Acta Crystallographica Section A
Foundations and Advances
ISSN 2053-2733

Received 16 September 2013
Accepted 14 January 2014

About systematic errors in charge-density studies

Julian Henn[a]* and Kathrin Meindl[b]

[a]Emil-Warburg-Weg 6, 95447 Bayreuth, Germany, and [b]Instituto de Biología Molecular de Barcelona (IBMB-CSIC), Barcelona Science Park, Baldiri Reixach 15, 08028 Barcelona, Spain. Correspondence e-mail: julian.henn@uni-bayreuth.de

The formerly introduced theoretical R values [Henn & Schönleber (2013). *Acta Cryst.* A**69**, 549–558] are used to develop a relative indicator of systematic errors in model refinements, $R^{\rm meta}$, and applied to published charge-density data. The counter of $R^{\rm meta}$ gives an absolute measure of systematic errors in percentage points. The residuals $(I_{\rm o} - I_{\rm c})/\sigma(I_{\rm o})$ of published data are examined. It is found that most published models correspond to residual distributions that are not consistent with the assumption of a Gaussian distribution. The consistency with a Gaussian distribution, however, is important, as the model parameter estimates and their standard uncertainties from a least-squares procedure are valid only under this assumption. The effect of correlations introduced by the structure model is briefly discussed with the help of artificial data and discarded as a source of serious correlations in the examined example. Intensity and significance cutoffs applied in the refinement procedure are found to be mechanisms preventing residual distributions from becoming Gaussian. Model refinements against artificial data yield zero or close-to-zero values for $R^{\rm meta}$ when the data are not truncated and small negative values in the case of application of a moderate cutoff $I_{\rm o} > 0$. It is well known from the literature that the application of cutoff values leads to model bias [Hirshfeld & Rabinovich (1973). *Acta Cryst.* A**29**, 510–513].

1. Introduction

In a former publication theoretical residual values (R values) were introduced (Henn & Schönleber, 2013). These are also called predicted R values as they give the attainable agreement factors from the experimental data and from the degrees of freedom of a model that is fitted to the data. The model parameters need not be specified; it is sufficient to know the number of parameters. For example, when N atoms are known to be in the structure and when it is planned to refine for each atom the coordinates and isotropic displacement factors, the number of parameters is $4N$.[1] It was shown that the weighted R value always increases when a weighting scheme is applied, which increases the experimental s.u.'s (standard uncertainties). It was also shown that the application of a weighting scheme leads to meaningless values of the goodness of fit (GoF) when the weighting-scheme parameters are adjusted to force values of the GoF to be close to 1 in applications to published experimental data. In the present work we develop these new measures further and apply them to charge-density studies.

[1] Symmetry considerations such as those implied on the coordinates and thermal motion parameters from a special position, and simplifications due to riding models for thermal motion of hydrogen atoms and other effects like disorder are neglected in this little illustrative example.

2. $R^{\rm meta}$, a measure of systematic errors

According to the definition of the IUCr there is no such thing as a *systematic error* in observations: '*Systematic error*: Contribution of the deficiencies of the *model* to the difference between an *estimate* and the true value of a quantity. [...] Note that, strictly speaking, observations are free of systematic error, in contrast to the *model* used for their interpretation' (Schwarzenbach *et al.*, 1989). Of course, 'model' comprises here all kinds of models used in the process, which includes models of absorption, extinction, of anomalous scattering and so forth. In the present work the term 'systematic error' is used when there is a systematic mismatch between model and data without assigning it to either. The important point is that the residuals are Gaussian distributed when no systematic errors apply.

Data descriptors and quality indicators. According to our understanding, many of the entities termed in the literature 'quality indicator' are actually 'data descriptors'. The difference between a quality indicator and a data descriptor is that in the latter case a numerical value is given without a reference, whereas in the former case a reference value is also specified or implicitly assumed. Therefore, all quality indicators are also data descriptors but not all data descriptors are quality indicators. The most prominent example are the R values from a refinement of model parameters against

Table 1
Definitions (left column) and predictions (right column) of R values.

F_o and F_c generally denote the non-negative structure-factor amplitudes. $\sigma(I_o)$ is abbreviated by σ to keep the notation short in equations (2), (4), (8) and (9); when referring to structure-factor amplitudes $\sigma(F_o)$ is used in equations (6), (11) and (12). The factor $\alpha = (N_{ref} - N_{par})/N_{ref}$, with the number of reflections used in the least-squares refinement, N_{ref}, and the number of independent model parameters, N_{par}, takes the degrees of freedom into account. See the text for more information.

Definition		Prediction	
$R_2 = \left[\frac{\sum (I_o - I_c)^2}{\sum I_o^2}\right]^{1/2}$	(1)	$R_2^{pred} = \left(\alpha \frac{\langle \sigma^2 \rangle}{\langle I_o^2 \rangle}\right)^{1/2}$	(2)
$R_{F^2} = \frac{\sum \lvert I_o - I_c \rvert}{\sum \lvert I_o \rvert}$	(3)	$R_{F^2}^{pred} = \left(\frac{2}{\pi}\right)^{1/2} \frac{\alpha \langle \sigma \rangle}{\langle \lvert I_o \rvert \rangle}$	(4)
$R_F = \frac{\sum \lvert F_o - F_c \rvert}{\sum \lvert F_o \rvert}$	(5)	$R_F^{pred} = \left(\frac{2\alpha}{\pi}\right)^{1/2} \frac{\langle \sigma(F_o) \rangle}{\langle F_o \rangle}$	(6)
$wR(F^2) = \left[\frac{\sum w(I_o - I_c)^2}{\sum w I_o^2}\right]^{1/2}$	(7)		
$wR(F^2)\lvert_{w=1/\sigma^2} = \left[\frac{\sum \frac{1}{\sigma^2}(I_o - I_c)^2}{\sum \frac{1}{\sigma^2} I_o^2}\right]^{1/2}$	(8)	$wR(F^2)\lvert_{w=1/\sigma^2}^{pred} = \left(\frac{\alpha}{\left\langle \frac{I_o^2}{\sigma^2} \right\rangle}\right)^{1/2}$	(9)
$wR(F) = \left(\frac{\sum w \lvert F_o - F_c \rvert^2}{\sum w F_o^2}\right)^{1/2}$	(10)		
$wR(F)\lvert_{w=1/\sigma^2(F_o)} = \left[\frac{\sum \frac{1}{\sigma^2(F_o)} \lvert F_o - F_c \rvert^2}{\sum \frac{1}{\sigma^2(F_o)} F_o^2}\right]^{1/2}$	(11)	$wR(F)\lvert_{w=1/\sigma^2(F_o)}^{pred} = \left[\frac{\alpha}{\left\langle \frac{\lvert I_o \rvert}{\sigma^2(F_o)} \right\rangle}\right]^{1/2}$	(12)

experimental data, which are always affected by noise. This R value is not a quality indicator in a strict sense, as it only measures the degree of agreement between experimentally and theoretically derived entities, but it does not tell which degree of agreement is to be expected in the first place. It is a quality indicator in a soft sense, however, because it is assumed that the actual R value gives the order of magnitude attainable for this data set. But this remains a working hypothesis as long as an attainable R value, given the data, is not specified.

