Acta Cryst. (1983). B**39**, 517–525

The Estimation of Average Molecular Dimensions from Crystallographic Data

By Robin Taylor and Olga Kennard

Crystallographic Data Centre, University Chemical Laboratory, Lensfield Road, Cambridge CB2 1EW, England

(*Received* 11 *January* 1983; *accepted* 28 *March* 1983)

Abstract

The estimation of average molecular dimensions from crystallographic data is examined from a statistical point of view. Given a sample of X-ray or neutron determinations of a particular molecular dimension, it is shown that the best estimate of the mean depends on the relative importance of experimental errors and environmental (*e.g.* crystal packing) effects. If experimental errors are more important, the observations should be weighted according to the precision with which they were determined. If environmental effects are more important, an unweighted mean is usually preferable. Alternatively, a 'semi-weighted' mean can be used. The importance of environmental effects can be assessed by a χ^2 test. Particular emphasis is placed on the estimation of average molecular dimensions using the Cambridge Structural Database. This database contains the atomic coordinates of over 30 000 organo-carbon crystal structures, but has only limited information about their e.s.d.'s. However, it is shown that the information can be usefully exploited in the estimation of average molecular dimensions.

I. Introduction

The determination of average molecular dimensions is one of the major objectives of crystallographic research. Apart from their intrinsic interest, compilations of 'standard' bond lengths and angles (Sutton, 1958, 1965) are invaluable as benchmarks in the evaluation of new structural data. Estimates of the average geometries of complete chemical residues [*e.g.* pyranoses (Arnott & Scott, 1972), nucleic acid bases (Taylor & Kennard, 1982*a*)] are useful in constrained least-squares refinement, normalization of E values and model building. In addition, they have a variety of uses in theoretical chemistry.

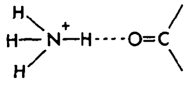
The mean value of a molecular dimension can be estimated in several ways. The simplest procedure is to calculate the unweighted average of all available observations. Alternatively, each observation can be weighted by a factor of $1/\sigma^2$, where σ is the e.s.d. of the observation, derived from the least-squares covariance matrix. The choice between these methods is not straightforward, because the observed values of a given molecular dimension in a series of related crystal structures are influenced by both experimental errors and environmental (*e.g.* crystal packing) effects. The optimum procedure for estimating the mean depends on the relative importance of these factors.

In this study, we apply statistical techniques described by Cochran (1954)* to the problem of estimating mean molecular dimensions. Particular emphasis is placed on the estimation of mean molecular dimensions using the Cambridge Structural Database (CSD; Allen *et al.*, 1979). CSD contains the atomic coordinates of over 30 000 organo-carbon crystal structures, but has only incomplete information about their e.s.d.'s.

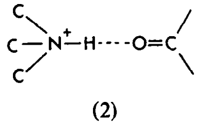
II. Preliminary considerations

II. 1. *How meaningful is a mean?*

It is always *arithmetically* possible to calculate the average value of a sample of crystallographic observations. It is not always useful, or even physically meaningful, to do so. Thus, we could not obtain a sensible mean value for the distance between bonded carbon atoms by averaging a mixture of C–C, C=C and C≡C bond lengths. The average value of a sample may be meaningful in some situations but not in others. For example, we recently found (Taylor, Kennard & Versichel, 1983) that the mean H···O distance of a sample of hydrogen bonds of type (1) was significantly longer than that of a sample of hydrogen bonds of type (2).



(1)



(2)

* Virtually all the statistical formulae discussed here are taken from Cochran's excellent paper, to which readers are referred for mathematical proofs, *etc.*

0108-7681/83/040517-09$01.50

Since this result is of some chemical interest, the estimation of the mean H···O distances was useful *in this case*. However, the same mean values would be of limited use in model building because the hydrogen bond is very sensitive to its environment. Thus, the H···O distance of any given hydrogen bond is unlikely to be close to the corresponding average value. In deciding whether an average value is meaningful, it is therefore important to consider the context in which it is to be used.

