

# Chapter 7.4

## PIXELstats

### Residual Density Analysis

Louis J. Farrugia  
Dept. of Chemistry  
University of Glasgow  
G4 9DS, Scotland, U.K.  
email: [louis@chem.gla.ac.uk](mailto:louis@chem.gla.ac.uk)

At the end of any crystallographic refinement, it is always appropriate to examine the residual electron density, to ensure that the applied model adequately describes all the information present in the experimentally measured structure factors  $F(\mathbf{H})_{\text{obs}}$ .

The residual electron density  $\Delta\rho(\mathbf{r})$  is defined as :

$$\Delta\rho(\mathbf{r}) = \rho_{\text{obs}} - \rho_{\text{calc}} = \frac{1}{V} \sum_{\mathbf{H}} (F(\mathbf{H})_{\text{obs}} - F(\mathbf{H})_{\text{calc}}) e^{-2\pi i \mathbf{H} \cdot \mathbf{r} + i\phi(\text{calc})} \quad (1)$$

The experimental amplitudes for  $F(\mathbf{H})_{\text{obs}}$  are used in the Fourier summation, but the phases  $\phi(\text{calc})$  are necessarily taken from the model. It is common practise to report just the extremes of the function  $\Delta\rho(\mathbf{r})$  over the whole unit cell (or if symmetry allows, a symmetry-unique sub-space of the unit cell). However the residual density should provide more information regarding the crystallographic analysis than just these two single numbers.

The experimentally determined residual density contains all the errors of the crystallographic experiment, from all sources, including (but not exclusively so)

- (i) all errors in the measured structure factor amplitudes, due to crystal defects (twinning, disorder), instrumental errors and misalignments, data processing errors including integration, absorption, extinction and thermal diffuse scattering errors.
- (ii) all model errors and inadequacies, such as scattering factors (spherical or aspherical), treatment of thermal motion, *etc.*
- (iii) errors in the computation of equation (1), such as Fourier truncation errors, inadequate sampling *etc.*

The program **PIXELstats** provides a number of statistical descriptors of the complete residual density function, which should be of diagnostic utility in deciding whether the applied model is appropriate. The program reads the grid files produced by the **WinGX** Fourier programs **SLANT-PLANE** and **FFT**, as well as XD grid files from **XDFOUR** or **XDFFT**. These should be difference Fourier 3D grid files, computed over the whole unit cell (this is automatically done in the case of the FFT algorithm). Some care should be taken to ensure that the computed map is sufficiently accurate, particularly :

- (i) the digital resolution of the map (*i.e.* the separation between grid points) should be sufficiently fine to satisfy the Nyquist-Shannon sampling condition [1]. A grid sampling of at least 1/2 that of the  $d$ -spacing resolution of the measured structure factors is necessary to avoid aliasing errors, and it is better to err on the side of caution. So if the resolution of the data is 0.6 Å, it is advisable to use a grid spacing of 0.2 Å or smaller along all axes. Once the Shannon limit has been reached however, there is no further advantage in reducing the grid size.

The program **PIXELstats** plots the binned distribution of the values of the pixels in the grid-file, and computes standard statistical descriptors, including the

population mean  $\mu$ , the population standard deviation  $\sigma$ , the sample skewness  $\mu_3$  and the kurtosis  $\mu_4$

