Using log-ratio-log plots to assess the association between chemical species in environmental samples.

Michael Ginevan

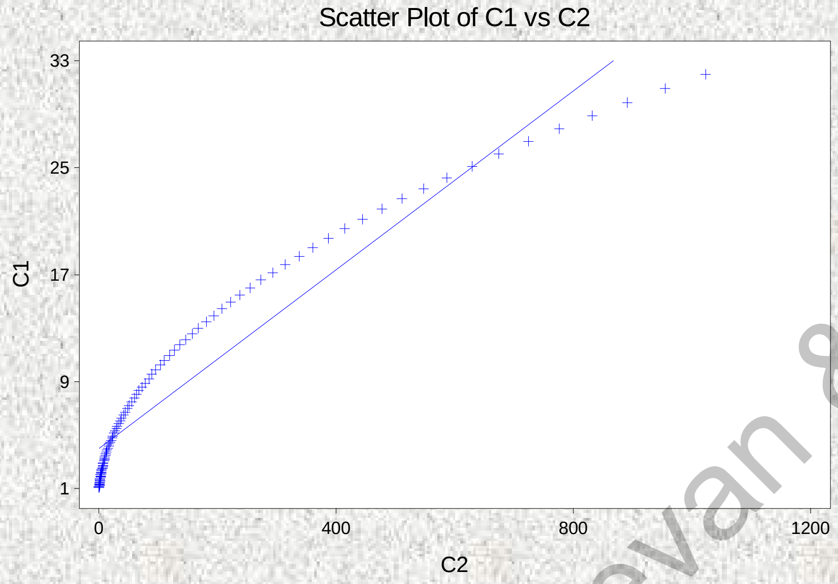
The problem

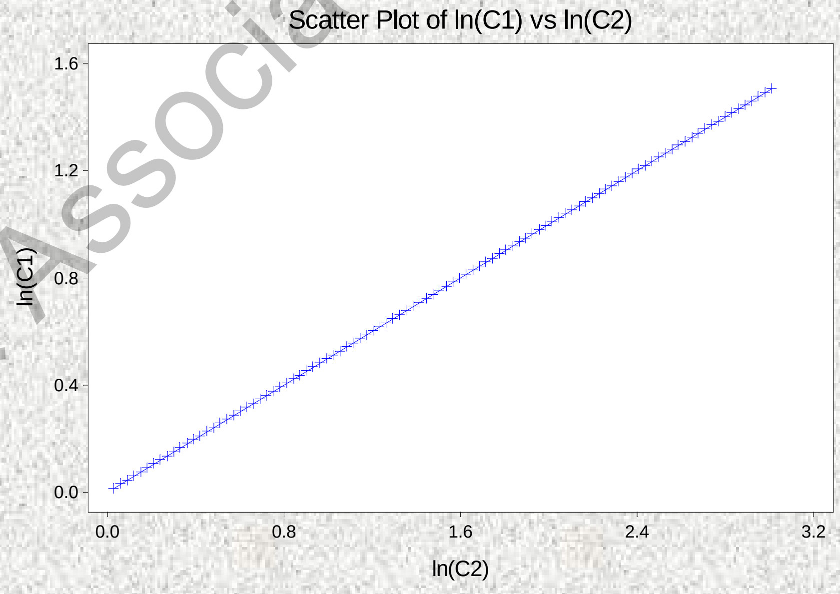
- Do a set of chemical species show a consistent composition across a set of environmental samples?
- If the species show constant relative concentration, one might infer that the chemical contamination has a common origin.
- Example: If one had a set of samples that include measurements for both lead (L) and cadmium (C), one might ask if the ratio C/L is consistent across samples.

Solution I

- Regress ln(C) against ln(L).
- If this regression fits the data well and the slope is not significantly different from one, one has evidence that the ratio of C/L is constant across samples.
- However, one may have regressions where fit is good but the slope of the log-log relationship differs from unity.
- Here log-log plots of concentration data suggest consistent relationship where one does not really exist. .

