Extracting, visualising and interpreting structure in geochemical data through compositional data analysis (CoDA)

Missing Values and Imputation

Eric Grunsky

Department of Earth and Environmental Sciences,
University of Waterloo, Ontario

egrunsky@gmail.com

2021-November-18

Missing Values

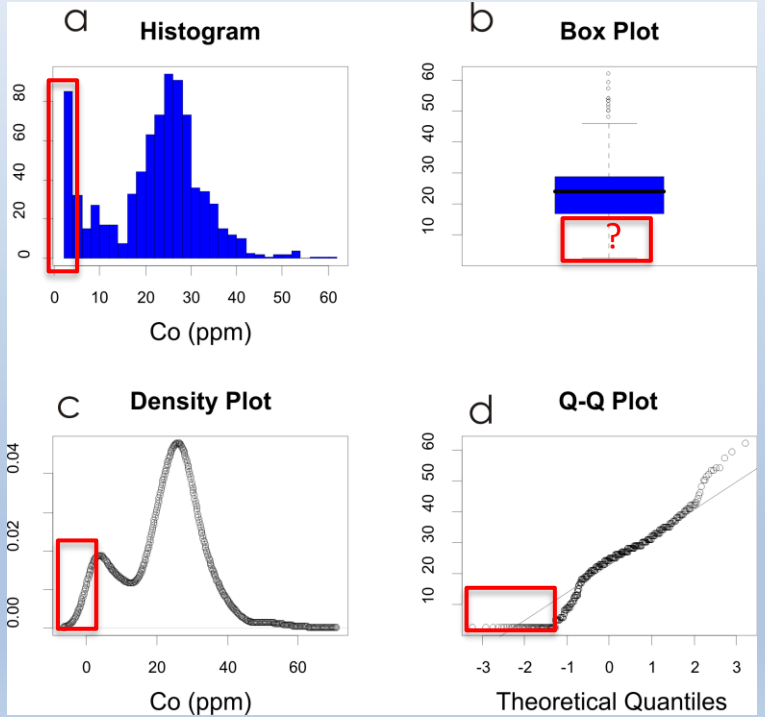
- Missing values are:
  - Censored values
  - Unknown (commonly reported as zero or some arbitrary number [e.g.-999])
- Estimates for replacements is more commonly done by regression methods.
- It is more difficult to estimate missing values or replacement for censored data if the data distributions are non-normal.

Censored Data Distributions

- Censored data very common in geochemistry (values < L.L.D.)
- Cannot accurately estimate mean and variance of censored data.
- Common practice is some fraction of L.L.D as replacement value. (i.e. 1/2, 1/3, 3/4). OK when only a few samples are censored.
- When many samples are censored special procedures are required.
- Replacement values are based on the characteristics of the data above the censored threshold (Sanford et. al, 1993).
- Approach as developed by Hron, Templ and Filzmozer (2010) [robCompositions] makes fewer assumptions about the normality of the distributions. Useful for missing values.
- EM algorithm approach by Palarea-Albaladejo & Martin-Fernandez [zCompositions] provides very good estimates of replacement values for censored data.

Dealing with Values Less than the Lower Limit of Detection

- Distribution of Co in metavolcanics collected during lithogeochemical sampling program in the Ben Nevis township area of Ontario. The analytical procedure for Co has a lower limit of detection of 5.0 ppm and 85 out of the 824 observations fall below that limit.
- The histogram shows a bar with a high frequency of observations at the lowest end of the scale. This bar represents the 85 values that are less than the detection limit.
- The Q-Q plot shows these values as a flat part of the distribution at the left side of the figure.
- The density and box plots do not show the censored values as clearly.
- Historically, censored data were handled by applying a substitute value; somewhere between 1/3 to 1/2 of the actual detection limit. As the number of observations below the lld (censored) increases, then this estimate will produce inaccurate estimates of the mean and variance.