Average molecular dimensions derived from crystallographic data may be systematically different from the equilibrium values of the corresponding dimensions in the gas phase. For example, the O···O hydrogen-bond distance in ice I is 2·75 Å (Kamb, 1968), compared with an equilibrium O···O separation of 2·98 Å in the gas-phase water dimer (Dyke, Mack & Muenter, 1977). The difference is partly due to cooperative interactions that occur in the solid state but not in the vapour phase, and partly due to the compressive effect of the crystal lattice. Furthermore, crystallographic observations may be significantly affected by libration, and therefore not comparable with dimensions determined by other physical methods. Thermal-motion corrections based on the segmented-body model (Johnson, 1970) can be as large as 0·02–0·04 Å for C—H and N—H bond lengths, even at 23 K (Jeffrey, Ruble, McMullan, DeFrees & Pople, 1981). There are, of course, other reasons why dimensions determined by different physical methods may not be comparable (Robiette, 1973); for example, X-ray observations may be influenced by the effects of non-spherical electron density.

II. 2. *Assumption of random sampling*

All the statistical methods outlined below are based on an assumption of random sampling; *i.e.* we assume that the available observations of the molecular dimension being studied are a random sample of all possible observations. In crystallographic research, this may not always be the case: crystal structure determinations are not performed at random, but for specific reasons (we hope!). For example, many structures have been determined by neutron diffraction because they contain very short, possibly symmetrical, O—H···O hydrogen bonds. Thus, an estimate of the mean H···O distance based on all available neutron data is likely to be biased towards short values.

II. 3. *Accuracy of experimental error estimates*

Many of the statistical formulae given below involve the e.s.d.'s of crystallographic observations. The accuracy of these quantities is therefore of some importance. Previous work (Hamilton & Abrahams, 1970; James & Williams, 1973; Taylor & Kennard, 1983; Verbist, Lehmann, Koetzle & Hamilton, 1972) shows that e.s.d.'s derived from least-squares covariance matrices are usually too small, because systematic errors in the diffraction experiment are not taken into account. Most studies suggest that the e.s.d.'s should be multiplied by a factor of about 1·3–2·0 in order to reflect the true experimental uncertainties in the observations. We therefore adopt the following procedure in this paper: the e.s.d.'s obtained from least-squares analysis are multiplied by 1·5, and the resulting 'corrected' values are regarded as exact estimates of the true experimental standard deviations of the corresponding observations. Of course, the true correction factor for any given e.s.d. is unlikely to be exactly 1·5. However, e.s.d.'s are usually only quoted to one or two significant figures, so small inaccuracies in the correction factor are immaterial.

III. Mathematical models: testing for environmental effects

We now assume that: (i) a total of k crystallographic observations of a given molecular dimension are available (x_i, $i = 1, 2, \ldots, k$), (ii) each observation, x_i, has associated with it a 'corrected' e.s.d., $\sigma(x_i)$, which can be regarded as an exact estimate of the experimental standard deviation of the observation, (iii) the average value of the observations is physically meaningful in the context in which it is to be used, (iv) the effects of libration, non-random sampling, *etc.*, are acceptably small.

In order to estimate the mean value of the dimension, we must first choose between two mathematical models for the data. In model A, we assume that the ith observation can be expressed as:

$$x_i = \mu + \varepsilon_i \qquad (1)$$

where μ is the true value of the molecular dimension and ε_i is the experimental error in its measurement. If we choose this model, we are effectively assuming that environmental effects are negligible compared with experimental errors, *i.e.* that the true value of the molecular dimension in the ith crystal structure is not significantly different from its true value in the jth structure, despite the difference in chemical environments. In model B, we assume that the true value of the dimension varies appreciably from one structure to the next because of crystal-packing forces, *etc.* Thus, the ith observation must be expressed as:

$$x_i = \mu_i + \varepsilon_i \qquad (2)$$

where

$$\mu_i \neq \mu_j, i \neq j. \qquad (3)$$

If this model is chosen, we are effectively trying to determine the overall mean, μ, of the μ_i. It is particularly important in this case to consider whether such a quantity is physically meaningful.