For this reason, the theoretical R values were developed. They may serve both as a quality indicator for the experimental data set and as a quality indicator for a fit of model parameters to this set. In the first case, the actual value is the theoretical R value and the reference value is zero, which corresponds to the limiting case of a data set with vanishing noise. In the second case the actual value is the R value from a refinement and the reference value is the theoretical R value, which takes into account the noise in the experimental data, as specified by the experimental s.u.'s.

The definitions of R values and the corresponding predicted R values are compiled in Table 1. The F-based values are only included for the principle of completeness and for enabling one to evaluate cases in which only F-based agreement factors are available or when refinements against F values were performed. The predicted values for F-based residual factors [equations (6) and (12)] are, however, more accurate when using the *XD* convention of negative values for F_o in the case of negative intensity observations (Volkov *et al.*, 2006) in the corresponding definitions [equations (5) and (11)], as this corresponds better to the assumed Gaussian distribution. A residual factor using the *XD* convention will be larger than or equal to a residual factor employing the usual convention of setting $F_o = 0$ for negative intensity observations.

As an example of the information content of the predicted R values, the $wR(F^2)\lvert_{w=1/\sigma^2}$ value is briefly discussed. Note that, according to its definition in equation (8), the *de facto* $wR(F^2)\lvert_{w=1/\sigma^2}$ value is independent of the scale of the experimental s.u.'s, *i.e.* it does not matter for this R value and for the model parameter values if all s.u. values are too small or too large by the same factor. An important consequence of this observation is that this R value cannot be changed by using a scaled set of s.u.'s; therefore it relies on the relative accuracy of the s.u.'s to each other, rather than on their absolute accuracy. The corresponding prediction in equation (9), however, is not independent of the scale of s.u.'s as can be seen from the mean-squared significance which enters the denominator.

For evaluating the degree of systematic errors in a refinement one may combine *de facto* R values according to the left column in Table 1 and the corresponding predicted R values from the right column in Table 1 in one entity, R^{meta}, in the following way:

$$R^{meta} = \frac{R^{de\,facto} - R^{pred}}{R^{de\,facto}}. \qquad (13)$$

That R value is chosen which corresponds to the minimized residual sum. R^{meta} is positive when $R^{de\,facto} > R^{pred}$, and indicates that either the σ values are too small or that systematic errors apply, or both. In this case it is recommended that one consults the GoF to verify that it is close to 1.[2] It is also helpful

[2] The GoF is meaningful only if no weighting scheme with parameters that are adjusted to the data in order to bring χ^2 close to 1 independent of intensity and resolution shells has been applied. A simple factor for multiplication of all s.u.'s, for example, is unproblematic.

to evaluate the distribution of residuals. If it is reminiscent of a Gaussian, a Gauss function may be fitted to the distribution, resulting in a more appropriate value σ_0 for the standard uncertainty of residuals. This value may be used to scale the s.u.'s from the reflection file. If, on the other hand, the distribution of residuals is not close to a Gaussian, this is evidence for the presence of systematic errors.

Negative values for $R^{\rm meta}$ appear when $R^{de\,facto} < R^{\rm pred}$. This indicates that the s.u.'s are too large or that overfitting takes place, or both. Again, the GoF may be consulted for the scale of s.u.'s and the residual distribution should be evaluated and, if close to a Gaussian, a Gauss function may be fitted to the residuals in order to obtain a realistic value for the experimental s.u.'s. If the distribution of residuals is not close to a Gaussian, this is again evidence for systematic errors.

In both cases it is recommended that one evaluates the distribution of residuals and its closeness to a Gaussian probability density function. This can be achieved with the help of a χ^2 test and is discussed in more detail in §5.

Under ideal circumstances, $|R^{\rm meta}|$ is a small number. The expected deviation of this small number from zero may be determined empirically with the help of a Monte Carlo simulation. Values of $|R^{\rm meta}|$ close to zero do not prove the data to be free of systematic errors. For this a Gaussian distribution of residuals and statistical independence of the residuals are additional necessary conditions. However, large values of $|R^{\rm meta}|$ disprove the data to be free of systematic errors and data free of systematic errors are expected to show small $|R^{\rm meta}|$ values.

Table 2
Statistical descriptors of the experimental data sets.

From left to right: No. of data set, No. of model parameters in the least-squares refinement (literature/cif), No. of all reflections and observed reflections in the data set as given in the literature, No. of all reflections as counted from the reflection files, conventional R factor as given in the literature, conventional R factor as calculated from the reflection file according to equation (5).

Data set	N(par)	$I_{\rm all}/I_{\rm obs}$ (literature)	$I_{\rm all}$	R (literature)	R
1	612	12801/9801	12803	0.0368	0.0364
2	975/288	15247/15219	15247	0.0298	0.0298
3	850	28457/26324	28457	0.0226	0.0226
4	130	3822/3603	3822	0.014	0.0140
5	176	5136/4288	5136	0.025	0.0238
6	64	5146/4101	5146	0.033	0.0298
7	64	3551/2707	3551	0.034	0.0289
8	754	10651/8057	8057	0.0145	0.0145
9	604	10651/8057	8070	0.0149	0.0149
10	754	12304/7986	7986	0.0166	0.0166
11	604	12304/7986	7981	0.0182	0.0182
12	859	5403/3943	3942	0.0100	0.0100
13	604	5403/3943	3942	0.0178	0.0178
14	79	8630/7006	8630	0.0308	0.0308
15	224	4737/	4737	0.0128	0.0128
16	219	4737/	4737	0.0127	0.0127
17	256	2652/2565	2652	0.0400	0.0400
18	256	2652/2565	2652	0.0220	0.0220
19	256	2652/2565	2652	0.0219	0.0219
20	74	8015/6332	8015	0.0328	0.0329
21	170	2409/1433	2409	0.0484	0.0487
22†	179	2409/1433	2409	0.0494	0.0705
23	170	2409/1433	2409	0.0484	0.0487