censored data

Imputation

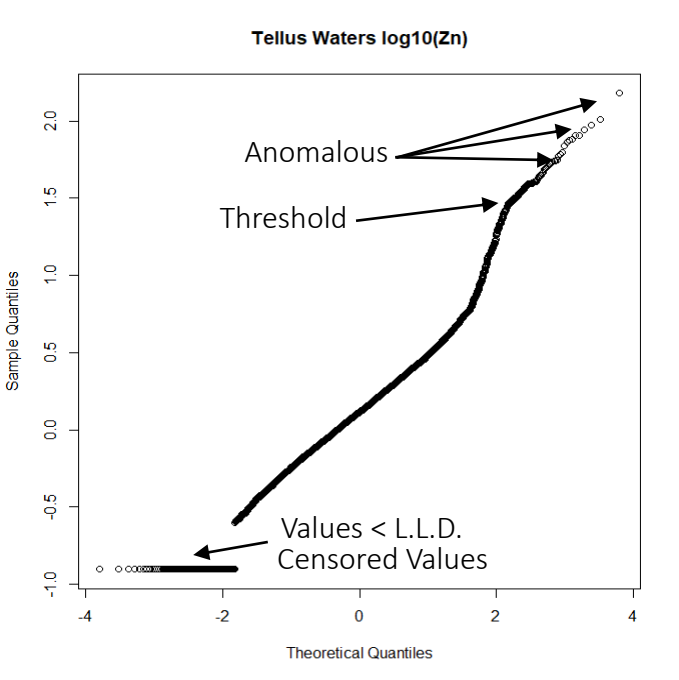
- Imputation - replacing missing data with substituted values (analytical values less/greater than the limits of detection).
- Commonly used methods:
  - Nearest neighbour (robCompositions)
  - Expectation/Maximum Likelihood (EM) (zCompositions)
  - Simple replacement by a constant
  - Multiple regression (robCompositions)

Software for Censored and Missing Values

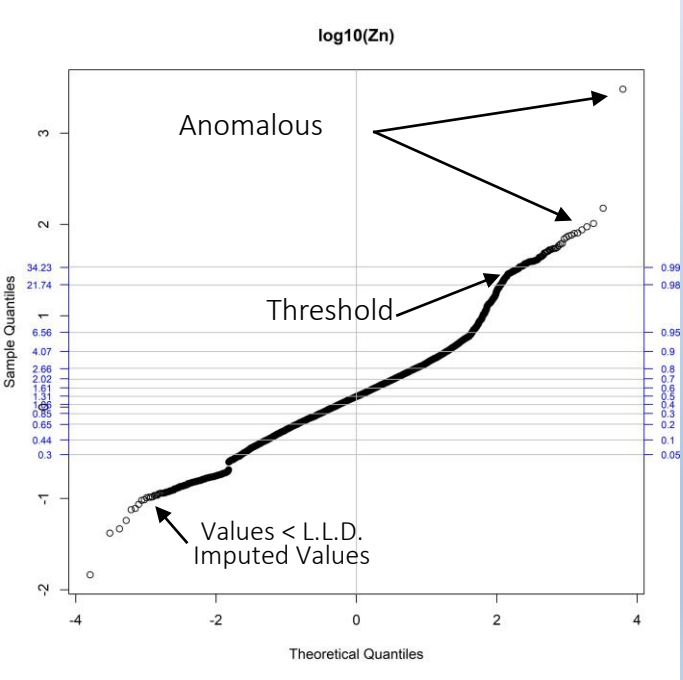
- R statistical environment
  - zCompositions (censored data only)
  - robCompositions (censored and missing data)
  - NADA (censored data) – used by the environmental community
- Medical (epidemiology) community has many procedures for imputing data.