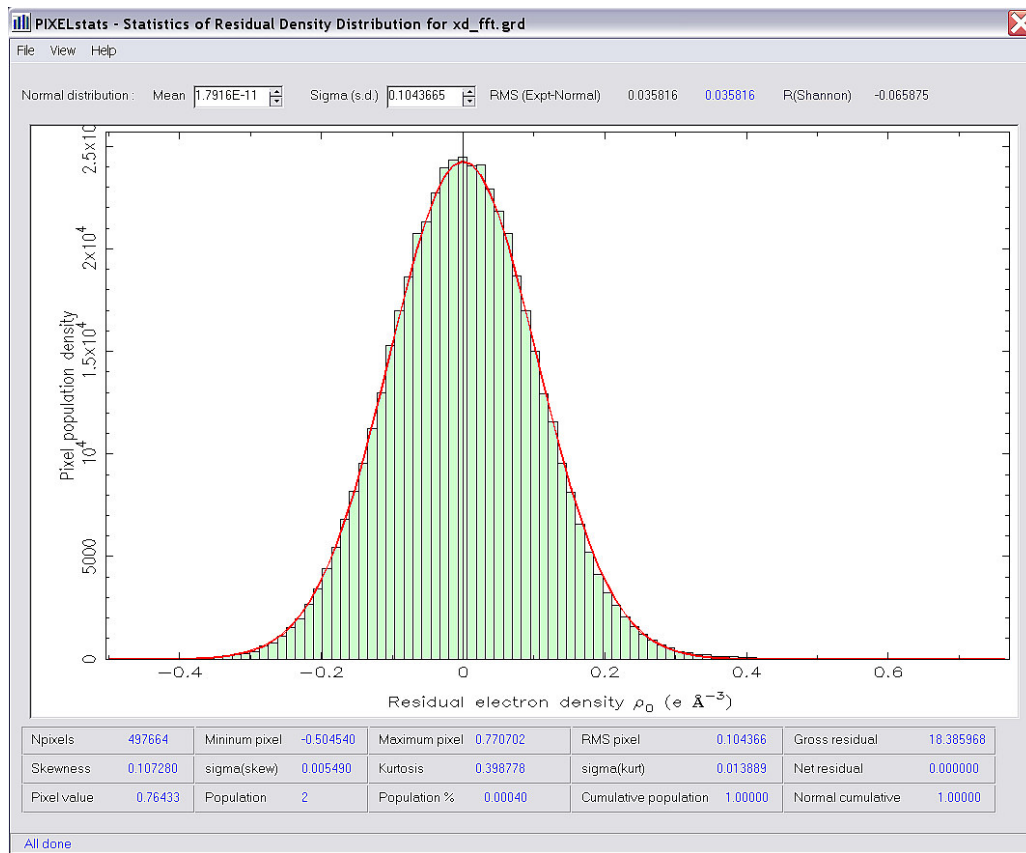
$$\mu = \frac{1}{N} \sum_{j=1}^N x_j \quad (2)$$

$$\sigma = \sqrt{\mu_2} \text{ where } \mu_2 = \frac{1}{N-1} \sum_{j=1}^N (x_j - \mu)^2 \quad (3)$$

$$\mu_3 = \frac{1}{N} \sum_{j=1}^N \frac{(x_j - \mu)^3}{\sigma^3} \quad (4)$$

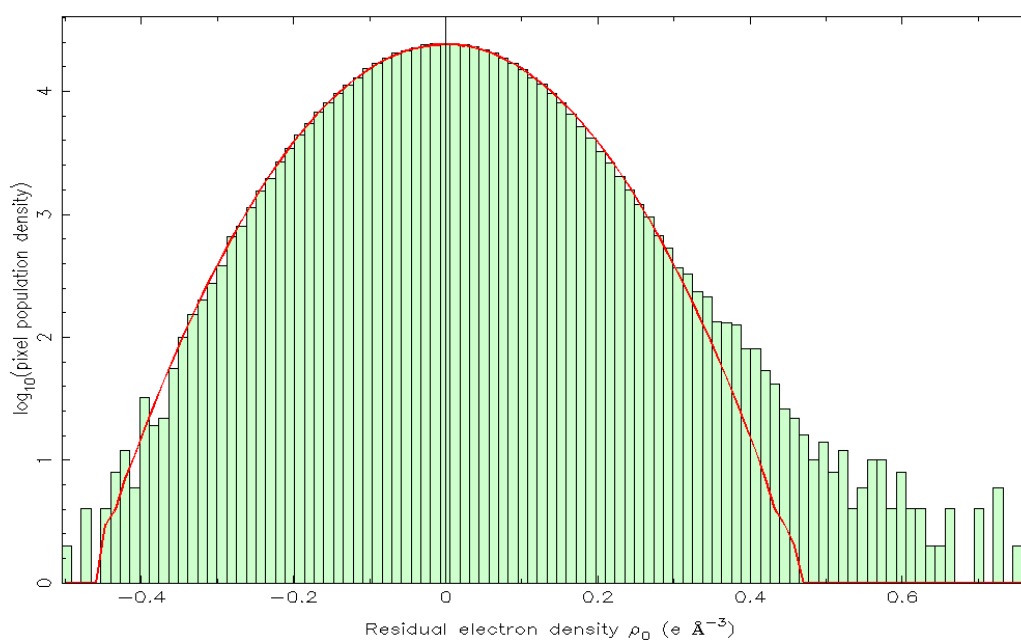
$$\mu_4 = \frac{1}{N} \sum_{j=1}^N \frac{(x_j - \mu)^4}{\sigma^4} - 3 \quad (5)$$

For a correctly calculated difference Fourier map, computed over the entire unit cell, the population mean should be exactly zero, regardless of any errors present in the data or model. Both the *XDFOUR* and *XDFFT* programs as well as *FFT* in *WinGX* give means which are normally  $< 10^{-10}$ , indicating this condition is met satisfactorily. Moreover, unless the contribution of  $F(000)$  is included, *any type* of Fourier map computed over the entire unit cell will have a mean of zero.



