



Data science in everyday geochemistry programs

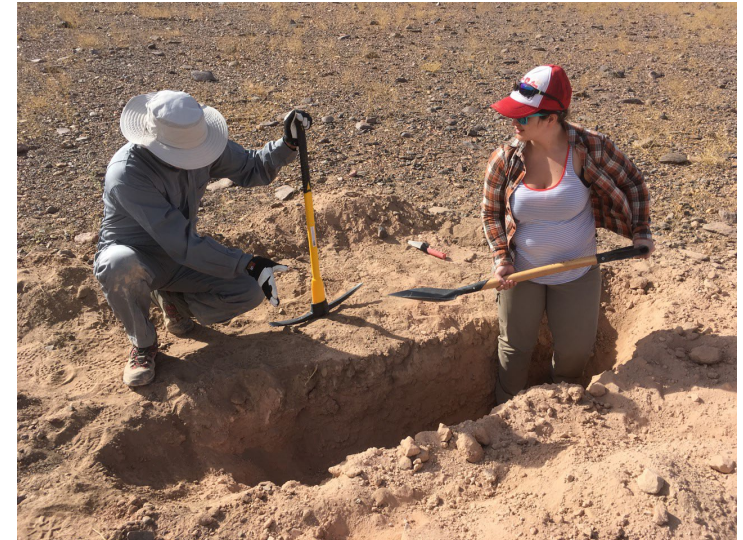
International Association of Applied Geochemistry 2024

October 12, 2024

Sam Scher | sscher@lkiconsulting.com
Tom Carmichael, Datarock

Oxides & media

Background information



Soils, streams and rock chips in exploration

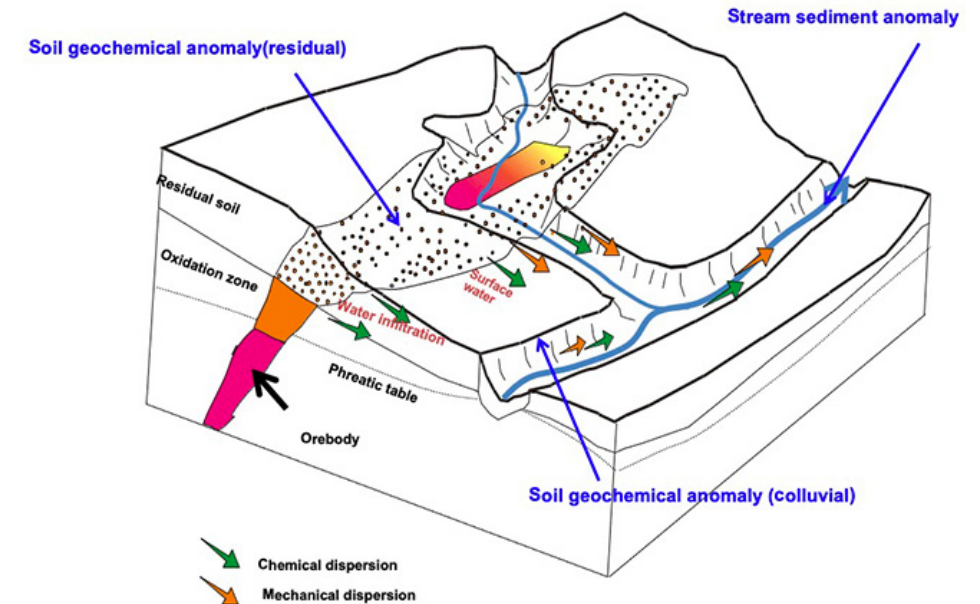
- **Stream Sediments:** identify broad geochemical anomalies and provide initial data for large-scale area coverage.
- **Soils:** refine target areas by detecting localized geochemical anomalies and element dispersion patterns.
- **Rock Chips:** confirm mineralization directly from the source, offering high-confidence data for ore potential.
- **Benefits:**
 - **Map Anomalies:** each method helps pinpoint geochemical anomalies and refine exploration targets.
 - **Optimize Drilling:** guides where to drill and conduct geophysical surveys, improving exploration efficiency.
 - **Reduce Risk:** combined data minimize drilling in non-productive areas, saving costs and resources.



Soil and salt sampling in faults, Atacama Desert, Chile.

Stream sediments

- Stream sediments are materials transported by rivers and streams. They provide a natural sample of the drainage basin upstream and can represent a mix of the bedrock and soil geochemistry of a large area.
- **Impact:**
 - **Broad Area Sampling:** stream sediments can cover large areas efficiently, allowing geochemists to identify regions with anomalous metal concentrations that may signal mineralization upstream.
 - **Pathfinder Elements:** they help identify pathfinder elements associated with specific ore types, guiding exploration to areas with higher potential.
 - **Downstream Geochemical Signature:** the sediments carry downstream signals of mineralization, helping to identify the source area of the mineralized material.

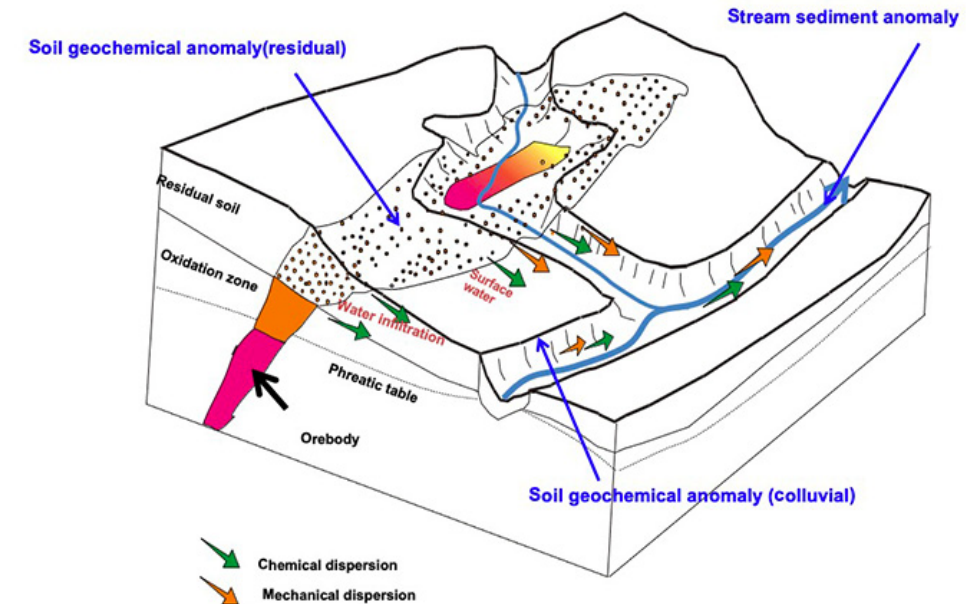


Dispersion: The key to success for geochemical exploration

[link](#)

Soils

- Soil samples are taken from the near-surface layer and represent the breakdown of underlying bedrock and transported materials.
- **Impact:**
 - **Targeted Anomalies:** Soil sampling is useful for delineating more localized geochemical anomalies, particularly in areas where outcrops are scarce or hidden by vegetation or weathering.
 - **Direct Reflection of Underlying Geology:** soil geochemistry may reflect the geochemical signature of underlying rocks, making it effective for detecting buried ore bodies or alteration zones.
 - **Geochemical Dispersion Halos:** soils can capture the geochemical dispersion halos that often extend beyond the mineralized zones, allowing for identification of broader target areas.



Dispersion: The key to success for geochemical exploration

[link](#)

Rock chips

- Rock chips are small fragments or chunks of rock collected directly from outcrops or from subsurface material during drilling or trenching.
- **Impact:**
 - **Direct Source Sampling:** rock chips provide a direct sample from the bedrock or mineralized zone, offering high-confidence data on the geochemistry of the source material.
 - **Lithogeochemistry:** analysis of rock chips gives a clear understanding of the chemical composition of the lithology and can help identify specific types of mineralization or alteration zones.
 - **Pinpointing Mineralized Zones:** while soils and stream sediments may indicate an anomaly, rock chips provide more precise information about the grade and extent of mineralization at a particular site.
 - **Alteration Signatures:** rock chips can also reveal the geochemical signatures of alteration zones, which are often associated with ore deposits, helping to narrow down exploration targets.



[Earth Science Australia](http://www.earthscienceaustralia.org.au)

Examples of ore bodies identified with geochemistry

- Hemlo Gold Deposit (Canada)
 - Deposit Type: Orogenic Gold
 - Sampling Method: Stream Sediments, Soils, and Rock Chips
- Red Chris Copper-Gold Deposit (Canada)
 - Deposit Type: Porphyry Copper-Gold
 - Sampling Method: Stream Sediments, Soils, and Rock Chips
- Voisey's Bay Nickel Deposit (Canada)
 - Deposit Type: Nickel-Copper-Cobalt Sulfide
 - Sampling Method: Stream Sediments, Soils, and Rock Chips



Voisey's Bay, Canada

Challenges of using these media

- Components of the sample medium have the ability to adsorb or otherwise trap mobile ions:
 - Organic matter (humic and fulvic acids, charcoal adsorption, chelation, complexation)
 - **Fe and Mn hydroxides (adsorption, co-precipitation)**
 - Clays (adsorption)
 - Carbonates (adsorption, co-precipitation, neutralization)
 - Silica (gels)
- The above host the exogenic signal (i.e., the geochemical signal of introduced ions or elements), which is very useful for exploration.

	Mn-oxides	Amorphous Fe-oxides	Goethite	Amorphous Al-oxides	Humic substances	
					(1)	(2)
Greatest	Cu ²⁺	Pb ²⁺	Cu ²⁺	Cu ²⁺	Ni ²⁺	Cu ²⁺
	Co ²⁺	Cu ²⁺	Pb ²⁺	Pb ²⁺	Co ²⁺	Ni ²⁺
	Mn ²⁺	Zn ²⁺	Zn ²⁺	Zn ²⁺	Pb ²⁺	Co ²⁺
	Zn ²⁺	Ni ²⁺	Co ²⁺	Ni ²⁺	Cu ²⁺	Pb ²⁺
	Ni ²⁺	Cd ²⁺	Cd ²⁺	Co ²⁺	Zn ²⁺	Ca ²⁺
	Ba ²⁺	Co ²⁺		Cd ²⁺	Mn ²⁺	Zn ²⁺
	Sr ²⁺	Sr ²⁺		Mg ²⁺	Ca ²⁺	Mn ²⁺
	Ca ²⁺	Mg ²⁺		Sr ²⁺	Mg ²⁺	Mg ²⁺
Least	Mg ²⁺					

* (1) Schnitzer and Hanson (1970); (2) Gamble and Schnitzer (1973).

In the secondary environment, metals move

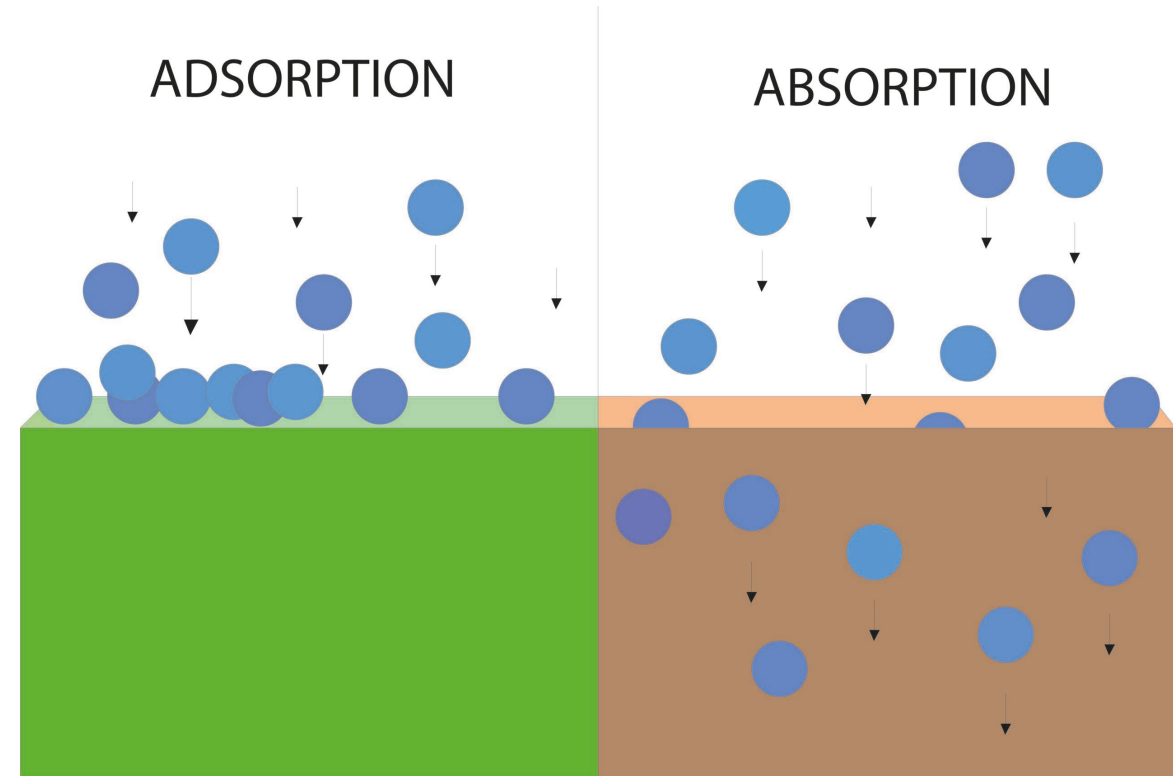
Approximate Eh range (volts)	Oxidising Acid <i>Very high mobility</i> B Br Cl I S He Rn <i>High mobility</i> Ca F Li Mg Na Cd Co Cu Ni U V Se Y Zn <i>Moderate mobility</i> Al Si Ag Au Hg Sb REE <i>Low mobility</i> K Li P Rb Si As Ba Be Bi Ge Mo Pb Se Cs Ti <i>Very low mobility</i> Fe Mn Ti Cr Ga Nb Pt Sn Ta W Zr	Oxidising Neutral to Alkaline <i>Very high mobility</i> B Br Cl I S He Rn <i>High mobility</i> F Na Mg F Sr Pd Se U V <i>Moderate mobility</i> Ca Na Mg Ag As Cd Cu Mo Ni Co Zn REE <i>Low mobility</i> K Li Mn P Si Rb Ba Be Bi Cs Pb Ge Hg Sb Ti <i>Very low mobility</i> Al Fe Ti Cr Te W Nb Ta Pt Sn Th Zr
	Reducing Acid <i>Very high mobility</i> Br Cl He I Rn <i>High to Moderate mobility</i> Ca F Fe K Mg Mn Na Sr Cd Cu Ni Pb Zn <i>Low to very low mobility</i> K Fe P Rb Si Ag As Ba Co Hg Li	Reducing Neutral to Alkaline <i>Very high mobility</i> Br Cl He I Rn <i>High to Moderate mobility</i> Ba Ca F Li Mg Mn Na Sr <i>Low to very low mobility</i> K Fe P Rb Si S Ag As Au Ba Bi Cd Co Cu Cr Hg Li Mo Ni Pb Pt REE Sb U V Zn
	Approximate pH range	
	<4 5 6-8	



Ferric and Al-hydroxide precipitation where acidic and neutral streams mix. Copper, Zn, Co, Ni, Mn, Ca, Mg, etc. are adsorbed onto Fe-hydroxide precipitates.

