Australian Government
Geoscience Australia

Exploring for the Future
minerals | energy | groundwater

Levelling Geochemical Data

P. T. Main, D. C. Champion

Workshop outline

Part 1: Levelling two datasets together

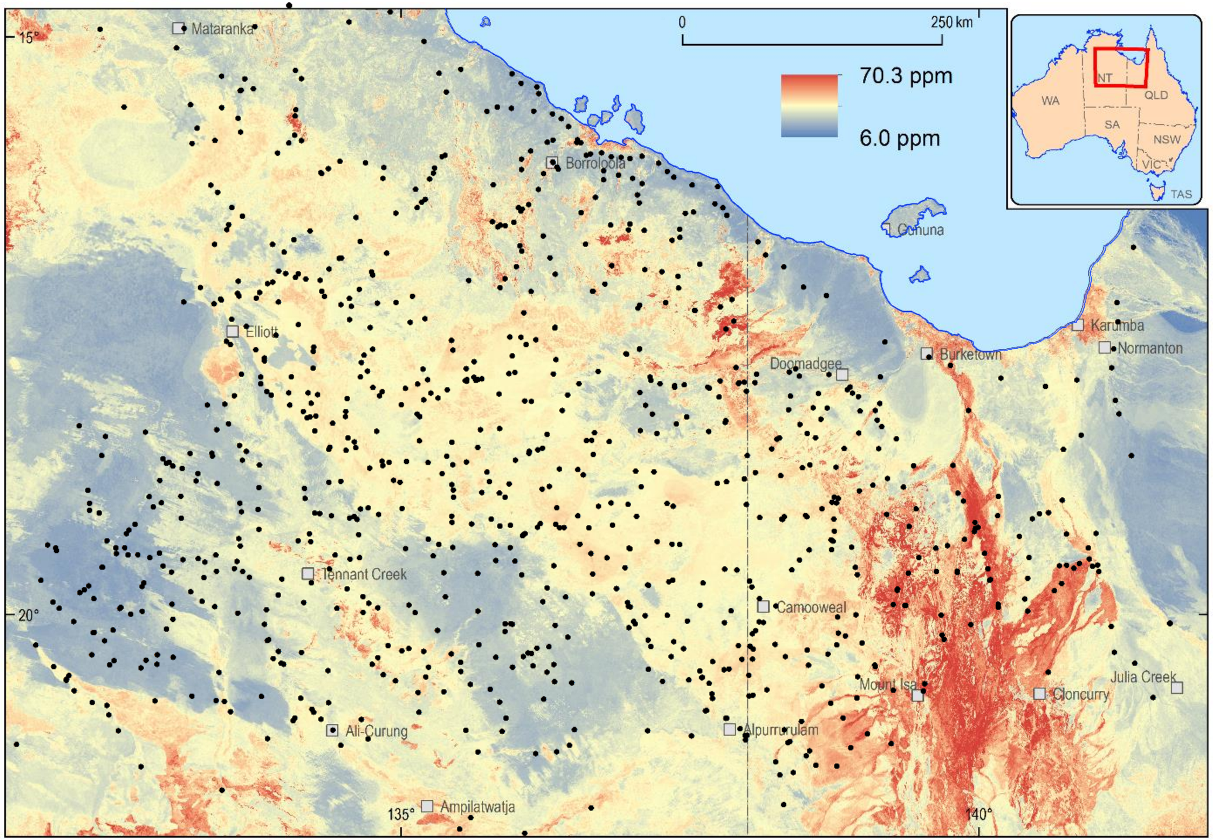
- Using standards to level data
- Using repeat analyses to level data
- Using R to undertake statistical analysis of geochemical data

Part 2: Levelling for lithology

- Using Z-score normalisation to correct for lithological effects on geochemical data

What can we used levelled data for

- Use a wide array of data holdings to get a bigger picture of the study area
- Can help highlight areas of anomalism that may otherwise have being masked