References: (1) Bibila Mayaya Bisseyou *et al.* (2012), (2) Paul *et al.* (2011), (3) Holstein *et al.* (2010), (4–7) Mondal *et al.* (2012), (8–13) Zhurov *et al.* (2011), (14–16) Janicki & Starynowicz (2010), (17–19) Domagała *et al.* (2011), (20) Dittrich *et al.* (2012), (21, 23) Li *et al.* (2010), (22) Palatinus *et al.* (2006). † Reflection file taken from Li *et al.* (2010), cif file taken from Palatinus *et al.* (2006) as the information given in the cif file corresponding to Li *et al.* (2010) was incomplete.

3. The experimental data

The experimental data were taken from a search on the IUCr webpage (http://journals.iucr.org). Those first nine publications presented by the search system for the search term 'charge density, cif' on the 17th June 2013 were taken that included theoretical and experimental structure-factor amplitudes or intensities. This resulted in the 23 data sets shown in Table 2.

The studies and data sets are of very different character: they include classical charge-density studies of a certain compound, often in combination with methodological developments like the application of charge-density database information in different forms in order to allow for charge-density studies in the case of low resolution (1, 17–19) or in the presence of disorder (3). A study of similarities of molecular moieties with the help of the $R_{\rm free}$ (2) is included, as well as a study on the dynamic density of multipole models at different temperatures (4–7), studies on the effect of anharmonic nuclear motion in explosives at different temperatures (8–13), studies on the chemical bond (14–16) of a phosphorus-containing compound, an experimental and theoretical multipole database transfer study applied to a plant flavonoid (17–19), studies on crystal-field effects (20) and on modulation functions of incommensurately modulated structures with the help of the maximum entropy method (21–23).

The number of refined model parameters was taken from the publication text and compared to the entry in the cif (crystallographic information file). From the structure-factor files the number of unique reflections contributing to the least-squares refinement and the conventional R factor ['_refine_ls_R_factor_all', see equation (5)] were calculated and compared to the entries in the cif files. For these numbers and references see Table 2. The cases in which these numbers disagree are briefly discussed below.

Number of parameters. There was one discrepancy in the number of parameters from the text (975) and from the cif file (288) for data set No. 2 (Paul *et al.*, 2011). From the description in the text it was concluded that the number of parameters of 975 is more likely. This number was used subsequently.

Number of reflections used in the least-squares refinement. There were deviations between the number of reflections used in the refinement given in the text or cif files and the number of reflections in the reflection files. Discrepancies occurred for data sets No. 1 (Bibila Mayaya Bisseyou *et al.*, 2012) and 8–13 (Zhurov *et al.*, 2011). In the first case the number of reflections as given in the publication text and in the cif file is 12801, whereas 12803 was found in the reflection file. For data set Nos. 8–13, the number of reflections used in the least-squares refinement and the number of observed reflections with $I_{\rm o} > 3\sigma(I_{\rm o})$ were given as 10651/8057 for data set Nos. 8

(RDX20KAnharmonic) and 9 (RDX20KHarmonic), 12304/7986 for data set Nos. 10 (RDX120KAnharmonic) and 11 (RDX120KHarmonic), and 5403/3943 for data set Nos. 12 (RDX298KAnharmonic) and 13 (RDX298KHarmonic). The total number of reflections in the structure-factor files were 8057 (data set No. 8), 8070 (No. 9), 7986 (No. 10), 7981 (No. 11) and 3942 (Nos. 12 and 13), close to or identical to the number of observed reflections. For those data set Nos. 8–13, the conventional R factors given in the cif files are identical to the recalculated conventional R factors from all reflections in the reflection files, such that it is concluded that the refinements corresponding to data set Nos. 8–13 were all done only with the observed intensities.

Conventional R factor. The conventional R factor as given in the cif files is defined according to equation (5) in Table 1 where the summation goes over all reflections used in the refinement. Differences between conventional R factors given in the cif files and recalculated conventional R factors occurred for data set No. 1 (cif file/recalculated: 0.0368/0.0364) and data set Nos. 5–7 (Mondal *et al.*, 2012) (5: 0.025/0.0238, 6: 0.033/0.0298, 7: 0.034/0.0289). The R values for data set Nos. 5–7 given in the corresponding cif file are according to the *XD* convention (Volkov *et al.*, 2006), instead of the conventional R factor, which is the cif standard. In the *XD* convention, a negative F value appears when the corresponding intensity observation is negative (Volkov *et al.*, 2006); therefore the conventional R value is smaller than or equal to the corresponding R factor in the *XD* convention. A distinct difference in the conventional R factor appeared in data set No. 22, where the reflection file was taken together with the reflection files No. 21 and 23 from Li *et al.* (2010). Since the corresponding cif file of Li *et al.* (2010) did not contain all relevant information for set No. 22, the cif file from the original publication (Palatinus *et al.*, 2006), which was cited in Li *et al.* (2010), was consulted. The conventional R factor for all 2409 reflections found in this cif file, 0.0494, is considerably smaller than the recalculated conventional R factor, 0.0705. The reasons could not be tracked down. Recalculation of the conventional R factor of the published reflection file in Palatinus *et al.* (2006) resulted in a value of 0.0506, which is closer to the value published in the corresponding original cif file. Minor differences in the conventional R values occurred for data sets No. 8 (0.0328/0.0329), No. 21 (0.0484/0.0487) and No. 23 (0.0484/0.0487).