Metal scavenging: adsorption

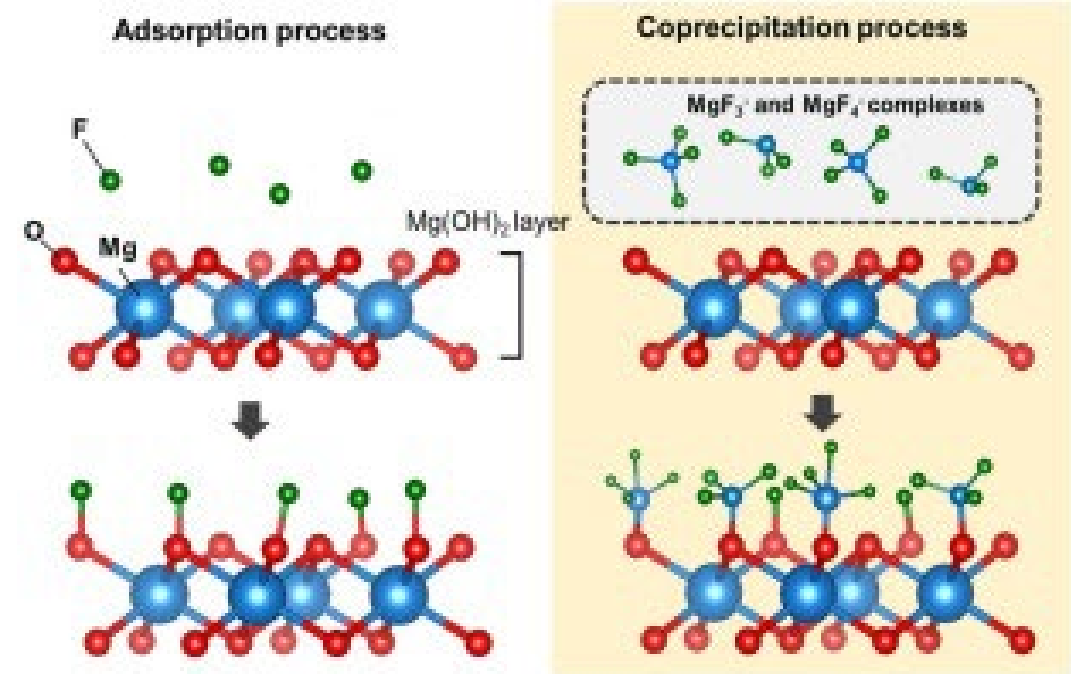
- Adsorption is the adhesion of metal ions to a surface.
- Fe and Mn hydroxides have high adsorption capacities.
- pH controls the adsorption capacity of clays:
 - At acid pH, H^+ ions are preferentially adsorbed or exchanged in place of metals.
 - At neutral to alkaline pH, clays can absorb high concentrations of metals.



[link](#)

Metal scavenging: co-precipitation and adsorption

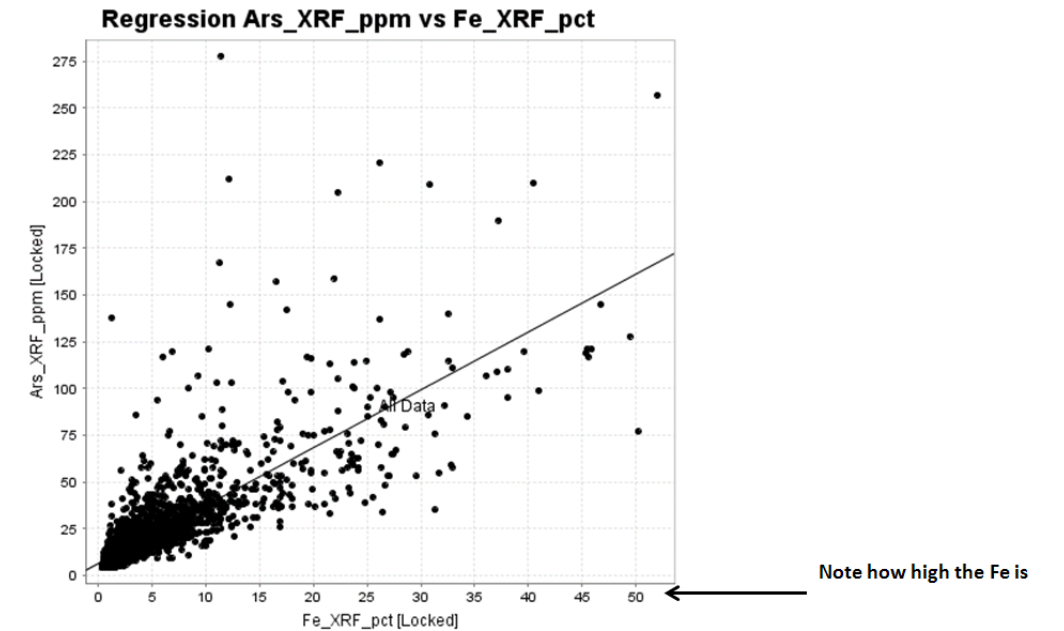
- Fe^{2+} and Mn^{2+} can remain in solution under strongly acidic or reducing conditions.
- Under oxidizing surface conditions, they precipitate as hydroxides.
- Precipitation can occur where reduced groundwater emerges at surface.
- Other elements may be incorporated into these hydroxides by co-precipitation.
- Mobile elements such as Zn and Cd can be removed from solution at pH below their pH of hydrolysis.
- Mo and As tend to be adsorbed onto Fe hydroxide precipitates; Co, Ni, and Cd on Mn hydroxides.



[Tsuchiya et al., 2020](#)

How I was taught to deal with the effects of scavenging

- To interpret the geochemical data correctly, it is important to differentiate between the concentration of the element that is naturally present due to the geologic background and the concentration of the element that is absorbed onto Fe-oxides due to scavenging.
- A basic normalization ratio doesn't fully account for the natural geochemical relationship between the element and Fe so for improved accuracy **residuals** are calculated to account for the expected relationship between the element and Fe under natural conditions.



Benn, 2014

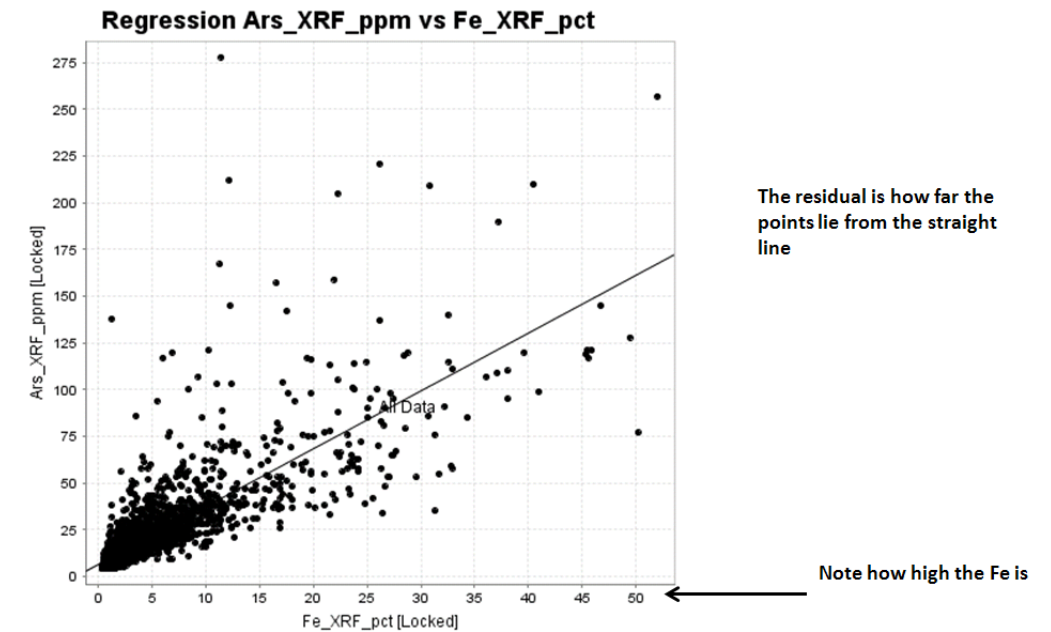
How I was taught to deal with the effects of scavenging

- First a regression line is calculated between the concentrations of Fe and the other element across all samples. This regression line represents the expected concentration of the element based on the amount of Fe, assuming a natural (non-scavenging) relationship.
- A **residual** is the difference between the observed concentration of the element and the value predicted by the regression model.

Residual = Observed concentration of X – Predicted concentration of X (from Fe)

Fe has a strong control on the As content – it is scavenging

As a result it is necessary to normalize the As by the Fe



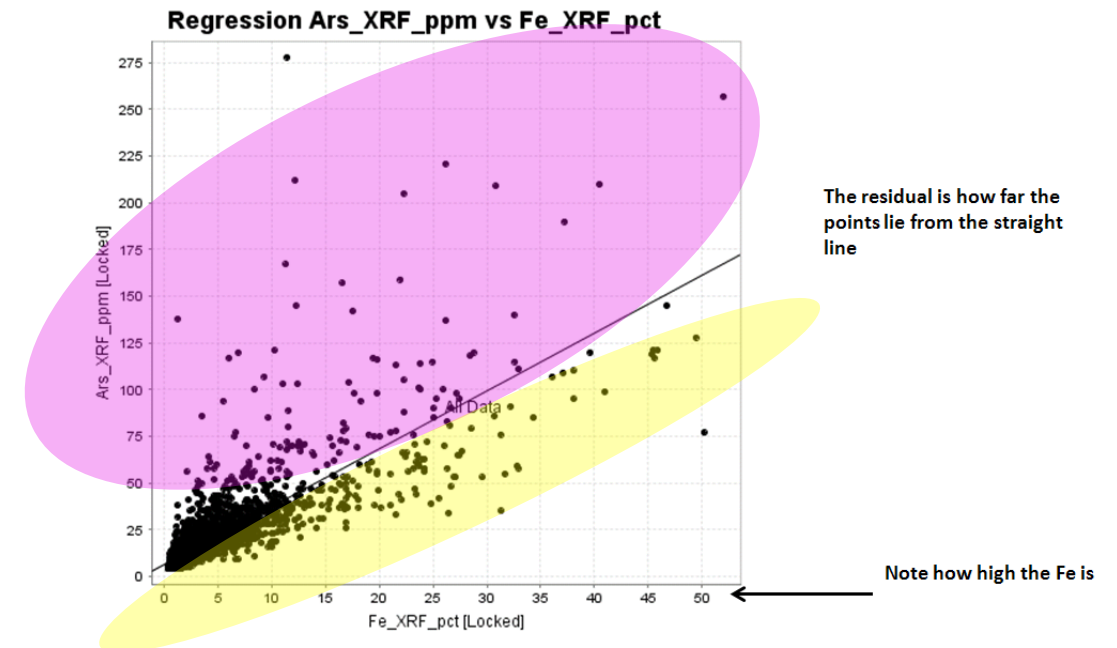
Benn, 2014

How I was taught to deal with the effects of scavenging

- A **positive** residual indicates that the element is present in higher-than-expected amounts, which may suggest an external input of the element beyond the influence of Fe-oxide scavenging
- A **negative** residual suggests that the element is present in lower-than-expected amounts, potentially indicating depletion due to Fe-oxide scavenging.

Fe has a strong control on the As content – it is scavenging

As a result it is necessary to normalize the As by the Fe

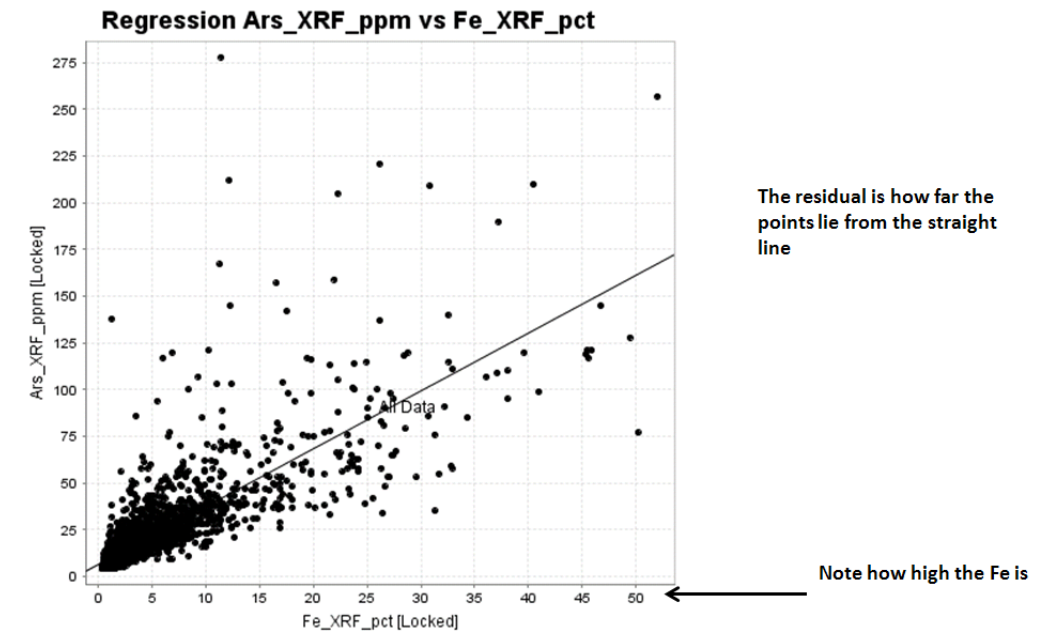


How I was taught to deal with the effects of scavenging

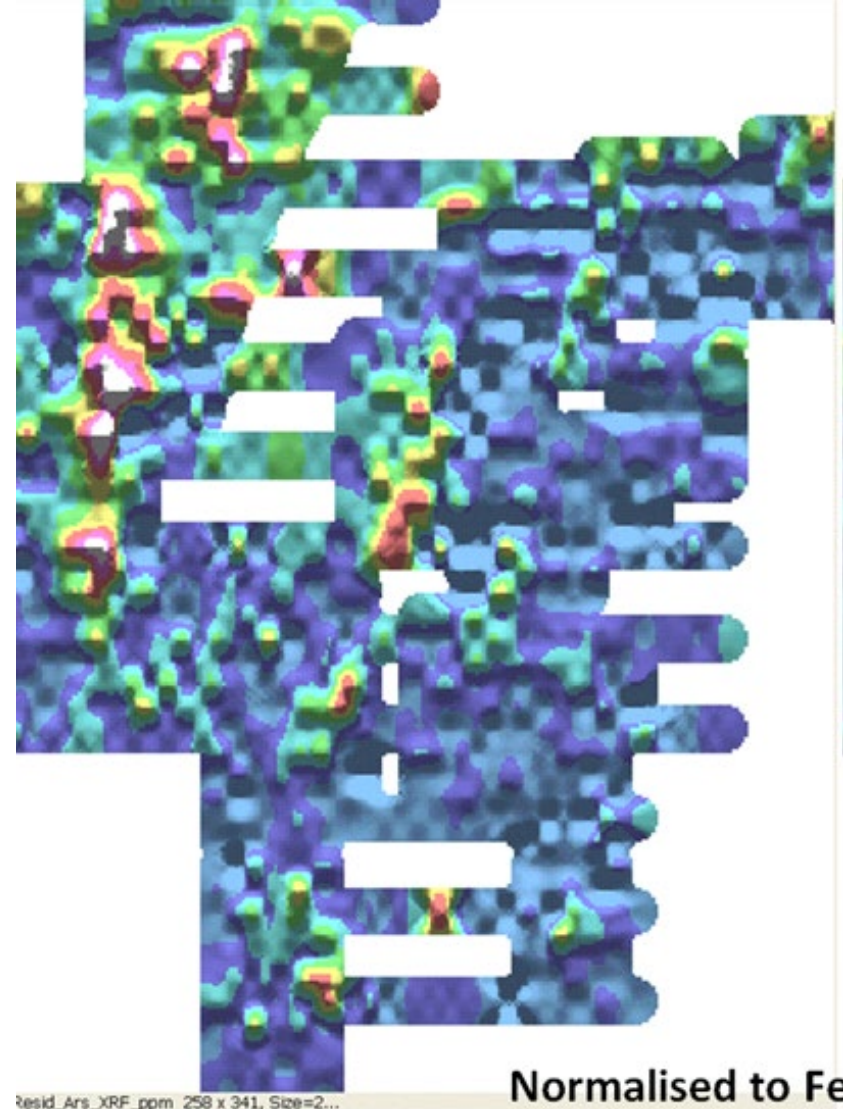
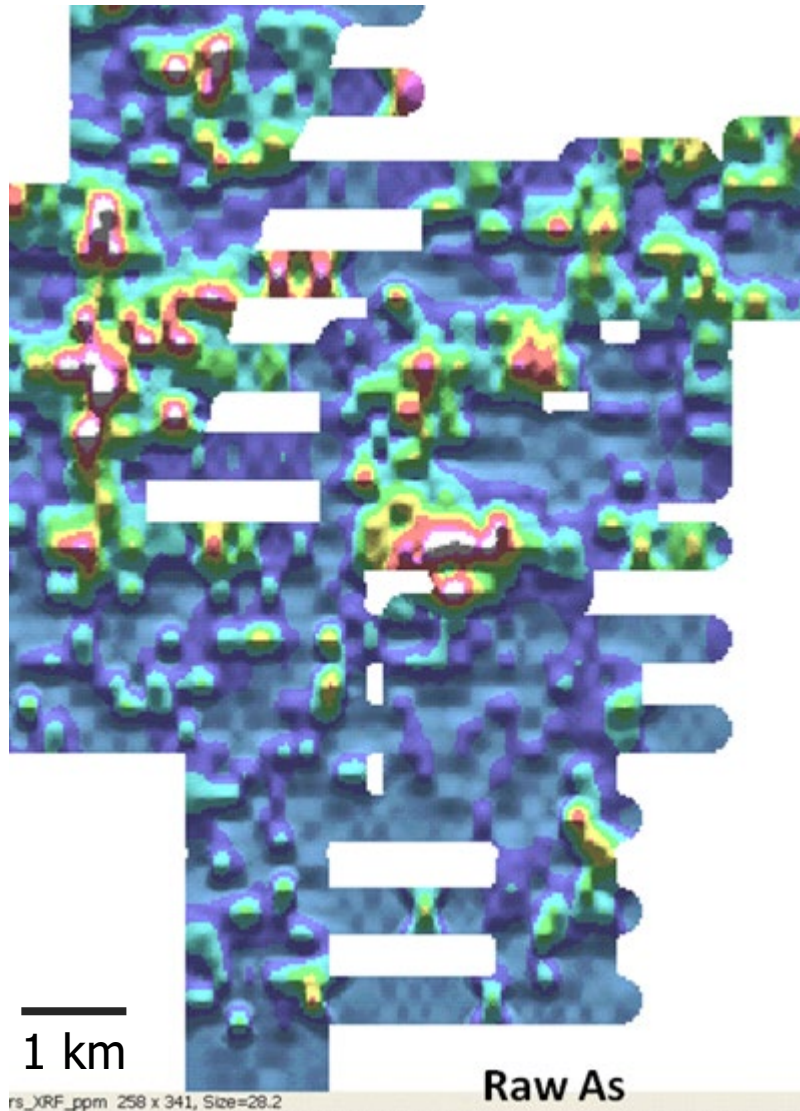
- A **positive** residual indicates that the element is present in higher-than-expected amounts, which may suggest an external input of the element beyond the influence of Fe-oxide scavenging
- A **negative** residual suggests that the element is present in lower-than-expected amounts, potentially indicating depletion due to Fe-oxide scavenging.
- Residuals can help distinguish areas where the element is being enriched or depleted independently of the Fe-oxide effect.