In order to decide which of the models is most appropriate, it is necessary to determine whether the observations agree with one another as closely as would be expected from their e.s.d.'s. This can be done by calculating the weighted sum of squares of deviations:

$$\chi^2 = \sum_{i=1}^{k} w_i(x_i - \bar{x}_w)^2 \tag{4}$$

where:

$$w_i = 1/\sigma^2(x_i) \tag{5}$$

and $\bar{x}_w$ is the *weighted mean* of the observations:

$$\bar{x}_w = \sum_{i=1}^{k} w_i x_i / \sum_{i=1}^{k} w_i. \tag{6}$$

If environmental effects are negligible, the weighted sum of squares of deviations follows (approximately) a χ^2 distribution with $(k - 1)$ degrees of freedom. Table 1 illustrates the calculation of χ^2 for: (i) the N(7)—C(8) bond lengths in twelve adenine derivatives, (ii) thirteen N–H bond lengths determined by neutron diffraction, (iii) the N(1)—C(2)—N(3) valence angles in thirteen cytosine rings, (iv) the C(5)—C(4)—O(4) bond angles in fifteen uracil derivatives, (v) the (C)O—P—O(H) valence angles of eight mono-anionic terminal phosphate groups (*i.e.* $R\mathrm{OPO_3H^-}$), (vi) the H···O distances in thirteen N—H···O hydrogen bonds determined by neutron diffraction.* χ^2 values that are statistically significant at the α = 0·05 level are marked. Also given in the table is the mean square value of the observational e.s.d.'s:

$$\overline{\sigma^2(x_i)} = \sum_{i=1}^{k} \sigma^2(x_i)/k \tag{7}$$

and the sample variance:

$$\sigma^2\,(sample) = \sum_{i=1}^{k} (x_i - \bar{x}_u)^2/(k-1) \tag{8}$$

where $\bar{x}_u$ is the *unweighted mean* of the observations:

$$\bar{x}_u = \sum_{i=1}^{k} x_i/k. \tag{9}$$

If environmental effects are small, $\overline{\sigma^2(x_i)}$ and $\sigma^2(sample)$ should be similar in magnitude.

The results in Table 1 show that environmental effects are negligible for the adenine N(7)—C(8) distances (3) and the cytosine N(1)—C(2)—N(3) intra-ring valence angles (4). Presumably, an appreciable amount of energy is required to distort these parameters from their equilibrium values. The χ^2 statistic indicates that environmental effects are significant for the N—H bond lengths and (C)O—P—O(H)

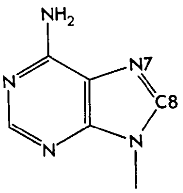
* These data are used throughout the paper to illustrate various calculations. They are subsets of data used in recent surveys of nucleoside geometries (Taylor & Kennard, 1982*b*) and N—H···O hydrogen bonds (Taylor & Kennard, 1983).

Table 1. *Calculation of χ^2 for selected crystallographic data*

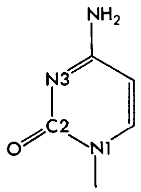
	Adenine N(7)—C(8) bond lengths (Å)		N—H bond lengths (Å) (neutron data)		Cytosine N(1)—C(2)—N(3) bond angles (°)		Uracil C(5)—C(4)—O(4) bond angles (°)		(C)O—P—O(H) bond angles of $R\mathrm{OPO_3H^-}$ groups (°)		H···O hydrogen bond lengths (Å) (neutron data)	
	x_i	$\sigma(x_i)$	x_i	$\sigma(x_i)$	x_i	$\sigma(x_i)$	x_i	$\sigma(x_i)$	x_i	$\sigma(x_i)$	x_i	$\sigma(x_i)$
	1·315	0·003	1·045	0·0015	119·0	0·3	125·6	0·3	105·5	0·15	1·814	0·0015
	1·311	0·003	1·041	0·0015	119·8	0·45	127·2	0·3	106·2	0·3	1·844	0·003
	1·322	0·012	1·054	0·003	118·8	0·6	125·1	0·3	99·3	0·3	1·728	0·003
	1·329	0·012	1·037	0·003	118·5	0·6	126·2	0·45	101·5	0·3	1·832	0·003
	1·347	0·021	1·025	0·003	119·7	0·6	125·2	0·45	105·5	0·3	2·121	0·003
	1·301	0·0225	1·035	0·006	119·6	1·05	125·9	0·45	106·1	0·45	1·997	0·0075
	1·378	0·0285	1·027	0·0075	118·4	1·05	123·5	0·6	104·3	0·45	1·808	0·0075
	1·325	0·030	1·030	0·0075	115·0	1·5	127·0	1·05	103·5	0·45	1·833	0·009
	1·314	0·030	1·048	0·0075	119·2	3·45	126·1	1·2			1·739	0·009
	1·333	0·0315	1·035	0·0075	118·3	5·4	124·4	0·45			1·772	0·009
	1·294	0·045	1·040	0·009	118·6	5·4	126·3	0·45			1·742	0·0105
	1·315	0·045	1·022	0·009	117·7	7·5	126·8	0·9			1·877	0·012
			1·027	0·0105	111·5	7·5	125·9	1·8			1·948	0·012
							125·6	2·1				
							129·9	2·25				
k	12		13		13		15		8		13	
$\overline{\sigma^2(x_i)}$	0·000742		0·0000434		14·50		1·18		0·12		0·00006	
$\sigma^2(sample)$	0·000492		0·0000923		5·33		2·07		6·06		0·01284	
χ^{2*}	11·6 (11)		70·5 (12)†		14·0 (12)		61·7 (14)†		503·8 (7)†		11245 (12)†	

