SHORT COMMUNICATION

Contributions intended for publication under this heading should be expressly so marked; they should not exceed about 1000 *words; they should be forwarded in the usual way to the appropriate Co-editor; they will be published as speedily as possible. Publication will be quicker if the contributions are without illustrations.*

Acta Cryst. (1975). B31, 1507

Effects of data set thresholds.

By R. E. STENKAMP and L. H. JENSEN, *Department of Chemistry and Department of Biological Structure, University of Washington, Seattle, Washington* 98195, *U.S.A.*

(*Received* 21 *March* 1974; *accepted* 8 *January* 1975)

The effect of different data thresholds on the conventional R, R_w, and the 'goodness of fit' have been investigated for two representative organic structures, one a dipeptide in the centrosymmetric space group $P2_1/c$, the other a dipeptide derivative in the noncentrosymmetric space group $P2_1$. For a given model and data of given quality, both R and R_w decrease as the threshold value increases, but R_w is less sensitive to the value chosen than R. On the other hand, R_w is more sensitive than R to the value of any term added to the counting error in the expression for the standard deviation. The effects of different thresholds on the 'goodness of fit' are also tabulated, and for the centrosymmetric structure the effects on the bond lengths and thermal parameters and on the precision of the positional parameters have been investigated.

The conventional index R is commonly used to monitor the refinement of a structural model. Although other indicators may be preferred, it is a useful index, and it is still the one most widely reported. Its value, or that of any other index, will depend not only on the precision of the data and the accuracy of the model, but also on the value of any threshold used to exclude weak reflections from the data set.

For a simple artificial test case, Hirshfeld & Rabinovich (1973) have demonstrated the systematic effect of a threshold on parameters such as thermal and scale. They recognize that in a typical real case the effect on 'structurally interesting parameters is rarely large enough to matter'. The results of Hirshfeld & Rabinovich are based on refinement on F^2.

We are interested in results from real data sets when refining on F. In refining on F, however, negative observations cannot be properly handled. Such observations may be omitted from the data or given zero weight, but this in effect introduces a threshold, albeit a small one. A more common practice, possibly with vestiges traceable to visually estimated photographic data, has been to use higher thresholds such as $\sigma(I)$ (O'Connor, 1973) or $2\sigma(I)$ (Hjortås, 1973; Frank & Degen, 1973) but even higher values have been reported (Verbist, Lehmann, Koetzle & Hamilton, 1972; Oskarsson, 1973). Thresholds may be given in terms of either F or I, but note that a threshold of $n\sigma(I)$ is equivalent to one of $2n\sigma(F)$.

In view of the variety of thresholds used in crystal-structure studies, it is of interest to know for some representative cases the extent to which the magnitude of the threshold affects R, R_w, and the 'goodness of fit' (GOF). Furthermore, it is important to know if bond lengths and thermal parameters differ significantly for refinements based on data sets with different thresholds and to know to what extent precision is affected.

Refinement with $\sigma(I) \simeq 2\sigma(F)$ threshold

We have investigated the effect of different threshold values for two structures: L,D-alanyl-D,L-methionine, ALMET ($C_8H_{16}N_2O_3S$: Stenkamp & Jensen, 1974), a centrosymmetric dipeptide structure in space group $P2_1/c$, and *N*-acetyl-L-phenylalanyl-L-tyrosine, NAPT ($C_{20}H_{22}N_2O_5$: Stenkamp & Jensen, 1973), a noncentrosymmetric dipeptide derivative structure in space group $P2_1$. Both structural models were refined on F by full-matrix or block full-matrix least-squares calculations with data sets limited to reflections with $\sigma(I) < I$ except that reflections which calculated greater than $\sigma(I)$ were included in the refinement with $\Delta F = (|F_{\text{thresh}}| - |F_c|) \exp(i\alpha_c)$. Weights were taken as $1/\sigma^2(F)$ which were based on the expression $\sigma(I_{\text{rel}}) = \sqrt{\sigma_c^2 + (kC)^2}$ where C is the scan count and $k = 0{\cdot}01$ for NAPT and 0·03 for ALMET.

In Table 1 we list for each threshold, $n\sigma(F)$, the number of reflections in the data set with F's greater than the threshold values and the percentage of the total this represents followed by R, R_w and the GOF. All values were calculated on the basis of the number of reflections greater than the threshold.

