International Journal of Materials Chemistry and Physics
Vol. 1, No. 3, 2015, pp. 417-430
http://www.aiscience.org/journal/ijmcp

Two Common Sources of Systematic Errors in Charge Density Studies

Julian Henn[1, *], Kathrin Meindl[2]

[1]Bayreuth, Germany
[2]Centre for Genomic Regulation (CRG), Barcelona, Spain

Abstract

Resolution independent residuals $\varsigma = (I_o - I_c)/\sigma(I_o)$ should not show any discontinuities, when plotted against the resolution $\sin\theta/\lambda$. The density of reflections per resolution range, however, changes with resolution, which limits the information value from these plots. Residuals plotted against *ranked* resolution values, should be uniform and are easy to interpret. These plots of individual ranked residuals show more details compared to binned data, and additionally the important spread of the data. Therefore, plots of individual ranked residuals, which are currently only very rarely applied (see for example Friese *et al.* (2013)), are superior over the plots of binned values and should be used for the detection of systematic errors. Plots against the ranked intensity, significance and standard uncertainties, respectively, can be chosen, too. This work concentrates on the resolution dependence. Any discontinuity is evidence for data processing errors *e.g.* from merging resolution batches. Applications to artificial data and to experimental data are given. Non-uniform features are observed in the experimental data. Characteristic features in the residuals resulting from distorted standard uncertainty (s.u.) values are discussed by means of normal probability plots, plots of observed *vs*. calculated intensities, and plots of residuals *vs*. resolution for 23 experimental data sets and 3 artificial data sets with accurate, overestimated and underestimated large s.u.s. Distorted s.u. values seem to be a very common source of errors as demonstrated here. The underlying cause for distorted s.u. values is that the integration software from detector data severely underestimates the s.u.s of the strong reflections. Underestimation of the s.u.s of the strong reflections also leads to artificially reduced $wR(F^2)$ values. When a weighting scheme is applied, there remains a tendency to not fully compensate for the underestimation. In virtually all data sets the s.u.s are underestimated as is seen by Goodness of Fit (*GoF*) values larger than one and slopes larger than one in the normal probability plots. Underestimated s.u. values artificially increase the number of rare events and manifest in form of outliers in other diagnostic tools based on the s.u.s, however, when no other systematic errors are present, the plots of observed *vs*. calculated intensities do *not* show outliers, even in the case of distorted s.u. values. The appearance of outliers in plots of residuals *vs*. resolution *not* accompanied by outliers in plots of I_o *vs*. I_c is therefore consequently taken as an indicator for distorted s.u. values. Individual outliers in plots of observed *vs*. calculated intensities are surprisingly also observed.

Keywords

Quality Indicator, Fit Data, Ranked Residuals, Systematic Errors, Charge Density

Received: September 15, 2015 / Accepted: December 9, 2015 / Published online: January 15, 2016

1. Introduction

In a former publication (Henn&Schönleber, 2013), the theoretical or predicted *R*-values were developed, which use experimental data to predict the agreement factor without the need to specify model parameters. In this approach it is assumed that no systematic errors are present and that the s.u.

* Corresponding author
E-mail address: Julian.Henn@uni-bayreuth.de (J. Henn)

values are accurate. A systematic error is a systematic mismatch between observed and model predicted data. According to the IUCr, systematic errors are *per definitionem* always on the side of the model (Schwarzenbach *et al.*, 1989). The mentioned conditions for the correct prediction of the agreement factor imply a Goodness of Fit (*GoF*) very close to one. Applications in standard structure determinations showed non-Gaussian distributions of residuals with a tendency of *GoF* values to be larger than one. Applications in more advanced and much more expensive charge density studies that are expected to be highly accurate surprisingly also showed very frequently non-Gaussian distributions of residuals (Henn & Meindl, 2014a) despite the large efforts that have to be coped with in such studies. Further analysis tools based on conditional probabilities were developed (Henn & Meindl, 2014b) and applied to the aforementioned high resolution data sets with the expectation to single out certain sources of errors with the help of graphical representations from conditional probability plots in the unit-square, the so-called BayCoN plots. Instead of showing large differences between different models applied to the same experimental data, more similarities rather than differences were seen among different models and even between certain different data sets. This led us to the hypotheses, that data processing steps are the dominant source of systematic errors. For a theoretical study, the s.u. values of an artificial data set were manipulated in order to show in one case too large and in the other case too small s.u. values for the strong intensities. The resulting BayCoN plots after a model refinement showed similarities with the BayCoN plots of many of the experimental high resolution data sets for the artificially lowered s.u. values, whereas the data set with artificially increased s.u. values showed to a much larger extent smooth and uniform BayCoN plots (Henn&Meindl, 2014b). The development of data descriptors and quality indicators for diffraction data is a fruitful and important topic, every stage of the data acquisition is involved: from the instrument parameter settings (see, *e.g.*, Sørensen& Larsen 2003), the data integration (Waterman & Evans, 2010), the data reduction (see, *e.g.*, Blessing, R. H. 1997; Evans & Murshudov, 2011, Weiss 2001) and density modelling (see, *e.g.*, Zhurov *et al.* 2011) to the post-refinement evaluation and global indicators of X-ray data quality (see, *e.g.,* Zhurov *et al.* 2008, Weiss 2001). The cited literature is by no means complete.

In this work we show that many experimental data sets suffer from data processing errors and from distorted s.u. values. It was earlier shown that underestimation of the s.u.s is a problem for CCD data (Waterman & Evans 2010). It persists to be a problem.

2. Distorted and Amplified Standard Uncertainty Values

When all s.u. values of a data set are too large or too small by a simple factor $x > 0$ we do not regard these as being distorted, but just by amplified. For simplicity we keep this terminology also in cases where the factor x is smaller than one.

2.1. Amplified Standard Uncertainties

When the s.u.s are amplified and no systematic errors are present this results in a Gaussian distribution of residuals with standard deviation of the Gaussian given by x and mean value close to zero. When the refinement is performed against intensity values and statistical weights are chosen, $1/s.u.^2(I_o)$, the resulting $wR(F^2)$ is unaffected as are the model parameters, *i.e.* they remain the same for *any* factor $x > 0$ (Henn & Schönleber, 2013). Application of a significance cut off leads to systematic errors, as the residuals necessarily cease to be identically distributed (Henn & Meindl, 2015), which is a requirement for the least squares procedure.

What *is* affected in this case are the s.u.s of the model parameters, which will be too large or too small by a factor x, the *GoF*, that will be $1/x$, the significance of reflections, which also changes with $1/x$ and the predicted $wR(F^2)$ - value, as it depends on the significance (Henn & Meindl, 2014a).

Therefore, when the goal of a model refinement is to obtain accurate model parameter estimates *and* accurate errors of model parameter estimates, it is indispensable to employ accurate error estimates of the reflection data and to exclude systematic errors.

2.2. Distorted Standard Uncertainties

A set of distorted s.u. values exists, if a single factor for the whole data set is not sufficient to correct these. An example would be when the full data set was merged from two individual data sets, *e.g.* from different detector positions, and when the s.u.s are not correctly scaled to each other. In that case a single factor might be sufficient to scale the s.u.s correctly in relation to each other, but the factor is not applied to the whole data set but only to a subset. Another example would be when an intensity- or resolution dependent function is needed to correct the s.u. values. Effects of distorted s.u. values are that the model parameter values as well as their s.u.s and the significance of intensities will be incorrect.

The refinement will generally yield a *GoF* different from one and the resulting distribution of residuals cannot be described by a Gaussian distribution. Previous results indicated that distorted s.u.s are more harmful, when the large s.u. values are underestimated rather than overestimated (Henn & Meindl, 2014b). Underestimated large s.u. values lead to artificially reduced *wR*-values at the expense of systematic errors in the data (Henn & Meindl, 2014b).

3. Error Models

Error models serve to obtain adequate s.u.s from those of the reflection file, which are known to be too small. As a consequence, it is appropriate to chose one error model for one set of experimental data. The same error model should be used, when different structure models are fitted to the same experimental data, as otherwise the error model is not exclusively employed to correct the precision of the experimental errors. The error model has the strongest impact on the low resolution data, where most of the strong observations reside. Different error models are in use in different refinement software packages.