* Degrees of freedom in parentheses.
† Significant at α = 0·05.

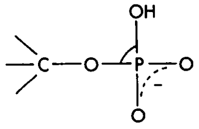
valence angles (5). These parameters are easily distorted and are found in a wide variety of crystallographic environments. The exocyclic C(5)—C(4)—O(4) bond angle in uracil (6) is also significantly affected by its environment, but probably not to the same extent as the (C)O—P—O(H) angle [for which σ^2 (*sample*) $\gg \overline{\sigma^2(x_i)}$]. The H···O distance is an extreme example of a 'soft' parameter which is very sensitive to changes in the environment.



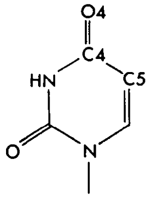
(3)



(4)



(5)



(6)

Our limited experience with the χ^2 test suggests that it is a sensitive indicator of the presence of environmental effects. However, a possible objection to its use is that the test will be invalidated by serious errors in the $\sigma(x_i)$. In case of doubt, a convenient rule-of-thumb is to assume that environmental effects are present unless $\sigma^2(\textit{sample}) < \overline{\sigma^2(x_i)}$. This is always a safe criterion because the statistical techniques appropriate to model *B* are also valid for model *A*, provided that $\sigma^2(\textit{sample}) \geq \overline{\sigma^2(x_i)}$ (see § V). More detailed discussion of methods for choosing between models *A* and *B* can be found in the statistical literature (Cochran, 1954, and references therein).

IV. Model *A*: environmental effects negligible

When environmental effects are negligible, the average molecular dimension is best (*i.e.* most precisely) estimated by using the weighted mean [$\bar{x}_w$ in equation (6)]. The standard error of $\bar{x}_w$ is given approximately by:

$$\sigma(\bar{x}_w) = \left[1/\sum_{i=1}^{k} 1/\sigma^2(x_i)\right]^{1/2} = (1/\sum_{i=1}^{k} w_i)^{1/2}. \quad (10)$$

Table 2 illustrates the calculation of $\bar{x}_w$ and $\sigma(\bar{x}_w)$ for the sample of thirteen cytosine N(1)—C(2)—N(3) valence angles referred to in the previous section. For comparison, the unweighted mean [$\bar{x}_u$ in equation (9)] is also given, together with its standard error:

$$\sigma(\bar{x}_u) = \sigma(\textit{sample})/\sqrt{k} \quad (11)$$

i.e.

$$\sigma(\bar{x}_u) = \left[\sum_{i=1}^{k} (x_i - \bar{x}_u)^2/k(k-1)\right]^{1/2}. \quad (12)$$

In this example, the precision of the weighted mean is better than that of the unweighted mean by a factor of more than 3 (0·64/0·19 ≃ 3·4). However, the unweighted mean of the seven most precise observations (those above the broken line in Table 2) is 119·11°, with a standard error of 0·22°. This suggests that when environmental effects are negligible, the unweighted mean of a subset of the most precise observations in the sample is often a good alternative to the weighted mean of the complete sample. Presumably, this is due to the sharp decrease in w_i as $\sigma(x_i)$ increases (see Table 2).

