UnitCell – What it is

UnitCell is a leastsquares refinement program to retrieve unit cell constants from diffraction data. The user supplies indexed reflections from crystal diffraction patterns.

The program is further described in:

T J B Holland and S A T Redfern (1997) "Unit cell refinement from powder diffraction data: the use of regression diagnostics". Mineralogical Magazine, 61: 65-77.

** New feature: a zeroshift (systematic small error in 2Theta or energy) may now be refined from the data, if desired.

Data file construction

Data input files have the form:

```
Enstatite
0.01
2 1 0  13.98
0 2 0  20.12
1 2 1  26.99
4 2 0  28.12
2 2 1  28.31
3 2 1  30.42
6 1 0  31.13
5 1 1  31.68
2 3 0  31.98
4 2 1  33.13
1 3 1  35.45
2 0 2  36.04
5 2 1  36.37
3 0 2  37.80
3 3 1  38.22
8 0 0  39.54
0 0 0
```

The first line is a title (<255 characters), the second line contains a representative uncertainty for the measurement (2theta, beam energy etc), and each subsequent line is the hkl and measurement (twotheta, dspace, or beam energy). The end of file is flagged by 0 0 0 for hkl.

••• A Note on Errors •••

Earlier versions of the program used weighted least squares, using a default uncertainty of 0.005 deg 2theta. However, in many cases this is too small, particularly with poorly crystalline or extremely fine-grained samples and values of sigmafit much greater than unity are calculated. In the past the user was advised to scale the uncertainties based on the value of sigmafit, but the current implementation is much more transparent in allowing the user to enter their best estimate of uncertainty in advance. If this has been estimated well then sigmafite should be close to 1.0 and no corrections are needed for the calculated uncertainties. It is the user's responsibility to interpret the uncertainties in the light of the value of sigmafit returned by the program, as indicated below.

How to evaluate uncertainties in cell parameters:

Examine the value of sigmafit in the output from UnitCell. If the fit is good and the uncertainties on the input are appropriate then sigmafit should be close to 1.0. If it is significantly larger than 1.0 (say 1.7 for example) then two possibilities exist:

a) there may be a poor fit to the data; in this case the residuals and regression diagnostics should be examined to see if any particular reflections are the source of the poor fit (see below)

b) the errors on the input data may be larger than the submitted estimate, in which case the errors on the cell parameters should be adjusted upwards in proportion. A doubling of the errors on 2theta leads directly to a doubling of cell parameter errors and a halving of sigmafit. Multiplying the errors by sigmafit yields the same result as an unweighted regression (and would return sigmafit to 1.0). There is no need to rerun the regression - with constant weights the cell parameters remain unchanged.

If sigmafit is less than 1.0, then the resulting uncertainties on cell parameters should not be adjusted downwards, unless you can robustly justify a case for input 2theta errors which are smaller than the uncertainties provided. It is probable that the estimate of input uncertainty has been overestimated.

••• A brief guide to using the Regression Diagnostics •••

The original reference on regression diagnostics is Belsley, Kuh and Welsh (1980) Regression Diagnostics: Identifying influential data and sources of collinearity. J Wiley. They were introduced to least squares (LSQ) problems in geology by Powell (1985) J. Met. Geol. 3, 231-243, and are briefly described there.

Regression diagnostics are numbers, calculated during the regression, which furnish valuable information on the influence of each observation on the least squares result and on the estimated parameters. Usually it is deletion diagnostics which are calculated, and these give information on the changes which would result from deletion of each observation from the regression. In the context of the least squares programs used here, the main diagnostics are briefly described below:

(in what follows n=number of observations and p=number of parameters)

- Hat. Hat values are listed for each observation and give information on the amount of influence each observation has on the least squares result. A hat value of 0.0 implies no influence whatever, whereas a hat of 1.0 implies extreme influence (that observation is effectively fixing one parameter in the regression). The sum of the hat values is equal to the number of parameters being estimated, so an average hat value is p/n. Hat values which are greater than a cutoff value of 2p/n are flagged as potential leverage points (highly influential).

- Rstudent. Ordinary residuals (y-ycalc) are not always very useful because influential data often have very small residuals. Rstudent is designed to take influence into account through division of the residual by sqrt(1-h). They are defined in Belsley et al 1980 and Powell 1985. A suitable cutoff for 95% confidence level is 2.0, any value of Rstudent above this magnitude may signify a potentially deleterious observation.

- Dfits. This is a deletion diagnostic involves the change in the predicted value of y upon deletion of an observation. The diagnostic printed gives the change in calculated y upon deletion of an observation as a multiple of the standard deviation of the calculated value. Values greater than the cutoff of 2sqrt(p/n) are to be treated as potentially suspicious.

- sig(i). This is simply the value that sigmafit would take on upon deletion of observation i. If this value falls significantly below the value for sigmafit (the standard error of the fit) then deletion of that observation would cause an improvement in the overall fit.

- DFbetai. The change in each fitted parameter upon deletion of observation i is flagged by this diagnostic. In the output it is given in terms of a percentage of the standard error. Observations which would cause any parameter to change by more than 30% of its standard error are flagged by the program as potentially suspicious.

The usefulness of diagnostics is that without re-running the regression it is possible to gain an understanding of which observations may be deleterious to the analysis. Outliers (large residuals) may not be a problem if they have a low influence (small hat). It may be a good strategy to remove the offensive observations sequentially until you are satisfied that the deleterious data have been removed. However, these are single-observation diagnostics and cannot detect deleterious effects of several observations acting together - there may be a masking effect.

Have fun!
Tim Holland. Updated 3 March 2021

(e-mail: tjbh@esc.cam.ac.uk)