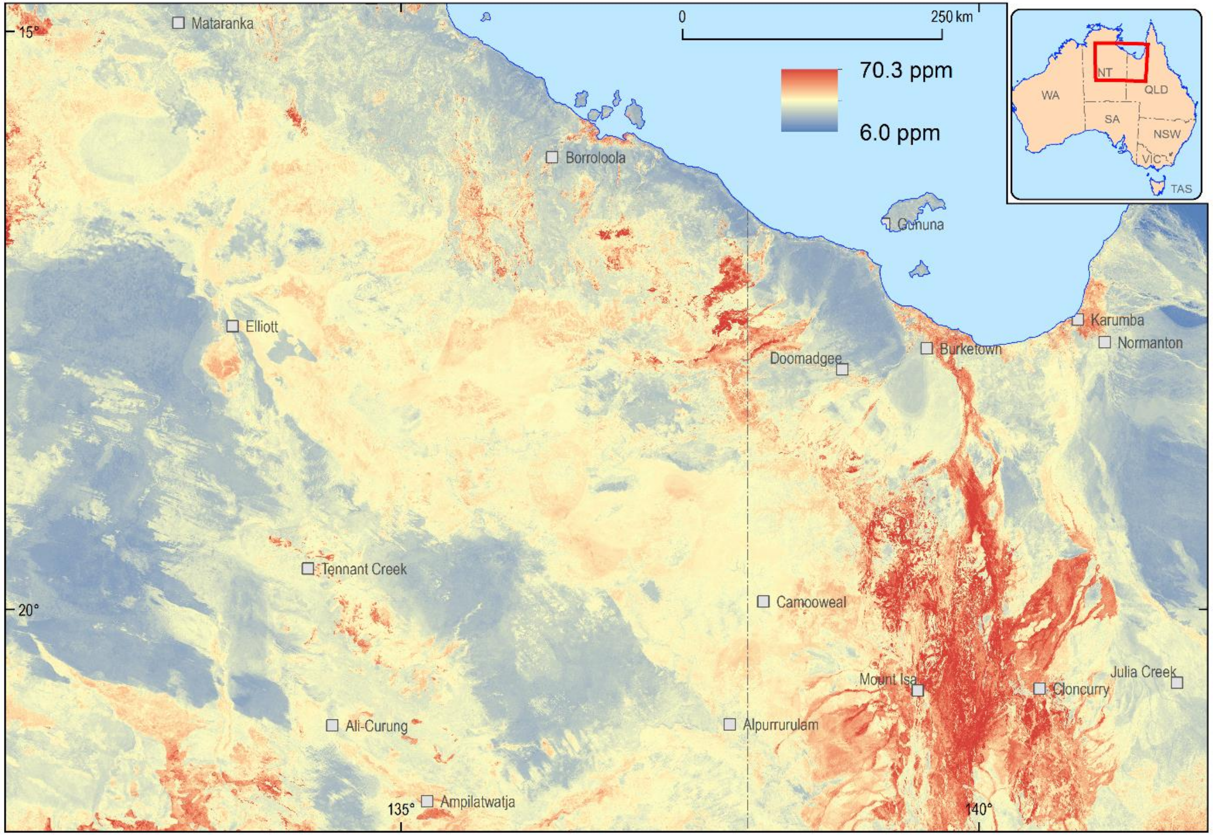


Two machine learning products for total digestion fine fraction Cu:

1. Using NAGS and NGSA data
2. Using NAGS, NGSA, and Hedley, Siegal, and Mammoth Mines data

What can we used levelled data for

- Use a wide array of data holdings to get a bigger picture of the study area
- Can help highlight areas of anomalism that may otherwise have being masked

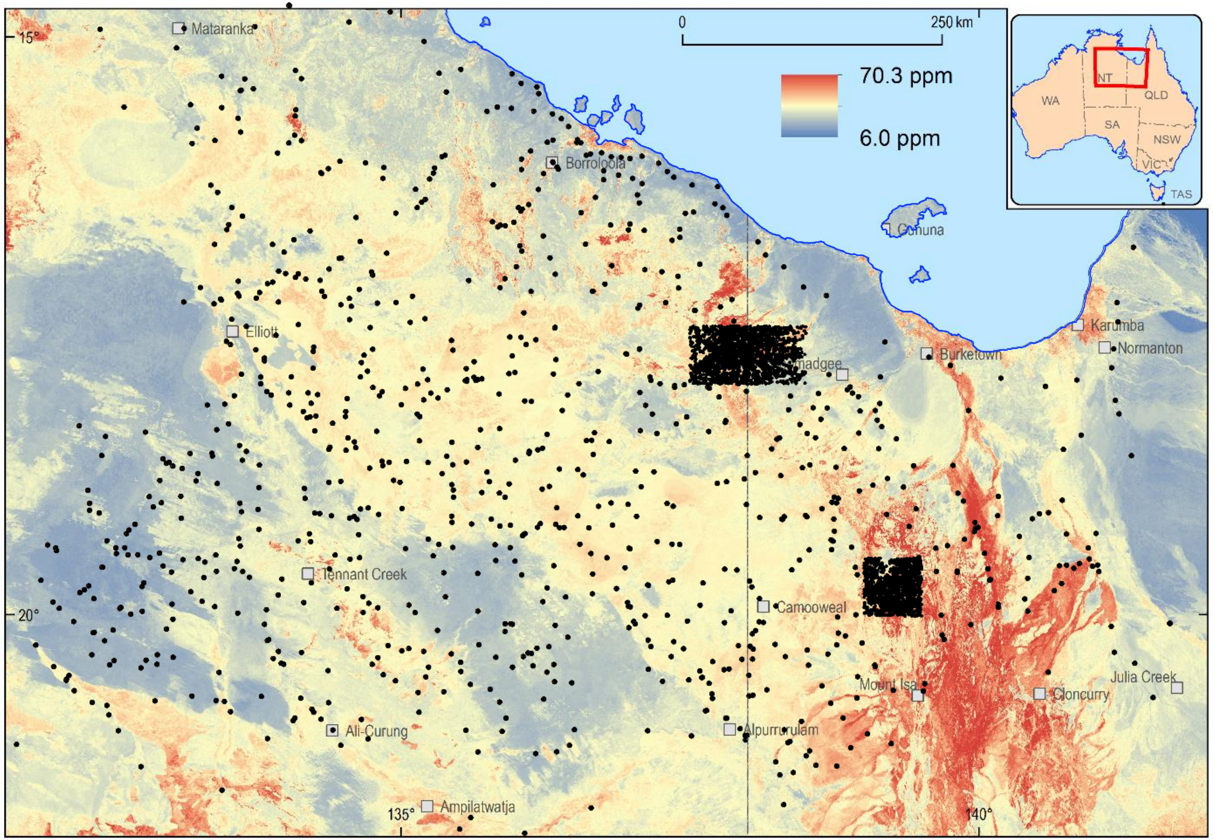


Two machine learning products for total digestion fine fraction Cu:

1. Using NAGS and NGSA data
2. Using NAGS, NGSA, and Hedley, Siegal, and Mammoth Mines data

What can we used levelled data for

- Use a wide array of data holdings to get a bigger picture of the study area
- Can help highlight areas of anomalism that may otherwise have being masked