Inspection of Table 1 shows that the higher the threshold the smaller R. This is expected, of course, because progressively more of the weak reflections are eliminated. R_w is also smaller for the data sets with higher thresholds, but it is much less sensitive than R to the threshold value used.

For a given threshold, R for the centrosymmetric structure is greater than that for the noncentrosymmetric one. This is consistent with the fact that the centrosymmetric structure has a larger proportion of weak reflections. As the threshold is raised, however, the difference between the R's for the two structures decreases.

For the lower thresholds, R_w for both structures is much less than R. This follows from the fact that the weak reflections with smaller weights contribute relatively less to R_w than to R. For the centrosymmetric structure, R and R_w are equal for the data set with threshold $4\sigma(F)$; and for higher thresholds, R is less than R_w. For the noncentrosymmetric structure, R is still greater than R_w for the highest thresholds used, but the trend suggests it would become less than R_w at thresholds beyond the limit investigated.

The GOF increases with increasing threshold for both structures although it appears to reach a plateau at the $3\sigma(F)$ threshold for ALMET. The increase is consistent

Table 1. *R*, R_w *and 'goodness of fit' as a function of the threshold for refinement based on data set with* $\sigma(I)$ *threshold [equivalent to* $2\sigma(F)$ *threshold]*

$R=\sum||F_o|-|F_c||/\sum|F_o|$, $R_w=[\sum w(|F_o|-|F_c|)^2/\sum w|F_c|^2]^{1/2}$, GOF $=[\sum w(|F_o|-|F_c|)^2/(\text{NREF}-\text{NVAR})]^{1/2}$ where NREF = number of reflections and NVAR = number of parameters. NVAR = 191 for ALMET and 312 for NAPT. Crystals of both structures were of equal volume, 3×10^{-3} mm^3.

	ALMET					NAPT				
$n\sigma(F)$	NREF	%	R	R_w	GOF	NREF	%	R	R_w	GOF
$0\sigma(F)$	2489	100	0·091	0·062	1·44	2957	100	0·055	0·038	2·25
1	2137	85·9	0·073	0·061	1·55	2803	94·8	0·049	0·037	2·30
2	1986	79·8	0·068	0·060	1·59	2721	92·0	0·047	0·037	2·33
3	1836	73·8	0·062	0·059	1·63	2625	88·8	0·046	0·037	2·37
4	1695	68·1	0·056	0·056	1·63	2513	85·0	0·044	0·037	2·41
5	1579	63·4	0·051	0·054	1·63	2411	81·5	0·042	0·037	2·46
6	1473	59·2	0·046	0·051	1·61	2315	78·3	0·040	0·036	2·49
7	1381	55·5	0·044	0·050	1·62	2205	74·6	0·038	0·036	2·53
8	1306	52·5	0·042	0·049	1·64	2125	71·9	0·037	0·035	2·56

Table 2. *R*, R_w *and 'goodness of fit' for ALMET as a function of threshold for refinements based on data sets with* $2\sigma(I)$ *and* $4\sigma(I)$ *thresholds [equivalent to* $4\sigma(F)$ *and* $8\sigma(F)$ *respectively]*

	Refined with $4\sigma(F)$ threshold				Refined with $8\sigma(F)$ threshold		
$n\sigma(F)$	NREF	R	R_w	GOF	R	R_w	GOF
$0\sigma(F)$	2489	0·092	0·062	1·44	0·092	0·062	1·45
1	2137	0·073	0·061	1·55	0·074	0·061	1·55
2	1986	0·068	0·060	1·59	0·068	0·060	1·60
3	1836	0·062	0·059	1·63	0·063	0·059	1·63
4	1695	0·056	0·056	1·63	0·056	0·057	1·64
5	1579	0·051	0·054	1·62	0·051	0·054	1·63
6	1473	0·046	0·051	1·60	0·046	0·051	1·60
7	1381	0·044	0·050	1·62	0·044	0·049	1·60
8	1306	0·042	0·049	1·63	0·042	0·048	1·62

with the observation that weak data sets with relatively larger random errors tend to give values for the GOF approaching more nearly the ideal value of unity.