As a consequence of the distinct effect of the weighting scheme on the residuals, the error model parameters must be chosen with care, as incorrect parameter values or an insufficient error model may lead to an incomplete correction of the s.u.s, resulting in underestimated s.u. values for the reflections in the lowest resolution shell.

4. Observations from a Study of Charge Density Data Sets

From Henn & Meindl, 2014b the following results were obtained: Virtually all 23 studied data sets are contaminated by systematic errors, as indicated by the differences between predicted and actual *R*-values and by non-uniform BayCoN plots. The R^{meta} values are all positive, which indicates presence of systematic errors and/or underestimation of s.u. values (Henn & Schönleber, 2013). All Goodness of Fit (*GoF*) values are larger than one except for those of data sets 8-11 (for a list of *GoF* values see Table 2 in the Appendix). *GoF* values larger than one also indicate too small s.u. values and/or presence of systematic errors. *GoF* values smaller than one generally indicate over fitting or too large s.u. values, however, small *GoF* values can be achieved also by underestimating the s.u.s of a few strong reflections and overestimating the s.u.s of the abundant remaining data.

Virtually none of the studied data sets showed a normal distribution of residuals, except data set 13, with sets 8-13 being generally close to a normal distribution as indicated by the normal probability plots and also by the χ^2-sums (see supplementary material of Henn & Meindl 2014b).

Table 1. Identical outliers in data sets 5-7.

Data set	h,k,l	I_c	I_o	s.u.	sinθ/λ	ζ
5	0,1,-1	3997.69	3141.19	64.8471	0.1221	-13.2079
	-2, 0 ,0	2886.02	2566.83	23.5159	0.097	-13.5735
6	0,1,-1	3967.08	3167.43	57.0966	0.1219	-14.0052
	-2,0,0	2871.86	2299.77	446.303	0.097	-1.28184
7	0,1,-1	3905.48	2741.07	33.0748	0.121	-35.2053
	-2,0,0	2761.15	2160.42	76.3217	0.0971	-7.87102

Table 2. Goodness of Fit values as given in the literature for data sets 1-23 and for the artificial data sets 24, 30, and 31.

Data set	GoF	Data set	Gof	Data set	GoF
1	1.908	11	0.86	21	1.58
2	1.069	12	1.25	22	1.53
3	3.5901	13	1.14	23	1.58
4	1.177	14	2.000	24	1.0031
5	1.172	15	1.357	30	3.0275
6	1.13	16	1.338	31	0.6802
7	1.282	17	2.190		
8	0.90	18	1.109		
9	0.86	19	1.112		
10	0.85	20	1.8986		

The finding that virtually all studied charge density data sets are affected by systematic errors is surprising because these studies are the most expensive and the most accurate ones. Systematic errors should therefore be very small, because expensive high resolution diffraction experiments are usually conducted to obtain more accurate models.

In addition, with high quality data and very advanced density models, those reflections affected by systematic errors should reveal themselves as outliers. If only a few individual reflections are affected by systematic errors, it is expected that the residual distribution is close to a Gaussian and that these individual reflections appear as outliers in both observed *vs*. calculated intensity (I_o, I_c) plots and in plots of residuals *vs*. resolution $(\varsigma, \sin\theta/\lambda)$. If, conversely, a whole range of data in one data set is affected by systematic errors such that frequent outliers appear for example in $(\varsigma, \sin\theta/\lambda)$-plots, those kind of systematic errors are most likely connected to data processing steps such as merging different batches or to the s.u.s and the error model.

In the following, the normal probability plots (npps), $(\varsigma, \sin\theta/\lambda)$ plots, and (I_o, I_c) plots of these data sets are discussed. The respective plots are found in the Appendix.

5. Normal Probability Plots

A complete list of npps including the weighting schemes, if used in the refinement, for all data sets is given in the Appendix. Statistical weights were applied in some data sets. Examples of npps for artificial data are given in Fig. 1 a, d, and g. Employing artificial data gives the opportunity to study the pure effects of under- and overestimation of large s.u. values under exclusion of other systematic errors and to compare these to the case of adequate s.u. values and to the experimental data.

Distinctly curved npps are according to Abrahams & Keve (1971) "necessarily caused by systematic error".

The artificial data sets 24 (no systematic error), 30 (underestimation of large s.u.s) and 31 (overestimation of large s.u.s) from Henn & Meindl (2014b) are taken to visualize the effects of distorted s.u. values in Fig. 1. The npp of data set 24 is very close to the diagonal line which represents the expected frequency of events under ideal circumstances. Small deviations in the periphery are due to the finite population. The npps of the other data sets immediately indicate deviations from the expected Gaussian distribution (Fig. 1 d, g). The case of underestimated large s.u. values appears as a line with slope larger than one (Fig. 1 g), whereas the case of overestimated large s.u. values appears as a line with slope lower than one (Fig. 1 d) and deviations from a linear behaviour in the periphery.

From the npps of the experimental data sets, not a single one is reminiscent of Fig. 1 d, instead all of them show a tendency to slopes larger than one corresponding to Fig. 1 g (see column A in Fig. 5, Appendix). This holds in particular for those data sets from area detectors and with statistical weights (1, 3, 5-7, 14-16, 20-23). For these data sets it is known from the literature that their s.u. values are very likely to be underestimated, in particular those from strong reflections. But also *e.g.* data sets 17-19, that include a weighting scheme, show this slope larger than one in the npp. This can even be observed for the high quality data sets 8-12, though only in the periphery of their npps and to a distinctly reduced extent.

From this common structural peculiarity of the npps of all experimental data sets it is postulated that they may have the same origin, which is that the s.u.s are distorted in a way that leads to an underestimation of the s.u. values.

The underestimation of large s.u. values may not be the only source of systematic error, as some npps show additional curvatures, for instance those for data sets 5-7 (see Fig. 2); however, that all data sets show this large slope is a striking feature.

It was discussed earlier (Henn&Meindl 2014b) that the underestimation of large s.u. values leads to reduced R values. The underestimation of the large s.u.s led from an adequate value $wR(F^2)$ = 3.60% (data set 24), which describes realistically the noise inherent in the data, to the artificially reduced value $wR(F^2)$= 0.98% (data set 30). The normal probability plots and residuals as a function of resolution of these data sets are shown in Fig. 1.

6. Residuals *VS.* Resolution, Relative Outliers

The artificial data sets are analysed first to study the effects of distorted s.u. values and to compare these to the ideal case of no systematic errors. In the case of adequate s.u.s the points of the $(\varsigma, \sin\theta/\lambda)$ plot (Fig. 1 b) arrange themselves smoothly and symmetrically around the zero line, and fade out gently from right to left. This smooth fade out appears because the corresponding Gaussian probability density function is naturally mapped out from the frequent events close to zero to the less frequent events in a distance of $\pm 1\sigma, \pm 2\sigma, \ldots, \ldots$ from the zero line. The scale for the distance from zero is given by σ, the standard deviation of the residual population. As the number of data points per resolution range becomes less for the low resolution range, the corresponding plot shows less frequently points in a distance from the zero line. To eliminate the dependency from the number of reflections per resolution range, the residuals are plotted against the ranking of resolution in Fig. 1 c. As expected for a data set without systematic errors, this plot shows a uniform distribution without discontinuities and without any trends to positive or negative values, with the overwhelming majority of data points in the limits given by $\pm 3\sigma$.

A very similar picture with respect to the residual plots is obtained for the case of overestimated large s.u.s (Fig. 1 e, f): The $(\varsigma, \sin\theta/\lambda)$ plot appears only a bit compressed in the vertical direction. This is because the residuals gather closer around the zero line due to the overestimation of the large s.u.s. It has been observed earlier that the overestimation of large s.u. values is less harmful with respect to the statistical independence of residuals (see 4.2.2 in Henn & Meindl 2014b). This is confirmed here.

