Goodness of fit metrics and automated source identification

Basil Coutant

Outline

- Why do we need GOF metrics?
- What do we want to measure?
- How do we identify sources?
- Our metrics for F, G, and X.
- Results for the Palookaville data.
- Automated profile matching against known profiles.
- General automated profile identification.

Why do we need GOF metrics?

- Give a specific mean to phrases like "this is a better profile."
- Quantify the confidence in the quality of the output of the models.
- Give focus to what needs improved.
- *Disclaimer:* The following are proposals! They may not measure items of interest. Better metrics may exist.

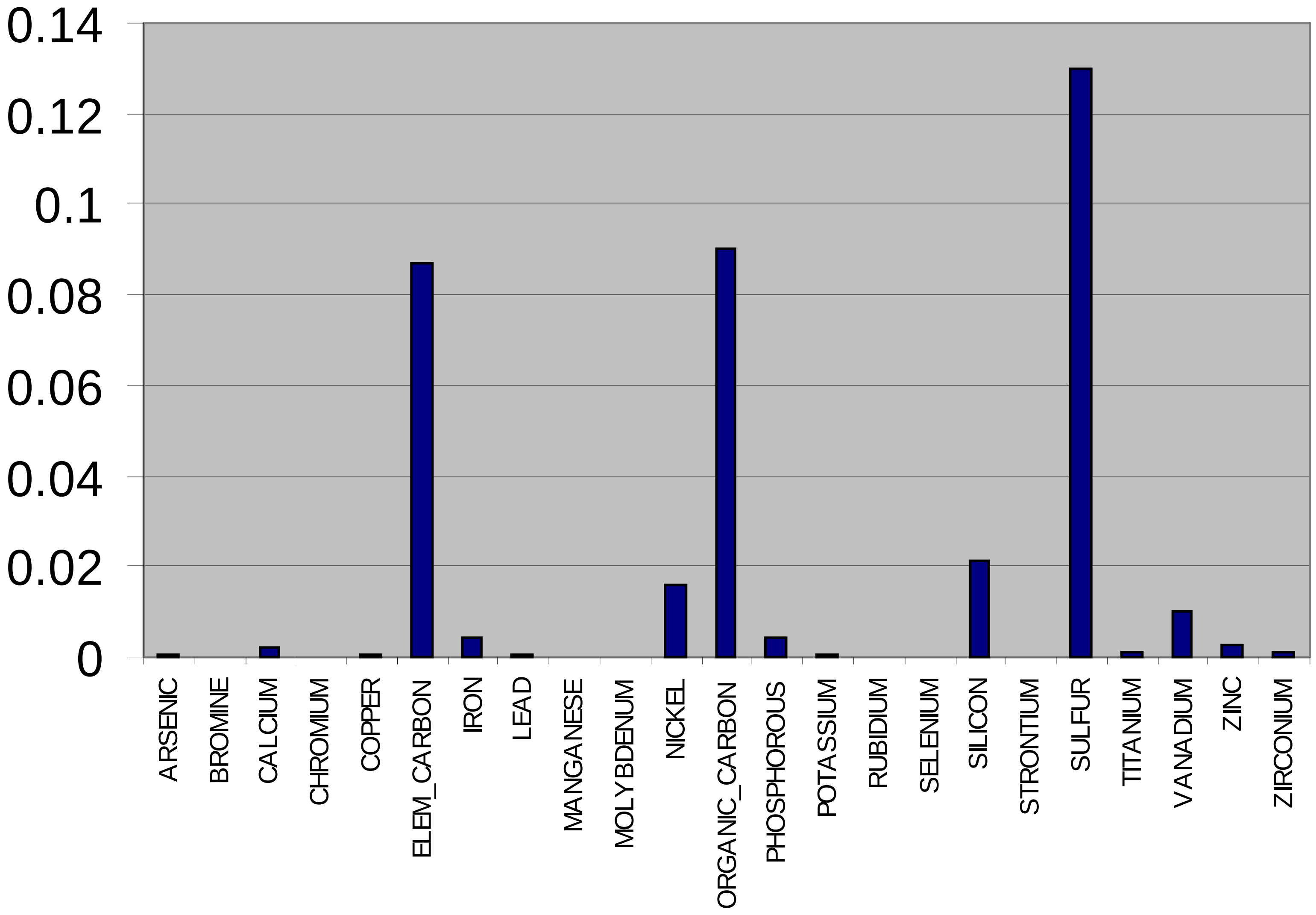
What do we want to measure?

