Post-refinement Determination of the Flack Parameter

The correlation between the absolute structure of a crystalline material and the measured Friedel (anomalous) pairs can be expressed as

$$Io \approx Ic = (1-x)Im^{+} + xIm^{-} \qquad (1)$$

where Im^{+} is the value of F^2 for one of a Friedel pair of reflections computed from the structural model, and *Ic* is the value of F^2 allowing for twinning by inversion. *x* is the Flack parameter which takes the value zero when the model corresponds to the correct absolute structure for an enantio-pure material, and unity if the model needs inverting. The Flack parameter and its su can be determined like other structural parameters during least squares refinement. Examination of the values of the Flack parameters determined for many materials of known enantio-purity and absolute configuration shows that the parameter tends to zero, and rarely gives a false indications. This has lead to a search for methods for determining x more robustly than simply including it in the main least squares refinement, especially in cases where the anomalous signal is likely to be weak.

Writing

$$Do = Io^{+} - Io^{-}$$

$$Dm = Im^{+} - Im^{-}$$

$$Ao = 0.5(Io^{+} + Io^{-})$$

$$Am = 0.5(Im^{+} + Im^{-})$$

$$Qo = Do / 2Ao$$

we get

$$\mathrm{var}(Do) = \mathrm{var}(Io^{+}) + \mathrm{var}(Io^{-})$$

$$\mathrm{var}(Ao) = 0.25 * \mathrm{var}(Do)$$

$$\mathrm{var}(Qo) = \left[\frac{2}{\left(Io^{+} + Io^{-}\right)^2}\right]^2 \left[Io^{-2} * \mathrm{var}(Io^{+}) + Io^{+2} * \mathrm{var}(Io^{-})\right]$$

for $Io^{+} \approx Io^{-} = I = Ao$ and $\sigma^2(Io^{+}) \approx \sigma^2(Io^{-}) = \sigma^2(Io)$, we get

$$\mathrm{var}(Qo) = \mathrm{var}(Io) \Big/ \left(2 * Io^2\right)$$

and hence

$$Qo / \sigma(Qo) = Do / \sigma(Do)$$

and

$$Qo / \sigma^2(Qo) = 2.I.Do / \sigma^2(Do)$$

Equation (1) can be recast into several different forms:

$$Do = (1 - 2x)Dm \qquad (5)$$

Using arguments explained in Flack or Parsons or Thompson we can define the "Quotient" equations.

$$Do / 2Ao = (1 - 2x)Dm / 2Am \quad (6)$$

$$Qo = (1 - 2x)Qm \qquad (7)$$

Equations (5) and (7) can be solved for (1-2x) - and hence x - by conventional least squares, with or without a non-zero intercept, assuming that there is no error in *Dc* or *Qc*. The code for the least squares best line $y = a + bx$ is based on Wolfram (http://mathworld.wolfram.com/LeastSquaresFitting.html) and was verified against Watts & Halliwell, Essential Environmental Science, Routledge, 1996. The standard deviations of the intercept and slope are based on Wolfram World but including weights, derived by DJW.

Define

nitem = number of data pairs x and y, each assigned a weight (wt) equal to the inverse of the appropriate variance, x is the independent variable (*Dc, Qc*) and y the dependent variable (*Do, Qo etc*).

ss = Σwt
sx = Σx*wt
sxx = Σ x*x*wt
sy = Σy*wt
syy = Σ y*y*wt
sxy = Σ x*y*wt
denom = ss*sxx - sx*sx

Gradient for a line with zero intercept = sxy/sxx

If the absolute value of the denominator (denom) is greater than zero
a = (sy*sxx - sx*sxy)/denom
b = (ss*sxy - sx*sy)/denom

The standard uncertainties in a and b are:

sqs = ((syy*sxx)-(sxy*sxy)) / ((nitem-2)*sxx)
sa = (sx*sx)/(nitem*nitem*sxx) + (1./nitem)
σ(a) = sqrt(sa * sqs)

sb = (((syy*sxx)-(sxy*sxy))/((nitem-2)*sxx*sxx))
σ(b) = sqrt(sb)