In the case of underestimated large s.u. values, the $(\varsigma, \sin\theta/\lambda)$ plot (Fig. 1 h) shows a characteristic shape with a fast decreasing envelop. Here, the frequency of so called "rare events" increases with decreasing resolution, which is in contrast to the cases of adequate, Fig. 1 b, and overestimated large s.u.s, Fig. 1 e. The term "rare event" refers to the expectation of events from a normal distribution, *i.e.* these need not be rare in real data sets. Rare events $|\varsigma| > 3$ appear in approximately 0.27% of all cases, rare events $|\varsigma| > 4$ in approximately 0.0063% of all cases and rare events with $|\varsigma| > 5$ in approximately 0.000057% of all cases, if a Gaussian Normal distribution applies. A list of rare events for the experimental data sets is given in Table 3 of the Appendix. Underestimation of s.u. values leads to an increased rate of rare events.

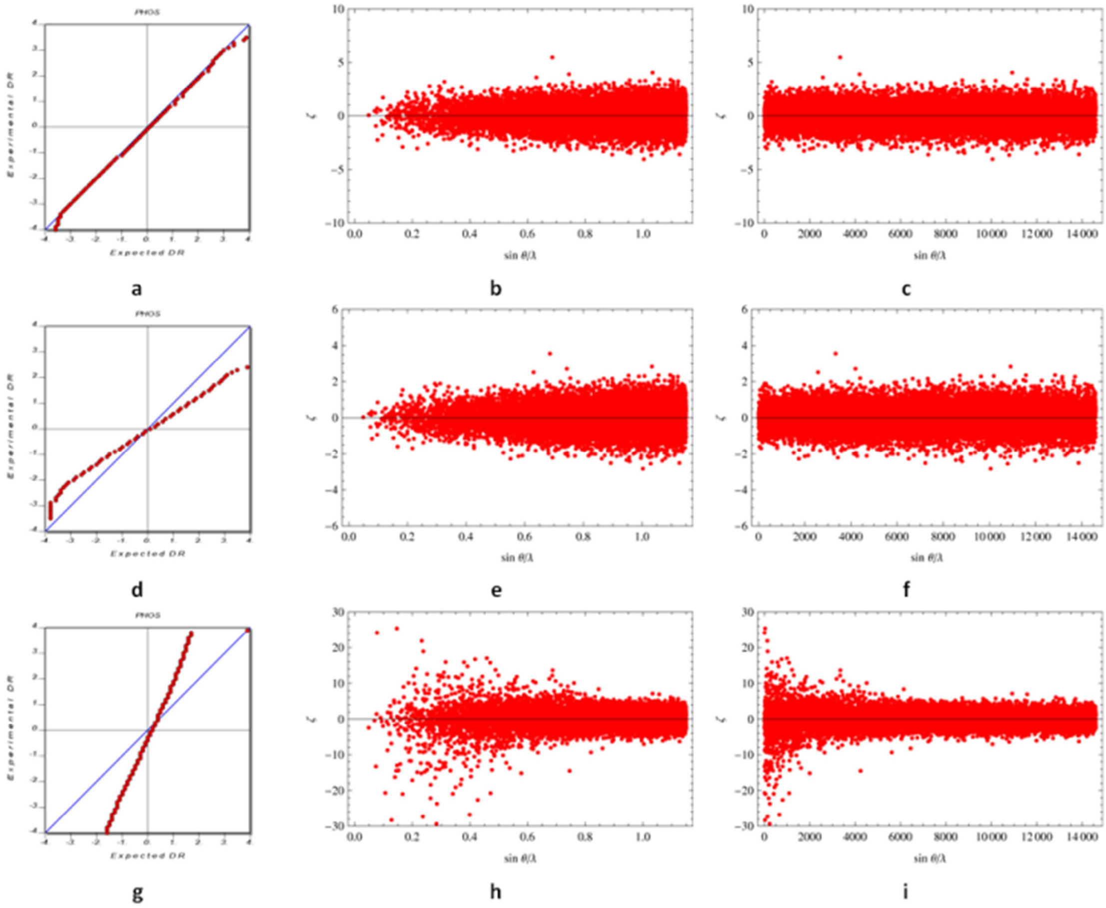


Fig. 1. Effects of over- (second row, d, e, f, data set 31) and underestimation (third row, g, h, i, data set 30) of large s.u. values compared to adequate s.u. values (first row, a, b, c, data set 24) for artificial data sets: Normal probability plots (first column, a, d, g), plots of residuals ς *vs*. resolution $\sin\theta/\lambda$ (b, e, h) and plots of residuals ς *vs*. ranking against resolution $\sin\theta/\lambda$ (third column, c, f, i).

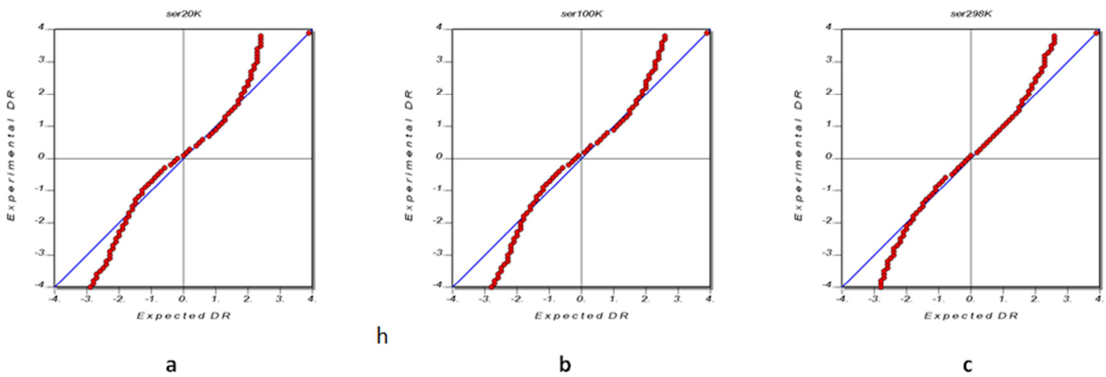


Fig. 2. Distinctly curved npps indicate presence of systematic errors. The npps of data sets 5 (a), 6 (b) and 7 (c).

Table 3. Number of rare events (5-σ) and the respective number of reflections of data sets 1-24, 20 and 31.

Data set	Number of rare events $\lvert\zeta\rvert > 5$/all events	Data set	Number of rare events $\lvert\zeta\rvert > 5$/all events	Data set	Number of rare events $\lvert\zeta\rvert > 5$/all events
1	436/12803	11	3/7981	21	183/2409
2	5/15247	12	0/3942	22	17/2409
3	2898/28457	13	0/3942	23	183/2409
4	5/3822	14	171/8630	24	1/14604
5	17/5136	15	37/8630	30	543/14604
6	15/5146	16	39/8630	31	0/14604
7	6/3551	17	1032/2652		
8	0/8057	18	456/2652		
9	1/8070	19	448/2652		
10	5/7986	20	173/8015		

Rare events are actually observed in crystallographic applications so frequently that it was even suggested to adapt the probability density function to the frequent appearance of rare events (Hooft*et al.* 2009). We suggest that at least a part of rare events in many crystallographic applications including charge density studies is caused by incorrect s.u.s. It is inappropriate in these cases to adapt the probability density function to the resulting distribution. An adequate error model is needed in these cases.

For an impression how plots of residuals *vs*. resolution appear in the case of systematic errors, those corresponding to Fig. 2 are depicted in Fig. 3

These plots suggest that there is a distinct dependence of the residuals for a resolution range from zero to approximately 0.35 $Å^{-1}$ followed by a discontinuity at approximately 0.35 $Å^{-1}$ in all three cases. The discontinuity appears in form of a very sharp increase of the moving average. An additional discontinuity appears at approximately 0.95 $Å^{-1}$ in the cases of data sets 5 (Fig. 3 a, d) and 6 (Fig. 3 b, e) and at approximately 0.75 $Å^{-1}$ for data set 7 (Fig. 3 c, f). Figure 3 f shows the large spread of residuals typical for underestimated large s.u. values for low resolution values.