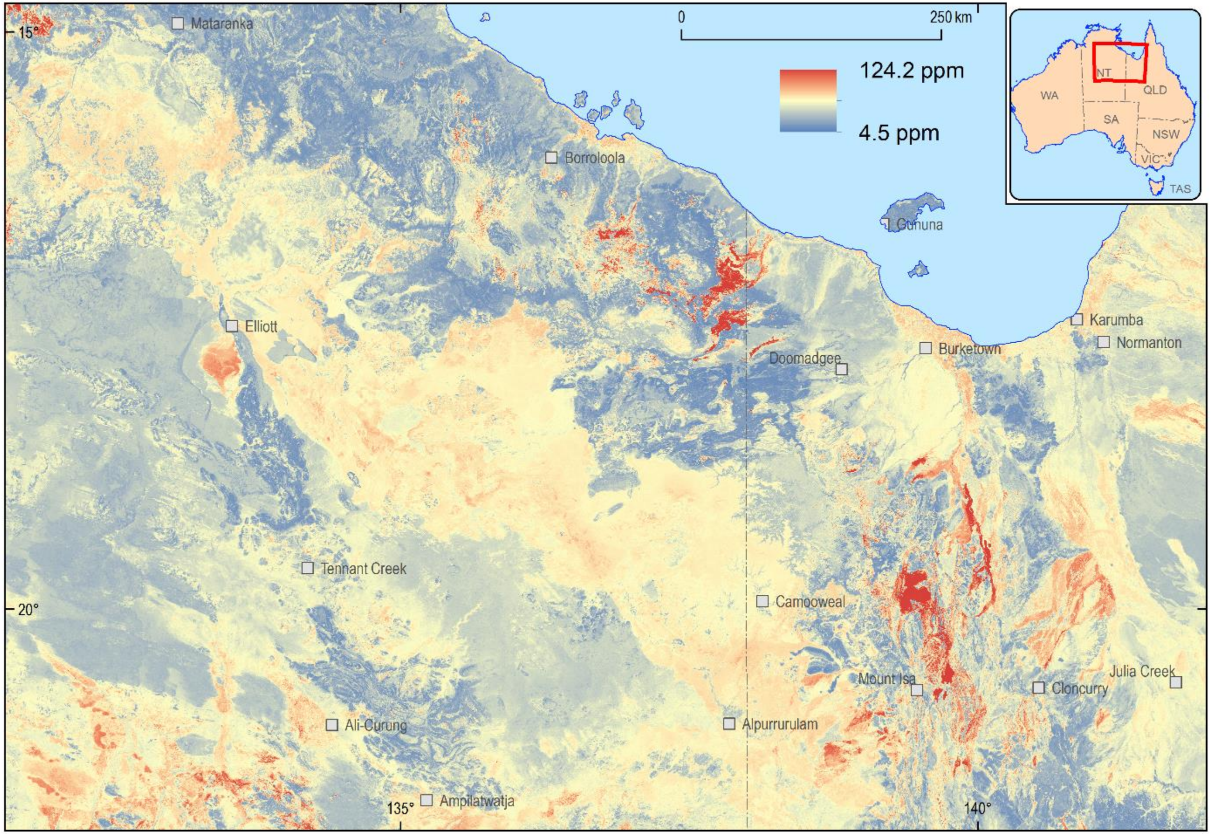


Two machine learning products for total digestion fine fraction Cu:

1. Using NAGS and NGSA data
2. Using NAGS, NGSA, and Hedley, Siegal, and Mammoth Mines data

What can we used levelled data for

- Use a wide array of data holdings to get a bigger picture of the study area
- Can help highlight areas of anomalism that may otherwise have being masked



Two machine learning products for total digestion fine fraction Cu:

1. Using NAGS and NGSA data
2. Using NAGS, NGSA, and Hedley, Siegal, and Mammoth Mines data

Levelling Geochemical data

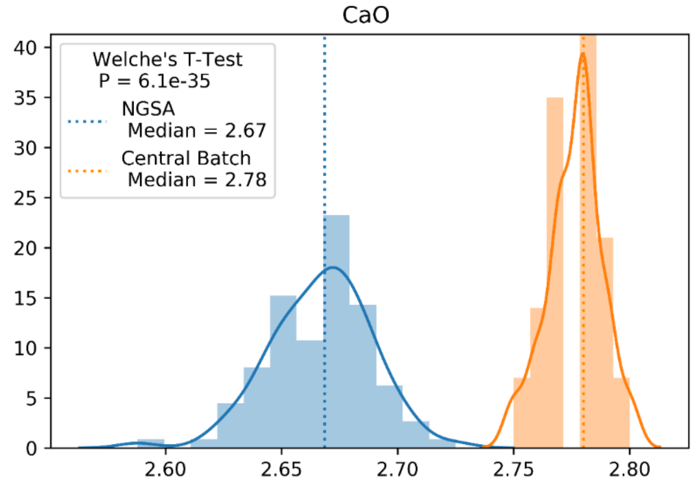
Levelling Process

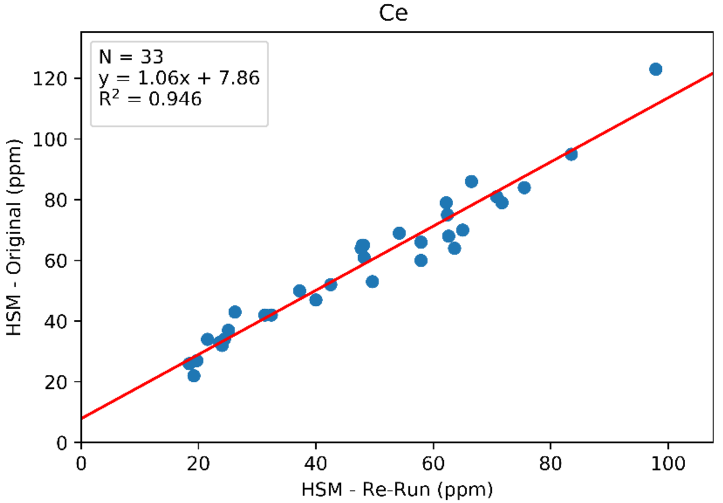
Multi-stage process depending on the survey

If there were enough standards analysed as part of the batch, and the same standards were analysed in the batch that is being used, the following process was used:

- Check the normality of the data
- Test the populations using a Welch’s t-test for normally distributed data and Wilcoxon signed-rank test for non-normal data
- If the populations are different (P < 0.05) then a correction factor is applied as a multiplier

For surveys where legacy samples were reanalysed, a linear regression was performed, and the samples corrected using $x = \frac{(y-intercept)}{slope}$