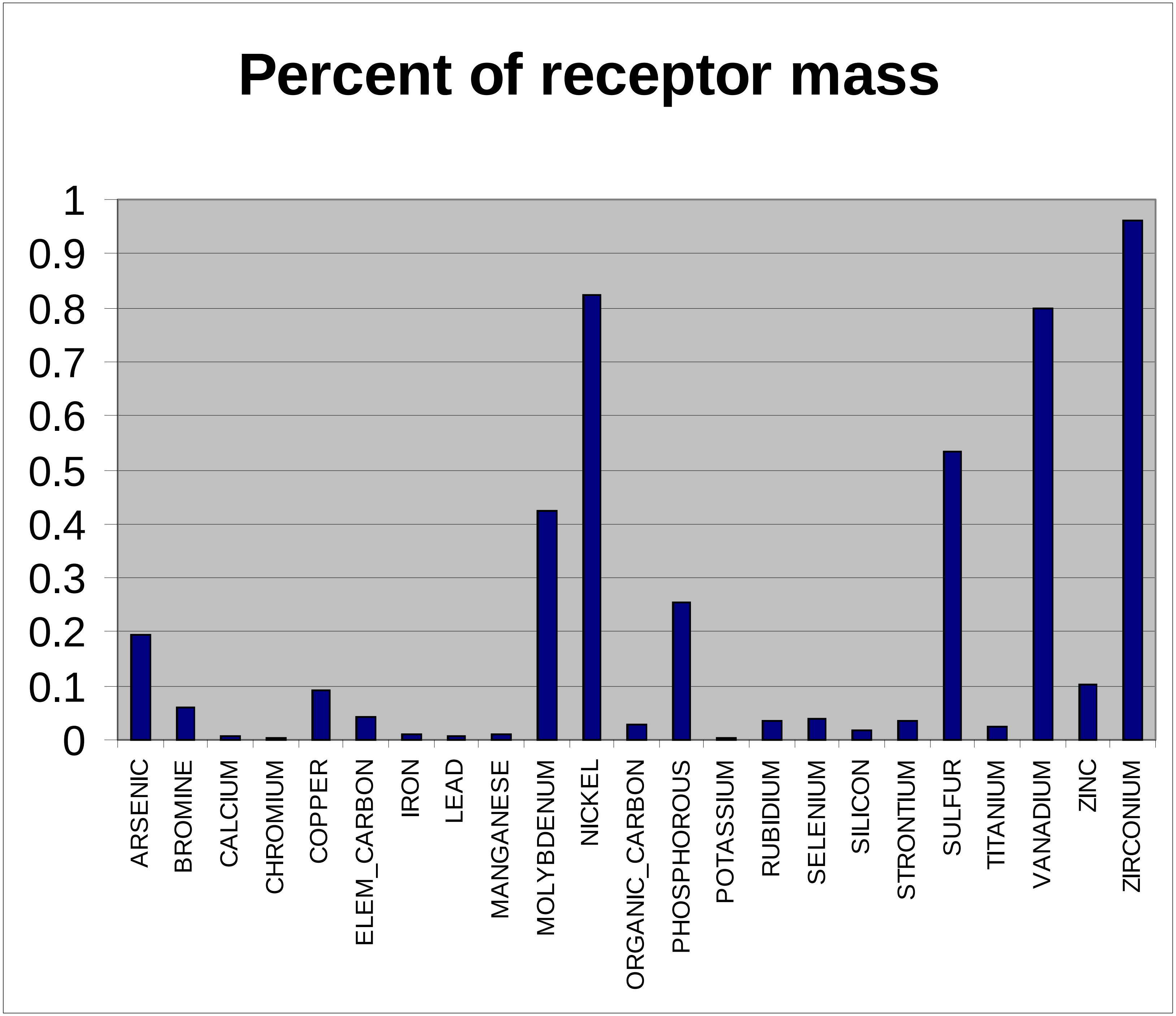
- Identifiability: We want a number such that something close to 0 means this is clearly identifiable as ...
- How close to: the profile matrix, a single profile, the contribution matrix, and/or the data matrix are we?
- Other?

How do we tell what a source is?

- Mathematical version: list / plot a source's make-up by the relative mass of each species. (The % source version.)
- Tracer version: list the important components of the source. But what is important?
- EPA version: list / plot percent of species mass due to a source. (The % species mass version.)

Percent of source mass





Percent of receptor mass

GOF for the profile matrix

- 2 versions: mean based / median based

  Both measure the relative error in the apportioned species mass from a source.

  = (Estimated species mass - true mass) over the average total mass of the species.

- F1= the root-mean-square of the these over the top 3 sources
- F2 = the median of the absolute values in relative error over the top 3 sources.

F1 - the mean based version

$$F1 = \sqrt{\frac{\sum_{\substack{\text{species i} \\ \text{top 3 sources j}}} \left[\frac{\left(\hat{F}_{i,j} - F_{i,j} \right)}{\text{the total average mass of species i}} \right]^2}{3(\text{the number of species})}}$$

Note: the F’s (the estimated and truth) are the mass of species i from source j.