Example, Log-log plot



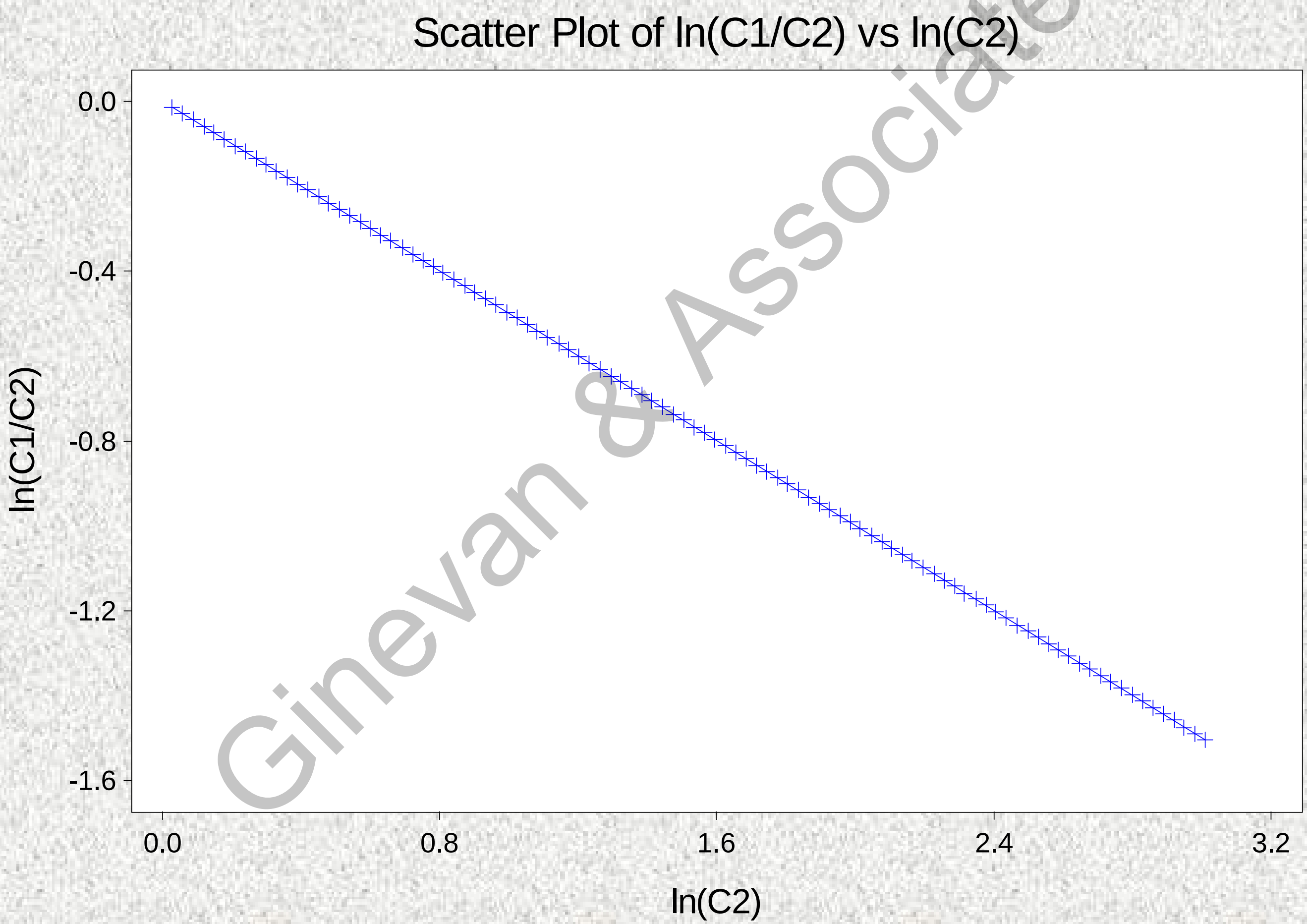


Slope in the second panel is significantly different from 1 but this is hard to see

