResid_Ag_Al_l...	+RStdResid_As_Al_log	+RStdResid_C...	8
+RStdResid_Cu_Fe_log*	+RStdResid_Mo_Fe_log*		+RStdResid_As_Al_log	+RStdResid_C...	6
+RStdResid_Cu_Fe_log*	+RStdResid_Mo_Fe_log*			+RStdResid_C...	5
+RStdResid_Cu_Fe_log*	+RStdResid_Mo_Fe_log*			+RStdResid_C...	5
+RStdResid_Cu_Fe_log*	+RStdResid_Mo_Fe_log*			+RStdResid_C...	5
	+RStdResid_Mo_Fe_log*	+RStdResid_Ag_Al_log	+RStdResid_As_Al_log*	+RStdResid_M...	5
+RStdResid_Cu_Fe_log	+RStdResid_Mo_Fe_log*	+RStdResid_Ag_Al_log		+RStdResid_C...	4
+RStdResid_Cu_Fe_log	+RStdResid_Mo_Fe_log*			+RStdResid_C...	4
+RStdResid_Cu_Fe_log*	+RStdResid_Mo_Fe_log*			+RStdResid_C...	4
+RStdResid_Cu_Fe_log*	+RStdResid_Mo_Fe_log*			+RStdResid_C...	4
+RStdResid_Cu_Fe_log*	+RStdResid_Mo_Fe_log		+RStdResid_As_Al_log	+RStdResid_C...	4
	+RStdResid_Mo_Fe_log*	+RStdResid_Ag_Al_log	+RStdResid_As_Al_log	+RStdResid_M...	4
	+RStdResid_Mo_Fe_log	+RStdResid_Ag_Al_l...	+RStdResid_As_Al_log	+RStdResid_M...	4
+RStdResid_Cu_Fe_log		+RStdResid_Ag_Al_log	+RStdResid_As_Al_log	+RStdResid_C...	3
+RStdResid_Cu_Fe_log		+RStdResid_Ag_Al_log	+RStdResid_As_Al_log	+RStdResid_C...	3
	+RStdResid_Mo_Fe_log*	+RStdResid_Ag_Al_log		+RStdResid_M...	3
		+RStdResid_Ag_Al_log	+RStdResid_As_Al_log*	+RStdResid_A...	3
	+RStdResid_Mo_Fe_log*		+RStdResid_As_Al_log	+RStdResid_M...	3
	+RStdResid_Mo_Fe_log*		+RStdResid_As_Al_log	+RStdResid_M...	3
	+RStdResid_Mo_Fe_log*			+RStdResid_M...	3
		+RStdResid_Ag_Al_log	+RStdResid_As_Al_log*	+RStdResid_A...	3
	+RStdResid_Mo_Fe_log	+RStdResid_Ag_Al_log	+RStdResid_As_Al_log	+RStdResid_M...	3
+RStdResid_Cu_Fe_log	+RStdResid_Mo_Fe_log			+RStdResid_C...	2
+RStdResid_Cu_Fe_log	+RStdResid_Mo_Fe_log			+RStdResid_C...	2
+RStdResid_Cu_Fe_log	+RStdResid_Mo_Fe_log			+RStdResid_C...	2
+RStdResid_Cu_Fe_log	+RStdResid_Mo_Fe_log			+RStdResid_C...	2
+RStdResid_Cu_Fe_log	+RStdResid_Mo_Fe_log			+RStdResid_C...	2
+RStdResid_Cu_Fe_log	+RStdResid_Mo_Fe_log			+RStdResid_C...	2
+RStdResid_Cu_Fe_log	+RStdResid_Mo_Fe_log			+RStdResid_C...	2
+RStdResid_Cu_Fe_log		+RStdResid_Ag_Al_log		+RStdResid_C...	2
+RStdResid_Cu_Fe_log*				+RStdResid_C...	2