4. Application to experimental data

Predicted and actual R values for all 23 data sets are depicted in Fig. 1. For refinements against F^2 values the corresponding F^2-values-based residual factors were used and for refinements against F values the corresponding F-values-based residual factors were used. Additionally, where refinements against F values were performed with *XD*, the corresponding *XD* convention was used (Volkov *et al.*, 2006). In all cases the predicted R values are based on $\sigma(I_o)$ and $\sigma(F_o)$, not on $\hat{\sigma}$ values [these are the result of a transformation of $\sigma(I_o)$ values by a weighting scheme, see equation (14) in Henn & Schönleber (2013)], as it is already known that application of a weighting scheme leads to reduced s.u.'s that are correlated to the residuals (Henn & Schönleber, 2013), which destroys the statistical independence of the residuals. Simple multiplicative factors for the s.u.'s are, however, unproblematic. Such a factor was applied only in one of the 23 studies (No. 2), where all s.u.'s were moderately increased by a factor of 1.101. This factor is taken into account. For $\sigma(F_o)$ values the usual convention was used:

$$\sigma(F_o) = \frac{1}{2}\frac{\sigma(F_o^2)}{|F_o|} \quad \text{for} \quad F_o^2 > 0, \tag{14}$$

$$\sigma(F_o) = \frac{1}{2}\sigma(F_o^2) \quad \text{for} \quad F_o^2 < 0. \tag{15}$$

Please note that this convention leads to a discontinuity in the $\sigma(F_o)$ values when approaching zero. The predicted and *de facto* R values are shown in Fig. 1 and are discussed in detail in the next paragraphs.

4.1. Predictions

The predicted R value (red line with squares in Fig. 1) is smaller than the actual R value (blue line with circles) for all data sets. For a large fraction of the data sets the predicted values are below the 1% level, which is indicated by a black line. Data set No. 3, for example, has a value of only 0.49%. All of these data sets below that line have the potential to reach an R value lower than 1%, provided there are no systematic errors and provided that the s.u.'s are correct. In this sense, the predicted R value is a quality measure of the data sets, as it gives the attainable R value and a low predicted R value is associated with high data quality, as a low content of noise is expressed in this way. Data set Nos. 5–7 correspond to a series of measurements of the same compound at different

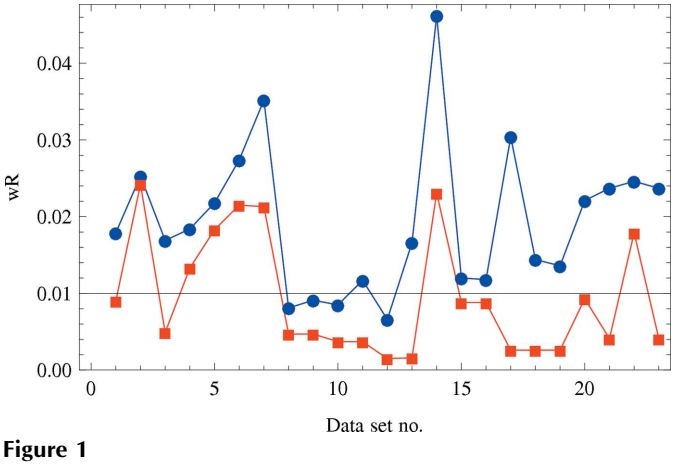

Figure 1
Predicted (red squares) and *de facto* (blue circles) *wR* values: for refinements performed against F values with *XD* (set Nos. 3–7, 15, 16, 20), the corresponding *XD* convention for the sign of F_o in the case of negative intensity observations was applied for the calculation of the *de facto* $wR(F)$ value in equation (11). Actual values were calculated according to equation (11) (data set Nos. 1, 2, 17–19, 21–23) and to equation (8) (data set Nos. 8–14). The predicted values were obtained from equations (9) and (12). The individual values are connected with a line as a guide for the eye.

temperatures (20, 100 and 298 K). The predicted R value initially increases with the temperature in this series but drops slightly for the 298 K data set. Data set Nos. 8–13 also correspond to measurements of one structure at different temperatures (20, 120, 298 K). The predicted R value is the same for data set Nos. 8 and 9 because the experimental data are the same, only the model changes from anharmonic to harmonic nuclear motion. The experimental data are the same for set Nos. 10 and 11 as well as 12 and 13, respectively. The predicted R value decreases with increasing temperature in this study, such that the lowest predicted R value is obtained for the data set with the highest experimental temperature in this series (data set Nos. 12, 13). This value, 0.16%, is the lowest of all data sets. A decreasing predicted R value for increasing temperature was not expected; it corresponds to increasing mean-squared significance of the data sets, such that the data set with the highest experimental temperature has the highest mean-squared significance of the data. However, due to the different number of reflections in the sets, the R values do not compare the same reflections. Data set Nos. 14–16 correspond also to the same compound; however, data set No. 14 is a conventional refinement against 8630 F^2 values, whereas set Nos. 15 and 16 are multipole refinements against 4737 F values. The R value corresponding to set No. 14 cannot be compared directly to those of set Nos. 15 and 16 due to the different number of reflections. Data set Nos. 17–19 again have identical values for the prediction because the experimental data are the same, only the models change. The predictions for model Nos. 21 and 23, 0.41%, are identical, whereas that for set No. 22, 1.79%, is considerably larger.

4.2. *De facto* values

The *de facto* R value is given by the blue line in Fig. 1. All predicted values are smaller than the actual ones. This need not be the case; it may be taken as a hint that experimental σ values are more likely to be underestimated than overestimated. This is indirectly confirmed by the fact that in a large fraction of the studies a weighting scheme was applied that increases the s.u.'s (set Nos. 8–13, 17–21, 23). It is also known from the literature that area-detector data tend to produce underestimated s.u.'s (see, for example, Waterman & Evans, 2010). *De facto* and predicted R values are closest for data set No. 2, followed by set Nos. 16, 15 and 8. Particularly far apart are the predictions and actual values for data set No. 17 with 2.78 percentage points, followed by data set No. 14 with 2.31 percentage points. It is interesting to see that data set Nos. 17 and 14 correspond to independent-atom model (IAM) refinements against high-resolution data. A large difference is expected in these cases.

4.3. Systematic errors as indicated by $R^{\rm meta}$

The systematic error as defined by equation (13) is shown in Fig. 2. The lowest contamination with systematic errors is given for data set Nos. 2 ($R^{\rm meta}$ = 0.04) and 5 (0.16). The values of both the predicted and the *de facto* R values were close for both sets in Fig. 1.