The 100(1-α)% confidence intervals can be computed using the 1-α/2 quantile of a t variate with (n-2) degrees if freedom:

$$c(a) = a \pm t^*_{[1-\alpha/2,n-2]}\ \sigma(a)$$
$$c(b) = b \pm t^*_{[1-\alpha/2,n-2]}\ \sigma(b)$$

In crystallography, n-2 is usually very large, so the appropriate values of t* are

5%	2.5%	1.25%	0.5%	0.25%
1.64	1.96	2.24	2.58	2.81

(Analysis of Straight Line Data, F.S. Acton, Dover 1966)

The correlation coefficient (R, r) and coefficient of determination (r-sq, r^2) are computed as in:

r = r in Watts & Halliwell, page 111
rsq = r-sq in Excel, & W&H, page 112

and

t = t in Watts & Halliwell page 113
tsq = F-test in Excel

Re-defining a denominator:

```
denom=(ss*sxx-sx*sx)*(ss*syy-sy*sy)
```

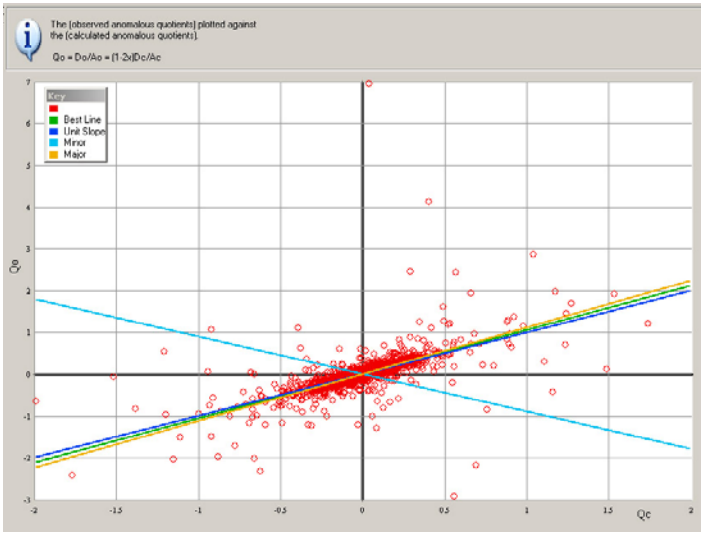
If the denominator is greater than zero

```
r   = (ss*sxy-sx*sy)/sqrt(denom)
rsq  = r * r
```

If ss is greater or equal to 2 and rsq less than or equal to 1,

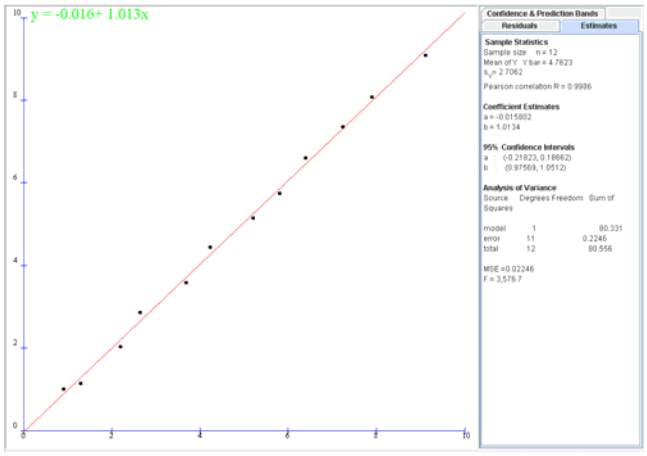
```
t  = r*sqrt(ss-2.)/sqrt(1.-rsq)
tsq =  rsq*(ss-2.)/(1.-rsq)
```