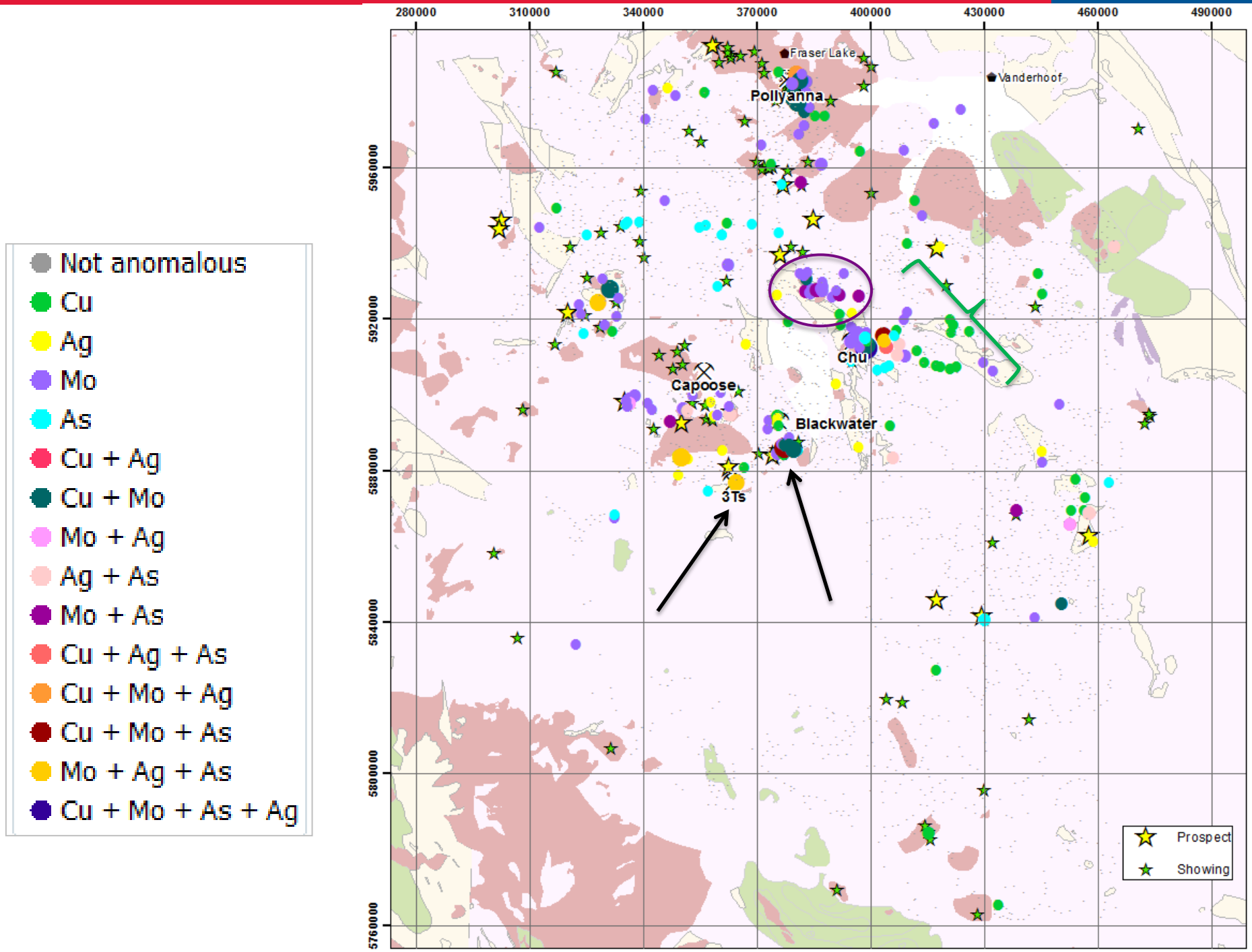
Far outlier*

Upper outliers only

Multivariate anomalies and known mineral occurrences

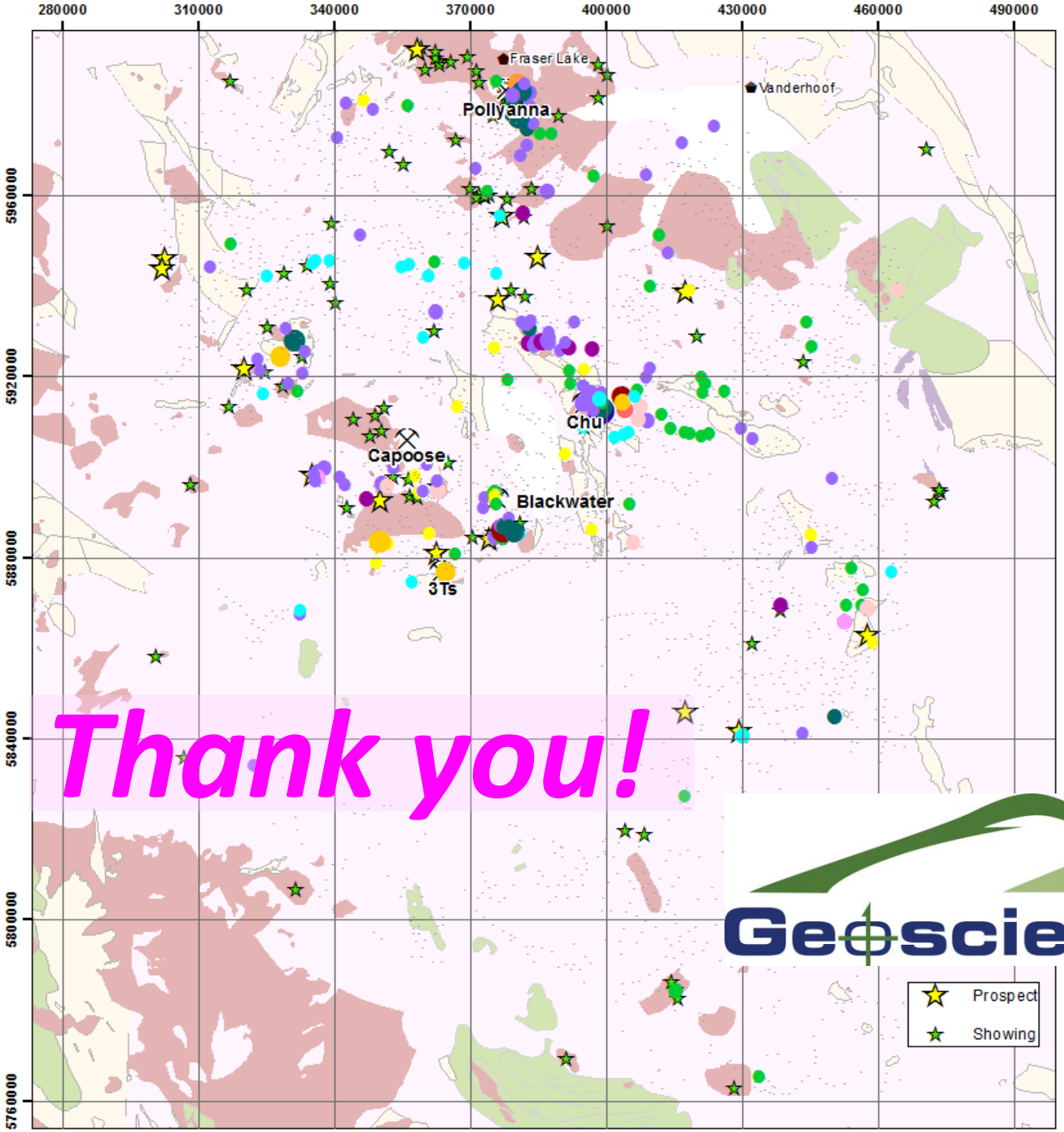





Not anomalous
Cu
Ag
Mo
As
Cu + Ag
Cu + Mo
Mo + Ag
Ag + As
Mo + As
Cu + Ag + As
Cu + Mo + Ag
Cu + Mo + As
Mo + Ag + As
Cu + Mo + As + Ag
280000
310000
340000
370000
400000
430000
460000
490000
5960000
5920000
5880000
5840000
5800000
5760000
Fraser Lake
Vanderhoof
Pollyanna
Chu
Capoose
Blackwater
3Ts
Prospect
Showing

Conclusions

- Simple geochemical interpretative methods (clustering, regression) allowed greater insight into the architecture of the geochemistry in the TREK area
- Second order derivative products (like those created in geophysics) which account for surficial processes, provide more confidence for targeting
- Samples with anomalous chalcophile element signatures were identified *per till group* and then quantified using robust regression analysis and combined using multivariate anomaly assessment
- The resulting targets are coincident with developed prospects, but also highlighted areas without known mineralization