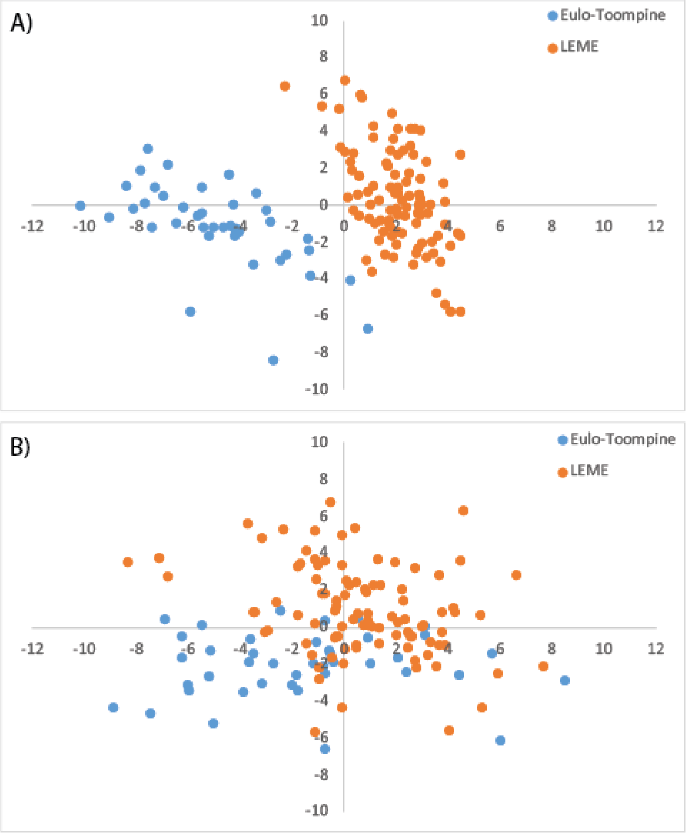
The Problem

Samples from the southern Thomson geochemical survey indicate a distinct difference in the PC 1 vs PC 2 plot (a)

The levelling procedure was applied to these samples as a test case for the NAC

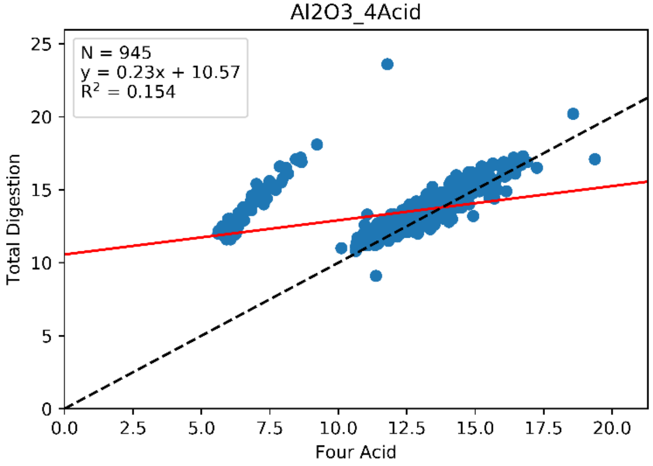
The levelling significantly reduces the effect of analysing the samples at different labs (b)

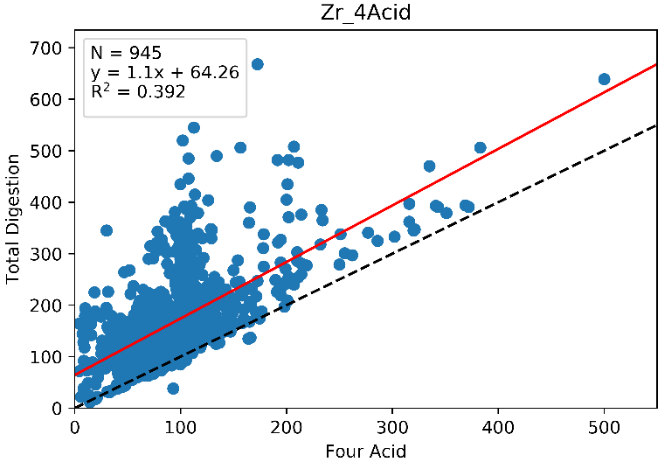
Levelling of samples is therefore required before any statistical analysis is done, particularly when samples have been analysed by different labs, different machines, or different calibration curves



Things to note

- Need to compare like with like, can’t compare different digestion types e.g. aqua regia with four acid
- Need to check the quality of the data being used, it may be necessary to QA/QC the data again
- Important to understand the technique used to acquire the data, i.e. is it XRF, ICP-MS, etc
- What medium was sampled i.e. stream sediments, rock, in-situ regolith

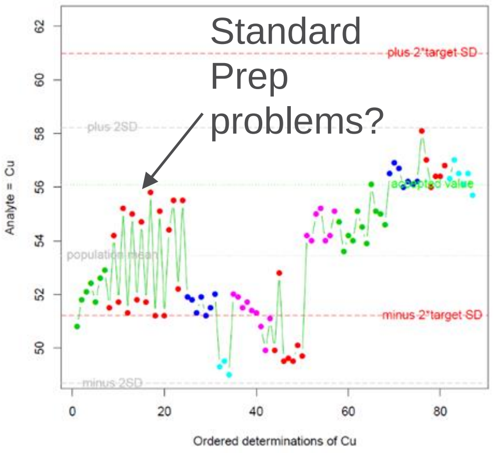


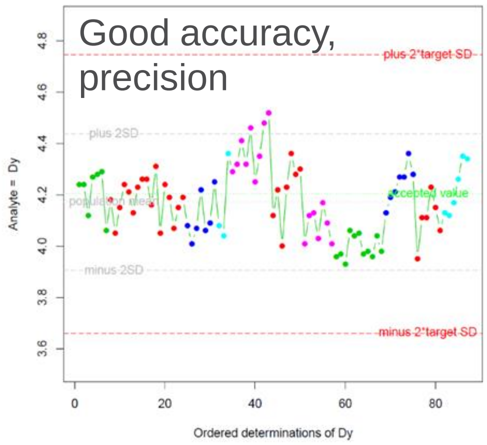


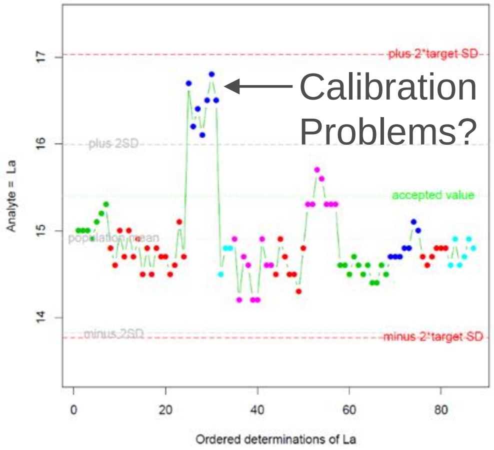
QAQC - An important step in all geochemistry work