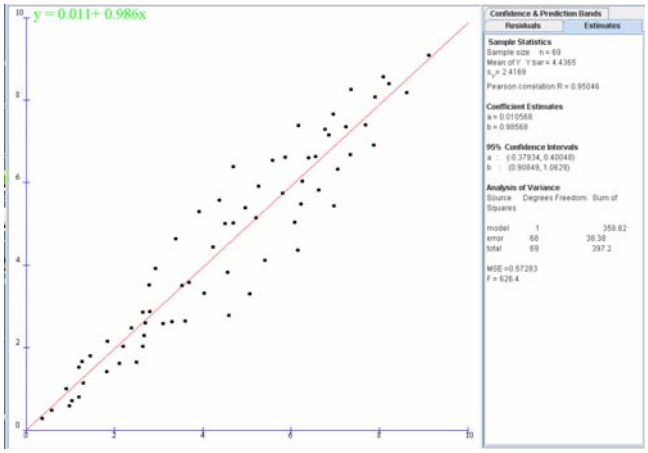
The diagram shows a scatter plot of *Qo* vs *Qm* for DECENT, and also shows the least squares best line and the line of unit gradient. The best line has a gradient of 0.940(5), the correlation coefficient is 0.956, and the coefficient of determination is 0.914.



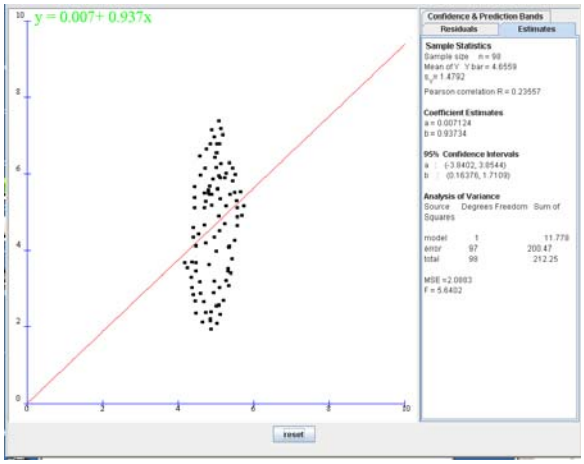
Except when the data points all lie on an exactly vertical line, it is always possible to fit a regression line. However, it may not always be a sound thing to do.

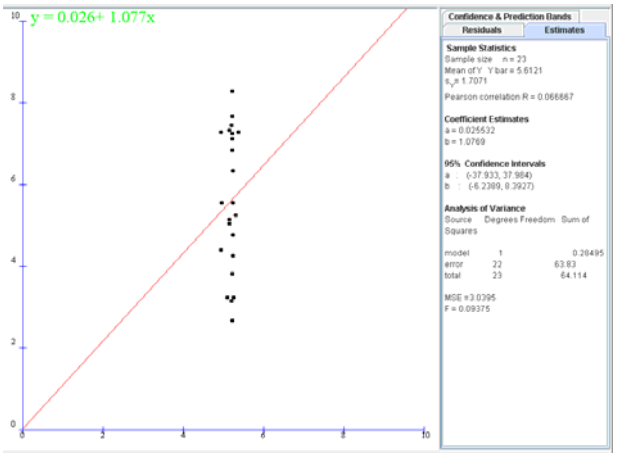
The following two plots, computed with http://www.math.csusb.edu/faculty/stanton/probstat/regression.html, correspond to a high degree of correlation with little noise (a), and slightly noisy data set (b).





The next pair show a very noisy data set (c), and one where it is unlikely that the 'dependent' variable actually depends on the control variable (d). In every case, the maths successfully 'fits' a straight line with approximately unit gradient.





Case	Pearson Correlation Coefficient	95% confidence levels on gradient	n
a	0.9986	±0.04	12
b	0.9505	±0.08	69
c	0.2356	±0.77	98
d	0.0667	±7.32	23

The correlation coefficient is independent of the number of observations, the standard uncertainty is proportional to 1/sqrt(n-2) so that the standard uncertainty can be reduced by adding in more "vanilla" data - the Emperor of China syndrome. The importance of a given datum is measured by its leverage (Prince ...). Since the mean values of *Do* and *Dm* (and the corresponding quotients) are close to zero, fitting a straight line can be regarded as a one-parameter model, so that the leverage of each data point is given by:

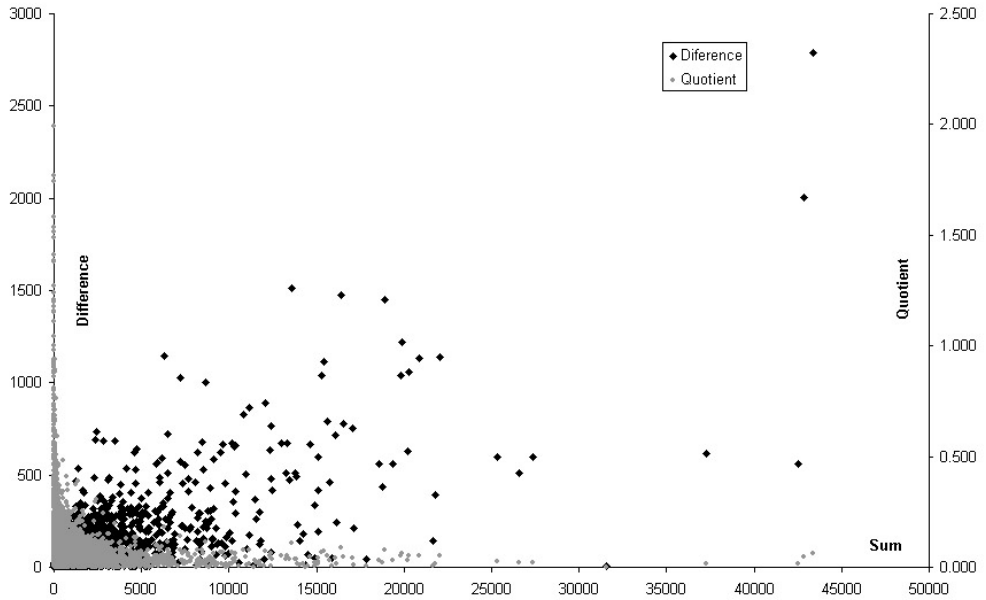
$$h_i = \frac{1}{n} + \frac{\left(x_i - \bar{x}\right)^2}{\sum\left(x - \bar{x}\right)^2} \qquad (8)$$

(The Statistical Sleuth, Ramsey & Schafer, Brooks/Cole, 2002, page 316), where x_i are the values of either *Dm* or *Qm* and $\bar{x}$ is the mean of *Dm* or *Qm*. The data with greatest leverage are those with large absolute values of *Dm* or *Qm*. Remember that *Dm* does not depend directly on *Am*

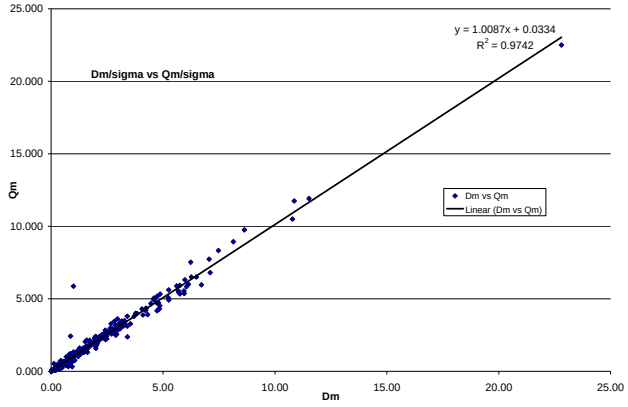
Excluding weak data can be particularly important when estimating the Flack *x* from quotients. Equation 6 can be recast as

$$Do = (1 - 2x)Dm.Ao / Am \qquad (9)$$

Since *Ao* = *Am* ±*error*, *Ao/Am* can take extreme values for small *Am*. In the usual case where <Ao> is very similar to <Am> (i.e. low conventional R factor) the *A* in equation 6 can be regarded as a weighting factor which down-weights strong reflections. This can be seen in plots of *Dm* vs *Am* and *Qm* vs *Am*



A plot of Dm/σ(Dm) vs Qm/σ(Qm) shows values lying on a fair straight line (equation 3) - the signal:noise is the same.



The least squares determination of (1-2x) from either differences or quotients is weighted by the inverse of the variances, so that the leverage, equation 8, has to be weighted. However, xpressing Q/var(Q) in terms of D and I (equation 4) we can see that the weighted quotient evaluation up-weights strong reflections compared to the difference evaluation.