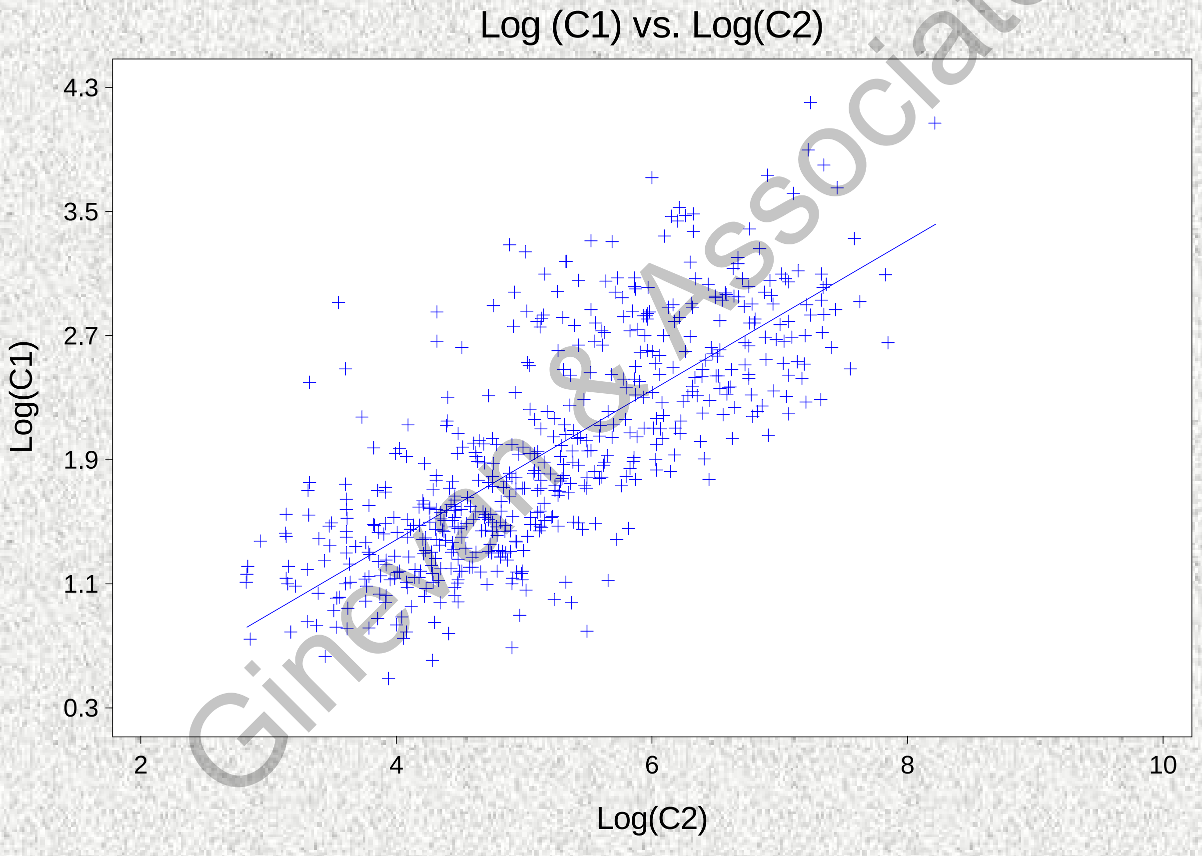
A Better Picture

- Plot the log of the ratio of the two components against the log of one member of the pair (e.g.for our example, ln(C/L) versus ln(L).)
- Such a plot will show a zero slope if the ratio is constant but will show non-zero slopes even for the case of complete independence of the two concentrations across samples.