Fe has a strong control on the As content – it is scavenging

As a result it is necessary to normalize the As by the Fe



Control by Fe concentration in soil



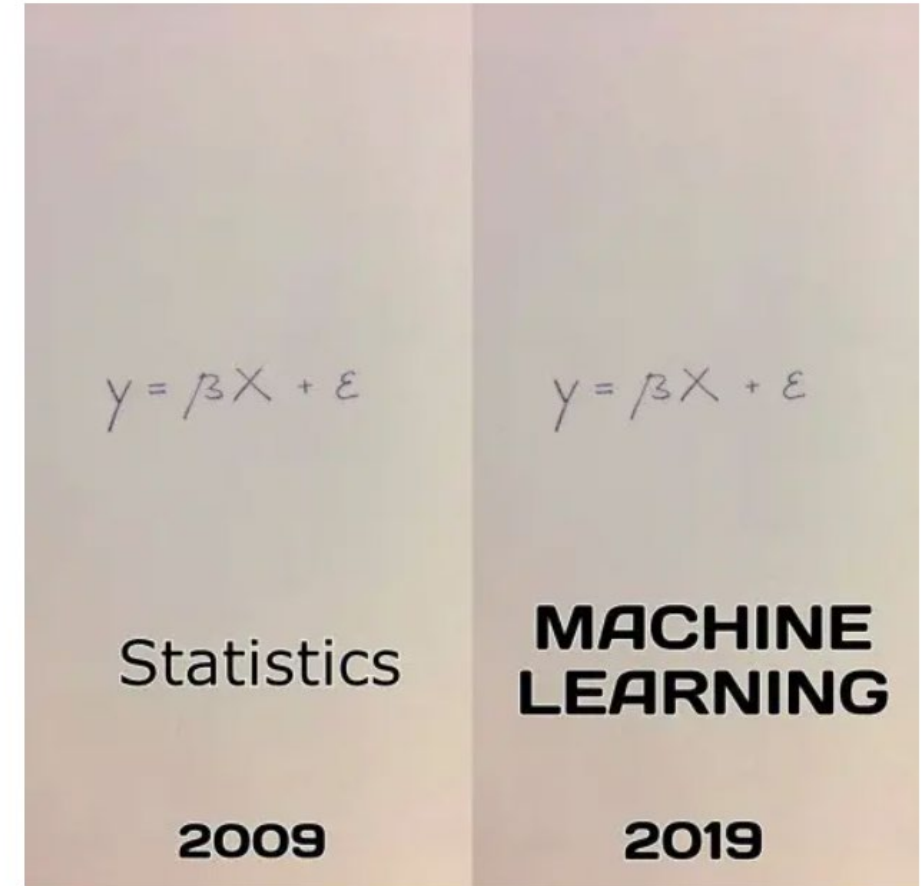
After normalizing for Fe, the pattern for As (residual values) becomes much clearer (Benn, 2014).

The Code-Along



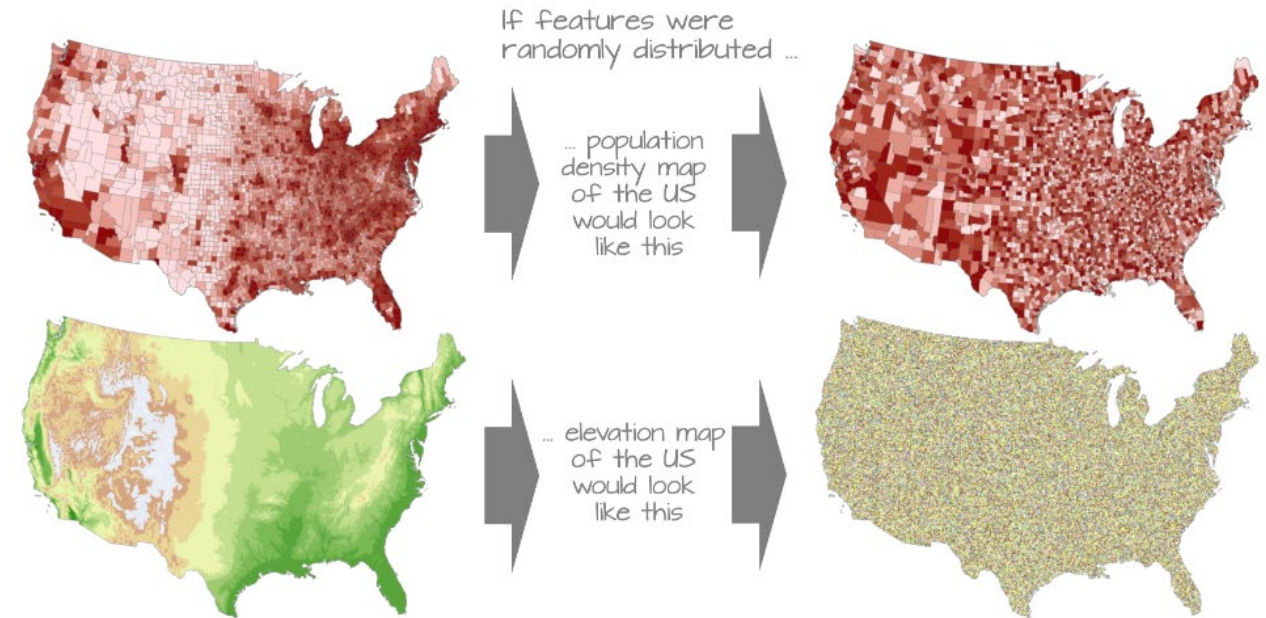
Spatial autocorrelation

...excuse me, what now?



Spatial autocorrelation, defined

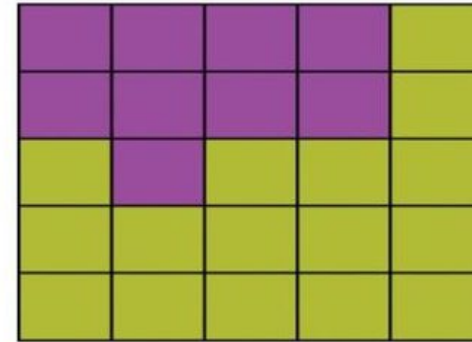
- Spatial autocorrelation refers to the concept in geography and spatial statistics where the spatial distribution of a variable demonstrates correlation based on its location.
- In simpler terms, it examines whether the occurrence of an attribute at one location is similar or different to the occurrence of the same attribute at nearby locations.
- In geochemistry, spatial autocorrelation can help in understanding the spatial distribution of certain elements in soil or rock samples.



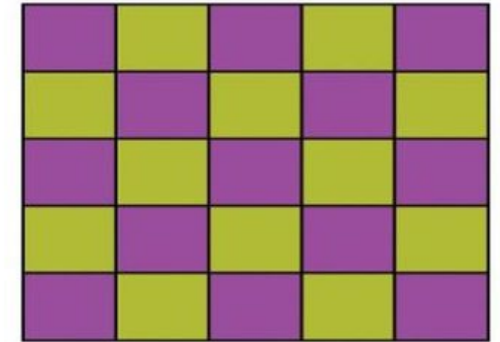
Spatial autocorrelation, defined

- There are two types of spatial autocorrelation:
 - **Positive Spatial Autocorrelation:** This occurs when similar values cluster together in space. For example, areas with high mineral content may be located near other high mineral content areas, and vice versa for low values.
 - **Negative Spatial Autocorrelation:** This happens when dissimilar values are located near each other. For example, high values may be surrounded by low values, creating a sort of checkerboard pattern.
- Spatial autocorrelation can be quantified using measures like Moran's I, Geary's C, or Ripley's K-function to help identify patterns and make predictions based on spatial relationships.

POSITIVE : Pattern of Similarity



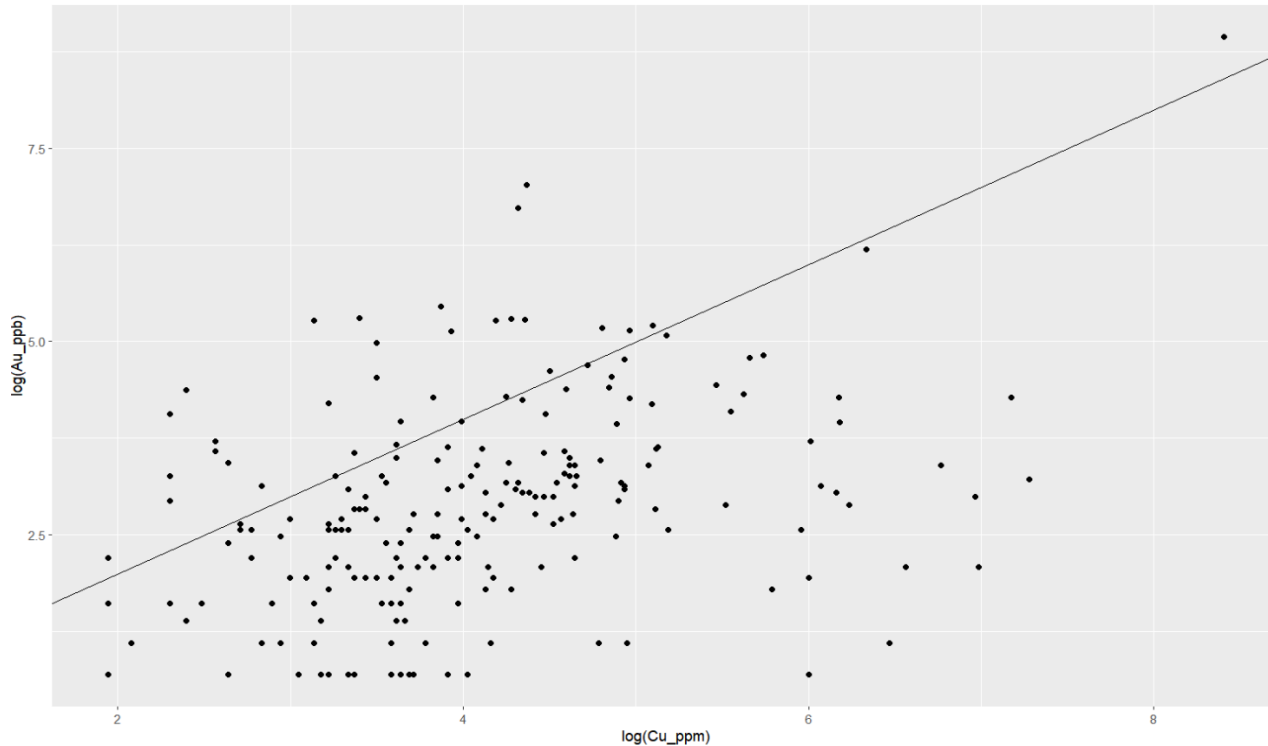
NEGATIVE : Pattern of Dissimilarity



[Campbell and Shin, 2011](#)

Is there correlation or is this effect due to spatial autocorrelation?

Are Cu (ppm) and Au (ppb) positively correlated or is this effect due to spatial autocorrelation?

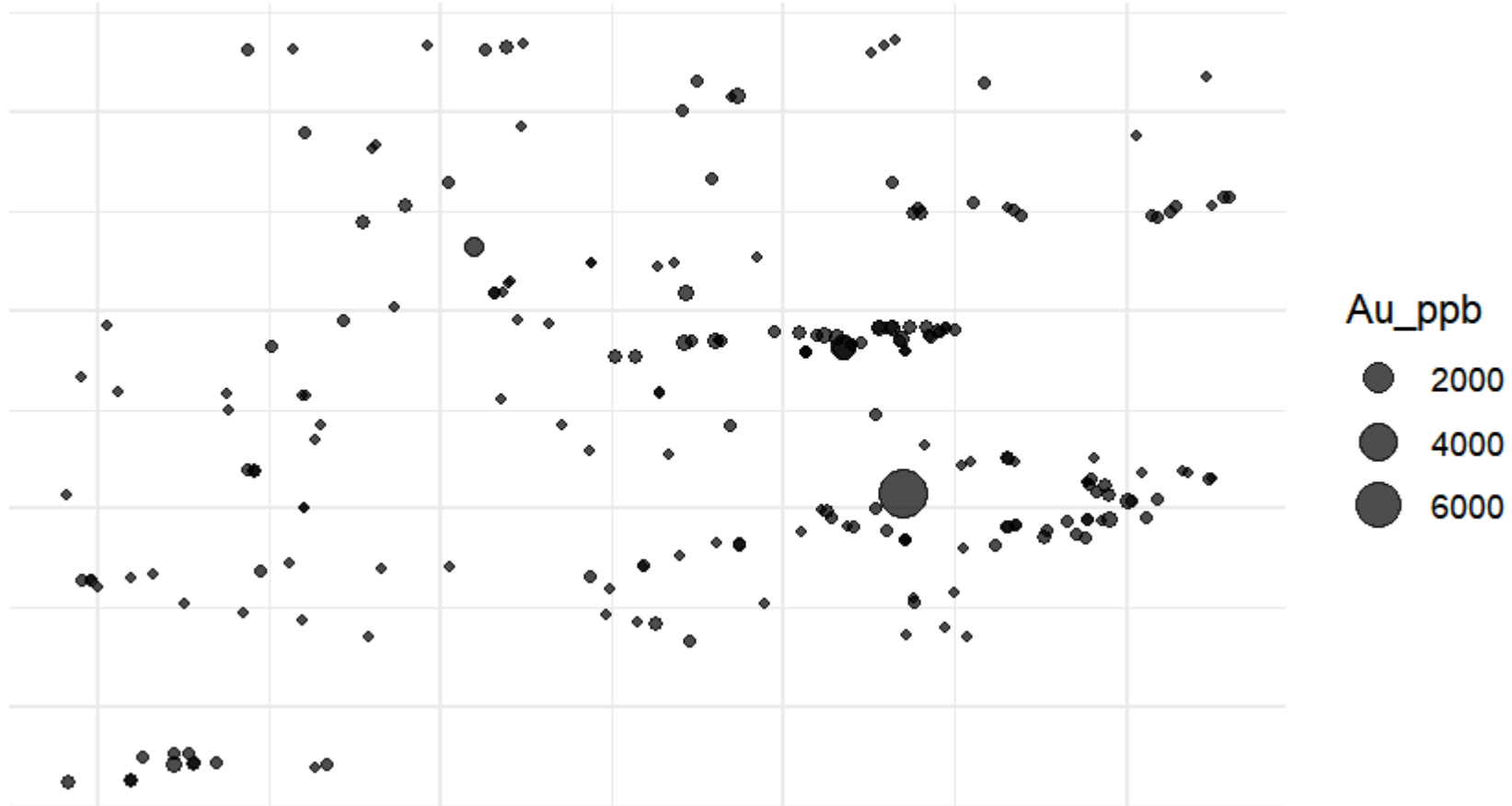


Plan:

1. **Load necessary packages.**
2. **Visualize the data** to identify potential spatial patterns.
3. **Prepare spatial data** and convert it into a spatial object.
4. **Create a spatial weights matrix** using a distance-based or neighbor-based approach.
5. **Calculate Moran's I** to assess global spatial autocorrelation.
6. **Use Local Moran's I** to explore localized clusters or outliers.