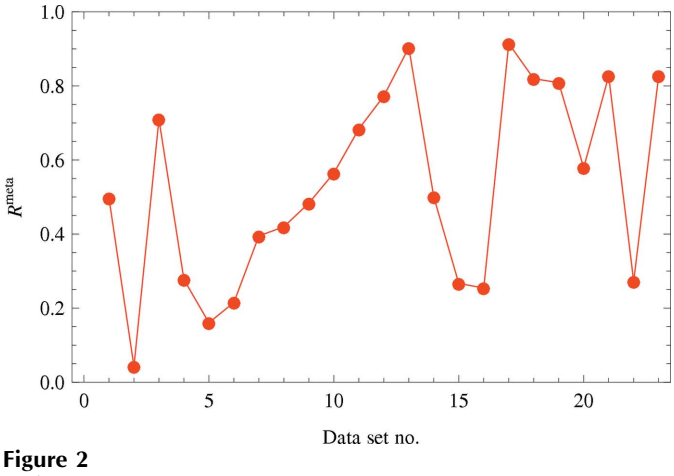

Figure 2
$R^{\rm meta}$ as defined in equation (13) with respect to predicted and *de facto* $wR(F^2)|_{w=1/\sigma^2}$ and $wR(F)|_{w=1/\sigma^2(F_o)}$ for all data sets.

The degree of contamination with systematic errors increases for the data sets (Nos. 5–7 and 8–13) with increasing temperature. In the series 8–13, the anharmonic models 8, 10 and 12 each show a reduced degree of contamination with systematic errors when compared to the respective harmonic model (9, 11, 13). The high degree of contamination with systematic errors for the data set No. 12 may be due to the exceptionally small *de facto* R value that enters the denominator in equation (13). In the series 14, 15, 16 and 17, 18, 19, the IAM refinements (14 and 17) each have the highest degree of contamination with systematic errors, which drops for the multipole refinements (15, 16, 18) and for the theoretical refinement (19); however, the drop in systematic errors from the IAM refinement to the multipole refinement appears small in the case of the series 17, 18 and 19 in view of the high level of contamination with systematic errors and in comparison to the drop from 14 to 15.[3]

The high degree of systematic errors as expressed by $R^{\rm meta}$ in data set Nos. 8–13 is surprising in view of the low *de facto* R values for these sets as shown in Fig. 1. This is because $R^{\rm meta}$ measures a relative systematic error with reference to the *de facto* R values [see denominator in equation (13)]. In order to give also an absolute rather than relative measure of systematic errors, one may take the counter of $R^{\rm meta}$ and interpret this as the difference between predicted and actual R values in percentage points.

The corresponding plot (Fig. 3) shows that the absolute differences are small indeed. For data set Nos. 8, 10 and 12 the difference between the *de facto* and the predicted R value is smaller than or equal to 0.5 percentage points (pp). Such low values result also for data set Nos. 2 (0.11 pp), 16 (0.30 pp), 15 (0.32 pp), 8 (0.34 pp), and 5 (0.35 pp) as well as 9 (0.44 pp) and 10 (0.48 pp).

The majority of charge-density studies presented here are based on a refinement against structure-factor amplitudes

[3] The comparability of $R^{\rm meta}$ values for data set Nos. 14 and 15 suffers, however, from the fact that in set No. 14 a refinement based on F^2 against 8630 reflections is performed whereas in set Nos. 15 and 16 a refinement based on F against only 4737 reflections is performed.

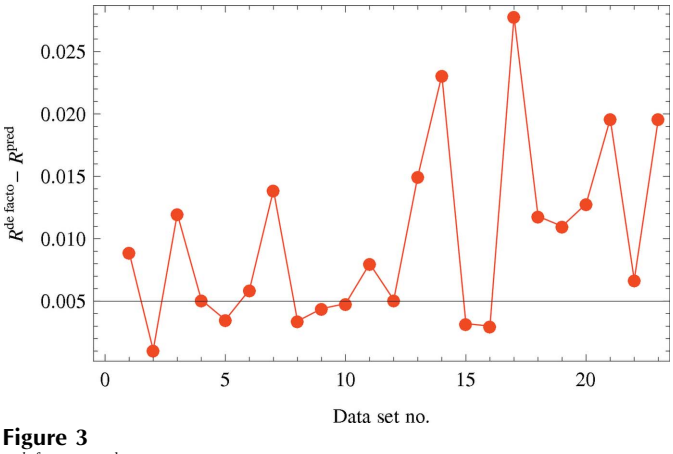

Figure 3
$R^{de\ facto}$–R^{pred} for all data sets.

rather than intensities. Refinements against structure-factor amplitudes always imply a significance cutoff, as at least the negative intensity observations are omitted, which corresponds to a significance cutoff $I_o/\sigma(I_o) \geq 0$. The effect of an intensity cutoff $I_o \geq 0$ will be discussed in §7.

5. Are residuals in charge-density studies distributed in a Gaussian fashion?

To analyze the distribution of residuals a χ^2 test was performed. The data were binned into $N_B = 10$ bins (bin ranges were chosen from $-\infty$ to $-2\sigma_0$, then in steps of 0.5 σ_0 from $-2\sigma_0$ to $2\sigma_0$, and from $2\sigma_0$ to ∞). To allow for deviations from zero means and unit variances in the presumed Gaussian distribution, the parameters

$$\mu_0 = \frac{1}{N_{ref}} \sum_{j=1}^{N_{ref}} \zeta_j \tag{16}$$

and

$$\sigma_0^2 = \frac{1}{N_{ref} - 1} \sum_{j=1}^{N_{ref}} (\zeta_j - \mu_0)^2 \tag{17}$$

were calculated from the list of residuals $\zeta_j = (I_{o,j} - I_{c,j})/\sigma(I_{o,j})$, $j = 1, \ldots, N_{ref}$ for each data set. This is especially important to allow for a scaling of σ values. The sum of squared differences between N_i observed and n_i expected events in bin number i, $i \in \{1, \ldots, 10\}$, each divided by the number of expected events for bin number i, was calculated, yielding the corresponding χ_S^2 value:

$$\chi_S^2 = \sum_{i=1}^{10} \frac{(N_i - n_i)^2}{n_i}. \tag{18}$$

The number of expected events was calculated from integrating the Gaussian function over the bin width and multiplying with the total number of events (structure factors) for each data set:

$$n_i = N p_i \tag{19}$$

Table 3
Statistical descriptors of the data sets.

From left to right: No. of data set, residual population mean value according to equation (16), square root of the residual population variance according to equation (17), χ_S^2 residual sum according to equation (18), minimum residual and maximum residual.