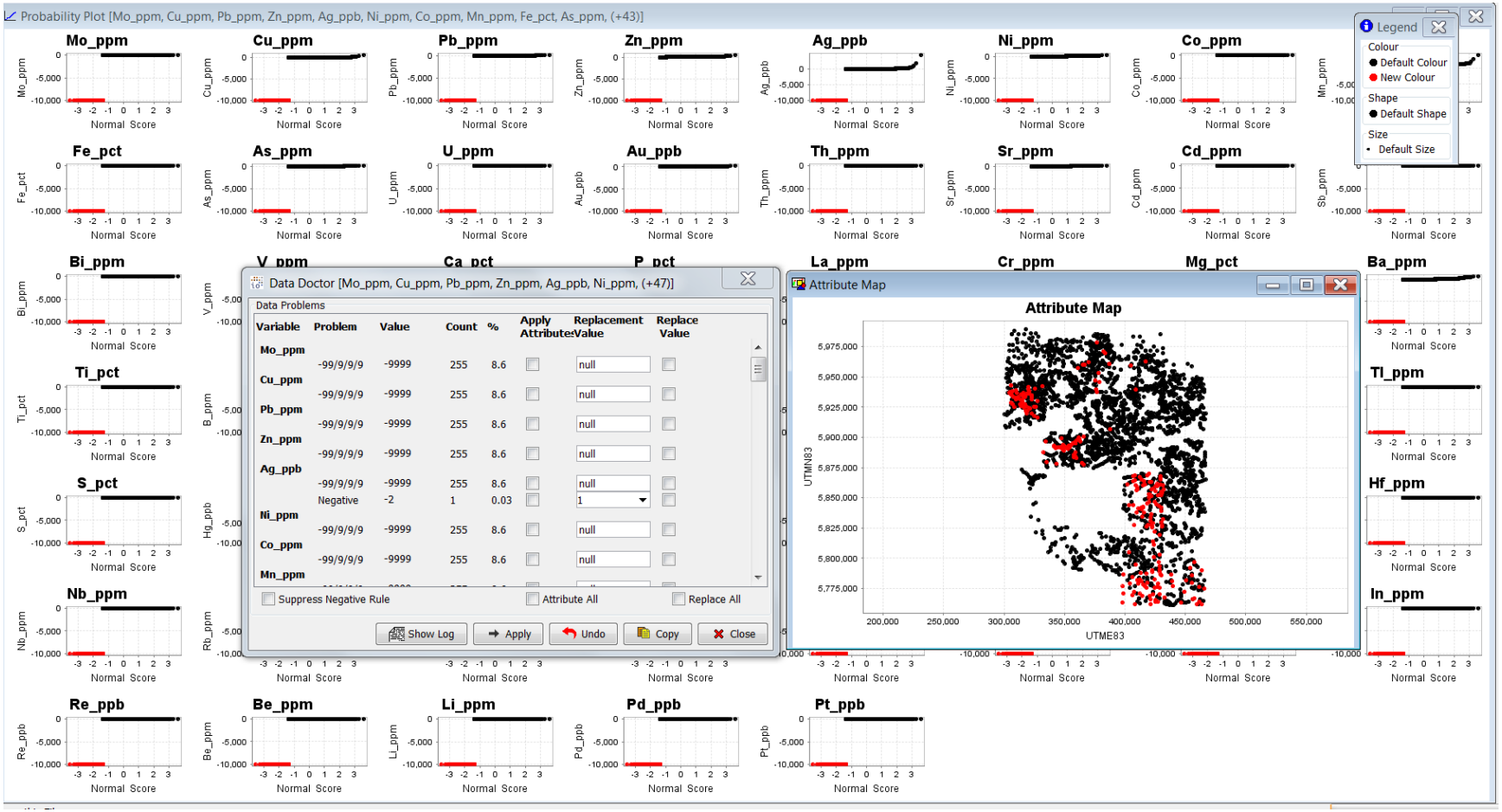


Thank you!

REFLEX® intelligence on demand

Geoscience BC

Data cleaning



Summary of previous work & Project Area

Cluster Analysis: Basics

- Cluster definition: a collection of data objects
  - Similar to one another within the same cluster.
  - Dissimilar to the objects in other clusters.

- Cluster analysis:
  - The main objective of clustering algorithms is to statistically represent distinct natural groupings of multidimensional data in as few groups as possible.
  - Based on a measure of distance between observations.
  - Used to discover structures in data without providing an explanation/interpretation.
  - Useful for partitioning a very large population, perhaps for data mining each sub-population separately.

Major Clustering Approaches

- 1. Hard partitioning algorithms, e.g. K-Means and K-Medoids
  - Construct various partitions and then evaluate them by some criterion. A fixed number of clusters, k, is generated (i.e. you select how many groups the data is to be partitioned into).
- 2. Hierarchy algorithms (not discussed further)
  - Create a hierarchical decomposition of the set of data (or objects) using some criterion.

K-Means: Hard Partitioning

- K-means algorithm
  - Breaks the dataset into groups.
  - Attempts to minimise the distance between points labelled to be in a cluster and a point designated as the centre of that cluster.
  - Distance used to measure a samples distance from the group's centroid is the Euclidean distance.
  - k, the number of groups, needs to be specified.

- Typical applications
  - As a stand-alone tool to get insight into data distribution.
  - As a pre-processing step for other algorithms.

K-Means

- Given k, the k-means algorithm uses a search heuristic and is implemented in 4 steps:
  1. **It begins with a random collection of k clusters**, i.e. objects partitioned into k non-empty subsets. Each cluster is represented by a vector of the mean values for each of the variables.
  2. **Compute seed points as the centroids** of the clusters of the current partition. The centroid is the centre (mean point) of the cluster.
  3. **Assign each object to the cluster** with the nearest seed point, i.e. Measure the distance between an observation and each of the k vectors of the mean values.
  4. **Recalculate the mean values based on the observations** that are now associated with each cluster, i.e. go back to Step 2. This often results in observations moving between clusters.
  - This **iterative process** is repeated until no more observations move from one cluster to another.

Objectives

1. Evaluate and assess all available till geochemical analyses in the TREK project area for comparability and utility on an element by element basis.
2. Rigorously interpret the till geochemical data selected in the first step and use exploratory data analysis (EDA) techniques (including cluster analysis) to determine till signature. Regression analyses were also employed to account for secondary surficial processes.
3. Create robust, second order, geochemical derivative products (ground-truthed to known mineral occurrences where possible) that delineate areas of increased mineral potential