Equation [5] can be rewritten as:

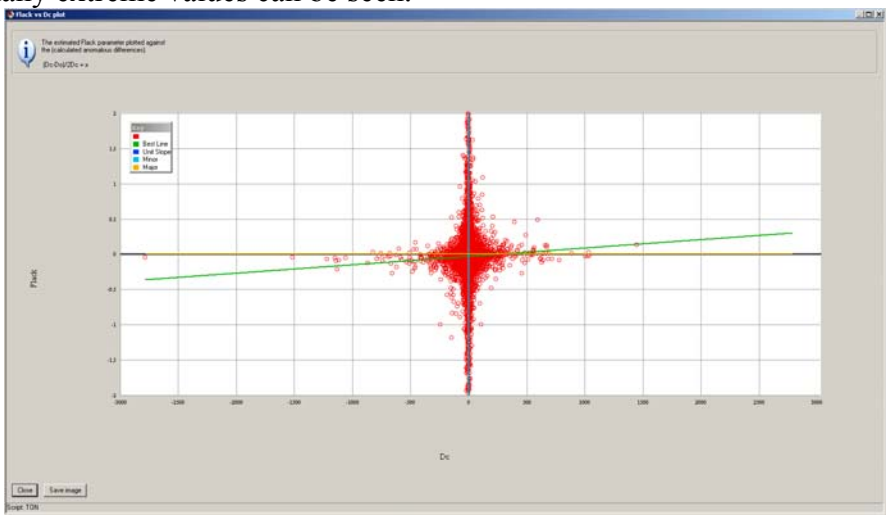
$$(Dm - Do) = 2xDm \qquad (10)$$

A plot of (*Dm*-*Do*) against *Dm* should be a straight line with gradient *2x*. If the model is of the correct hand for an enantio-pure material, the line should lie on the x-axis. The y-axis now shows the residuals as a function of *Dm*, and for a good model they should be largely noise. The quotient version (with small *Ao* and *Am* excluded) shows the residuals scaled inversely as the reflection intensity.

Finally, equation [5] can also be cast as:

$$x = (Dm - Do) / 2Dm$$

Plotting x against the right hand side should give a horizontal line at the value of the Flack parameter. If /*Dm*/ is very small compared to /*Do*/, the value of *x* can take extreme values. For a structure with low anomalous scattering, individual *x* can be massive, and even for good data many extreme values can be seen.



An alternative way of looking at the *Do-Dm vs Dm* scatter-plot is to treat it as a 2D collection of (possibly weighted) objects. The inertial tensor computed from these objects can be decomposed into a major and minor principal axes. For a good anomalous scatterer, the major axis will point roughly in the *x-axis* direction. For a poor anomalous scatterer, the major axis will point in the noise direction.