That the frequency of outliers increases with decreasing resolution instead of being independent from the resolution is observed for many experimental data sets (in which the resolution information was available), for more information see column C of Fig. 5 in the Appendix. It should be stressed again, however, that *any* discontinuity in plots of residuals *vs*. ranked resolution values indicates *per definitionem* an error, as in the case of absence of systematic errors these plots are featureless.

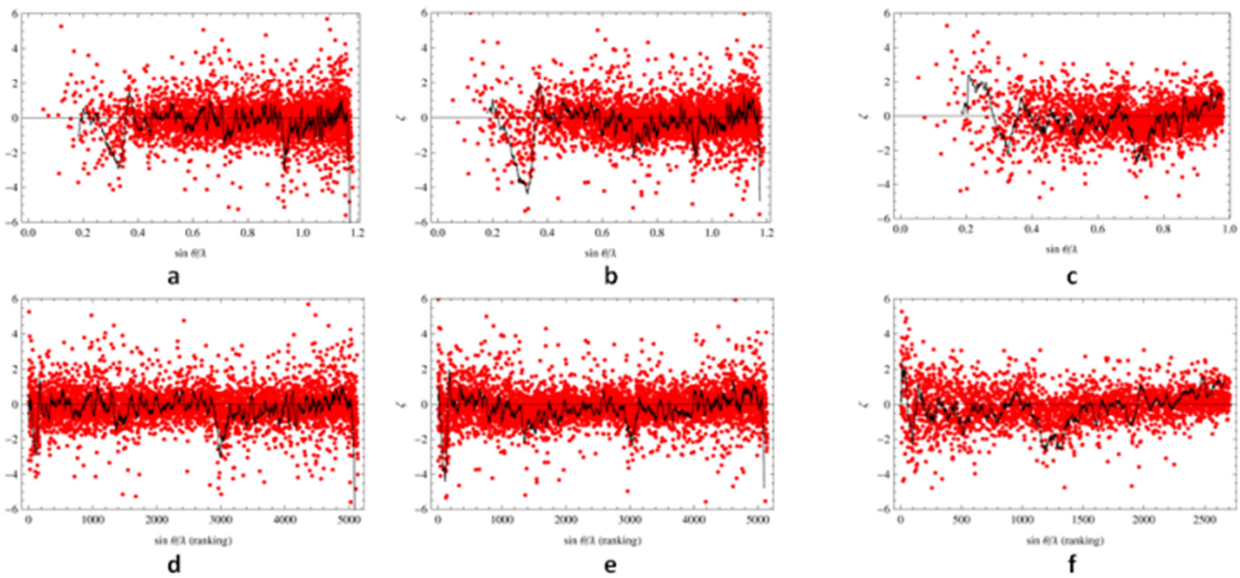


Fig. 3. Residuals *vs.* (ranked) resolution for data sets 5 (a, d) (-14.56/5.69), 6 (b, e) (-14.45/6.25) and 7 (c, f) (-6.66/6.93). The values in brackets give the minimum and maximum residual for each data set. The black line indicates a moving average value over 50 consecutive data points. For clear visibility, the moving average was amplified by a factor 3.

More of these discontinuities are found for data sets 3 and 20 at $\sin\theta/\lambda \approx 0.35$ Å^{-1}. This discontinuity in the behaviour of the residuals is most likely caused by merging different resolution batches. The residuals between zero and 0.35 Å^{-1}seem to be organized so as to produce predominantly positive residuals in the beginning and negative residuals towards the end of this range. This pattern is repeatedly observed in different data sets (see Appendix).

Data sets 15 and 16 each show a markedly downturn beyond approx. 1.2 Å^{-1}, that lead to 37 (set 15) and 38 (set 16) outliers with $|\varsigma| > 5$ in a resolution range starting from 1.165 Å^{-1}.

Plots of residuals *vs*. ranked s.u. values are shown in Fig. 4 to give more evidence for the presence of underestimated (large) s.u.s. and to study their influence. The sorting is in ascending order, *i.e.* the weak s.u. values are found to the left of each plot and the strong s.u. values to the right.

All three plots in Fig. 4 (a - c) show a slight tendency to get broader to the right. This is most clearly seen from Fig. 4 c. In plots Fig. 4 a and b, this tendency is obstructed by the appearance of many large residuals in the middle part of each plot with a tendency to negative residuals (this is more clearly seen for plots with a larger plot range in the direction of the ordinate, see for these plots Fig. 6 in the Appendix). The plots of observed *vs*. calculated intensities show systematically reduced strong observed intensities in Fig. 4 (d - f). The (100-data points moving) average values for the residuals are negative for almost all reflections but those with the largest s.u. values, where the average becomes suddenly positive in all three cases. When many negative residuals are counterbalanced by a small number of positive residuals, the distribution of residual is dis-balanced, as each reflection, regardless of weak or strong, regardless of significant or insignificant should have the same probability to produce a positive or a negative residual. If this does not apply, the residuals are not identically and not independent distributed.7. Outliers in Plots of Observed *VS.*Calculated Intensities

The concept of outliers is not easy to specify rigorously as large deviations are expected for data sets being only large enough. In the present context we take a soft attitude to the concept of outliers and apply this term to individual reflections that seem to scatter more distinctly than others. This of course remains a matter of subjectivity. Those outliers in plots of observed *vs*. calculated intensities, (I_o, I_c) plots, reveal immediately individual reflections affected by systematic errors. The opposite of the above sentence is not true: the appearance of systematic errors, such as from distorted s.u. values need not be visible in (I_o, I_c) plots. This is because the model may still predict the observed intensities accurately as measured by absolute deviations between (I_o, I_c). If the s.u. values are underestimated, however, those predictions appear as outliers in npp and other analysis tools that rely on the relative accuracy of the s.u.s, but they need not appear as absolute outliers. The distorted s.u.s from Fig. 1 for example do not lead to outliers in plots of observed vs. calculated intensities (data not shown).

As was seen in the preceding paragraph, it is a characteristic effect of underestimated large s.u.s that these lead to an increase of rare events in the low-resolution part of the data as measured by the residuals. If this is the only source of systematic errors, it does obviously not lead to any absolute outliers. This is because the experimental intensities are still predicted accurately by the model, only by the exaggerated standards of the severely underestimated s.u.s these predictions appear as outliers. The appearance of rare events accompanied by a non-appearance of absolute outliers may therefore be taken as a characteristic of underestimated large s.u.s.

The appearance of absolute outliers is rare in the experimental data sets, however, for example in sets 5, 6, and 7, which are from the same publication (Mondal *et al.*, 2012) the reflections (0,1,-1) and (-2,0,0) appear repeatedly as absolute and, with one exception, as relative outliers (for more information see Fig. 4 (d - f) and Table 1 in the Appendix).

7. Summary

All experimental data sets show the common feature of large slopes in the npps, none shows the opposite behaviour of slopes smaller than one. This is interpreted as an effect of underestimation of s.u. values in almost all data sets. This should be true for the data sets with statistical weights anyway but it seems to be true also for the data sets employing a weighting scheme. This assumption is also in accordance with most of the *GoF* values being larger than one, only those of data sets 8-11 are smaller than one. Despite their small *GoF* values, also these data sets show the characteristic deviations in the periphery of npps and relative large spreads of residuals in the lowest resolution shells. These exceptional data sets need to be analysed separately.

It was discussed and exemplified with artificial data that the underestimation of the large s.u.s leads to an increase of rare events in the low resolution part of the data. The observation of frequent rare events is omnipresent in crystallographic applications. The underestimation of strong s.u.s is a simple explanation for this omnipresence. If no other sources of error are present, the underestimation of strong s.u. values is characterized by a large slope >1 in the (periphery of the) normal probability plots, by a *GoF* > 1, the pronounced appearance of outliers $|\varsigma| > 3$, by a broadened low resolution end in plots of residuals *vs*. ranked resolution, and by the *absence* of absolute outliers in plots of observed *vs*. calculated intensities. These signs of underestimated large s.u. values are present in virtually all data sets but one. In some data sets additional signs of systematic errors appear, such as absolute outliers or markedly curved npps. Even in the high quality data sets 8-12, the deviations in the periphery of the normal probability plots are all such that larger slopes than one are produced and never in the opposite direction.