As well as showing the pixel distribution, *PIXELstats* also displays the (normalised) normal (Gaussian) distribution, corresponding to the computed sample  $\mu$  and  $\sigma$  [(2)

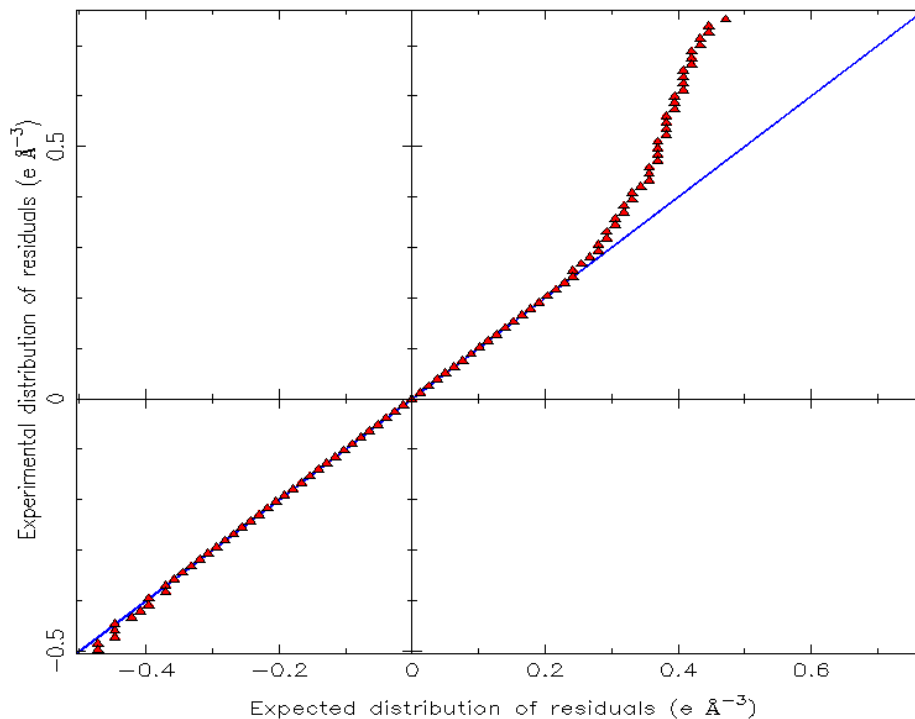
and (3)], in red. In the ideal case, where all features of the density have been modelled (apart from the random noise in the data), then the pixel distribution should correspond to a normal distribution. If this situation pertains, then it can be said that the data contains no further information than is implied by the model. Of course, since a model may incorporate physically non-meaningful features, this condition can never be a *sufficient* condition for a satisfactory crystallographic refinement, but can be construed as a *necessary* one.



In all experimental difference maps examined so far, it is observed that tails, out with the normal distribution are observed. However, the integrated electron population in these regions is normally very low, especially for those pixels at the extremes of the distribution, and their significance may be questionable. It is generally easier to examine these tails by plotting the  $\log(\text{population density})$ , as shown above. A right mouse press at any position to the right of the central zero line will show the integrated electron population *in excess of a normal distribution*, from that point onwards, to the *end* of the distribution (a similar mouse press to the left of the zero line shows the same, but to the *start* of the distribution).

There are several standard statistical tests that can be used to quantify how close an actual distribution is to a normal distribution. Two of the best known are the  $\chi^2$  test and the Kolmogorov-Smirnov test. Both of these are implemented in **PIXELstats**, but are of limited use, since the sample size (*i.e.* the number of pixels) is generally extremely large, typically 0.5 - 1.5 Mpixels, and because of this, both methods typically estimate a zero probability for a normal distribution. Another visual method is the so-called normal-probability plot of Abrahams and Keve [2]. The deviations of the experimental cumulative distribution are compared with the (expected) normal distribution. In the ideal case of a perfect normal

distribution, all points should lie on the line with slope = 1, with a zero intercept. Deviations of the experimental points from this line show how well the approximation to a normal distribution is achieved. It should be noted that the test is very sensitive in the tail regions away from the central zone, since a normal (Gaussian) distribution rapidly falls to extremely low probabilities. In addition, as emphasised above, the density of points in these regions is very low, and their significance should not be overestimated. The plot illustrated below shows in fact an exceptionally good agreement for a normal distribution, apart from a minor disagreement for residual densities greater than  $0.3e\text{\AA}^{-3}$