F2 - the median based version

$$F2 = \underset{\substack{\text{species i} \\ \text{top 3 sources j}}}{median} \frac{\left|\hat{F}_{i,j} - F_{i,j}\right|}{\text{the total average mass of species i}}$$

Note: the F’s (the estimated and known) are the mass of species i from source j.

Profile GOF metric results

PMF

	Area	Roads	Residual Oil	Overall
F1	0.21173	0.077373	0.15582	0.1582
F2	0.02965	0.020977	0.00702	0.0147

UNMIX

	Area	Roads	Residual Oil	Overall
F1	0.22982	0.14356	0.052937	0.1594
F2	0.13709	0.12594	0.028796	0.0582

** The UNMIX fit is based on the expanded profile and contribution. The “expansion” is OLS not weighted!

Comments

- F1 is very sensitive to the largest relative errors (the worst part of the fit). Changes in the those can make a big difference.
- F2 is often representative of the first 3 quartiles.
- All species are treated equally.
  - No weighting! (We have seen that the errors tend to be correlated.)
- Estimates >100% of the average species mass are replaced with the average.

GOF for the contributions.

- Since the GOF for the profile is mass based. G1 measures the time series fit.
  - The contribution matrix is scaled to have a mean of 1 in each column. Each entry measures the sources contribution relative to that sources average.
  - Again only the top 3 sources are considered.

Contribution GOF

$$G = \frac{\sum_{\substack{\text{measurement periods i} \\ \text{top 3 sources j}}} \left(G_{i,j} - \hat{G}_{i,j} \right)^2}{3\left(\text{the total number of measurement periods}\right)}$$

The G’s are the relative (Estimated / known) source contributions = measurement period mass divided by the average for that source.

Additional check on contributions

- Each of the top 3 predicted scaled time series are regressed against the time series of the source that best matches.
  - The intercept and slope measure any bias,
  - The intercept should be ~ 0,
  - The slope should be ~ 1, and
  - r-squared is an alternate measure of GOF.

GOF to the raw data.

- The main object function for PMF measures the GOF of the model solution versus the raw data.
  - We modify it slightly by dividing by its expected value to make the number comparable across different problems and solutions.
  - This is clearly biased toward PMF.

The raw data GOF

$$Q = \sum_{i,j} \left(\frac{X_{i,j} - \hat{X}_{i,j}}{\sigma_{i,j}} \right)^2 \qquad X = \frac{Q}{\mathrm{df}}$$

$\sigma_{i,j}$ = the standard error of the $\mathrm{X}_{\mathrm{i,j}}$ measurment

df = the number of data points – the number of estimated parameters.

The X’s are the measured / predicted species mass seen at the receptor.