The fact that underestimated large s.u.values lead to smaller, but unfortunately unrealistic *R*-values, and the fact that the s.u.s come too small from the integration routines and have to be corrected, supports the hypothesis that the correction of s.u. values approaches the adequate values from the underestimated side and that it is more likely that some strong s.u.s remain underestimated rather than overestimated. Underestimated large s.u.s are of course not the only source of systematic errors, as can be seen from the discontinuities (see *e.g.* data set 6 at approx. 0.35 Å^{-1} or data set 9 at approx. 0.30 Å^{-1} in the Appendix) and slopes (see *e.g.* the resolution range starting from 1.00 Å^{-1} of data sets 15 and 16 in the Appendix) in resolution ranges of data in the $(\varsigma, \sin\theta/\lambda)$ - plots.

As a helpful complementary tool for the analysis of the residuals these were plotted against the ranking of resolution and of s.u. values, instead of against the resolution value or s.u. value only and instead of plots against binned values, which omit the important information about the spread of the data. Plots of individual residuals against ranked data lead to the same density of points in each region of the plot and therefore to plots that are much easier to interpret: they should appear in the same way for any resolution (intensity, s.u., etc.) range. Discontinuities in these plots are a sign of systematic errors and point towards data processing errors, *e.g.* from merging different resolution batches. These plots were helpful in identifying the two common sources of systematic errors in charge density studies: the underestimation of large s.u. values and data processing errors, that lead to discontinuities in these plots. We suggest that normal probability plots should be mandatory for each publication in the field of charge density studies.

Appendix

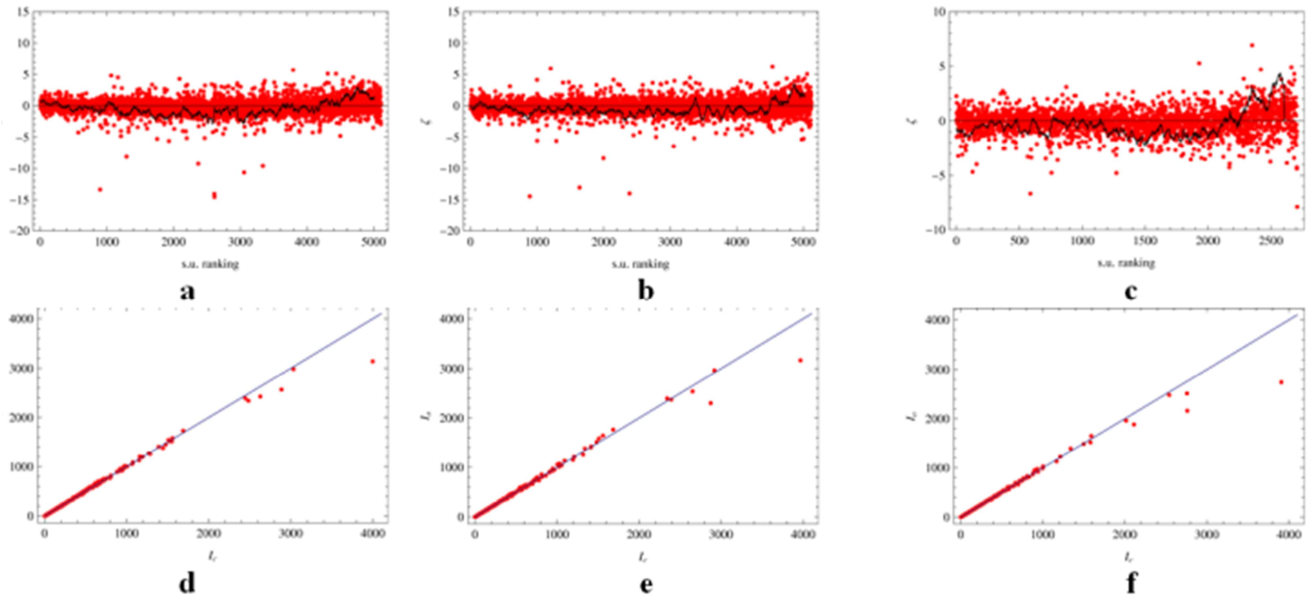
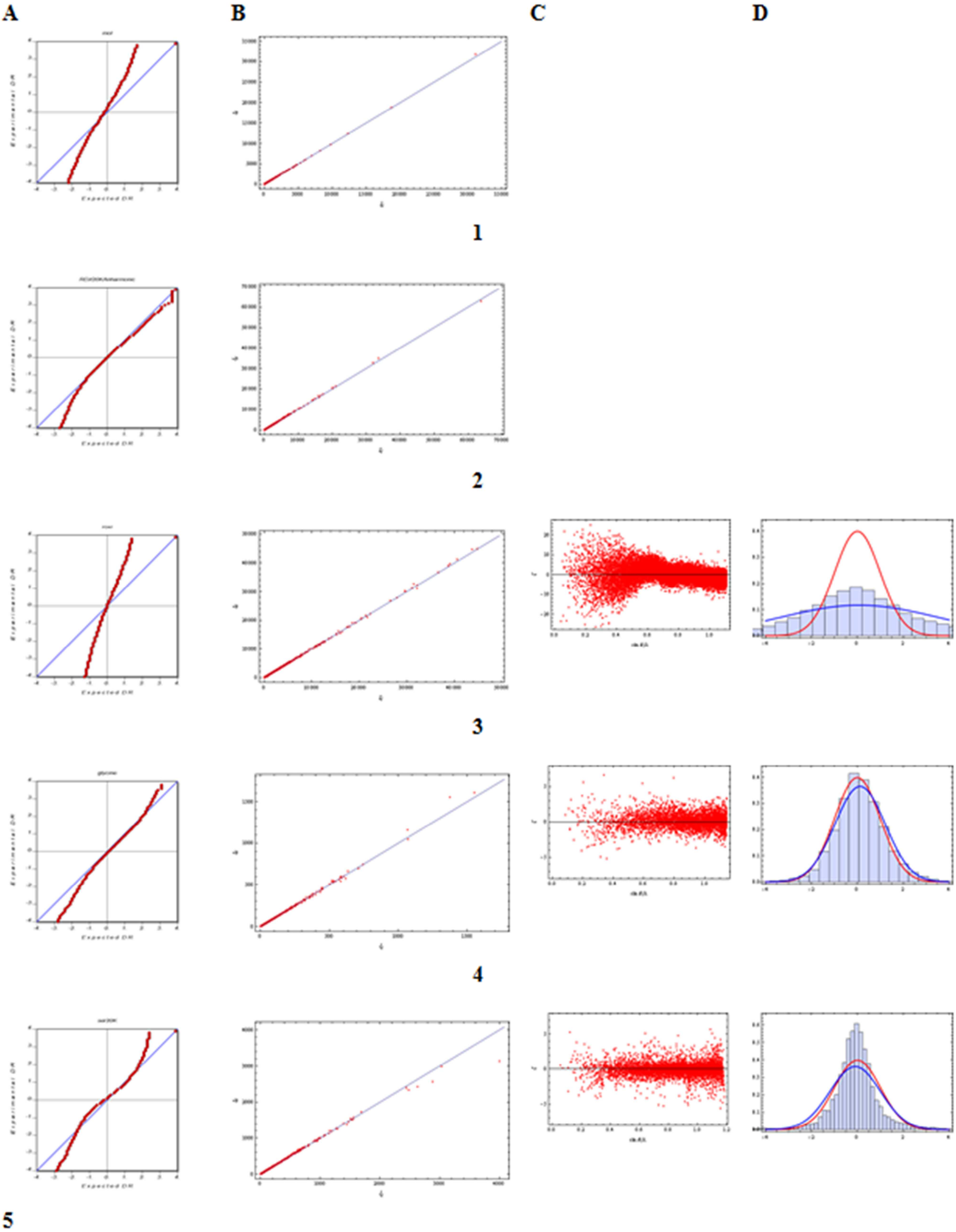
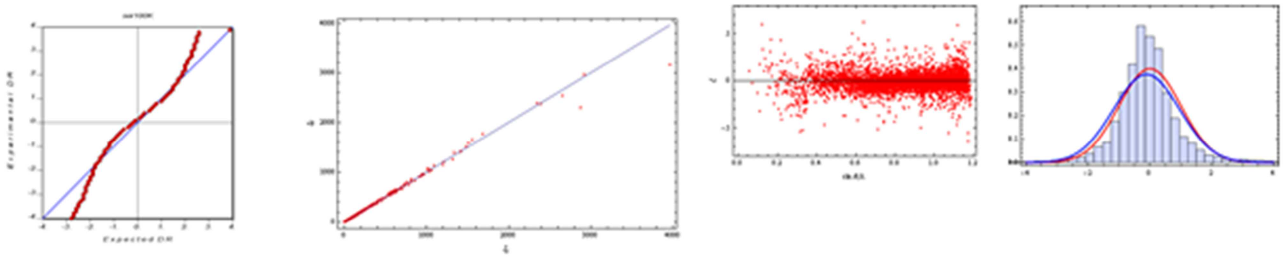
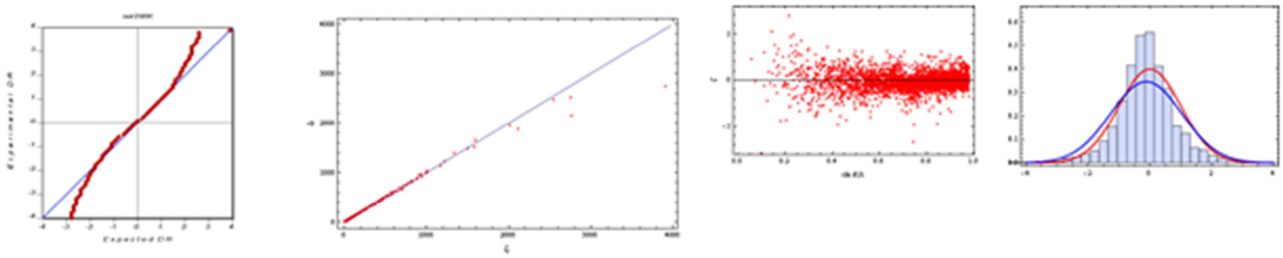