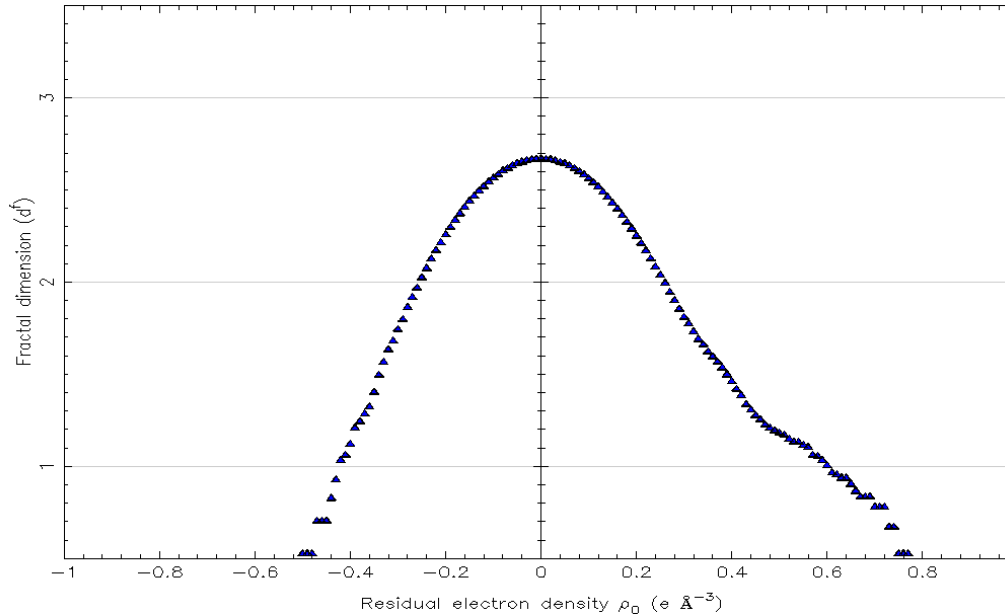


### Fractal Dimension

The Minkowski-Bouligand fractal dimension of the residual iso-density surface at a constant value  $x$  is defined as

$$d^f(x) = \lim_{\varepsilon \rightarrow 0} \frac{\log N(x, \varepsilon)}{\log(1/\varepsilon)} \quad (6)$$

This function has been proposed by Meindl & Henn [3] as a pertinent characteristic of the residual density distribution. Expression (6) is evaluated in *PIXELstats* using a modified box-counting algorithm, where  $\varepsilon$  is the *characteristic length* of the box used. This length  $\varepsilon$  is effectively a scaling factor, to give a maximum value for  $d^f$  of 3.0 for a 3-dimensional map and 2.0 for a 2-dimensional map. The fractal dimension plot visually provides essentially the same information as the log(probability density) distribution plots, and an ideal normal distribution appears as a paraboloidal curve.



### Shannon Information Entropy

The Shannon Information Entropy  $H(x)$  [4] of a distribution is given by

$$H(x) = -\sum_{i=1}^n p(x_i) \log_b p(x_i) \quad (7)$$

It provides a measure of the *information content* of the distribution, and is here used as a simple figure of merit to describe and compare the experimentally observed pixel distribution with the assumed normal distribution (based on the *mean*  $\mu$  and *standard deviation*  $\sigma$  of the experimental distribution). The  $p(x_i)$  used in (7) above are the normalised probabilities for the observed distribution in each bin and the corresponding normalised Gaussian probabilities for that bin. The base  $b$  of the logarithm used in **PIXELstats** is 10. The  $R(\text{Shannon})$  index quantifies the comparison, and is defined as

$$R(\text{Shannon}) = \frac{100 \times \{H(\text{Expt}) - H(\text{Normal})\}}{H(\text{Normal})}$$

Typically, good fits have absolute values of  $R(\text{Shannon})$  indices less than 0.1%

### References

1. H. Nyquist (1928) *Trans AIEE*, **47**, 617-644
2. S. C. Abrahams and E. T. Keve (1971) *Acta Cryst* **A27**, 157-165.
3. K. Meindl and J. Henn (2008) *Acta Cryst*, **A64**, 404-418.
4. (a) C. E. Shannon (1949) *Proc. Institute Radio Engineers*, **37**, 10-21; (b) C. E. Shannon (1948). *Bell System Tech. J.*, **27**:379-423, 623-656. <http://cm.bell-labs.com/cm/ms/what/shannonday/paper.html>.