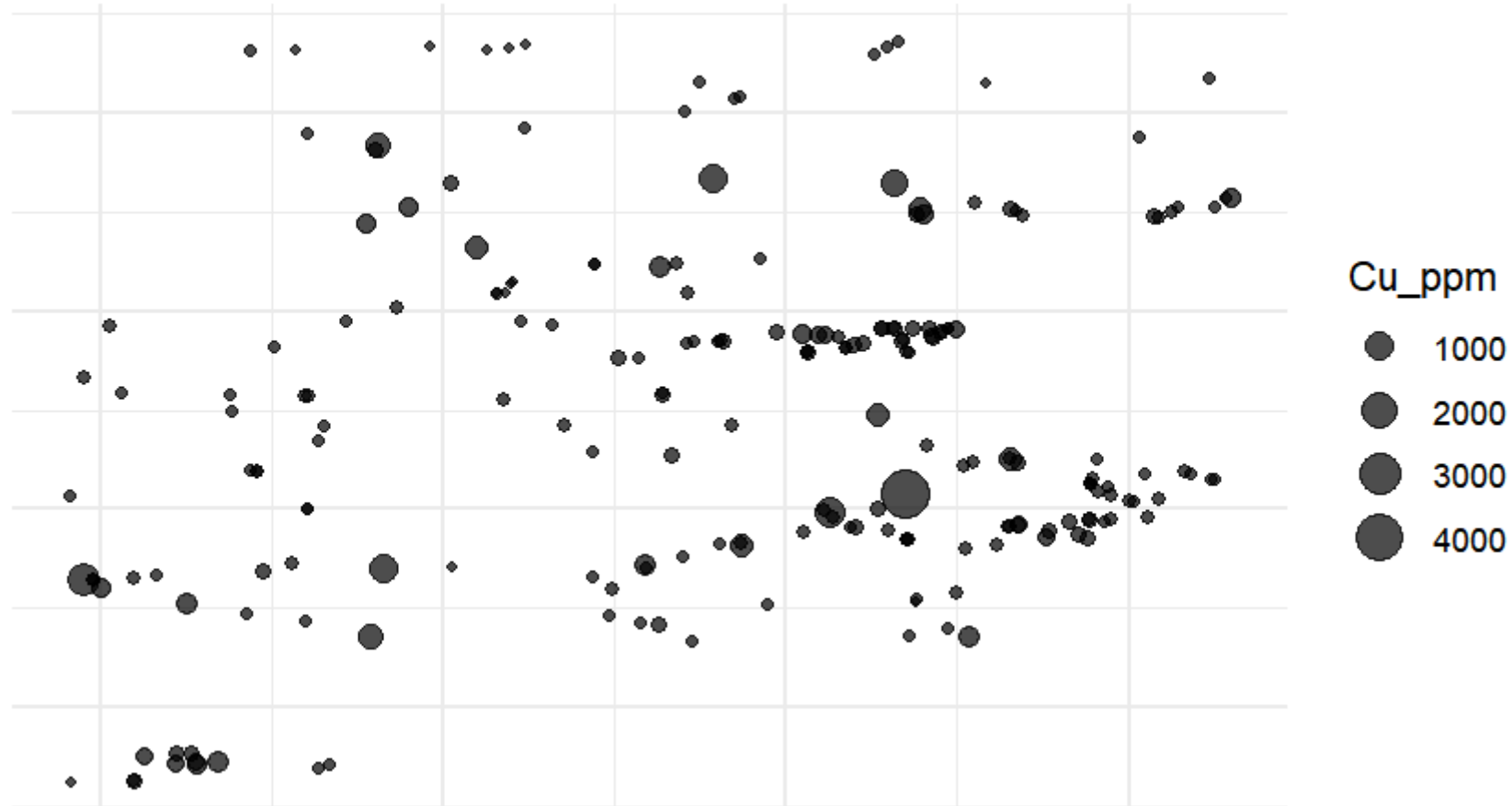
Visualizing spatial distribution: Au

Spatial Distribution of Au (ppb)



Visualizing spatial distribution: Cu

Spatial Distribution of Cu (ppm)



Interpreting the results

Moran's I values:

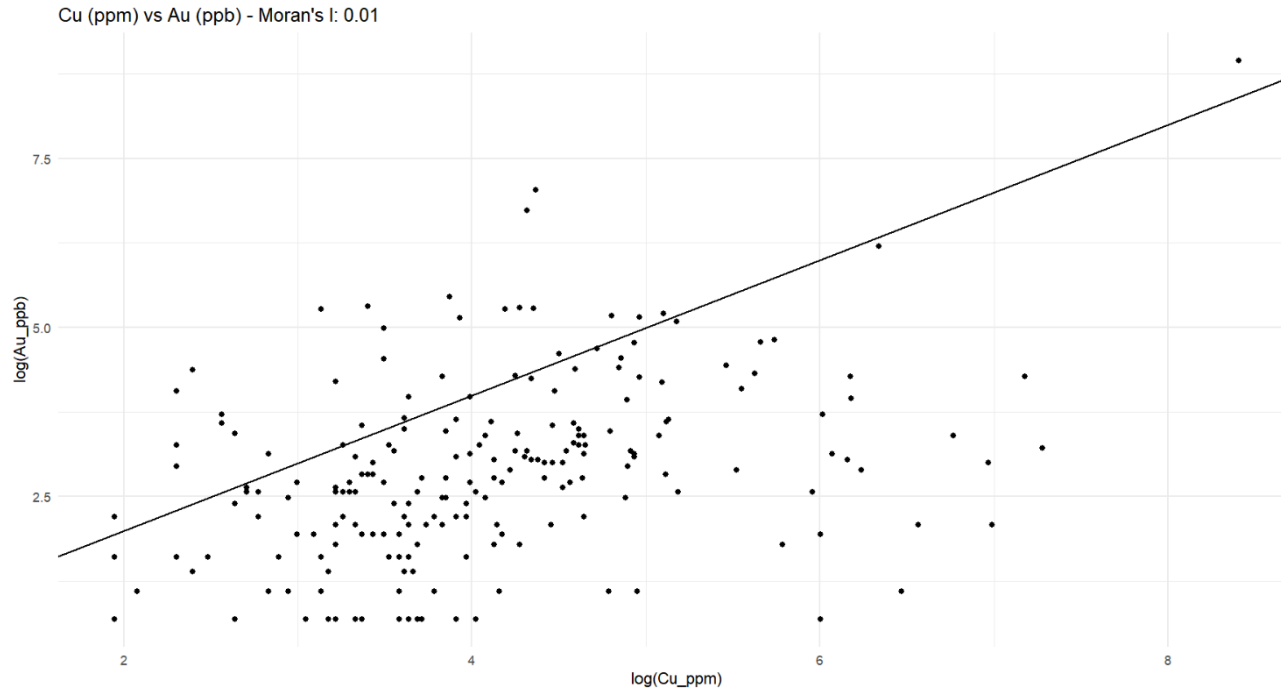
- Close to +1 indicates strong positive spatial autocorrelation, i.e., similar values cluster together in space.
- Near 0 suggests no spatial autocorrelation, i.e., the values are randomly distributed in space.
- Close to -1 indicates negative spatial autocorrelation, i.e., dissimilar values are adjacent.

Significance

- If the p-value > 0.05 , the Moran's I values are not significant and the variable does not show significant spatial autocorrelation.

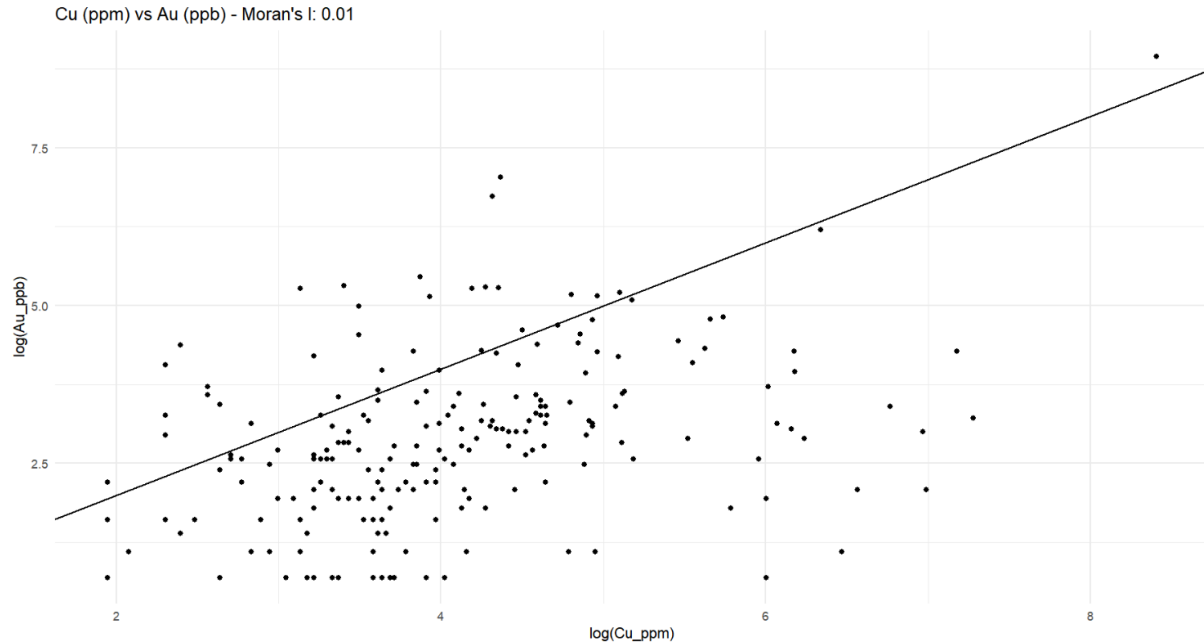
Variable	Observed	p-value
Moran_Au	0.022	0.016
Moran_Cu	0.005	0.728

How do we interpret this plot?



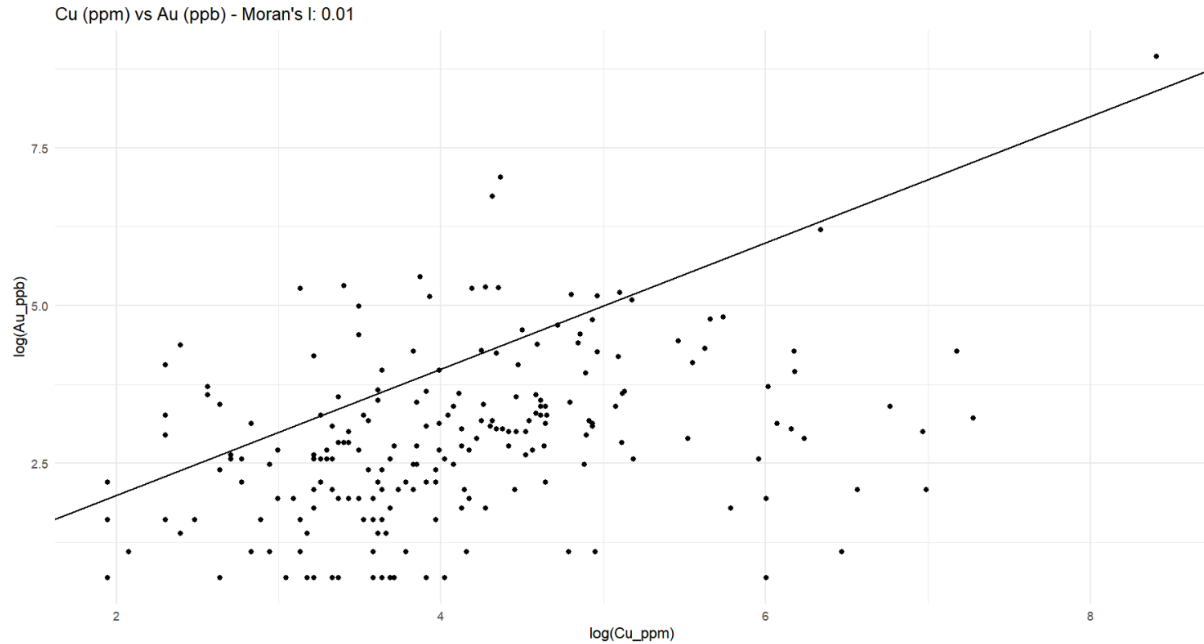
- Looking at the bivariate plot, it shows that while there is a general trend of increasing gold with increasing copper, the relationship is not perfectly linear, as there is significant scatter around the regression line.

How do we interpret this plot & Moran's I?



- Moran's $I = 0.01$ suggests very weak positive spatial autocorrelation. The Moran's I value ranges from -1 (indicating strong negative spatial autocorrelation) to +1 (indicating strong positive spatial autocorrelation).
- A value near 0, such as 0.01, implies that the spatial arrangement of Cu and Au values is close to random.

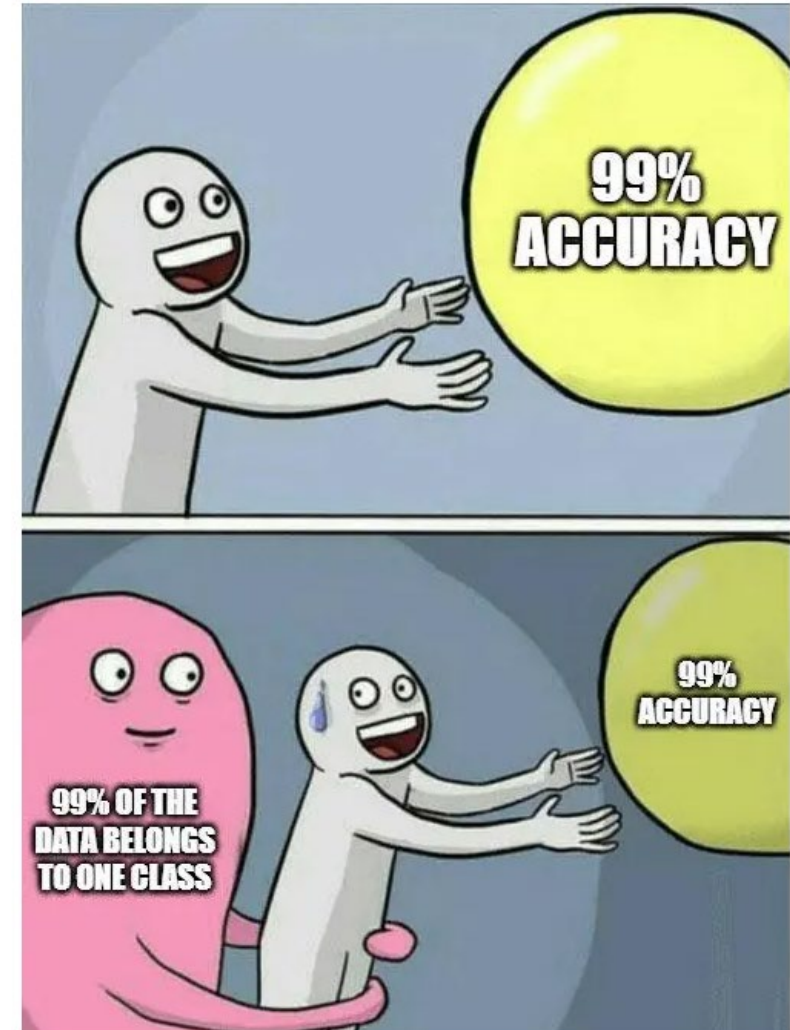
How do we interpret this plot & Moran's I?



- Here, any spatial clustering or pattern in the distribution of Cu and Au is negligible.
- The correlation observed between them is likely due to their inherent relationship between their concentrations rather than spatial autocorrelation.

Using data science to identify whether there are Fe- and/or Mn-oxides affecting Au concentration

Data science for the seemingly mundane



Our code-along



- In the code, we use **SHAP values** (Shapley Additive exPlanations) in conjunction with an **XGBoost model** to explore and quantify the influence of certain elements, specifically **Fe** and **Mn** (think their oxides), on the concentrations of **Cu** (copper) or **Au** (gold).
- This approach allows us to investigate whether there is **scavenging** of Au (or any other element) by Fe or Mn oxides in the geochemical dataset.
- By using this SHAP-based approach with XGBoost, we are able to quantify and visualize the contribution of Fe and Mn to the prediction of Au concentrations, which helps determine whether **Fe- or Mn-oxide scavenging** is playing a role.
- Look for strong positive SHAP values or clear dependencies of Au on Fe/Mn as evidence of such scavenging.
- In the following slides we are going to break down the key steps and how they relate to detecting Fe- or Mn-oxide scavenging of Au.