Fig.4. Residuals *vs*. ranked s.u.s (a - c) and plots of observed *vs*. calculated intensities (d - f}) for data sets 5 (a, d), 6 (b, e) and 7 (c, f). The black line indicates a moving average over 100 consecutive data points multiplied by 5 for clear visibility.

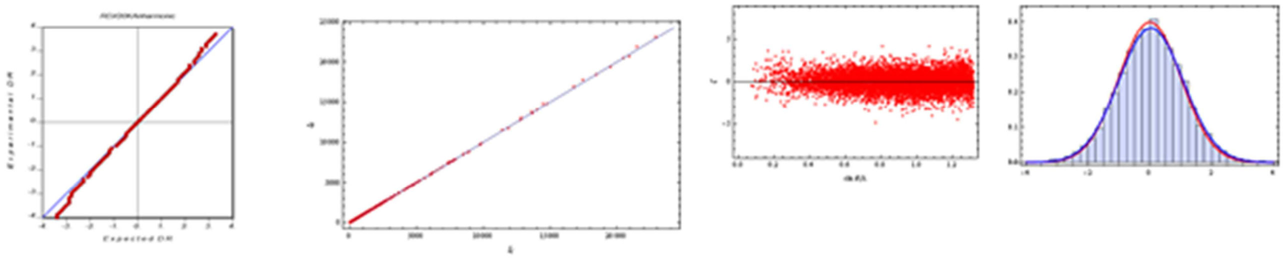

A
B
C
D
1
2
3
4
5



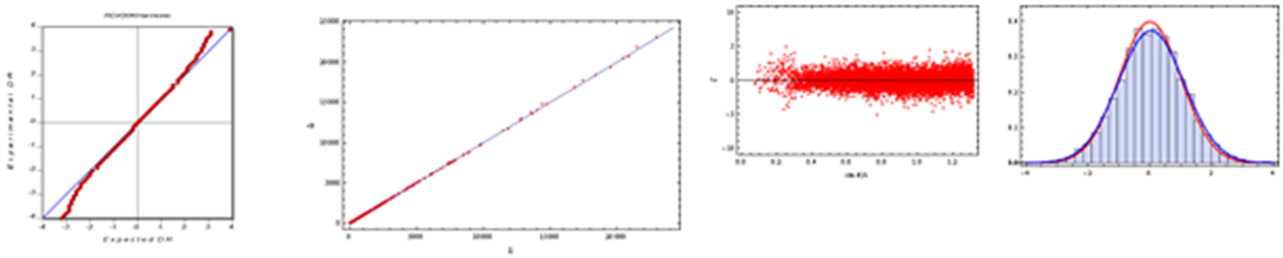
6



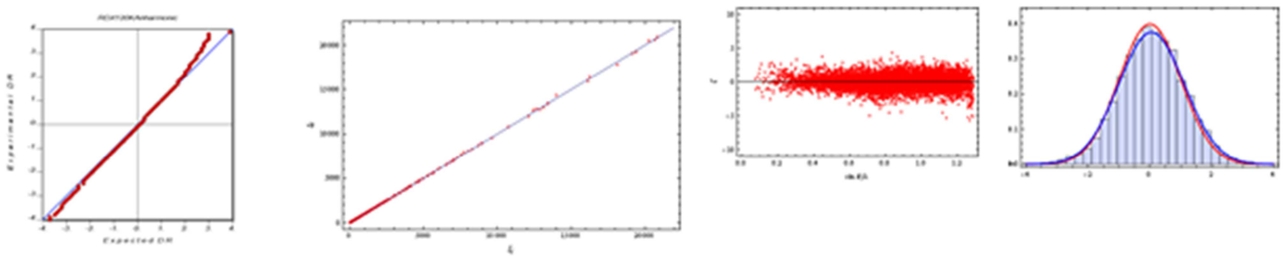
7



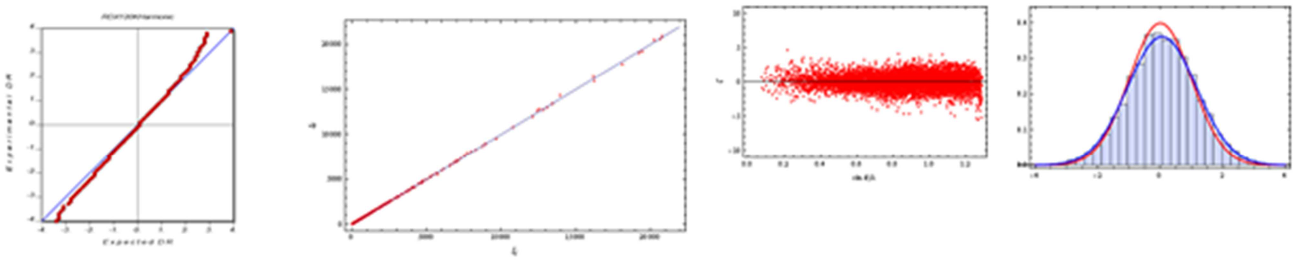
8



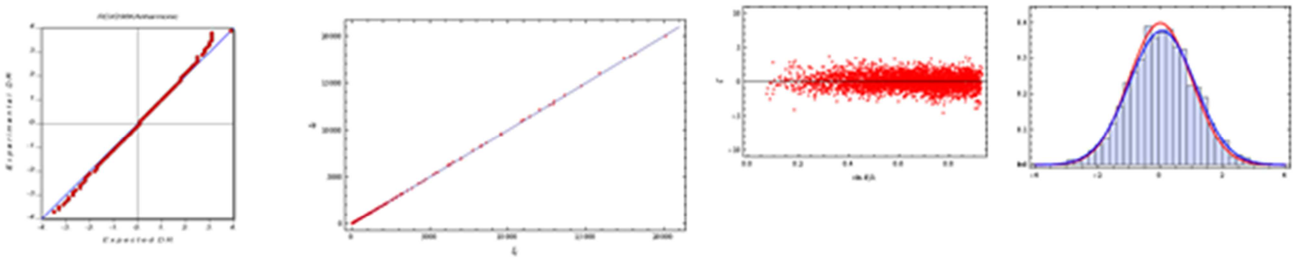
9



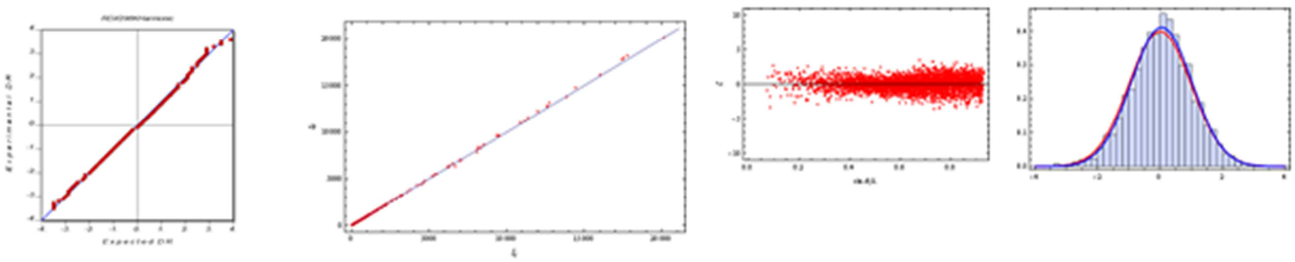
10



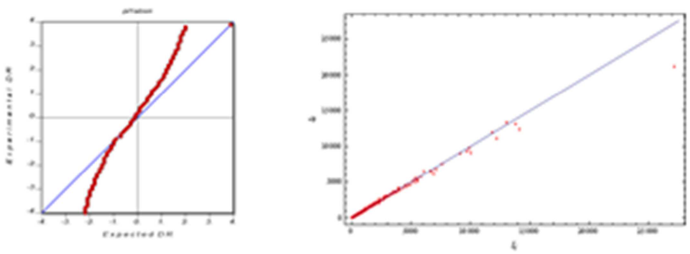
11



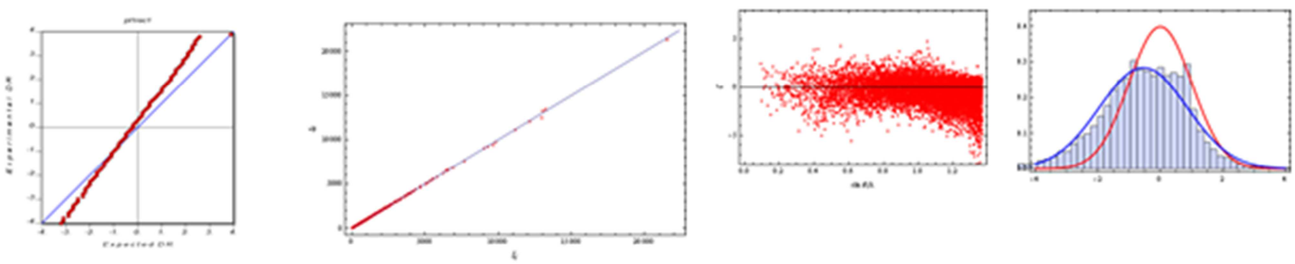
12



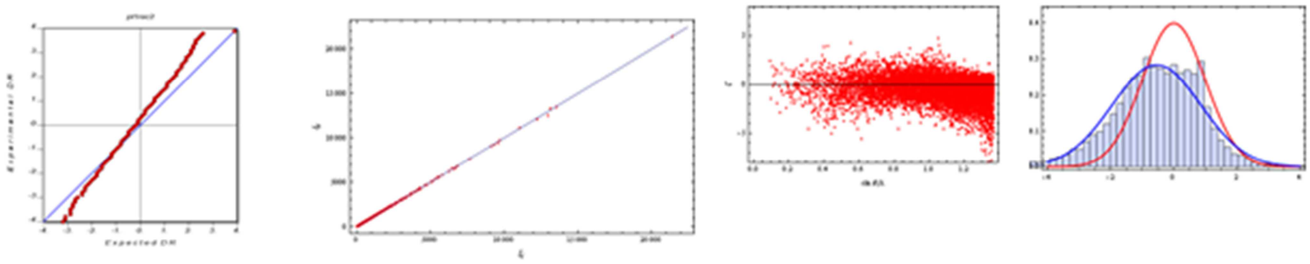
13



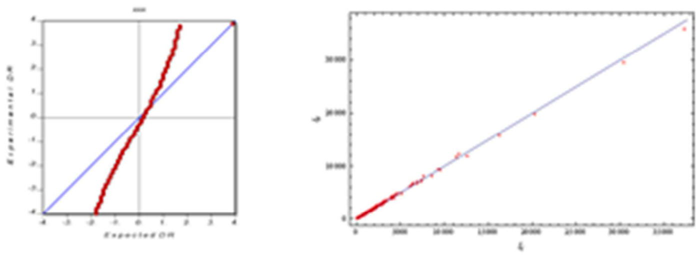
14



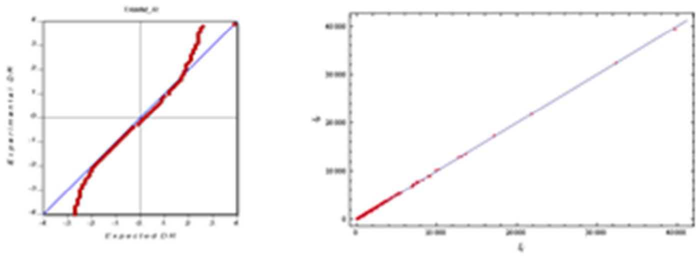
15



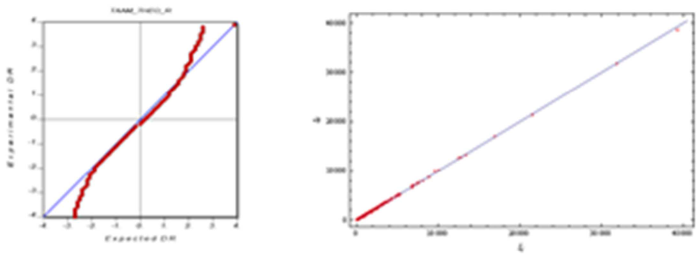
16



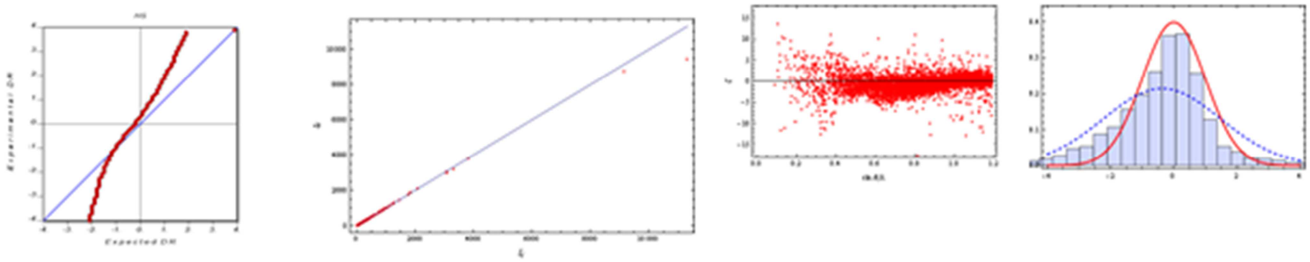
17



18



19



20

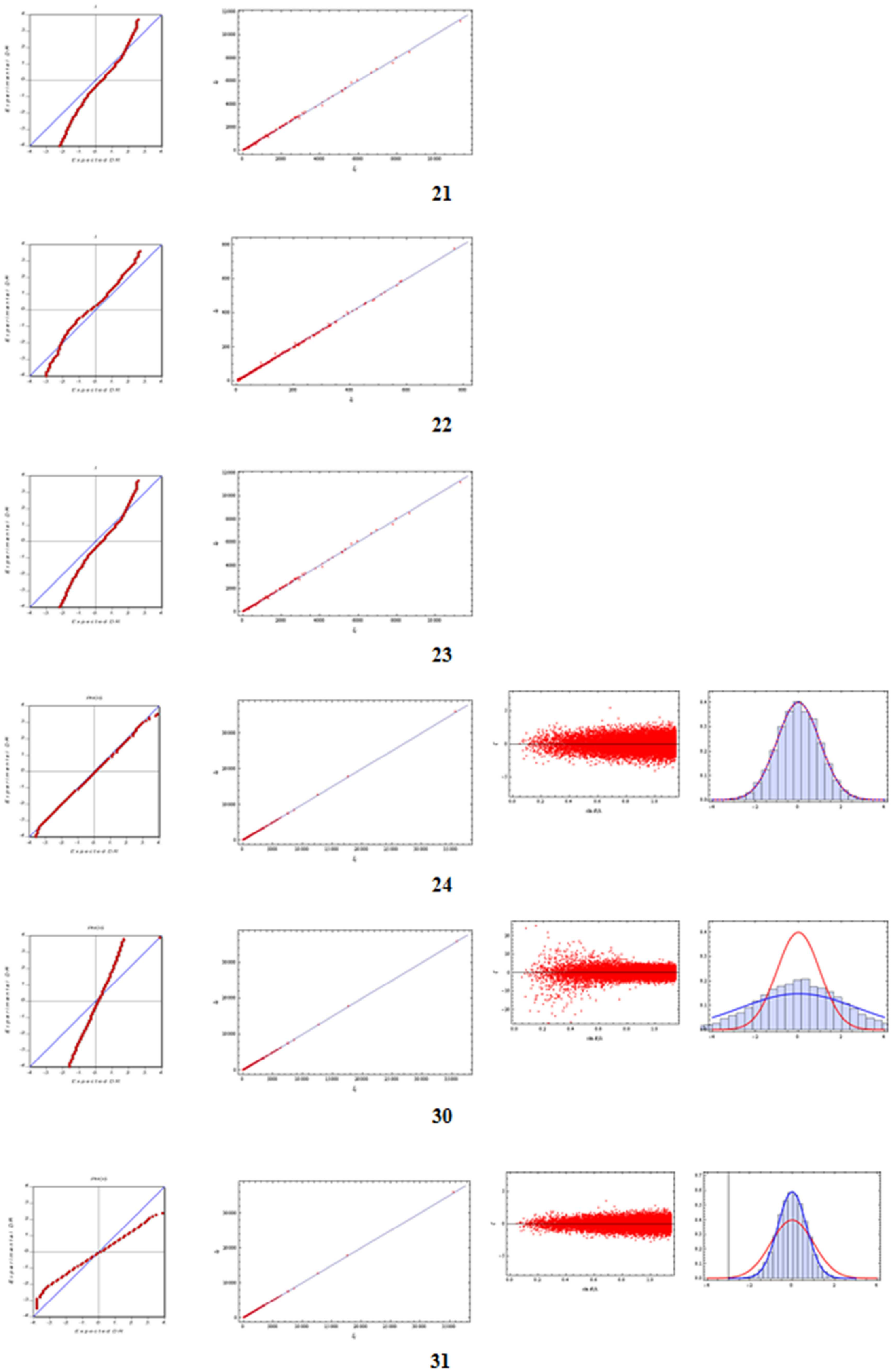


Fig. 5. Normal probability plots (column A), Io *vs.* Ic plots (column B), residuals (ζ) *vs.* sinθ/λ plots (column C) and probability density histograms (column D) with a Normal distribution ($\mu = 0$, $\sigma = 1$) indicated by a red line and a Normal distribution with mean value and standard