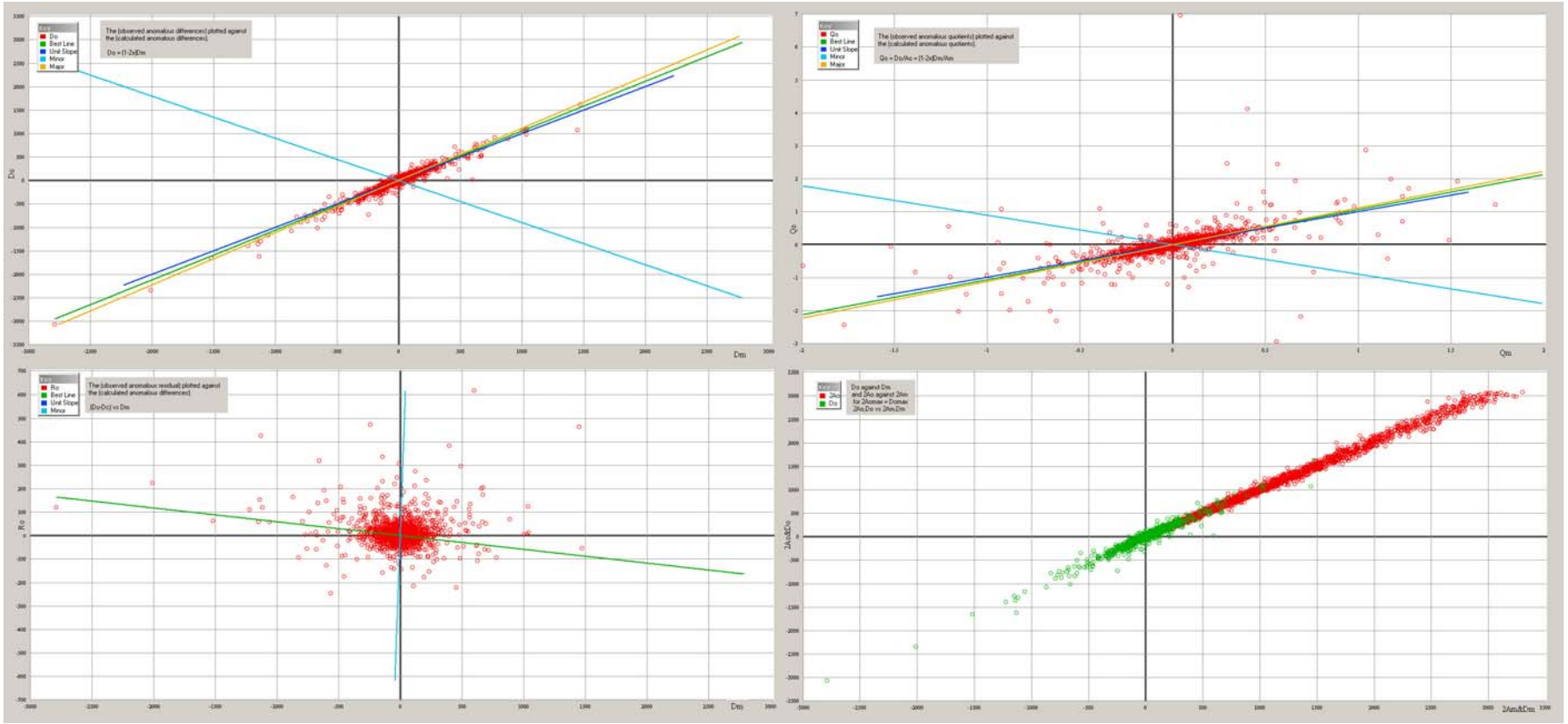
The current (March 2013) version of CRYSTALS implements equations 5, 7 and 10, together with Flack's *Do vs Dm* and *2Ao vs 2Am* plot (Analysing Friedel averages and differences. Simon Parsons, Phillip Pattison and Howard D. Flack. Acta Cryst. (2012). A68, 736–749) and the Normal Probability Plot (Abrahams, S.C.; Keve, E.T., Acta Cryst. 1971, A27, 157-165).

Five filters are provided (with default values) to exclude reflections which may either introduce instability into the calculations (very small denominators) or are suspected of being in serious error.
Reflections are accepted if:
/*Do*/ less than CRITER * /*Dm*/ (Ton's original filter for excluding unreasonable differences)
Ao is within *Am* ± Filter_1 * *Dm*/2 (Ao is reasonably similar to *Am*)
Ao is within *Am* ± Filter_2 % (*Ao* is within Filter_2 % of *Am*)
Am is greater than Filter_3 * sigma(*Ao*) (exclude small *Am* denominators)
Dm is greater than Filter_4 * sigma(*Do*) (exclude small differences)

In the following diagrams, the plots are:

(a) *Do* vs *Dm*	(b) *Qo* vs *Qm*
(c) Residual *Ro* vs *Dm*	(d) 2*Ao* vs 2*Am* (red) and *Do* vs *Dm* (green)