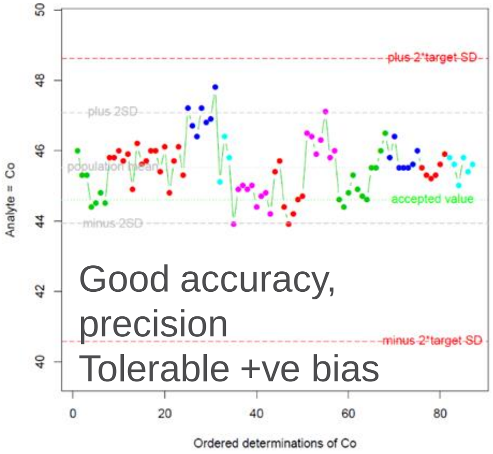
Standards

- Used to check
  - accuracy
  - precision
  - within/between batch variation
- Problems may reflect drift, calibration errors, poor standard preparation
- Readily seen by plotting analysis by batch and date
- Also by stats


Standard Prep problems?

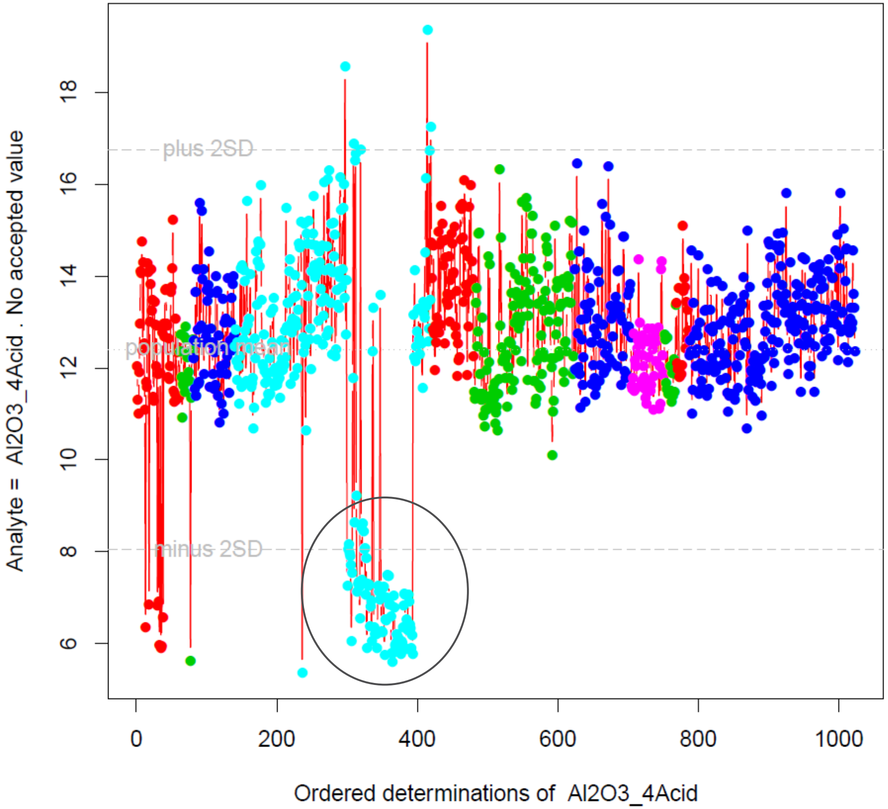

Good accuracy, precision


Calibration Problems?