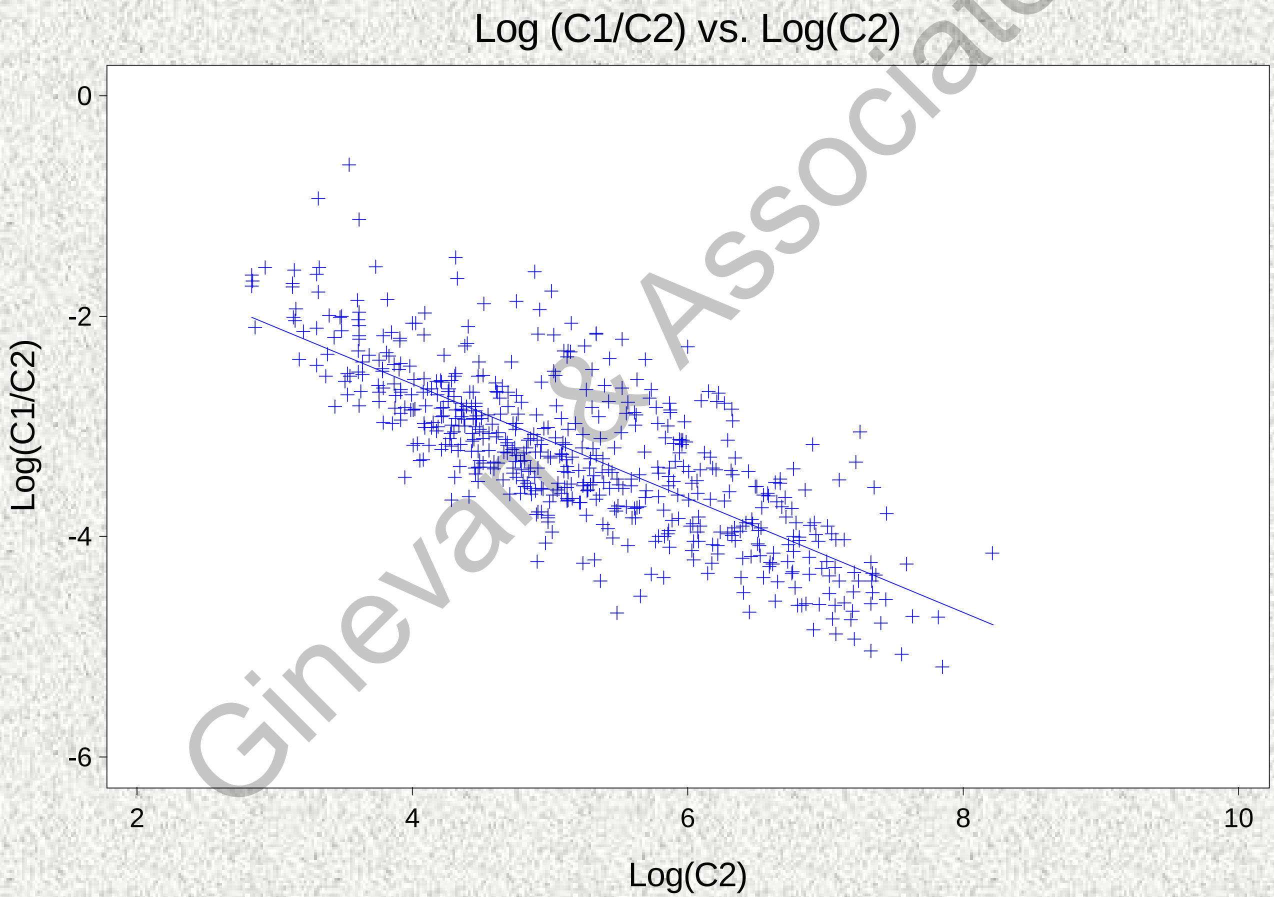
The Same Data as in the log-log plot



Consistent Association?



No!