Good Anomalous Scatterer

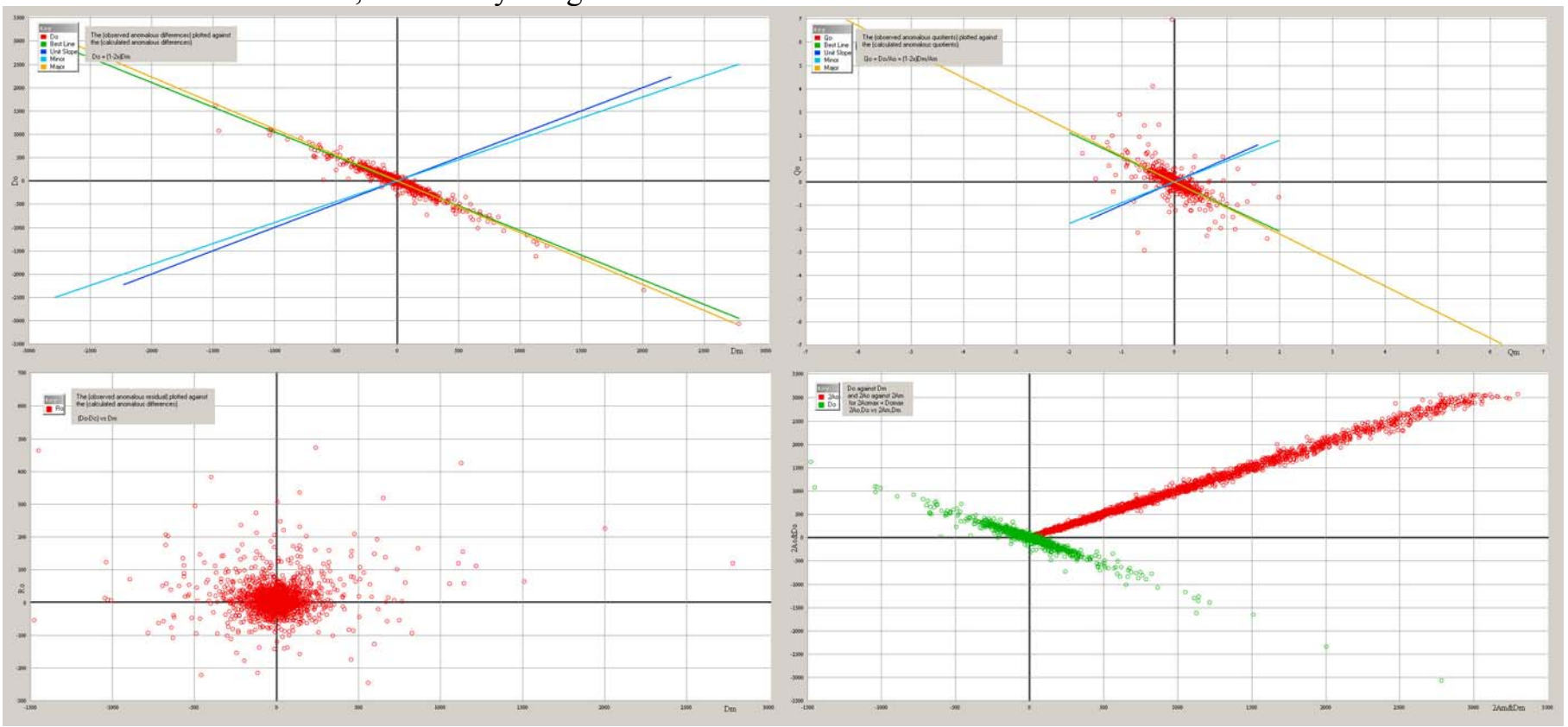


Formula: $C_{21}\ H_{23}\ Br_1\ N_2\ O_4$, Space Group $P2_1\ 2_1\ 2_1$
No of Reflections processed = 8640
No of Friedel Pairs found = 3831 No of Friedel Pairs used = 3824
No of Unpaired Reflections = 18 No of Centric Reflections = 960

RA	RD	wRA2	wRD2	Friedif	Flack	esd	Hooft y	esd	Do/Dm	esd	Qo/Qm	esd
2.8	28.0	3.8	29.7	497.51	-0.03	0.01	-0.025	0.002	-0.029	0.003	-0.031	0.003

Note that in (a) and (b) the best line (green) and major axis if inertia (orange) lie close to the unit gradient (blue). In (c) the residuals are small compared to *Dm*. Plot (d) shoes the anomalous differences(green) behaving much as the anomalous averages (red)