deviation from the residuals indicated by a blue line. The underestimated large s.u.s (data set 30) cannot be described adequately by a Gaussian, in contrast to the overestimated large s.u.s (data set 31). For the calculation of the npp, the residual vs. resolution plots and for the probability histograms the weights given in the literature were used.

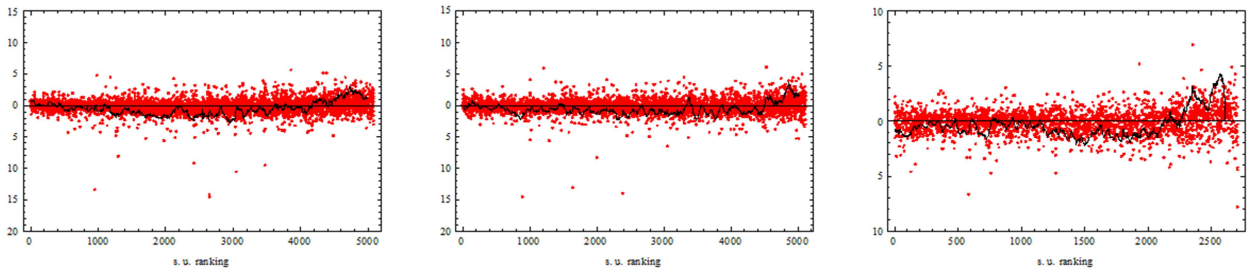


Fig. 6. Residuals vs. ranking of s.u. values for data set 5, 6, and 7.

References

[1] Abrahams, S. C. &Keve, E. T. (1971). ActaCrystallographica Section A, 27(2), 157-165. URL: http://dx.doi.org/10.1107/S0567739471000305

[2] Blessing, R. H. (1997). Journal of Applied Crystallography 30 (4), 421-426.

[3] Evans, P. R. and Murshudov, G. N. (2013). ActaCrystallographica Section D, 69(7), 1204-1214.

[4] Friese, K., Grzechnik, A., Posse, J. M. & Petricek, V. (2013). High Pressure Research, 33(1), 196-201. URL: http://dx.doi.org/10.1080/08957959.2012.758723

[5] Henn, J. & Meindl, K. (2014a). ActaCrystallographica Section A, 70(3), 248-256. URL: http://dx.doi.org/10.1107/S2053273314000898

[6] Henn, J. & Meindl, K. (2014b). ActaCrystallographica Section A, 70(5), 499-513. URL: http://dx.doi.org/10.1107/S2053273314012984