Data preparation: subsetting

- First, we prepare a subset of the geochemical data, selecting the variables that we want to include in the model.
- Elements like **Fe**, **Mn**, **Cu**, and **Au**, as well as other potentially relevant elements (V, Zn, Ba, Cr, etc.).

```
pr_geochem_subset <- pr_geochem %>% select(Fe_pct, Mn_ppm, Au_ppb,  
V_ppm, Zn_ppm, Ba_ppm, Cr_ppm, Cu_ppm, As_ppm, Pb_ppm, Na_pct,  
Ti_pct, K_pct, Ca_pct, Mg_pct)
```

Data preparation: preprocessing using bestNormalize

- **Handles Skewed Data:** Geochemical data (e.g., element concentrations) are often skewed, and normalization helps make distributions more symmetric for analysis.
- **Improves Model Performance:** Many statistical models assume normality. Normalization enhances model accuracy by making data more suitable for these algorithms.
- **Automates Best Method Selection:** Automatically chooses the best normalization technique (e.g., log, Yeo-Johnson, Box-Cox) based on each variable's distribution.
- **Mitigates Outliers:** Compresses the scale of extreme values, reducing the impact of outliers.
- **Ensures Comparability:** Makes variables with different units or scales comparable in analysis.
- **Meets Statistical Assumptions:** Helps fulfill the normality and homogeneity requirements for statistical tests and models.

```
## - Best normalize to help with skewed data, always set all_orderNorm = FALSE  
pr_geochem_subset_norm <- bestNormalize(as.matrix(pr_geochem_subset),  
                                         allow_orderNorm = FALSE)  
  
pr_norm_df <- data.frame(pr_geochem_subset_norm$x.t)
```

Modeling to predict element concentrations

- We need to train an XGBoost model to predict the log-transformed Au concentrations based on the other geochemical variables in the subset.
- **This model allows us to quantify the contribution of different variables, like Fe and Mn, to the prediction of Au concentrations.**
- Later, we can replace Au_ppb with other elements in the label to analyze the influence of Fe and Mn on them instead of Au.

```
xgb_model <- xgboost(data = pr_norm_df %>%  
  select(-Au_ppb) %>%  
  as.matrix(), label = log(pr_norm_df$Au_ppb),  
  nrounds = 200,  
  early_stopping_rounds = 10  
)
```

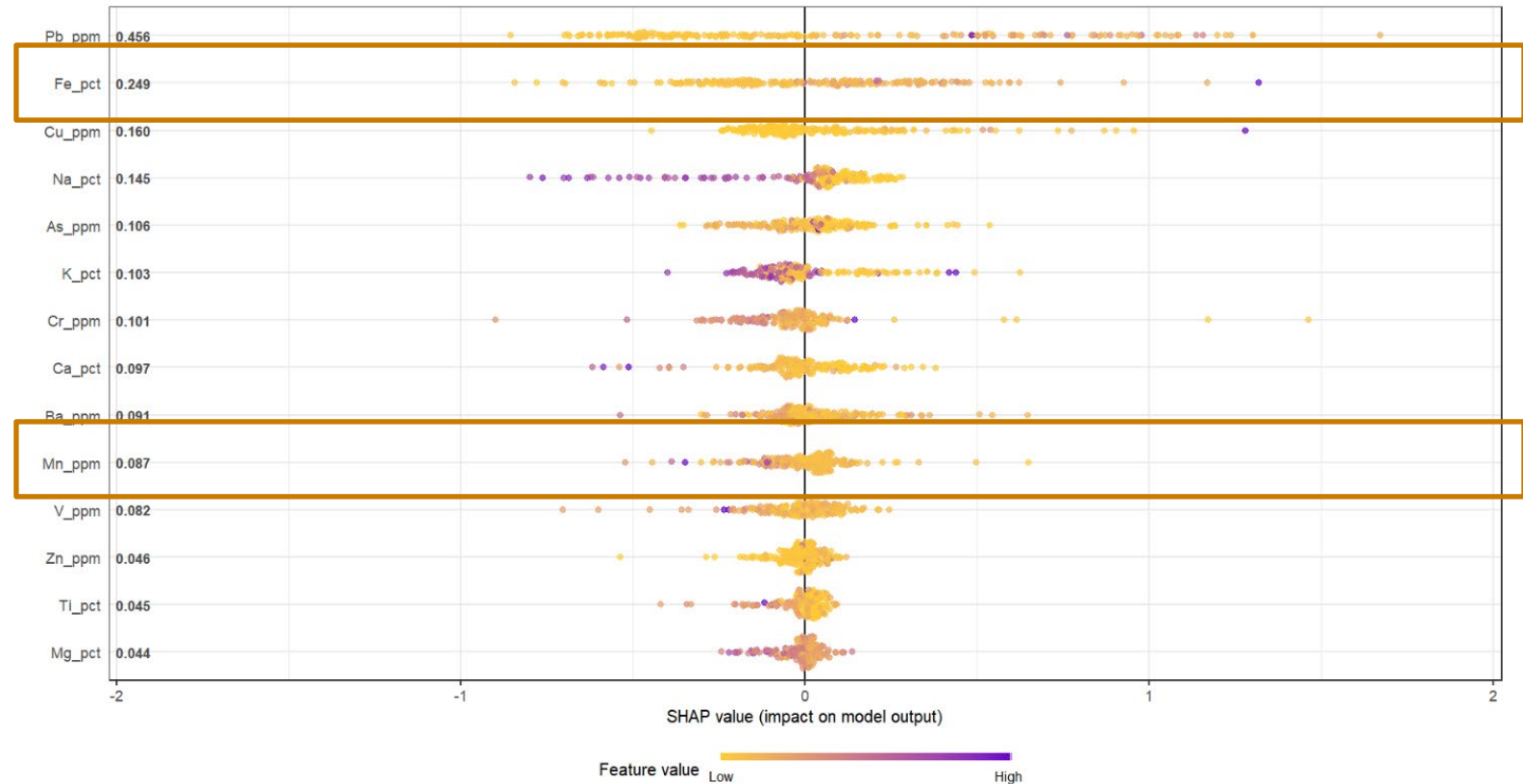
Understanding partial dependence with SHAP values calculation

- SHAP values are calculated to explain the contribution of each feature (e.g., Fe_pct, Mn_ppm) to the predictions made by the XGBoost model.
- SHAP values provide insights into how much each variable contributes to predicting $\log(\text{Au_ppb})$.
- This helps in understanding the partial dependence of Au on Fe, Mn, and other elements.

```
shap_values <- shap.values(xgb_model = xgb_model,  
                           X_train = pr_norm_df %>%  
                           select(-Au_ppb) %>%  
                           as.matrix()  
                           )  
  
shap_long <- shap.prep(xgb_model = xgb_model,  
                       X_train = pr_norm_df %>%  
                       select(-Au_ppb) %>%  
                       as.matrix()  
                       )
```

SHAP – Impact on model output

- The SHAP summary plot shows how each feature contributes to the prediction across all samples in the dataset.
- You can observe if Fe_pct or Mn_ppm are important features, which would indicate that they have a significant effect on Au concentrations.



```
shap_values <- shap.values(xgb_model = xgb_model1,
  X_train = pr_norm_df %>%
    select(-Au_ppb) %>%
    as.matrix()
)

shap_long <- shap.prep(xgb_model1 = xgb_model1,
  X_train = pr_norm_df %>%
    select(-Au_ppb) %>%
    as.matrix()
)

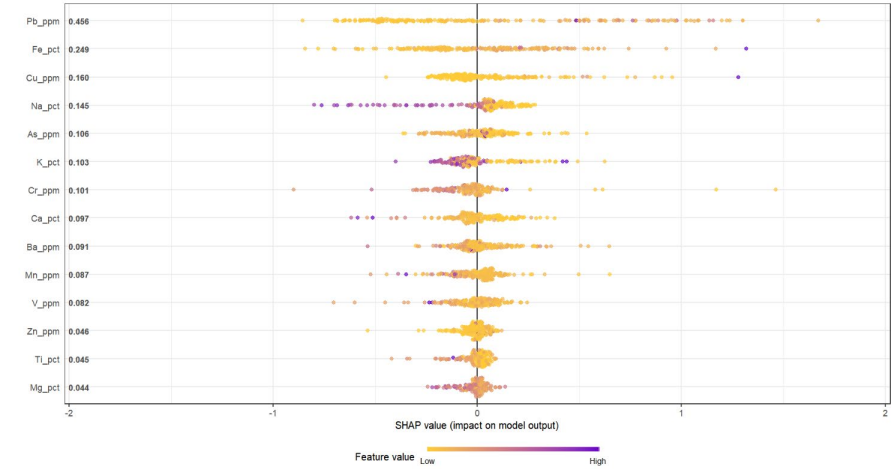
# - shap plot looking at what elements are affecting Cu/Au and how much
shap.plot.summary(shap_long)
```

SHAP – Impact on model output

- **Key Findings:**

- **Fe:** has a significant positive impact on predicting gold (Au) concentrations, supporting the hypothesis of Fe-oxide scavenging for Au. Higher Fe values correlate with higher Au predictions.
- **Mn:** shows a weaker influence, suggesting that Mn-oxide scavenging may occur but is less impactful compared to Fe.
- **Pb:** is the most influential feature, possibly indicating a geochemical association between lead and gold, suggesting shared mineralization processes.
- Other elements like Cu, Na, and As have moderate impacts but contribute less to Au predictions.

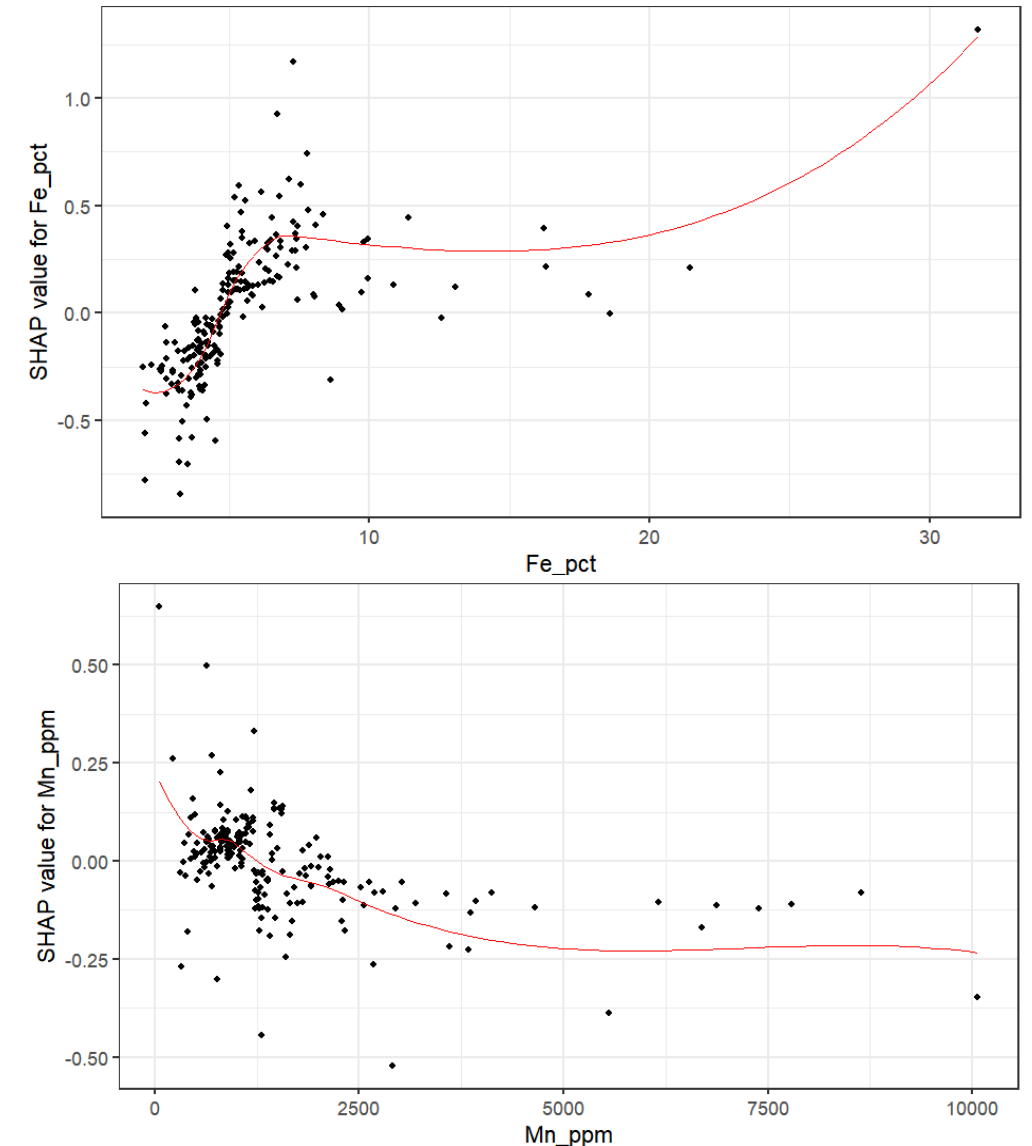
- **Conclusion:** the strong influence of Fe_pct highlights Fe-oxide scavenging as a key process controlling gold concentration in the dataset, while Mn-oxide scavenging plays a less important role.



Partial dependence plots for Fe and Mn

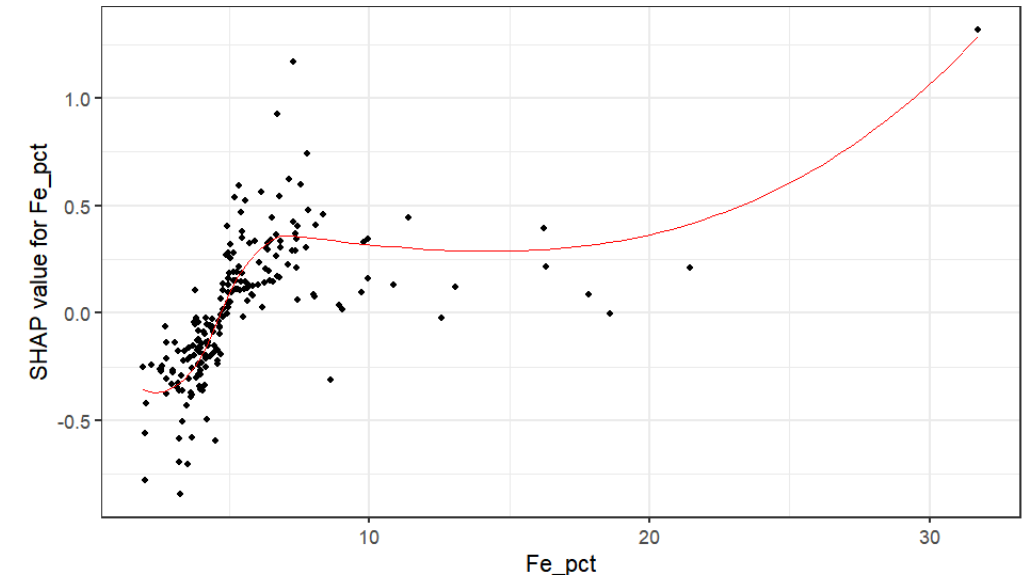
- These plots are particularly important for assessing oxide scavenging.
- If there is a clear dependence of Au on Fe or Mn, this could suggest that these oxides are scavenging or associating with the element (Au).
- The shape and spread of the dependence plot shows how changes in Fe or Mn concentration affect the predicted Au concentration.
- In these plots we are looking for trends where higher Fe or Mn values correspond to higher Au concentrations. This would be evidence of Fe or Mn-oxide scavenging, as these oxides tend to adsorb trace metals, leading to higher concentrations where Fe or Mn is abundant.

```
shap.plot.dependence(data_long = shap_long, x="Fe_pct")  
shap.plot.dependence(data_long = shap_long, x="Mn_ppm")
```



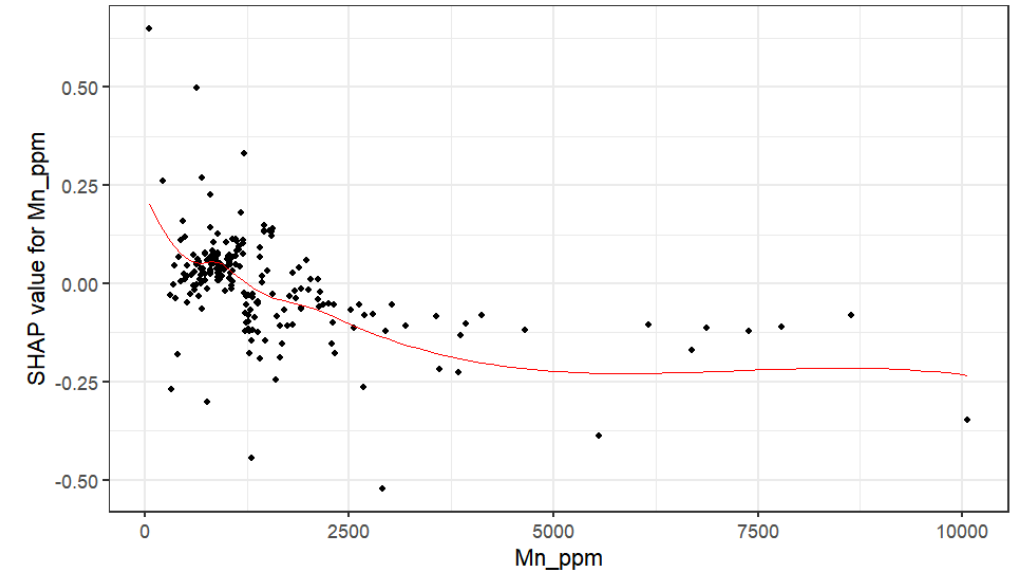
Partial dependence plot interpretation

- **Fe_pct**
 - **Positive SHAP Values:** SHAP values increase as Fe_pct increases, especially for Fe_pct values above 10%. This indicates that higher Fe concentrations have a **positive contribution** to the prediction of higher gold concentrations.
 - **Curved Trend:** The relationship is nonlinear, with a marked increase in the SHAP value at **Fe_pct > 20%**, suggesting that **Fe-oxide scavenging** becomes more pronounced at higher Fe levels. This supports the hypothesis that Fe-oxides play a key role in concentrating gold.