The Anatomy of an Element Distribution

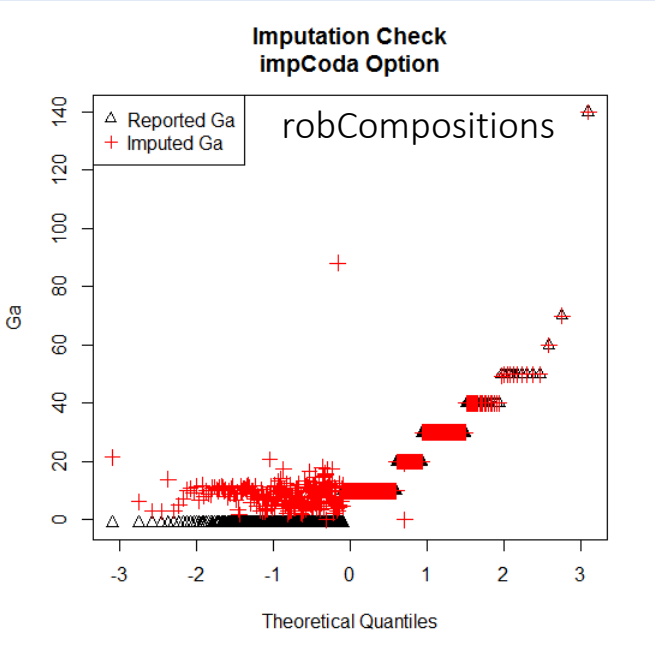
Before Imputation

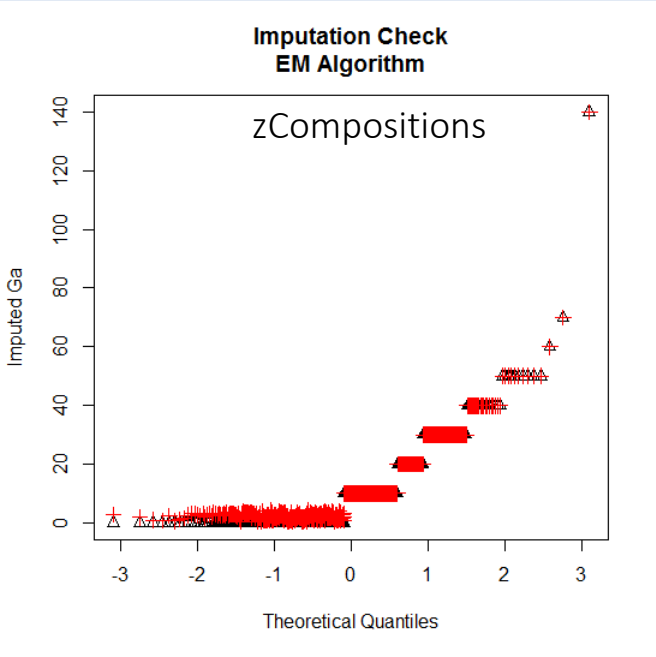


After Imputation



Imputation for ICP Analyses – Lake Sediments





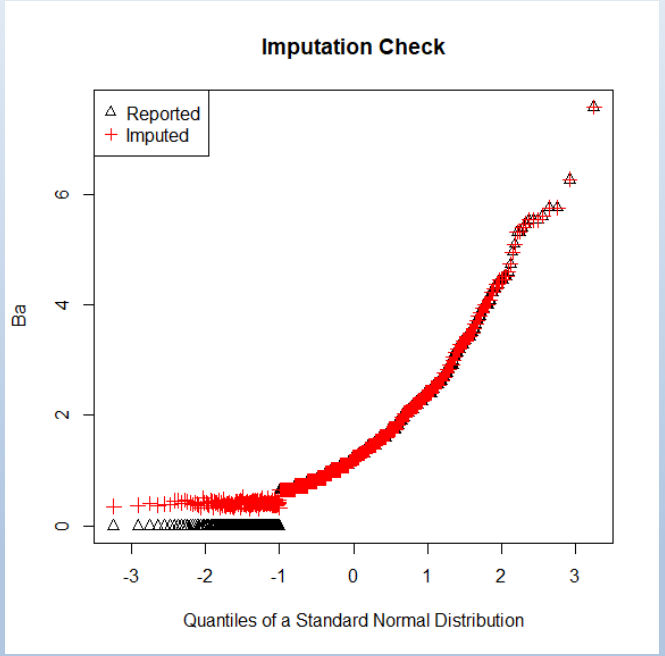
430/505 samples < LLD

robCompositions – nearest neighbour analysis

zCompositions – EM algorithm

zCompositions

- There are a number of functions to calculate replacement values for censored data in compositional datasets.
- Th function lrEM implements model-based ordinary and robust Expectation-Maximisation algorithms to impute left-censored values (e.g. values below detection limit, rounded zeros) via coordinates representation of compositional data which incorporate the information of the relative covariance structure. Alternatively, it can be used to impute missing data.



Effects of Censored Data

- When there are many censored values for a given element the estimate of the mean and variance is less than the “true” mean and variance (assuming a known distribution).
- When many elements are censored, the combined effect will distort the inter-element relationships and create difficulty in effectively interpreting multi-element signatures associated with background and target populations or processes of interest.
- When possible – correct for censoring!