Data set	μ_0	σ_0	χ_S^2	Min. ζ	Max. ζ
1	−0.60	5.47	13122.15	−444.00	9.69
2	0.02	1.04	135.92	−4.77	6.57
3	0.08	3.41	3337.01	−99.35	29.81
4	0.12	1.09	110.10	−5.20	6.62
5	−0.10	1.10	915.27	−14.56	5.69
6	−0.11	1.06	710.92	−14.45	6.25
7	−0.12	1.15	597.96	−35.21	10.52
8	0.04	1.64	667.54	−18.55	27.24
9	0.05	1.84	1054.10	−25.73	29.67
10	0.08	2.21	1592.40	−26.78	44.28
11	0.09	3.03	2871.86	−51.61	53.05
12	0.01	3.78	1537.89	−47.72	33.68
13	0.13	9.52	2177.94	−99.88	86.59
14	−0.23	1.97	1280.65	−24.09	24.85
15	−0.28	1.31	231.00	−4.75	4.88
16	−0.28	1.29	235.56	−5.55	4.76
17	0.92	11.12	933.54	−67.18	81.05
18	0.50	5.22	795.15	−51.49	55.05
19	0.44	4.94	631.22	−50.67	36.45
20	−0.40	2.29	2702.51	−92.00	19.63
21	0.51	5.75	2678.35	−147.35	53.53
22	−0.48	1.48	351.00	−11.76	13.55
23	0.51	5.75	2678.35	−147.35	53.53

$$N = \sum_i n_i \tag{20}$$

$$p_i = \int_a^b \frac{1}{(2\pi)^{1/2}\sigma_0} \exp\{-[(x - \mu_0)^2/2\sigma_0^2]\}\, dx \tag{21}$$

with (a, b) being the limits of the bins.

Large values of χ_S^2 indicate that it is unlikely that the residuals correspond to a Gaussian distribution with mean μ_0 and variance σ_0^2. The χ^2 probability function $Q(\chi^2, \nu)$, an incomplete gamma function, gives the probability that the sum of squares of ν random Gaussian variables with $\mu = \mu_0$ and $\sigma = \sigma_0$ will be greater than χ^2. There are $N_B = 10$ bins and three constraints [*i.e.* equation (20), $\mu = \mu_0$, $\sigma = \sigma_0$] resulting in seven degrees of freedom $\nu = N_B - 3$. A level of significance α is chosen, which gives the probability that the empirically determined distribution of values, which is assumed to correspond to a Gaussian distribution, appears by chance. It is recommended in the literature to be tolerant of low probabilities due to outliers and a level of 0.001 is mentioned (Press *et al.*, 1992). Choosing this level of significance $\alpha = 0.001$ results in $\chi^2_{\alpha=0.001;\nu=7} = 24.32$. For $\chi_S^2 > 24.32$ the assumption that the residuals are distributed according to a Gaussian distribution must be rejected at the given significance level, because the probability for the residuals stemming from a Gaussian distribution with $\mu = \mu_0$ and $\sigma^2 = \sigma_0^2$ is smaller than or equal to α. As mean values and variances are taken from the population, the test performed here is about the form of the residual distribution, *i.e.* whether or not the residuals can be described as Gaussian distributed. The

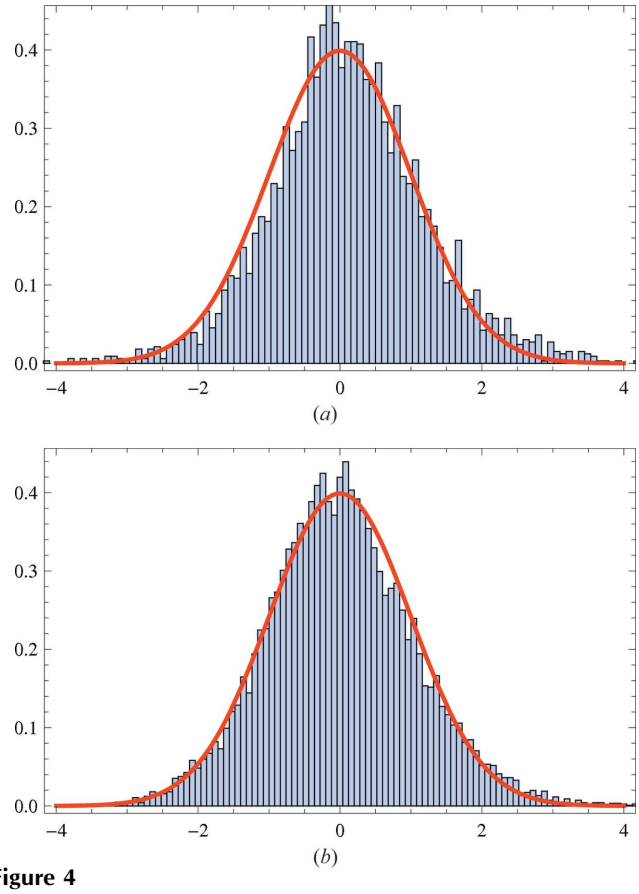

Figure 4
Probability histograms of the distribution of residuals ζ of the data sets with smallest χ^2_S values. (*a*) Data set No. 4; (*b*) data set No. 2.

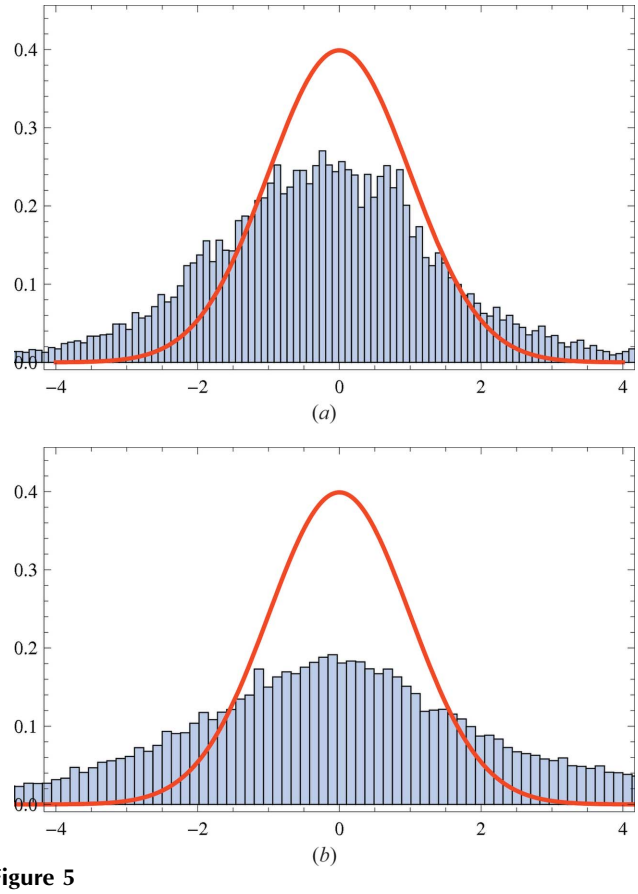

Figure 5
Probability histograms of the distribution of residuals ζ of the data sets with largest χ^2_S values. (*a*) Data set No. 1; (*b*) data set No. 3.

parameters of the Gaussian distribution play only a minor role. The estimates μ_0, σ_0 and the sum χ^2_S are listed in Table 3.