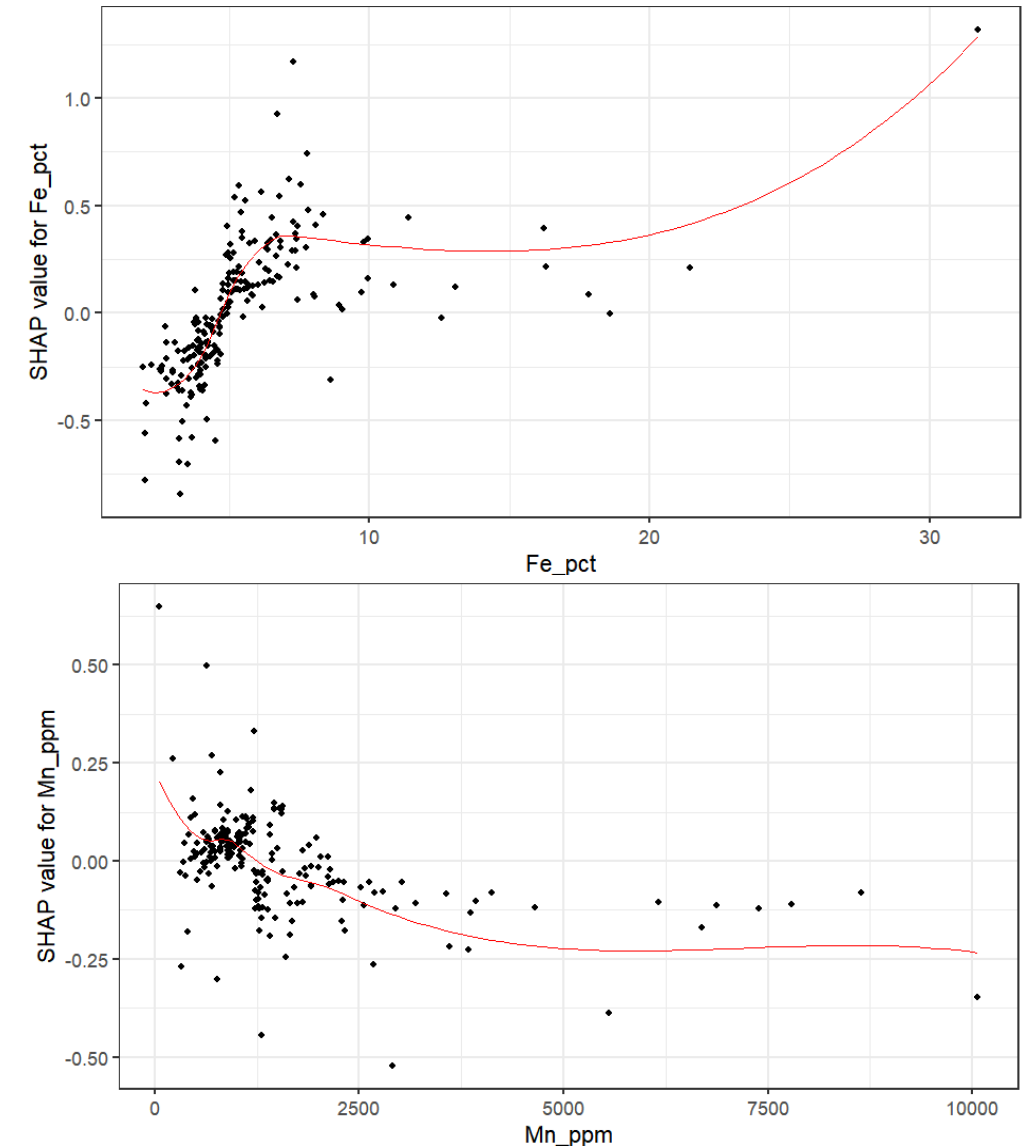
Partial dependence plot interpretation

- **Mn_ppm**
 - **Weaker Influence:** the SHAP values for Mn_ppm decrease slightly as Mn concentration increases, especially for concentrations above **2000 ppm**. This suggests that higher Mn concentrations may have a **small negative or neutral impact** on Au concentrations.
 - **Negative Contribution:** the generally lower SHAP values for Mn indicate that Mn-oxide scavenging has **less of an effect** on Au than Fe-oxide scavenging.



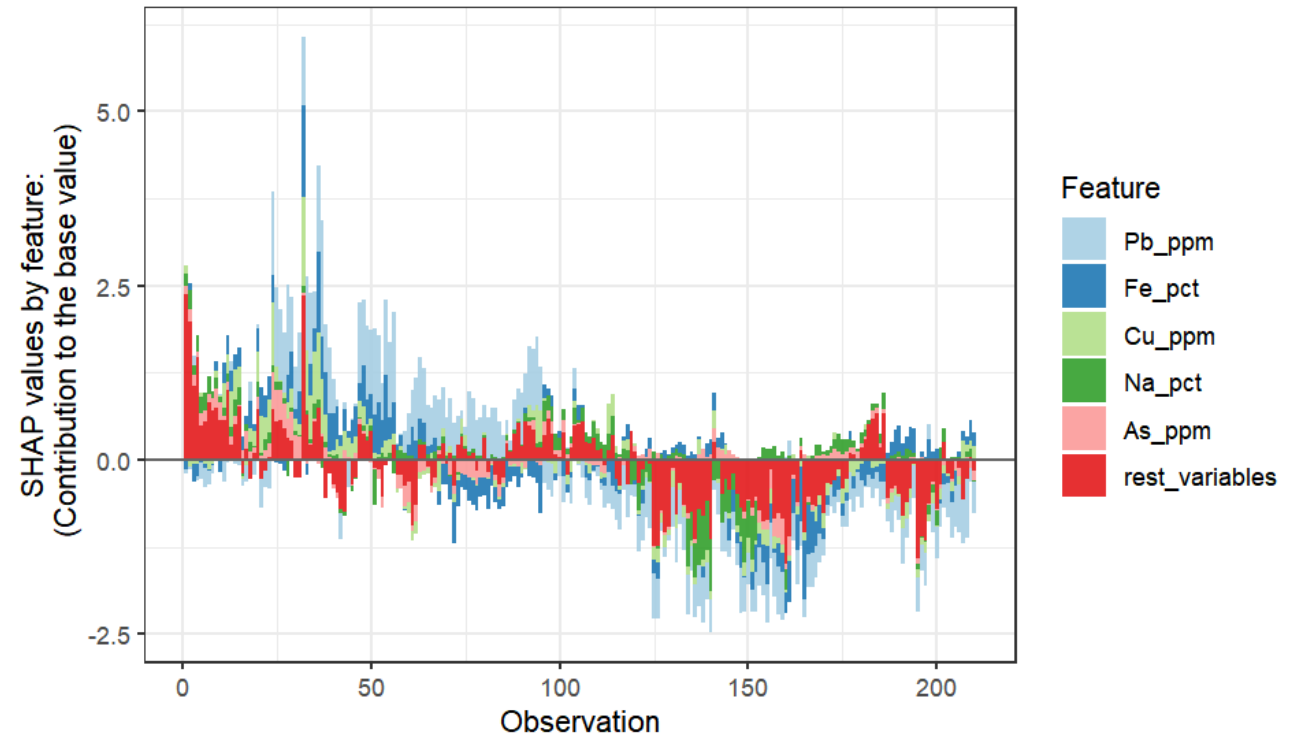
Partial dependence plot interpretation: summary

- The **Fe_pct** plot shows a clear, positive relationship between Fe concentration and Au, reinforcing the idea of **Fe-oxide scavenging** for Au.
- The **Mn_ppm** plot shows a more neutral to negative relationship, suggesting that **Mn-oxide scavenging** is either less significant or potentially inhibitive to Au concentration compared to Fe oxides.
- In the context of this study, these plots visually confirm that **Fe-oxide scavenging** is likely a key process influencing Au distribution, while **Mn-oxide scavenging** plays a lesser role.



Force or waterfall plots

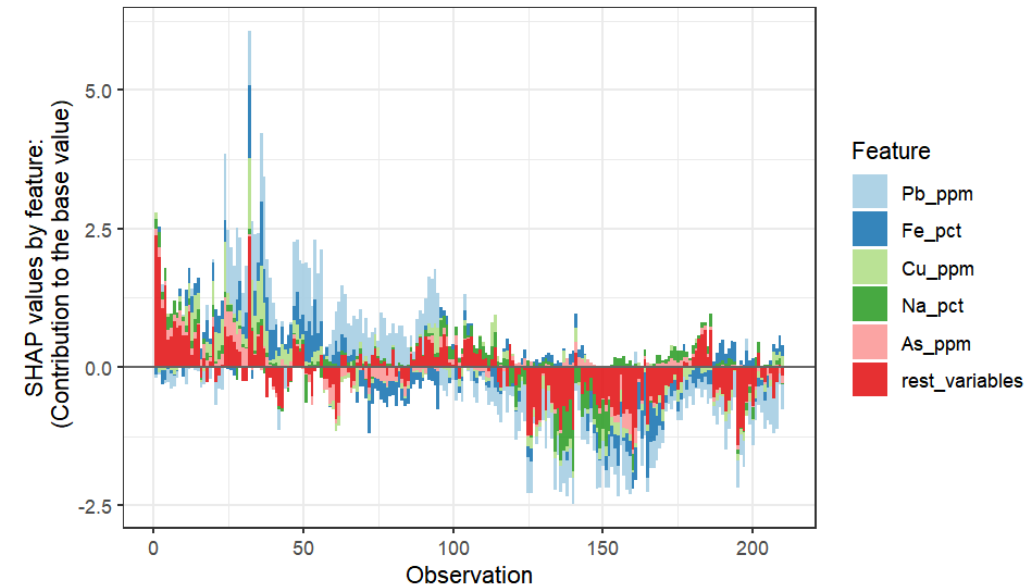
- These plots shows the SHAP values for each feature, illustrating how each feature contributes to the prediction of Au for individual samples.
- They visualize the cumulative contribution of Fe, Mn, and other elements to the predicted Au concentration for specific data points, providing insight into how much Fe or Mn affects the prediction.



```
## - prepping to look at the partial dependence of elements on  
values of Au/Cu  
plot_data <- shap.prep.stack.data(shap_contrib =  
shap_values$shap_score, top_n = 5, n_groups = 10)  
  
## - waterfall plot of partial dependence  
shap.plot.force_plot(plot_data, zoom_in = FALSE)
```

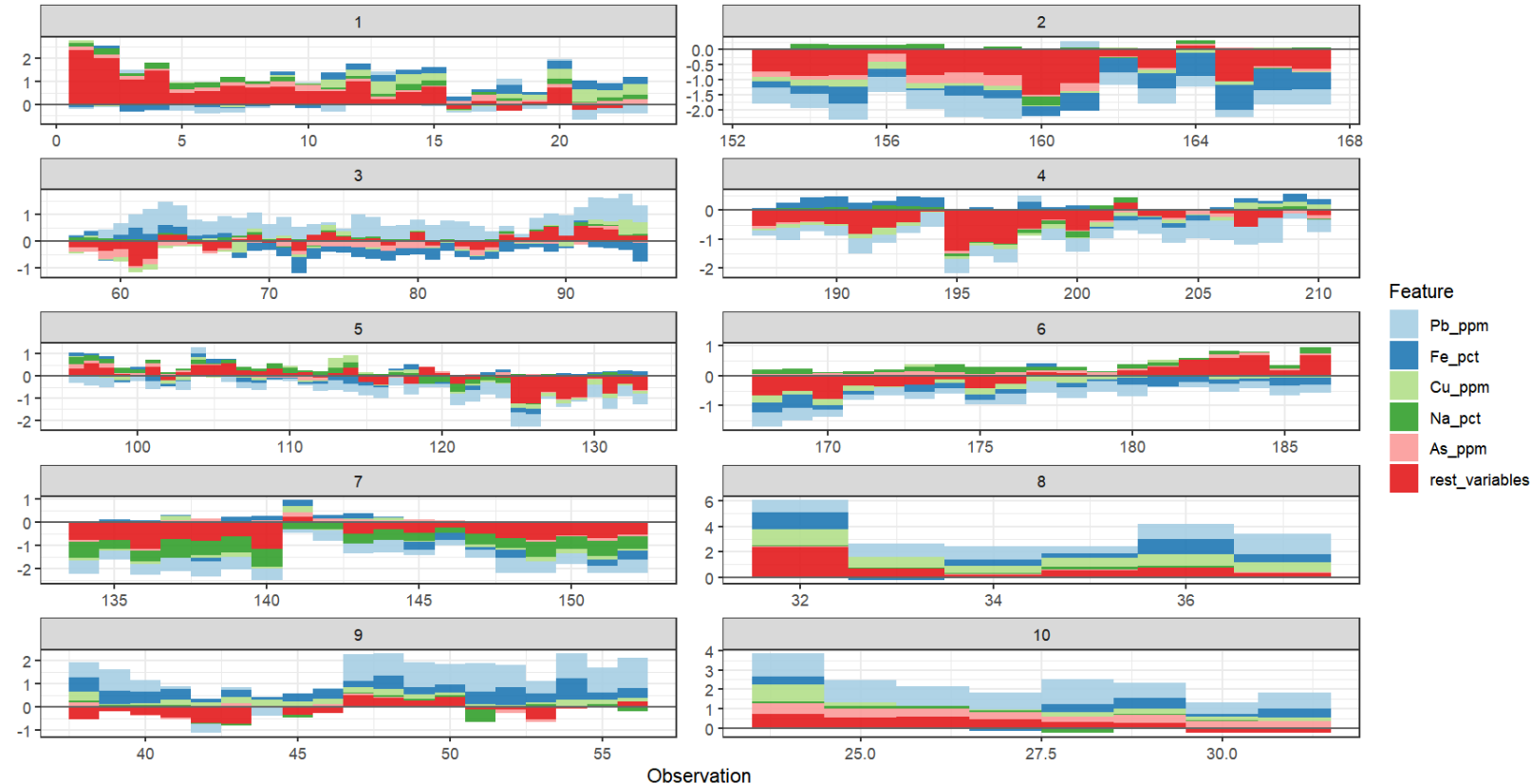
Force or waterfall plots: interpretation

- **Pb_ppm:** strong positive contribution in many samples (light blue). High lead concentrations are key drivers for predicting higher gold concentrations.
- **Fe_pct:** consistent positive impact across most samples (dark blue), suggesting Fe-oxide scavenging of gold. Some variability with negative contributions in certain samples.
- **Cu_ppm:** Positive but smaller contribution compared to Pb and Fe (green). Likely involved in mineralization (can also represent different systems / erosional level); there is a less pronounced effect.
- **Na_pct:** modest contribution (light green), generally neutral or slightly negative.
- **As_ppm:** variable influence (pink), with both positive and negative contributions, indicating its role may depend on local geochemistry.
- **Rest of Variables:** (Red) these elements contribute smaller, more varied effects.
- **Pb_ppm and Fe_pct are the most influential features,** supporting the hypothesis that Pb and Fe (especially Fe-oxide scavenging) are key factors in controlling gold distribution in this dataset.