Good accuracy, precision
Tolerable +ve bias

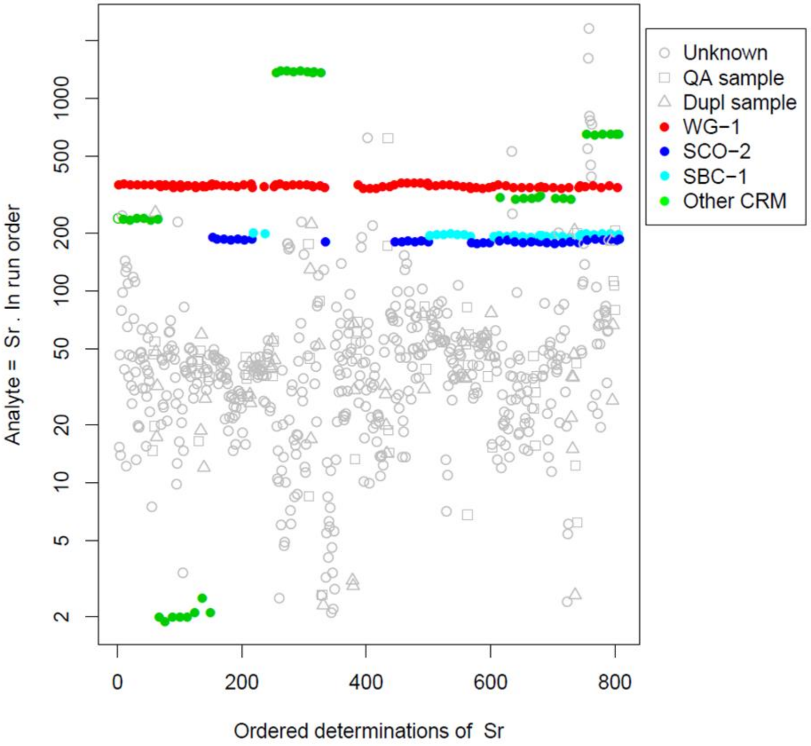
Can just use sample results

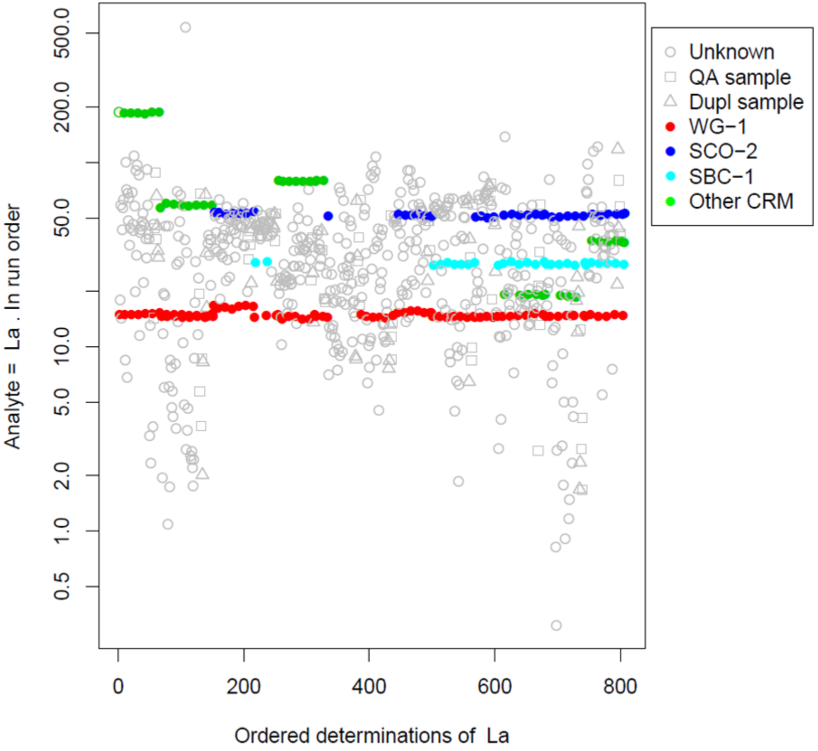
- Can check analysis simply using unknowns if samples or individual analytes are similar
- Example shows Al2O3 by date for >1000 felsic igneous rock analyses
- Can clearly identify batch with some problem, e.g., calibration error, sample prep (e.g., dilution) problems



Suitability of Standards

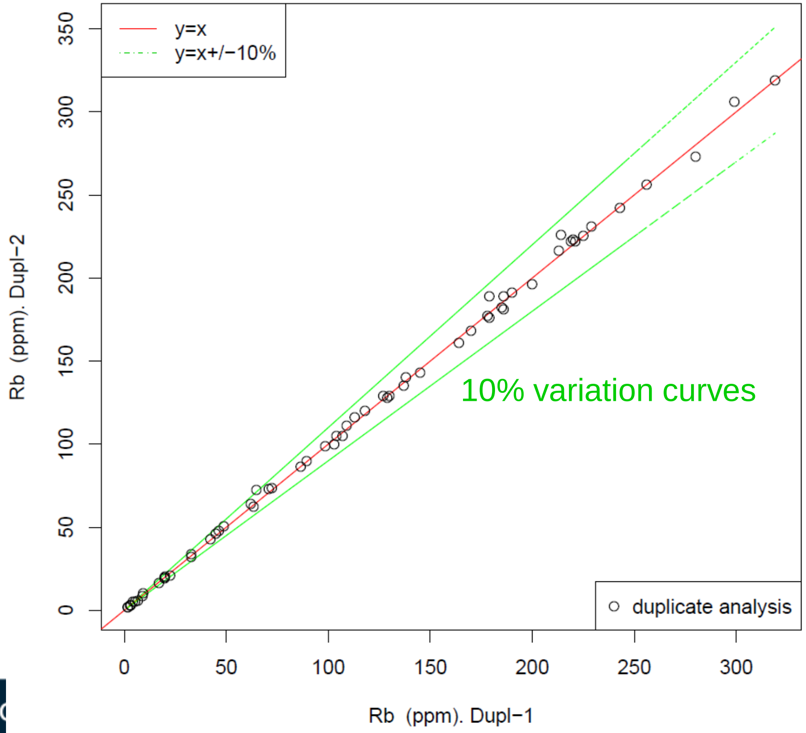
Should cover expected range of values; Should be of similar matrix





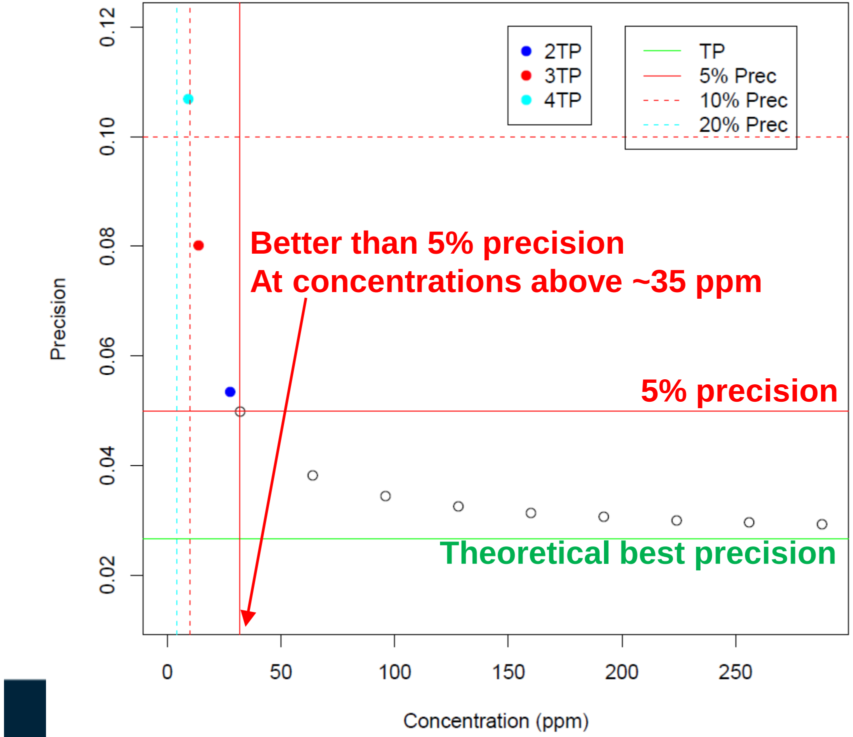
Use of duplicates

- Can use duplicate unknowns to calculate precision-concentration curves
- also detection limits (requires raw data)



Duplicate pairs. Below detection not plotted.