By far the smallest χ^2_S residual sums in Table 3, 110.10 (No. 4) from point-detector data and 135.92 (No. 2) from charge-couple device (CCD) data processed with *SORTAV* (Blessing, 1987), are still much larger than the critical value of 24.32. Therefore, the hypothesis that any of these 23 residual distributions is close to a Gaussian distribution must be rejected at the significance level of $\alpha = 0.001$. The list of minimum and maximum residuals given in columns 5 and 6 already indicates that a normal distribution of residuals is unlikely for most of the data sets, because the absolute values are much too large, probably with the exceptions of data set Nos. 2, 4, 15 and 16. Set Nos. 15 and 16 come closest to a Gaussian distribution after set Nos. 4 and 2 [see also the normal probability plots and probability histogram plots in the supporting information[4]; for more information on normal probability plots, see Abrahams & Keve (1971)]. Histograms of the residual distributions for the data sets with smallest and largest χ^2_S values are shown in Figs. 4 and 5 together with a red line indicating the probability density function of a normal distribution with zero mean, $\mu_0 = 0$ and unit variance $\sigma_0^2 = 1$.

[4] Supporting information for this paper is available from the IUCr electronic archives (Reference: KX5024).

6. Why are residuals not distributed in a Gaussian fashion?

The role of the Gaussian distributions of residuals cannot be overemphasized, as the whole procedure of least-squares fitting is based on this assumption. The Gaussian distribution of residuals needs to be confirmed *a posteriori* to justify the method and the numerical values of the parameters and their s.u.'s. The analysis of the preceding paragraph shows, however, that it is very unlikely for most of the data sets that the distribution of residuals is in accordance with the assumption of a Gaussian distribution. This leads to the question of what factors might prevent a least-squares refinement from resulting in a Gaussian distribution of residuals?

The reflection data need to be statistically independent, which is not the case with area detectors (see *e.g.* Waterman & Evans, 2010), due to correlations among detector pixels and due to data-processing steps. There is also, however, a data set included not employing an area detector (No. 4). The corresponding refinement has the lowest χ^2_S value, but the residual distribution is still unlikely to be Gaussian. Also the set closest to a Gaussian distribution (No. 2) employs CCD data. We infer that correlation problems related to area detectors are at least not the only source responsible for deviations from a Gaussian distribution of residuals. Zhurov *et al.* (2008) emphasize the importance of the beam-conditioning devices

Table 4
Influence of correlations introduced by the structure model and of data truncation on χ^2_S in IAM refinements against type-one consistency data sets with s.u.'s according to equation (22) (for more information see text).

Set	p_1	p_2	Cutoff	$N_{\rm ref}$	μ_0	σ_0	χ^2_S	$\underline{wR(F^2)\ de\ facto}$ $w = 1/\sigma^2$	$\underline{wR(F^2)\ \text{pred}}$ $w = 1/\sigma^2$	$R^{\rm meta}$
24	0.01	1.5	No	14604	0.00	1.00	11.60	0.0360	0.0358	0.01
25	0.01	1.5	Yes	13385	0.10	0.96	150.89	0.0334	0.0343	−0.03
26	0.025	1.5	No	14604	0.00	1.00	9.91	0.0665	0.0663	0.00
27	0.025	1.5	Yes	13377	0.10	0.96	149.84	0.0617	0.0634	−0.03
28	0.05	1.5	No	14604	0.00	1.00	11.92	0.1091	0.1088	0.00
29	0.05	1.5	Yes	13355	0.10	0.96	155.26	0.1011	0.1040	−0.03

and integration software, and show normal probability plots of refinements with point detectors and with area detectors. Unfortunately, it still remains exceptional that normal probability plots are shown in charge-density studies.

Another mechanism introducing correlations occurs from the structure model. It was also mentioned that intensity or significance cutoffs lead to systematic errors. These effects are investigated in the next section.

7. Influence of the structure model and of intensity and significance cutoffs

To study the extent to which the structure model introduces correlations among residuals, it is helpful to employ artificial data. This gives total control not only over the model parameters but also over the s.u.'s. Knowing the true model parameters also ensures that the model is capable of describing the data adequately. In refinements against experimental data this is only a working hypothesis, that is justified *a posteriori* by small residual factors. The artificial data were generated according to the following procedure.

(*a*) From a model refinement against high-resolution data the calculated intensities were extracted and written as $I_{\rm o}$ to a reflection file together with their s.u.'s. This results in a set of known true noiseless intensities of a precision given by the respective variances and known true model parameters. A control least-squares refinement on this data set yields correspondingly only zero R values. The s.u.'s given in this set are of only formal character, as the intensities themselves are noise-free. We call this kind of set a 'type-zero consistency set'. This type of set has its own applications; however, it is not a good simulation of real data, as the reflections do not contain any noise.

(*b*) An error model was chosen according to

$$\text{s.u.} = I_{\rm o} \times p_1 + p_2, \qquad (22)$$

i.e. it was assumed that the s.u.'s are in proportion to the intensity with a factor of proportionality p_1 and a constant term p_2. This emulates the background signal and prevents s.u. values from approaching zero.

(*c*) Gaussian noise was then added to the intensities with mean value set to zero and σ according to the respective s.u.'s. from the error model. We call this kind of set a 'type-one consistency set'. It may serve as simulated data. The noise contained is not only formal but realized.

(*d*) The type-one consistency set was subjected to a least-squares refinement. The resulting residuals were analyzed.