Grouping SHAP contributions

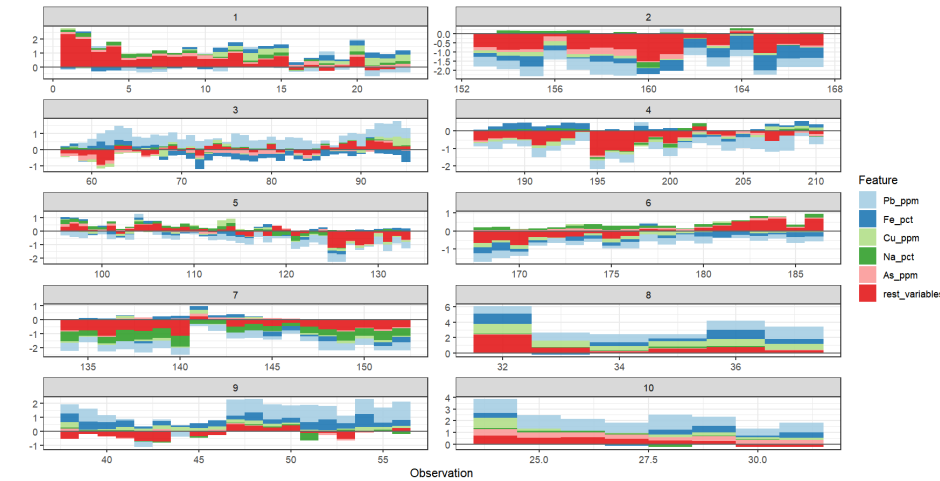
- This step groups samples based on SHAP values, showing how different combinations of features (such as high Fe or Mn concentrations) influence the predictions.
- Samples with higher Fe or Mn SHAP values would likely show a strong dependence of Au on these oxides.



```
shap.plot.force_plot_bygroup(plot_data)
```

Grouping SHAP contributions: interpretation

- **Fe_pct**: consistently contributes **positively** across multiple sample groups, reinforcing the role of **Fe-oxide scavenging** in concentrating gold.
- **Pb_ppm**: major positive contributor in certain groups, indicating a strong geochemical association between **lead** and **gold**.
- **Cu_ppm** and **Na_pct**: show smaller, more variable contributions, with copper sometimes positively impacting predictions.
- **As_ppm**: typically shows a **negative** contribution, suggesting an inverse relationship with gold concentrations in certain samples.
- **Group variability**: contributions change across different sample groups, reflecting geological heterogeneity in the dataset.
- **Conclusion**: Fe and Pb are key drivers in predicting higher gold concentrations, with **Fe-oxide scavenging** potentially being a dominant process.



Let's look at our results spatially

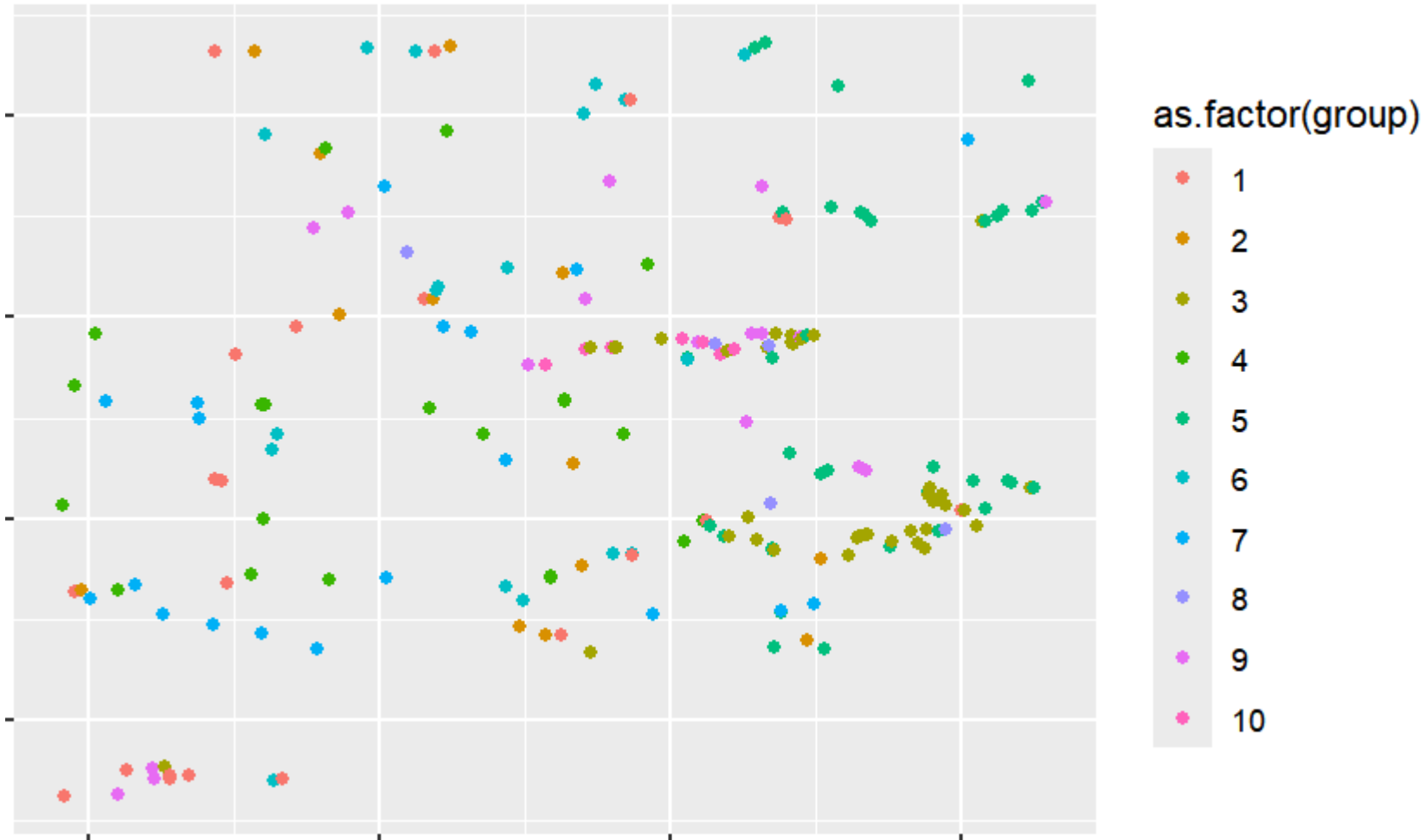
Because geochemistry is spatial.



Spatial Distribution of SHAP-Derived Groups

- **Visualization:** Plot the spatial distribution of the 10 groups derived from SHAP force plots, using geographic coordinates to understand whether there is any spatial clustering.
- **Spatial Patterns:**
 - Fe-oxide scavenging: check if groups with strong Fe-oxide associations (from SHAP force plots) are clustered in Fe-rich zones.
 - Mn-oxide associations: similarly, evaluate if groups with Mn-oxide influence correspond to Mn-rich areas.
- **Analysis:**
 - Overlay geochemical data (Fe, Mn, Au, etc.) on the map to investigate potential elemental correlations with groups.
 - Use Moran's I to quantify spatial autocorrelation of groups and identify significant clustering patterns.
- **Conclusion:** Mapping the groups spatially helps visualize elemental associations (e.g., Fe-oxide scavenging) and determine spatial clustering of mineralization processes.

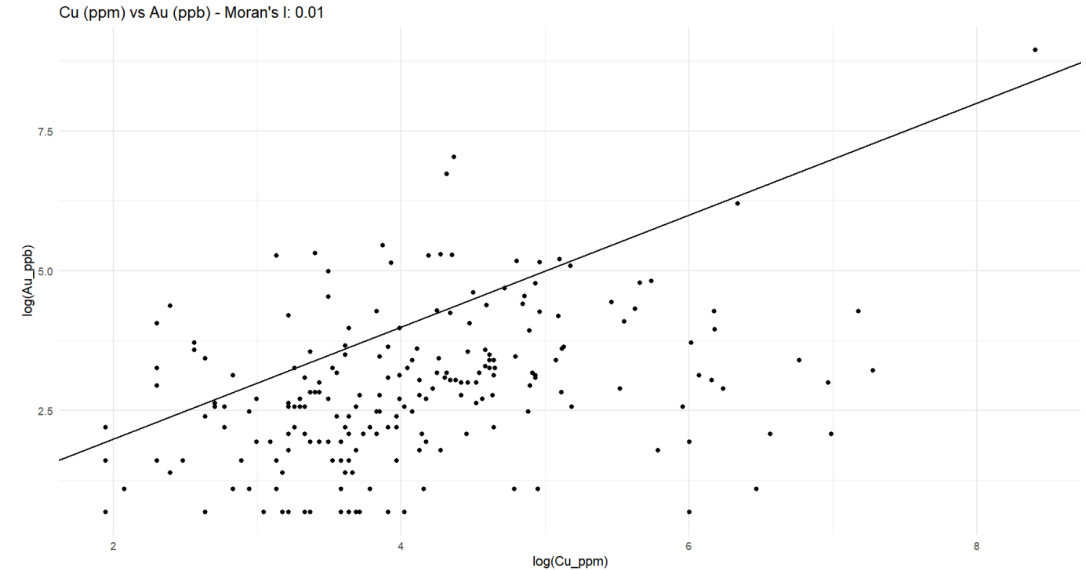
Plotting groups spatially



- By mapping these groups based on their spatial coordinates, we can understand how the groupings are distributed across the geographic area.
- This may help reveal any spatial patterns or clustering.
- Importantly, and in the context of this study, we should use this spatial map to **correlate group locations with Fe and Mn concentrations** to confirm if Fe-oxide or Mn-oxide scavenging is occurring in specific regions.

Check for spatial autocorrelation

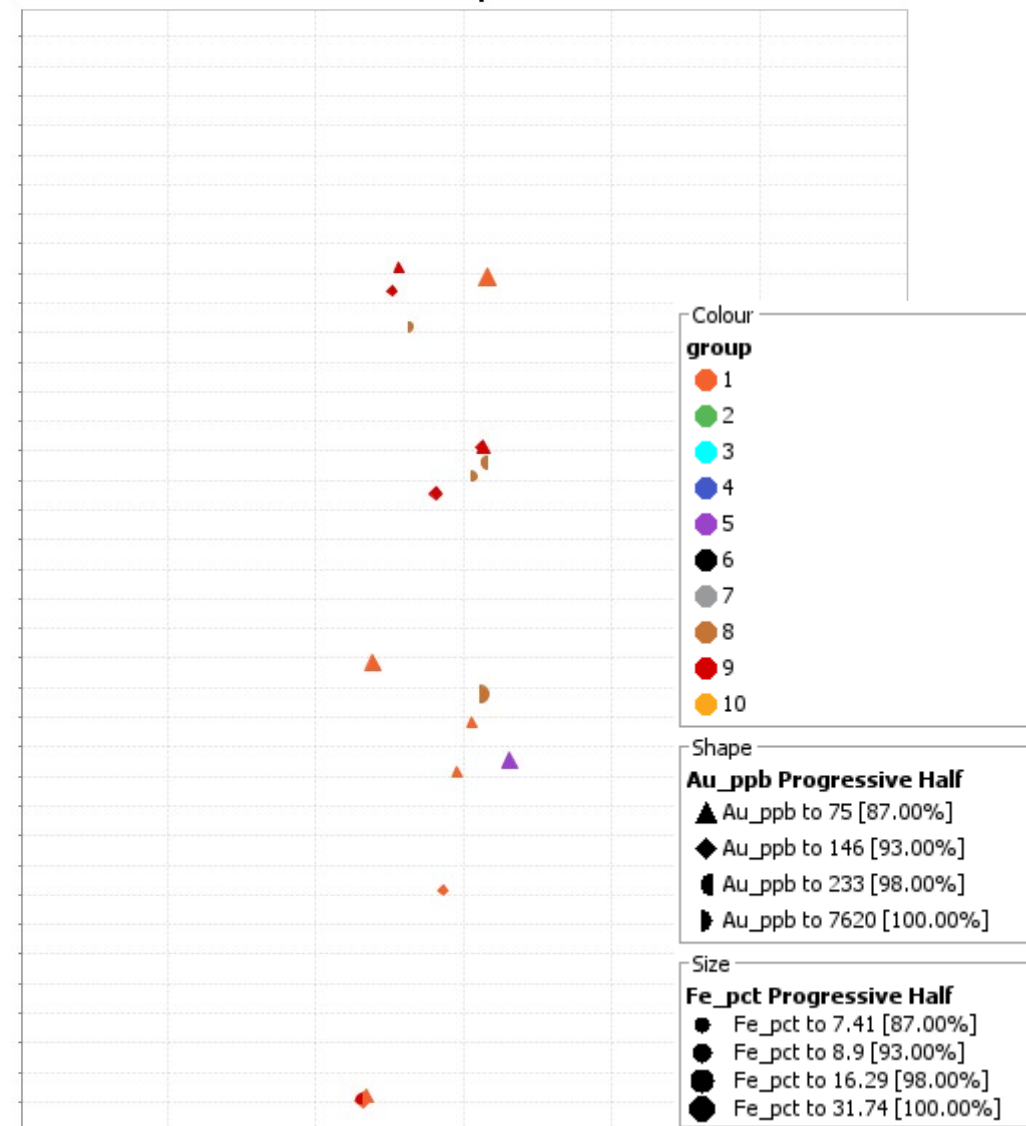
- To further understand the spatial relationship of these groups, you can calculate **Moran's I** or other spatial autocorrelation metrics on the groupings. This will provide a statistical measure of whether the groups show significant spatial autocorrelation.
- In other words, this will help quantify whether the spatial distribution of the groups is random or exhibits clustering.



The positive relationship between Cu and Au observed in the plot is likely real and not primarily influenced by spatial autocorrelation, given the low Moran's I value.

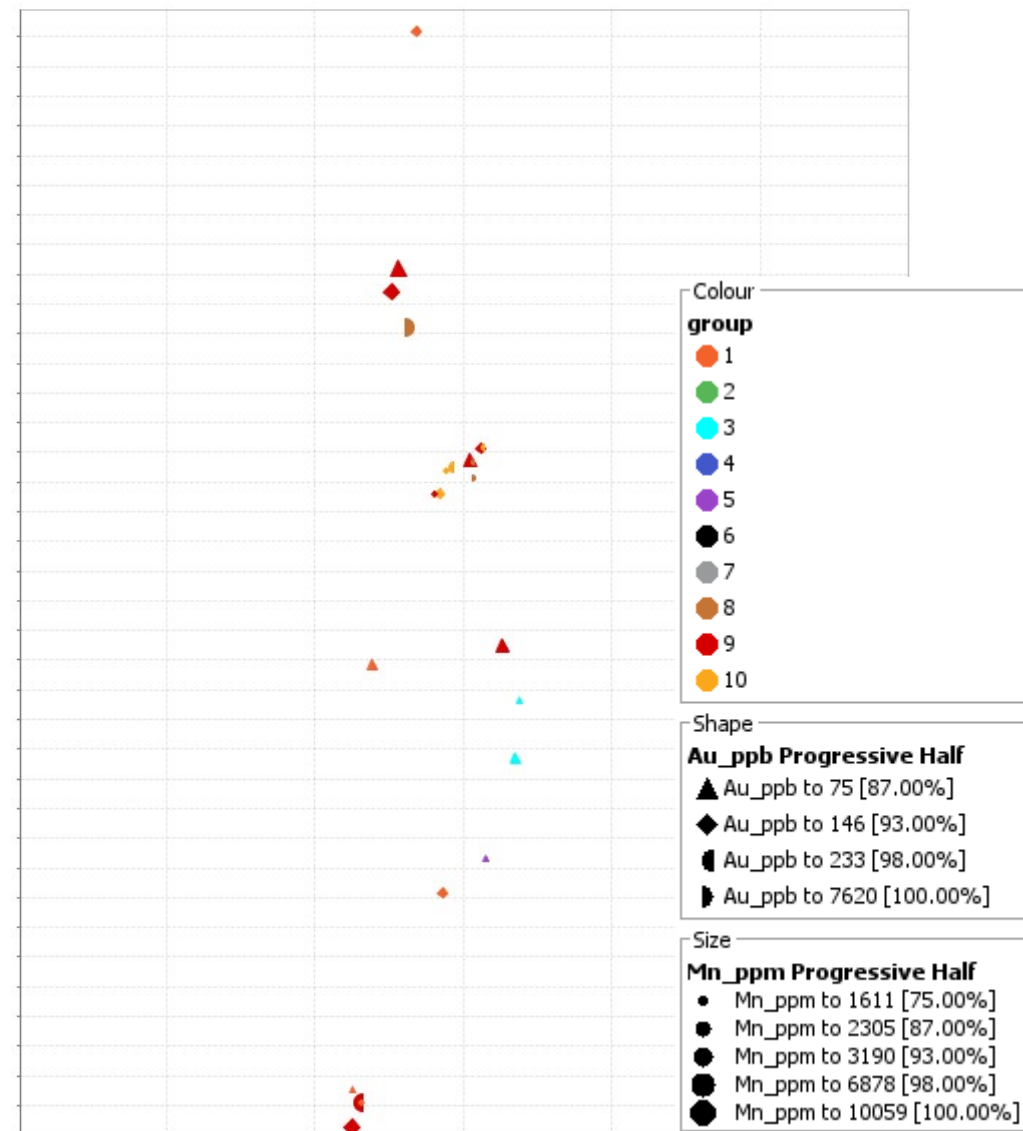
Spatial patterns: Fe

- Are clusters enriched in Fe, plotting where there is high Fe, or in our interpretation: do **Fe-rich areas** correspond with groups showing strong Fe-oxide scavenging?
- Groups 1, 5, 8, 9 with Au > global background.