[7] Henn, J. &Meindl, K. (2015). ActaCrystallographica Section A, 71(2), 203-211. URL: http://dx.doi.org/10.1107/S2053273314027363

[8] Henn, J. & Schönleber, A. (2013). ActaCrystallographica Section A, 69, 549-558.

[9] Hooft, R. W. W., Straver, L. H. & Spek, A. L. (2009). ActaCrystallographica Section A, 65(4), 319-321. URL: http://dx.doi.org/10.1107/S0108767309009908

[10] Mondal, S., Prathapa, S. J. & van Smaalen, S. (2012). Acta Crystallographica Section A, 68(5), 568-581. URL: http://dx.doi.org/10.1107/S0108767312029005

[11] Waterman, D., Evans, G. (2010), Journal of Applied Crystallography, 43, 1356-1371.Weiss. M. (2001). Journal of Applied Crystallography, 34, 130-135

[12] Schwarzenbach, D., Abrahams, S. C., Flack, H. D., Gonschorek, W., Hahn, T., Huml, K., Marsh, R. E., Prince, E., Robertson, B. E., Rollett, J. S. & Wilson, A. J. C. (1989). ActaCrystallographica Section A, 45(1), 63–75.

[13] Sørensen, H. O. & Larsen, S. (2003). Journal of Applied Crystallography, 36(3), 931-939.

[14] Zhurov, V. V., Zhurova, E. A., Stash, A. I., Pinkerton, A. A. (2011), ActaCrystallographica Section A, 67(2) 160-173.

[15] Zhurov, V. V. and Zhurova, E. A. and Pinkerton, A. A. (2008). Journal of Applied Crystallography 41(2), 340-349.