G and X GOF Results

PMF

	Area	Roads	Residual Oil	Overall
G	0.01	0.01	0.01	0.01
			Q	0.1610 x 11994

UNMIX

	Area	Roads	Residual Oil	Overall
G	0.33	0.57	0.20	0.36
			Q	1.9202 x 11994

** The UNMIX fit is based on the expanded profile and contribution. The “expansion” is OLS not weighted!

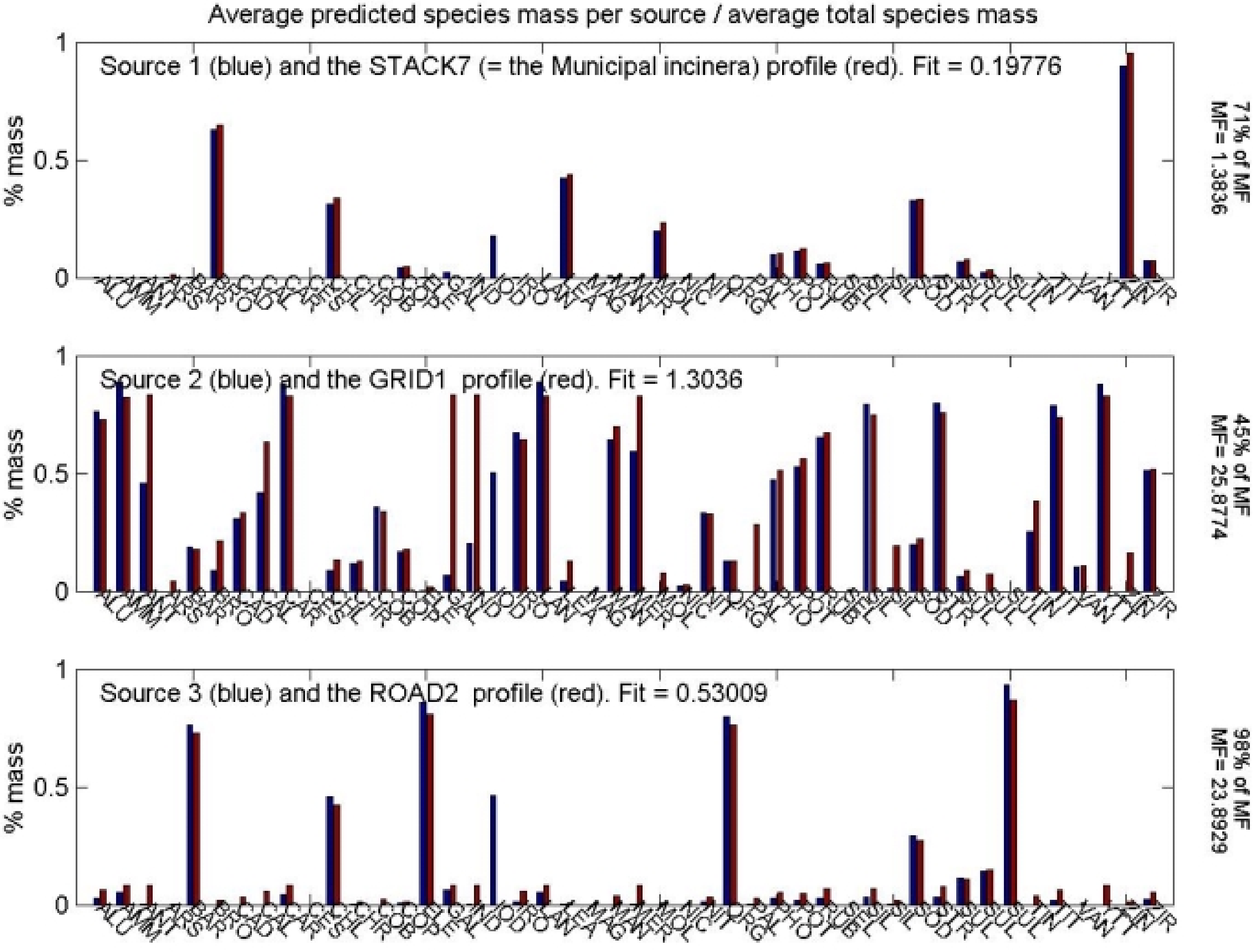
** Q is naturally broken down by species, not source.

Automated profile matching against known profiles.

- All permutations of the 3 largest sources are compared against the 3 largest predicted source profiles. The least overall F1 (F2) is used to declare the matching and the overall measurement of fit.
- # of matched profiles = # of time series that have an r^2 > .9 with a true time series.
- # matching based on r^2 is sometimes better.

General automated profile identification.

- Goal: Find an automated procedure that identifies the the output from one of these tools.
  - The smaller the #, the more likely that we have correctly identified a source. (There is no need to standardize.)
- Idea: Modify F1 to match individual profiles against a list of potential profiles


Average predicted species mass per source / average total species mass
Source 1 (blue) and the STACK7 (= the Municipal incinera) profile (red). Fit = 0.19776
% mass
1
0.5
0
71% of MF
MF= 1.3836
Source 2 (blue) and the GRID1 profile (red). Fit = 1.3036
% mass
1
0.5
0
45% of MF
MF= 25.8774
Source 3 (blue) and the ROAD2 profile (red). Fit = 0.53009
% mass
1
0.5
0
98% of MF
MF= 23.8929

The algorithm

- Speciate profiles can’t have an total mass. The predicted total mass is used as truth. Potential identifications are made assuming a source with the known profile has the predicted total mass.
  - species with estimates >100% the average species mass are lowered to the species average.
  - Unlike matching against known profiles, duplicate matches are allowed.
- List all the source types that have a fit that is within 20% of the best fit.

How well does it do?

- Sources 1-5 of the PMF solution are given the same identification Dr. Hopke.
- Source 6 is identified as a very poor fit to several alternatives, including Dr. Hopke’s identification as the lime kiln.
- Source 7 is very strongly identified as the missing source. (Not an area.)
- Sources 8 & 9 are given weak fits to several alternatives, none the same as Dr. Hopke’s solution.

Possible variations

- Weighting with
  - SE's from the tools.
  - MDL's (time below) and/or species uncertainties.
  - Species "importance."
  - Correlation within the errors may make this a bad idea. (Positive, not negative as implied by constraints.)
- Use medians or quartiles to reduce sensitivity to any outliers.

Conclusions

- The profile metrics have worked well.
  - They let one objectively identify sources without a knowledge of the chemistry.
  - They provide a systematic way of measuring the overall quality of the fit.
- The data metric has clearly been valuable for PMF.
- Other simulation results suggest that that we should pay more attention to correlation within the time series solutions.