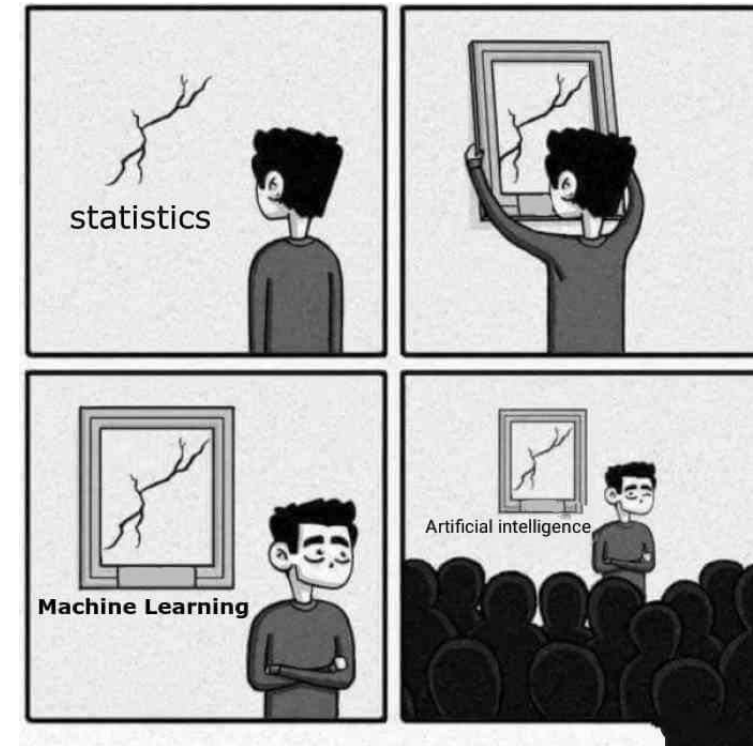
Spatial patterns: Mn

- Are clusters enriched in Mn, plotting where there is high Mn, or in our interpretation: do **Mn-rich areas** correspond with groups showing secondary Mn-oxide scavenging?
- Groups 1, 3, 5, 8, 9 with Au > background.



Let's quantify this a bit better...

Residuals



Code Along



- We need to predict the amount of **Au** that should be present **if not for Fe-oxide scavenging**. The idea is to apply a combination of spatial analysis (to account for spatial autocorrelation) and machine learning techniques like SHAP (to quantify the impact of Fe on gold concentration predictions):
 - Train a model with Fe_pct to predict gold concentrations and capture the influence of Fe-oxide scavenging.
 - Train a second model without Fe_pct to predict gold concentrations without the influence of Fe.
 - Compare the predictions to quantify how much of the gold concentration is due to Fe-oxide scavenging.
 - Analyze spatial autocorrelation in the residuals to ensure that other spatial effects are accounted for.
 - Interpret the results to make conclusions about the impact of Fe-oxide scavenging and identify any remaining patterns in the gold distribution.

Next steps

- We have already quantified how much **Fe_pct** contributed to the prediction of **Au_ppb** in your dataset by using **SHAP values** with from the XGBoost model.
 - This allows the capture of Fe contribution to Au prediction.
 - From this, we can estimate how much of the predicted Au concentration is due to **Fe-oxide scavenging**.
- **Now we need to use a model without Fe_pct:**
 - To predict the gold concentration **if Fe-oxide scavenging were not present**, we should retrain the model **without Fe_pct** as a predictor and compare the results to the original model.

Train the XGBoost model without Fe_pct

- Train a model that excludes **Fe_pct** from the predictors. This model should predict the concentration of gold **without the influence of Fe-oxide scavenging**.

```
xgb_model_without_Fe <- xgboost(data = pr_norm_df %>%  
  select(-Au_ppb, -Fe_pct) %>%  
  as.matrix(),  
  label = log(pr_norm_df$Au_ppb),  
  nrounds = 200  
)
```

Compare the predictions

- Compare the predictions from the two models:
 - The original model includes the impact of Fe-scavenging
 - The new model excludes the impact of Fe, providing an estimate of what Au concentration would look like without Fe-oxide scavenging.

```
# Predictions with Fe in the model
predicted_with_Fe <- predict(xgb_model, pr_norm_df %>%
  select(-Au_ppb) %>%
  as.matrix()
)

# Predictions without Fe in the model
predicted_without_Fe <- predict(xgb_model_without_Fe, pr_norm_df %>%
  select(-Au_ppb, -Fe_pct) %>%
  as.matrix()
)
```

Estimate the contribution of Fe-oxide scavenging

- The difference between the predicted Au concentrations from the model with **Fe_pct** and the model without **Fe_pct** will give an estimate of how much of the Au is attributable to **Fe-oxide scavenging**.

```
# Difference between predictions (contribution of Fe-oxide scavenging)  
Fe_oxide_contribution <- predicted_with_Fe - predicted_without_Fe
```

Spatial autocorrelation for residuals

- **Moran's I statistic:** 0.0116 (weak positive spatial autocorrelation)
- **p-value:** 0.1856 (not statistically significant)
- **Conclusion:** No significant spatial autocorrelation in Au concentrations, meaning the observed spatial distribution of gold is likely random.

```
# Calculate residuals for the model without Fe influence
residuals <- pr_norm_df$Au_ppb - exp(predicted_without_Fe)
moran_residuals <- moran.test(residuals, listw)
```

```
Moran I test under randomisation

data: pr_geochem$Au_ppb
weights: listw

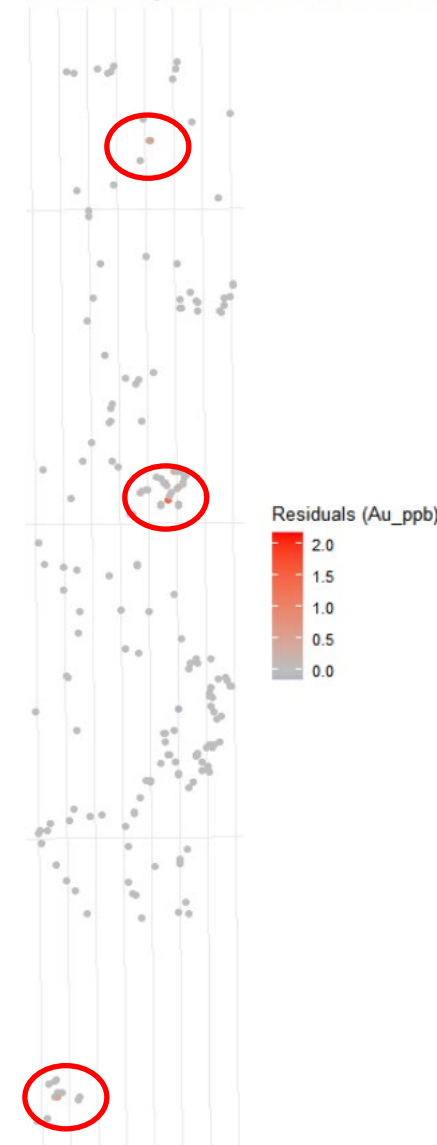
Moran I statistic standard deviate = 0.89411, p-value = 0.1856
alternative hypothesis: greater
sample estimates:
Moran I statistic      Expectation      Variance
      0.0115824332      -0.0047846890      0.0003350925
```

Residual interpretation

Residual Map: Map the residuals from the model without Fe to see which areas might have anomalous gold concentrations not explained by Fe-oxide scavenging.

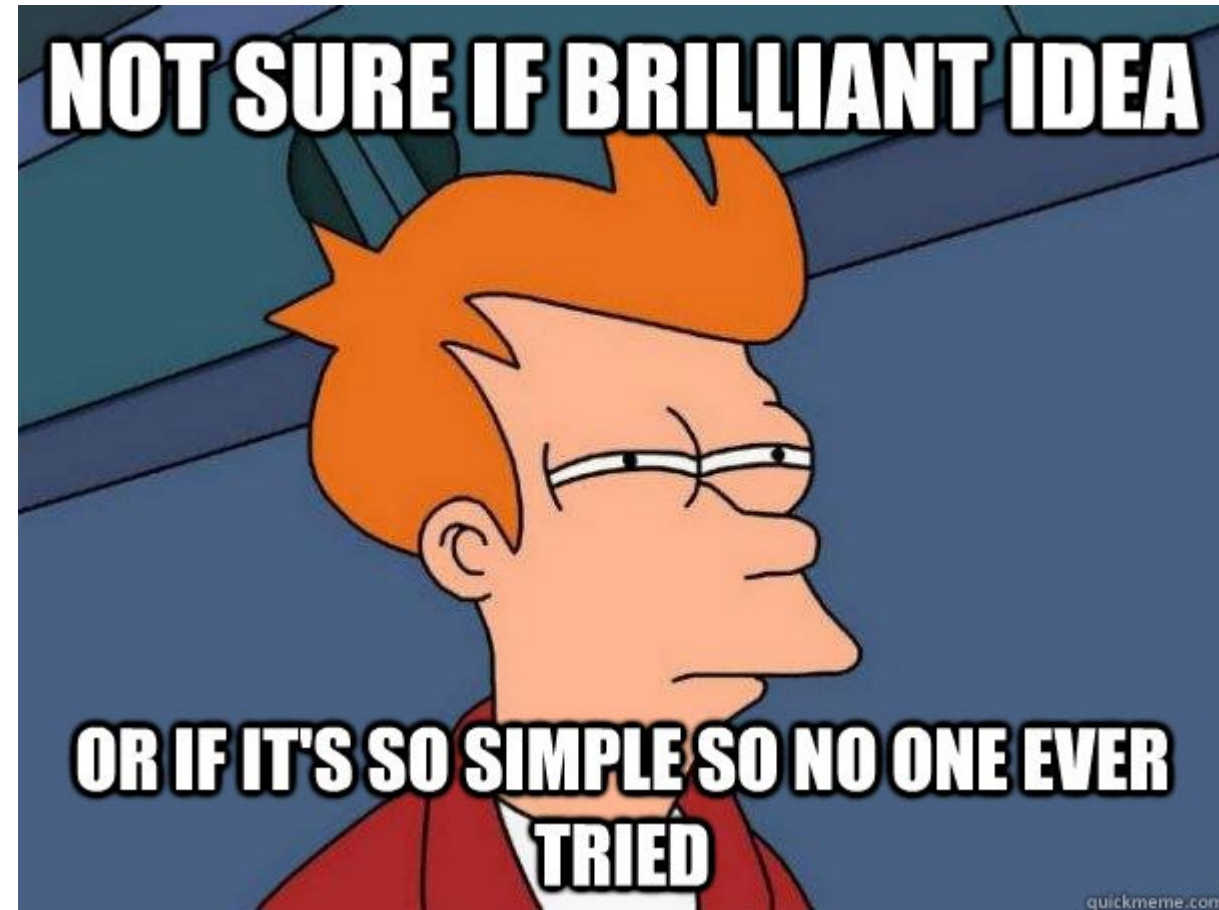
- Red areas indicate potential regions of Fe-oxide scavenging that were not accounted for by the model, as the exclusion of Fe_pct led to underestimations of gold concentration.
- **These regions likely have anomalous gold concentrations that are not explained by other geochemical variables, making them good targets for further investigation.**

Residual Map: Anomalous Gold Concentrations (without Fe-oxide scavenging)



Let's finish out this story

Combining our ML with geology



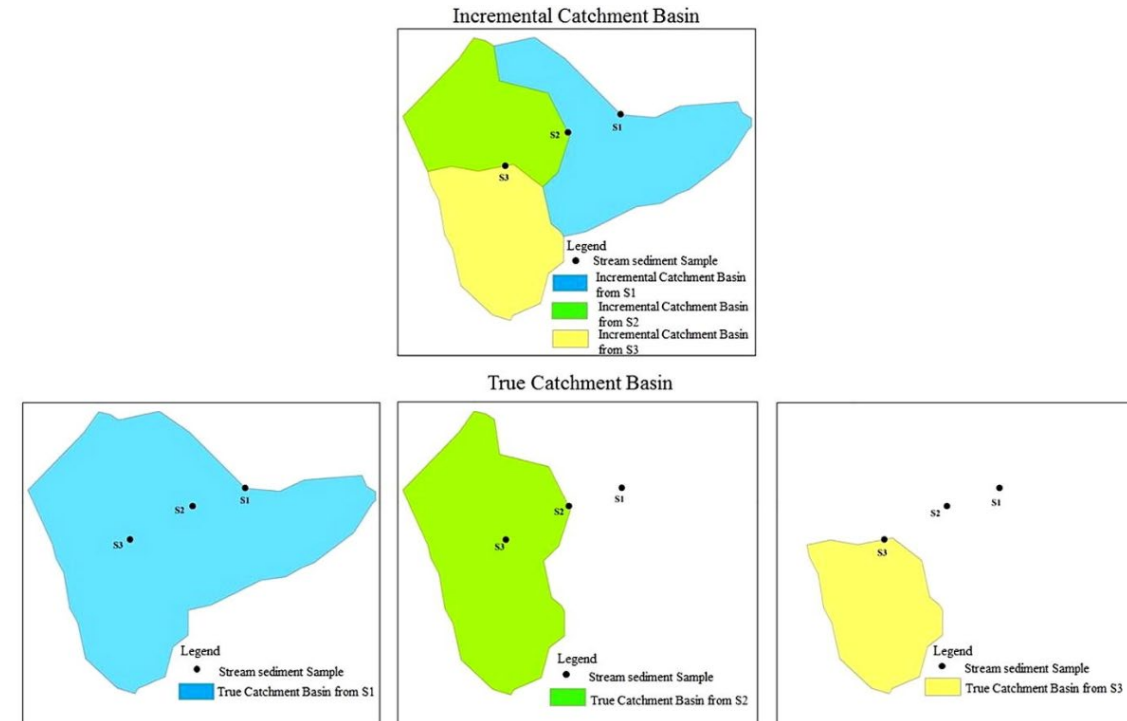
Have we checked the Clarke Values?!

- In geology and geochemistry, a Clarke value refers to the average concentration or abundance of a chemical element in the Earth's crust; it is a benchmark for assessing how enriched or depleted an element is in geological samples.
- In exploration, comparing sample concentrations to Clarke values helps geologists and geochemists identify potential mineralization and guide further investigation.
- For example, if a geochemical sample has 500 ppm copper, and the Clarke value for copper is 55 ppm, this indicates a nearly tenfold enrichment relative to the Earth's crust average, suggesting a potential copper deposit.
- Note, that this should be AOI specific! Know your rocks 😊

Element	Clarke Value
Si	~27.7%
Al	~8.1%
Fe	~5%
Cu	~55ppm
Au	~4ppb

Have you made your catchment basin map?

- Catchment maps are used to collect stream sediment samples for geochemical analysis.
- A catchment basin is the area where surface runoff flows into a river or stream, collecting sediments and geochemical signals.
- It's important to understand catchment basins in geochemical anomaly detection that indicate upstream mineralization.



Nezhad et al., 2017

Targets to follow-up

- Using a combination of the geology (think favorable host rocks and using knowledge of the lithologies to determine Clarke values), catchment basin map, and your now 'residual' geochemical data you can reinterpret your stream sediment data to generate follow-up targets.
- Good luck!





Data science in everyday geochemistry programs

International Association of Applied Geochemistry 2024

October 12, 2024

Sam Scher | sscher@lkiconsulting.com
Tom Carmichael, Datarock