10% variation curves



Element = Rb . Precision−concentration

Better than 5% precision
At concentrations above ~35 ppm

5% precision

Theoretical best precision

Levelling using standards

Using R to check the data

- First step involves checking the normality of the two input datasets
- Shapiro-Wilk test can be used to determine if a population has a normal distribution
- shapiro.test(X)
- The results of this test can help determine which population statistic test to use
- Things to remember:
    - H_0: Null hypothesis, if $P > 0.05$ we accept the null hypothesis
    - H_1: Alternate hypothesis, if $P < 0.05$ we reject the null hypothesis and accept the alternate
    - The null hypothesis for the Shapiro-Wilk test is that the population is of a normal distribution

Using R to check the data

Welche's T-test:

- Unpaired students t-test
- Used to determine if two populations are statistically similar
- Assumes normality of the input data
- t.test(X,Y, paired = FALSE)

Wilcoxon rank sum test:

- Also used to determine if populations are statistically similar
- No assumption of normality in the input dataset
- wilcox.test(X, Y, paired=FALSE)

H_0 for both test: there is no statistical difference between the two populations

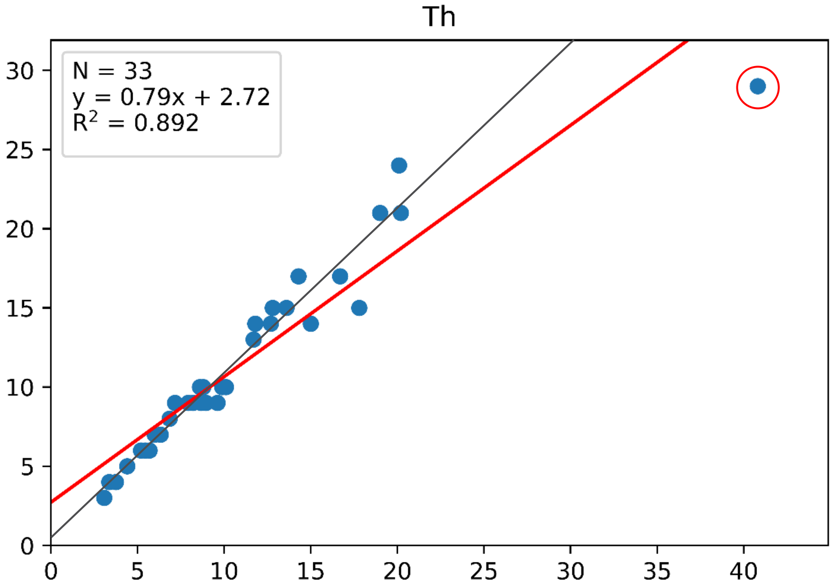
Next Step for levelling via standards

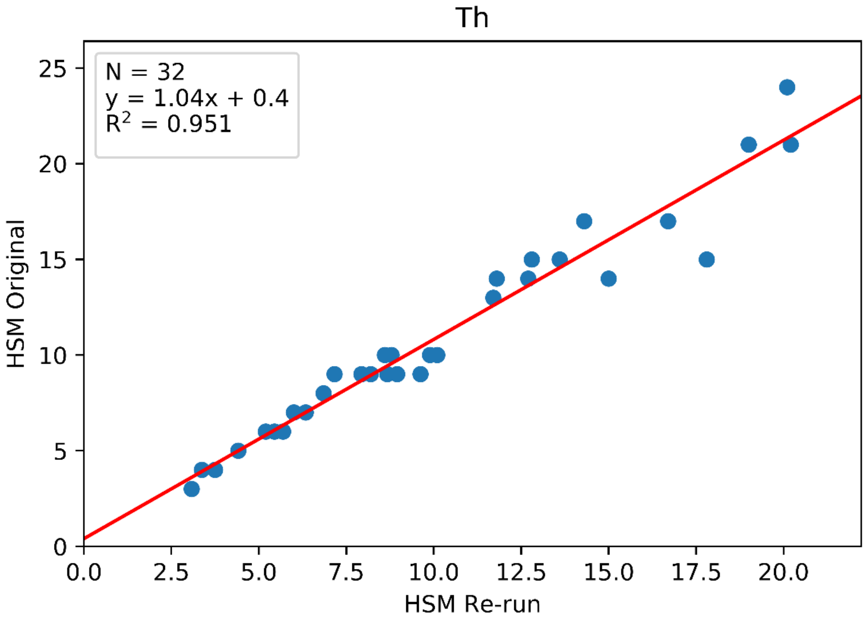
- If the population statistics indicate that the standards are of the same population (we have accepted H_0) then no levelling needs to occur
- If we accept H_1 then we need to create a correction factor in order to level that element
- By taking the median of each population we can calculate the difference (note: in situations where the standards are skewed or contain a number of outliers, the median will not be a good approximation of the peak)

Levelling using linear regression

Linear Regression

- In the cases where standards aren’t reported, available, or were run we can re analyse samples in order to compare the rerun data with the original.
- Once the new data has being acquired a simple XY cross plot can be used with a linear regression to find the line of best fit
- The equation for the line of best fit (y = mx + b) can then be rearranged to corrected the values to a 1 to 1 line
- $x = \frac{y - b}{m}$ where b is the intercept and m is the slope
- Need to check the plots to ensure any outliers aren’t having a significant impact on the regression line