To simplify the procedure, only an IAM model with anisotropic displacement parameters (ADPs) was generated from a refinement against experimental data. More about the experimental data can be found in Henn *et al.* (2010). The χ^2_S residual sum was calculated according to equation (18). In the case where the model effectively introduces correlations among the residuals, this should express itself in a larger value of χ^2_S. To study the effect of intensity and significance cutoff, the refinement was repeated with a cutoff $I_{\rm o} > 0$ and residuals were analyzed again. The results are shown in Table 4.

Table 4 shows that correlations introduced by the structure model are not a factor that drives the χ^2_S residual sum over the limit of 24, which corresponds to the level of significance $\alpha = 0.001$, as in all cases where truncation was not applied (set Nos. 24, 26 and 28 in Table 4), the residual sum stays well below this threshold value. That indicates that the hypothesis of the corresponding residual distributions following a Gaussian distribution cannot be rejected. If the intensity truncation $I_{\rm o} \geq 0$ is applied, however, the corresponding χ^2_S residual sums exceed the threshold value by far. From this we can directly conclude that, for example, the χ^2_S values for data set Nos. 8–13 will fall when the cutoff is abandoned. The truncation also leads in all cases to a positive shift in μ_0 due to the missing negative parts of the error distribution for low- and zero-intensity observations as well as to a negative shift in σ_0. The model parameters resulting from the least-squares refinement are partially able to describe the truncated distribution and the same number of model parameters is applied to a reduced number of observations, thereby leading to weak over-fitting as indicated by the negative $R^{\rm meta}$ values, whereas all other $R^{\rm meta}$ values are zero or very close to zero, as expected for data sets that do not contain systematic errors.

8. Conclusion

The formerly introduced measures of attainable R values for a given data set are used to assess the quality of a given experimental data set and to assess the degree of systematic errors in a given set of observed and calculated intensities. The most important conclusion from this work is that the requirement of a Gaussian distribution of residuals is in most cases not fulfilled. This was already the case for standard structures (Henn & Schönleber, 2013) and it is surprisingly

also the case in high-resolution charge-density studies, which are much more expensive in terms of expenditure of time, know-how and financial support. Indeed, only one of these publications offered a test on the validity of this important assumption by employing normal probability plots, although this minimum assessment of the distribution of residuals is easily feasible for example with the *WinGX* (Farrugia, 1999, 2012) suite. A normal distribution of residuals is still not a guarantee for absence of systematic errors, but a necessary condition. Further steps have to be taken to ensure that the model refinement yields statistically independent residuals. This is the topic of a subsequent paper. Only two data sets come close to a normal distribution of residuals: one is from point-detector data (set No. 4), the other from CCD data processed with *SORTAV* (set No. 2). We suggest including a threshold value for the probability of the data stemming from a normal distribution of $\alpha = 0.001$ in future charge-density studies, in order to direct and focus research activities into this important question. The present work shows how application of a cutoff criterion, which is mandatory in *F*-based refinements, may prevent the residual density distribution from following a Gaussian.

References

Abrahams, S. C. & Keve, E. T. (1971). *Acta Cryst.* A**27**, 157–165.

Bibila Mayaya Bisseyou, Y., Bouhmaida, N., Guillot, B., Lecomte, C., Lugan, N., Ghermani, N. & Jelsch, C. (2012). *Acta Cryst.* B**68**, 646–660.

Blessing, R. H. (1987). *Crystallogr. Rev.* **1**, 3–58.

Dittrich, B., Sze, E., Holstein, J. J., Hübschle, C. B. & Jayatilaka, D. (2012). *Acta Cryst.* A**68**, 435–442.

Domagała, S., Munshi, P., Ahmed, M., Guillot, B. & Jelsch, C. (2011). *Acta Cryst.* B**67**, 63–78.

Farrugia, L. J. (1999). *J. Appl. Cryst.* **32**, 837–838.

Farrugia, L. J. (2012). *J. Appl. Cryst.* **45**, 849–854.

Henn, J., Meindl, K., Oechsner, A., Schwab, G., Koritsanszky, T. & Stalke, D. (2010). *Angew. Chem.* **122**, 2472–2476.

Henn, J. & Schönleber, A. (2013). *Acta Cryst.* A**69**, 549–558.

Hirshfeld, F. L. & Rabinovich, D. (1973). *Acta Cryst.* A**29**, 510–513.

Holstein, J. J., Luger, P., Kalinowski, R., Mebs, S., Paulman, C. & Dittrich, B. (2010). *Acta Cryst.* B**66**, 568–577.

Janicki, R. & Starynowicz, P. (2010). *Acta Cryst.* B**66**, 559–567.

Li, L., Schönleber, A. & van Smaalen, S. (2010). *Acta Cryst.* B**66**, 130–140.

Mondal, S., Prathapa, S. J. & van Smaalen, S. (2012). *Acta Cryst.* A**68**, 568–581.

Palatinus, L., Dušek, M., Glaum, R. & El Bali, B. (2006). *Acta Cryst.* B**62**, 556–566.

Paul, A., Kubicki, M., Jelsch, C., Durand, P. & Lecomte, C. (2011). *Acta Cryst.* B**67**, 365–378.

Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery, B. P. (1992). *Numerical Recipes in C: the Art of Scientific Computing*. New York: Cambridge University Press.

Schwarzenbach, D., Abrahams, S. C., Flack, H. D., Gonschorek, W., Hahn, Th., Huml, K., Marsh, R. E., Prince, E., Robertson, B. E., Rollett, J. S. & Wilson, A. J. C. (1989). *Acta Cryst.* A**45**, 63–75.

Volkov, A., Macchi, P., Farrugia, L. J., Gatti, C., Mallinson, P. R., Richter, T. & Koritsanszky, T. (2006). *XD2006. A Computer Program Package for Multipole Refinement, Topological Analysis of Charge Densities and Evaluation of Intermolecular Energies from Experimental and Theoretical Structure Factors*. University at Buffalo, State University of New York, USA; University of Milan, Italy; University of Glasgow, UK; CNRISTM, Milan, Italy; and Middle Tennessee University, USA.

Waterman, D. & Evans, G. (2010). *J. Appl. Cryst.* **43**, 1356–1371.

Zhurov, V. V., Zhurova, E. A. & Pinkerton, A. A. (2008). *J. Appl. Cryst.* **41**, 340–349.

Zhurov, V. V., Zhurova, E. A., Stash, A. I. & Pinkerton, A. A. (2011). *Acta Cryst.* A**67**, 160–173.