The GOF exceeds unity for both structures, but by a much wider margin for NAPT than for ALMET. About half the difference is accounted for by the different value of k used in the two refinements (Stenkamp & Jensen, 1974). Part of the remaining difference can be accounted for by the presence of the heavy S atom in ALMET, but an additional reason for the more nearly ideal value of the GOF follows from the larger proportion of weak reflections for the centrosymmetric structure. Possibly a more important reason than either of the above can be found in the lack of precision of the ALMET data. Although the ALMET crystal was essentially the same size as the one for NAPT, it was of poorer quality and the data, therefore, are poorer. Random errors tend to be dominant, masking systematic errors in the data and model and leading to a more nearly ideal value of the GOF. Indeed, if the data have large random errors and if the weights are properly chosen, the GOF will approach its ideal value, but R and the standard deviations in the model parameters will be poor.

Refinement with other thresholds

Since the results in Table 1 are based on refinements of data sets with $\sigma(I)\simeq2\sigma(F)$ thresholds, we considered it essential to test for possible differences in the results when refining data sets with other thresholds. For these tests we used the ALMET data with thresholds of $2\sigma(I)\simeq4\sigma(F)$ and $4\sigma(I)\simeq8\sigma(F)$ and coordinates from the next-to-last least-squares cycle with the $\sigma(I)\simeq2\sigma(F)$ threshold. The results are summarized in Table 2. Comparison of the corresponding entries in this table with those in Table 1 shows them to be essentially the same irrespective of the threshold used for the refinement.

Table 3. *Positional standard deviations for the* S *atom*, C, N *and* O *atoms, and* H *atoms for refinements of ALMET data based on different thresholds*

Threshold	S	R.m.s. σ's C, N, O	H
$2\sigma(F)$	0·00120 Å	0·00319 Å	0·0482 Å
$4\sigma(F)$	0·00121	0·00331	0·0496
$8\sigma(F)$	0·00121	0·00335	0·0495

Although there are fewer reflections in the data sets with higher thresholds, the increase in the positional σ's is small, emphasizing the very small contribution the weak reflections make. The results are summarized in Table 3 where the r.m.s. σ's for the S atom, the mean value for the C, N and O atoms, and the mean for the H atoms are tabulated.

Conclusions

The ALMET and NAPT structures may be considered as representative of small organic structures which can be crystallized and investigated by the methods of X-ray diffraction. These tests show the pronounced dependence of R on the threshold used, particularly for the ALMET structure with its relatively less intense data set, and serve to emphasize the importance of the magnitude of the threshold when comparing R values from different refinements. R_w is found to be much less sensitive than R to the magnitude of the threshold used but it has been shown to be dependent on the value of k in the expression for $\sigma(I_{rel})$ (Stenkamp & Jensen, 1974).

The results from refining the ALMET structure with different thresholds shows that for a *sufficiently extensive* data set there is little loss in precision on eliminating the weak reflections, even when the data set was reduced by almost half. This suggests that in lengthy calculations on such data sets, substantial computational economies can be achieved without loss of significant precision by use of relatively high thresholds. In fact, in the case of the ALMET refinements with different thresholds, none of the bond lengths or thermal parameters were found to differ by as much as one standard deviation.

The results reported here should not, however, give license for the indiscriminate use of high thresholds, particularly so since their use tends to eliminate a disproportionate number of the higher-angle reflections. Instead, each data set should be considered on the basis of its extent in reciprocal space, the precision of the observations, and the number of observations relative to the number of parameters to be determined. The weighting function must also be considered since use of the full data set does not necessarily lead to more reliable parameters unless the validity of the weights has been established. Finally, the results reported here shed no light on the effect of thresholds on convergence to false minima.

We are grateful to Professor V. Schomaker for helpful discussions. This work was supported under Grant GM-10828 from the National Institutes of Health.

References

FRANK, G. W. & DEGEN, P. J. (1973). *Acta Cryst.* B**29**, 1815–1822.
HIRSHFELD, F. L. & RABINOVICH, D. (1973). *Acta Cryst.* A**29**, 510–513.
HJORTÅS, J. (1973). *Acta Cryst.* B**29**, 767–776.
O'CONNOR, B. H. (1973). *Acta Cryst.* B**29**, 1893–1903.
OSKARSSON, Å. (1973). *Acta Cryst.* B**29**, 1747–1751.
STENKAMP, R. E. & JENSEN, L. H. (1973). *Acta Cryst.* B**29**, 2872–2878.
STENKAMP, R. E. & JENSEN, L. H. (1974). *Acta Cryst.* B**30**, 1541–1545.
VERBIST, J. J., LEHMANN, M. S., KOETZLE, T. F. & HAMILTON, W. C. (1972). *Acta Cryst.* B**28**, 3006–3013.