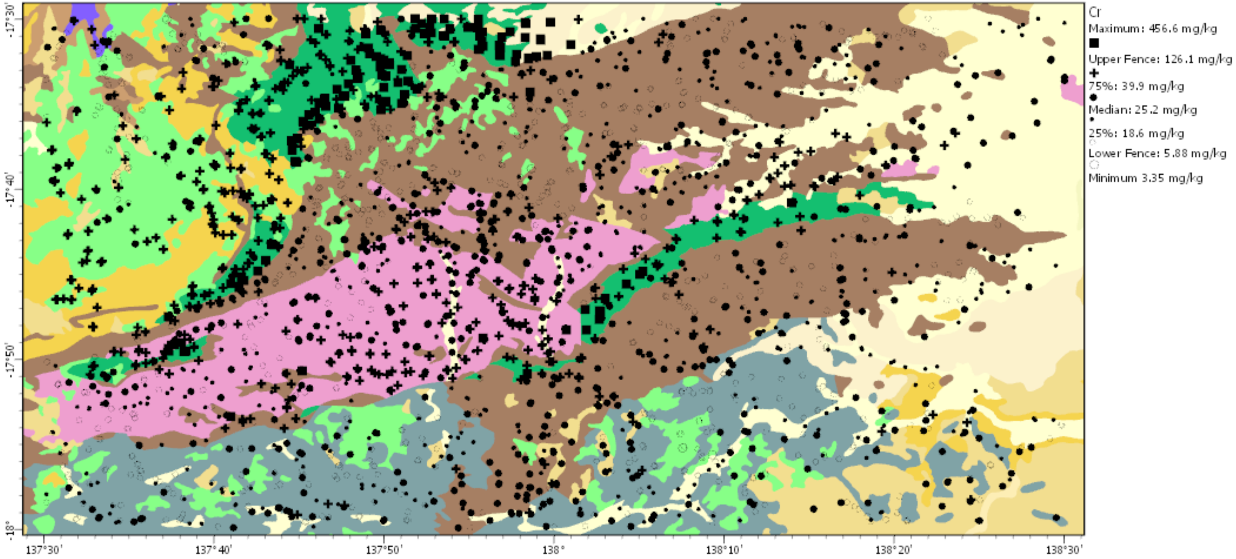


GEOSCIENCE AUSTRALIA

Levelling for lithology/regolith

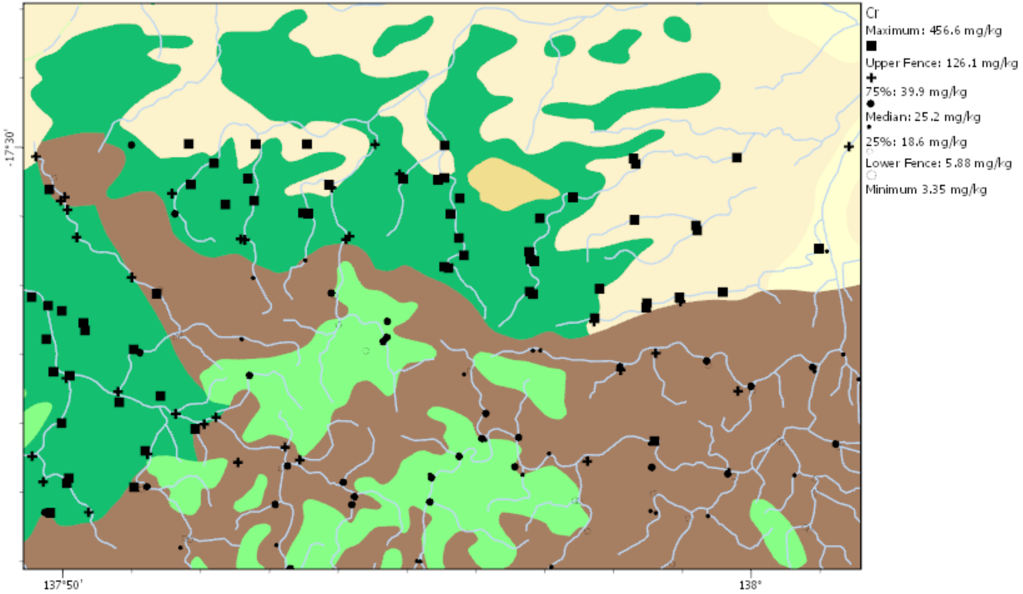
Why level for lithology

- Lithology (including regolith types) can have a strong effect on the geochemical response that is measured
- This can lead to swamping of the signature and may mask outliers
- By correcting for lithological influences we can pull out the outliers from both the areas of a natural high and a natural low



Choosing our groups

- Need to define the groups based on both lithology and geochemistry
- Need to make sure that the group isn't too small or the statistics won't be valid
- For stream sediment data care needs to be take to ensure that transported material is attributed to the correct group



Transforming our data

- Before we can level the data for lithology we need to use a log ratio transformation

- For the levelling approach to work we are relying on the assumption of normality

- We need to transform the data using a log ratio transform owing to the multivariate nature of geochemical data

- The transformation is achieved using a Centered Log-ratio (CLR) transformation

- $clr(z) = ln\left(\frac{x_i}{\sqrt[n]{x_1 \times x_2 \times \cdots \times x_n}}\right)$ Geometric mean (excel → =geomean())

Z-Scores

- The Z-Score will normalise each group to similar levels
- The score represents how many standard deviations away from the median a value is
- The data is presented in the form of a Z-score with no information on concertation given
- It represents a good method of outlier detection that is hypothetically blind to lithology
- $Z_i = \frac{x_i - \bar{x}}{\sigma}$

Thank You!

Questions?

Phone: +61 2 6249 9111

Web: www.ga.gov.au

Email: feedback@ga.gov.au

Address: Cnr Jerrabomberra Avenue and Hindmarsh Drive, Symonston ACT 2609

Postal Address: GPO Box 378, Canberra ACT 2601