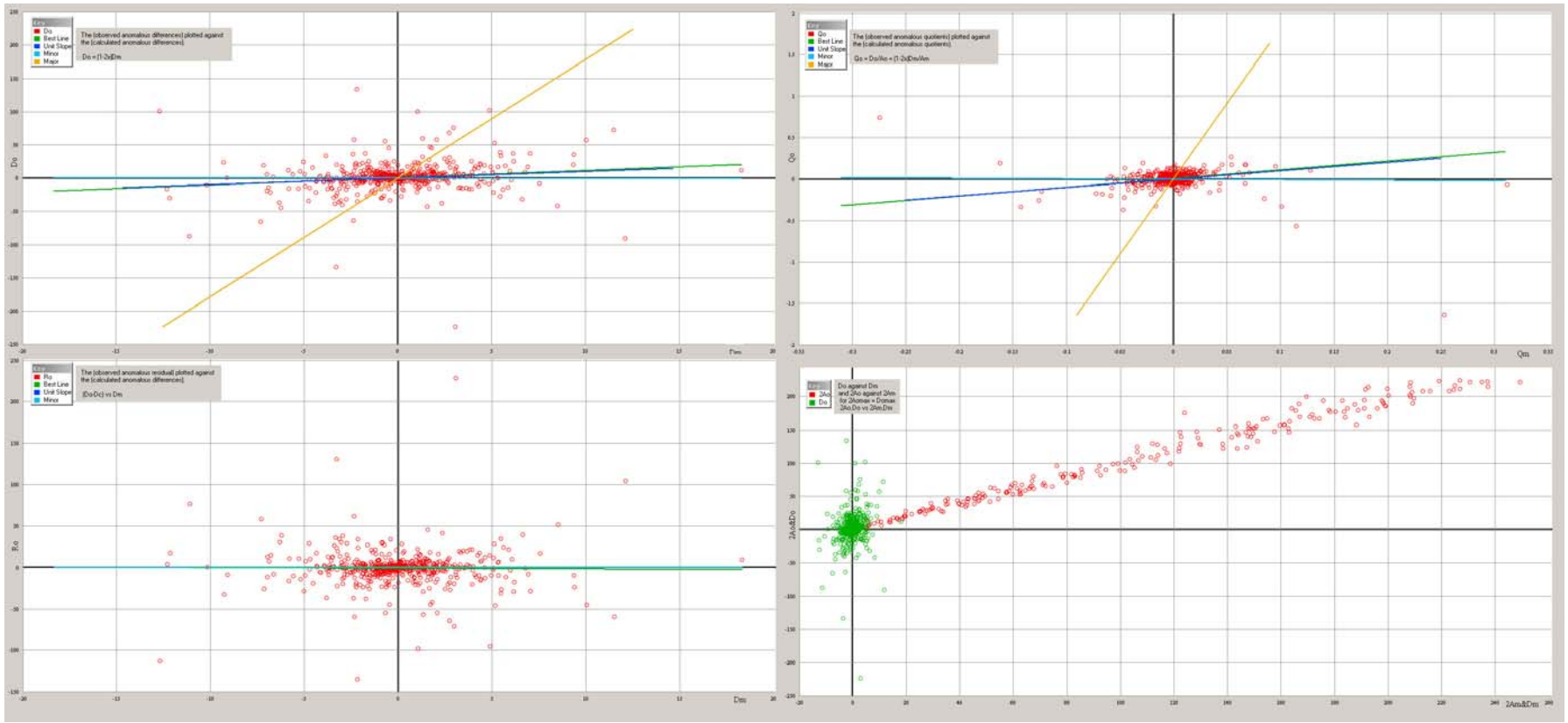
Good Anomalous Scatterer, incorrectly assigned



RA	RD	wRA2	wRD2	Friedif	Flack	esd	Hooft y	esd	Do/Dm	esd	Qo/Qm	esd
2.8	187.9	3.8	188.3	497.51			1.029	0.017	1.029	0.003	1.031	0.003

From plots (a), (b) and (d) it is evident that the model needs inverting.

Poor Anomalous Scatterer.



Formula: $C_8\ H_{16}\ N_1\ O_{4.50.}$ Space Group: C 2

No of Reflections processed = 1476

No of Friedel Pairs found = 650 No of Friedel Pairs used = 635

No of Unpaired Reflections = 50 No of Centric Reflections = 126

RA	RD	wRA2	wRD2	Friedif	Flack	esd	Hooft y	esd	Do/Dm	esd	Qo/Qm	esd
4.2	97.5	5.2	97.7	35.51	-0.003	0.164	-0.027	0.083	-0.040	0.086	-0.023	0.087

In plots (a) and (b) the vertical range (±250 and ±2.0) greatly exceeds the horizontal range (±20 and ±0.35). Plotted on equally scaled axes, the major axis (orange) is almost vertical - there is much more noise than potential signal.