The case of independence I

- If ln(C1) is Y and ln(C2) is X, the correlation coefficient (r) between ln(C1/C2) and ln(C2) is:

$$\frac{\sum X (Y-X)}{[\sum X^2 \ \sum (X-Y)^2]^{1/2}}$$

- If ln(C1g) and ln(C2) are independent, then this simplifies to:

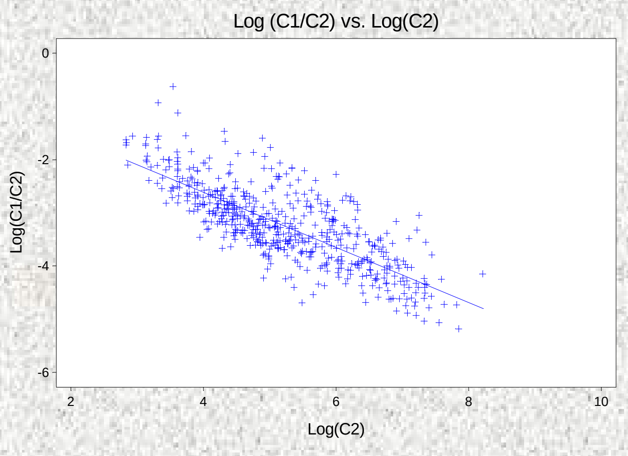
$$\frac{-\sum X^2}{[\sum X^2 (\sum(X^2 + Y^2)]^{1/2}}$$

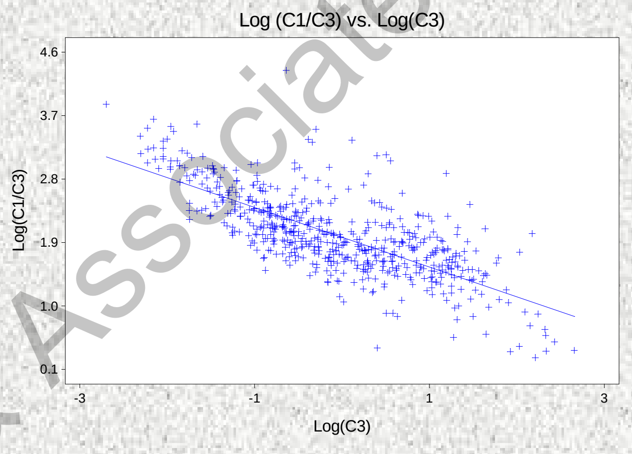
- Which, when the variance of ln(C1) equals the variance of ln(c2) is equal to $-1 / (2^{1/2})$ or -0.707.

The case of independence II

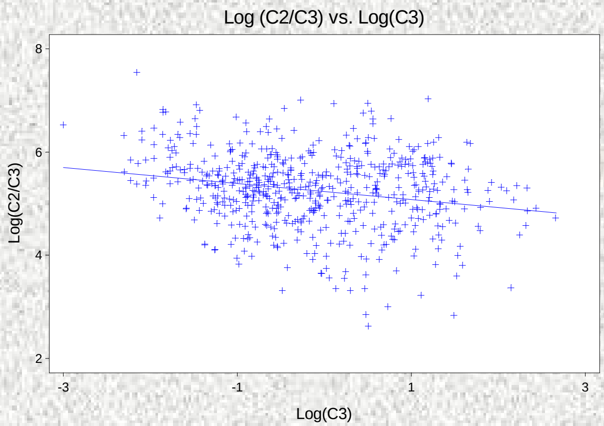
- If the variance of ln(C1) is much greater than the variance of ln(C2) the correlation tends toward zero.
- If the variance of ln(C1) is much less than the variance of ln(C2) the correlation tends toward 1.
- Neither case is interesting because one of the two variables is essentially a constant.
- Note that we are working on logs so two chemicals in the same samples usually have similar variability.
- Standardizing the log transformed concentration data to unit variance is a bad idea because it in effect applies a power transformation in the original variable space, and can create an impression of a constant ratio for data that have a non-constant ratio.

A LRL Scatter Plot Matrix

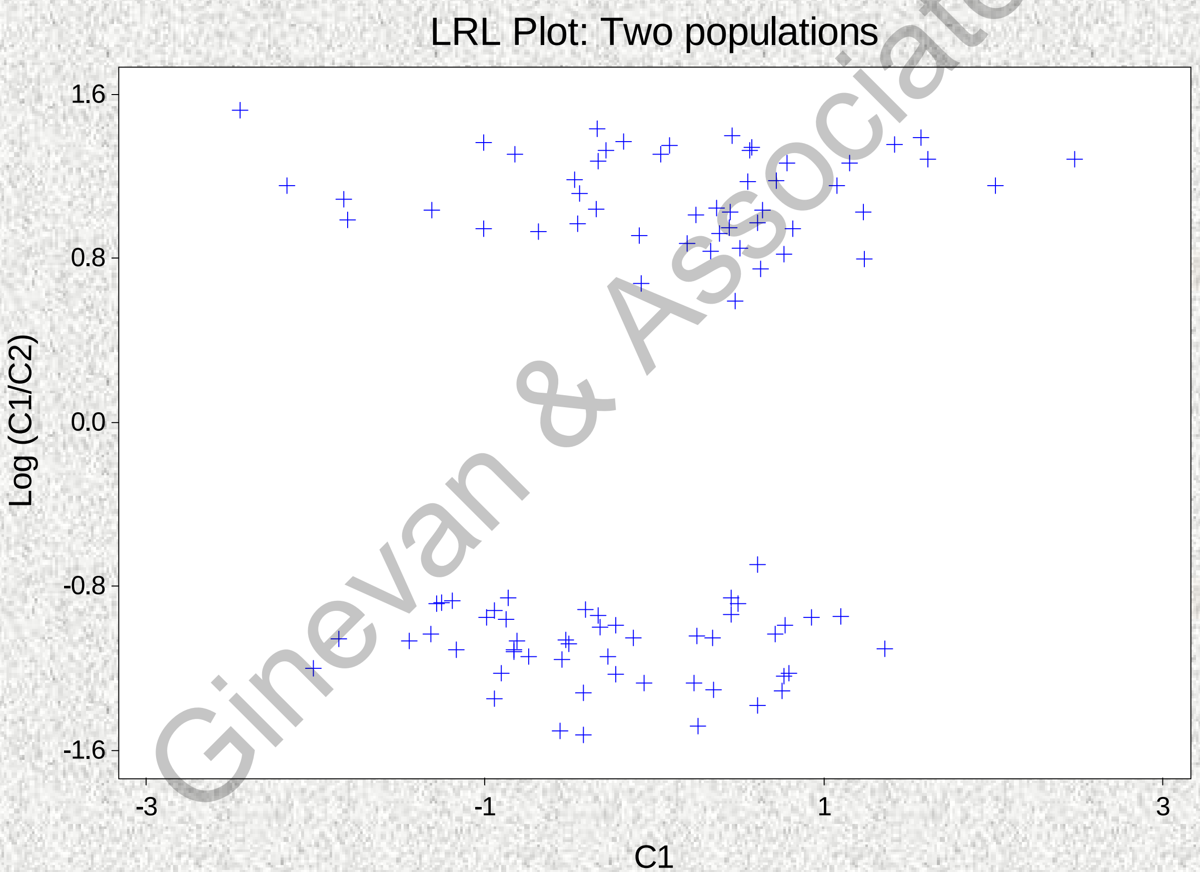




C2 and C3 are closest to a constant ratio



Finding Groups



Hypothesis Testing

- One can use the regression lines in LRL plots as a basis for hypothesis tests.
- But simply regressing $\ln(C_1)$ vs $\ln(C_2)$ results in a more straightforward statistical evaluation.
- One can answer two questions.
  - Is the relationship between C_1 and C_2 strong (high correlation)?
  - Does the regression coefficient for C_1 on C_2 differ significantly from one?
- If there is a strong correlation and the regression coefficient does not differ significantly from one, we have good evidence of a constant ratio.

Conclusions

- LRL plots are a simple, but useful, way of looking at chemical concentration data.
- Can be used to dispute the idea that a good log-log regression fit implies a constant ratio relationship.
- Can allow one to screen a data set to determining which pairs of chemical concentrations do exhibit constant relationships.
- Can also help identify multiple groups of observations with constant within group but different between ratios for a given pair of chemical concentrations
- In summary, LRL plots are a tool that provides a unique perspective for the search for patterns in environmental data.