Equation (10) is based on the assumption that the $\sigma(x_i)$ are exact estimates of the experimental standard deviations of the corresponding x_i (an assumption often made in crystallographic studies). When this is not a good approximation, the value of $\sigma(\bar{x}_w)$ obtained from equation (10) will be too small. The extent to which $\sigma(\bar{x}_w)$ is underestimated in any given case cannot be determined precisely. However, Cochran & Carroll (1953) studied the effect of errors in the $\sigma(x_i)$ (and hence in the weights, w_i) on the variance of the weighted mean. Their results suggest that the neglect of uncertainties in the $\sigma(x_i)$ is unlikely to cause serious underestimation of $\sigma(\bar{x}_w)$ unless equation (10) is used with a very large sample of observations, or the observations are taken from structure determinations with unusually low (reflection/parameter) ratios. In these cases, it might be better to use an unweighted mean.

Table 2. *Calculation of weighted mean: cytosine* N(1)—C(2)—N(3) *bond angles* (°)

x_i	$\sigma(x_i)$	$\sigma^2(x_i)$	w_i [$=1/\sigma^2(x_i)$]
119·0	0·3	0·0900	11·111
119·8	0·45	0·2025	4·938
118·8	0·6	0·3600	2·778
118·5	0·6	0·3600	2·778
119·7	0·6	0·3600	2·778
119·6	1·05	1·1025	0·907
118·4	1·05	1·1025	0·907
115·0	1·5	2·2500	0·444
119·2	3·45	11·9025	0·084
118·3	5·4	29·1600	0·034
118·6	5·4	29·1600	0·034
117·7	7·5	56·2500	0·018
111·5	7·5	56·2500	0·018

$\bar{x}_w = 119{\cdot}07$ $\quad$ $\bar{x}_u = 118{\cdot}01$

$\sigma(\bar{x}_w) = 0{\cdot}19$ $\quad$ $\sigma(\bar{x}_u) = 0{\cdot}64$

V. Model B: environmental effects not negligible

When environmental effects are not negligible, the ith observation is given by:

$$x_i = \mu_i + \varepsilon_i = \mu + (\mu_i - \mu) + \varepsilon_i \qquad (13)$$

where μ is the overall mean of the μ_i, *i.e.* the quantity we wish to estimate (see § III). Part of the variation in the x_i is due to experimental errors and part is due to differences between the μ_i:

$$\sigma^2(sample) = \sigma^2(\mu) + \sigma^2(exptl). \qquad (14)$$

An estimate of $\sigma^2(exptl)$ is provided by the quantity $\overline{\sigma^2(x_i)}$ in equation (7), and $\sigma^2(sample)$ is given by equation (8). Thus, $\sigma^2(\mu)$ can be estimated as:

$$\sigma^2(\mu) = \sigma^2(sample) - \overline{\sigma^2(x_i)} \qquad (15)$$

i.e.

$$\sigma^2(\mu) = \sum_{i=1}^{k} (x_i - \bar{x}_u)^2/(k-1) - \sum_{i=1}^{k} \sigma^2(x_i)/k. \qquad (16)$$

The quantities $\bar{x}_w$ and $\sigma(\bar{x}_w)$ [equations (6), (10)] *are no longer suitable as estimates of the mean value of the sample and its standard error.* This is because they do not take into account the variance due to environmental effects, $\sigma^2(\mu)$. Instead, the average value of the sample is best estimated by the *semi-weighted mean*, $\bar{x}_s$:

$$\bar{x}_s = \sum_{i=1}^{k} W_i x_i \Big/ \sum_{i=1}^{k} W_i \qquad (17)$$

where:

$$W_i = 1/[\sigma^2(\mu) + \sigma^2(x_i)]. \qquad (18)$$

The standard error of $\bar{x}_s$ is given approximately by:

$$\sigma(\bar{x}_s) = \left(1/\sum_{i=1}^{k} W_i\right)^{1/2}. \qquad (19)$$

The semi-weighted mean is a compromise between the weighted ($\bar{x}_w$) and unweighted ($\bar{x}_u$) means. When environmental effects are large compared with experimental errors [*i.e.* $\sigma^2(\mu) \gg \sigma^2(x_i)$ for all i] the semi-weights are all given approximately by:

$$W_i \simeq 1/\sigma^2(\mu) = \text{a constant}. \qquad (20)$$

Thus, $\bar{x}_s$ tends towards $\bar{x}_u$. When environmental effects are small compared with experimental errors [$\sigma^2(\mu) \ll \sigma^2(x_i)$] the semi-weights are approximated by:

$$W_i \simeq 1/\sigma^2(x_i) = w_i \qquad (21)$$

and $\bar{x}_s$ tends towards $\bar{x}_w$. If we are very reluctant to assume that environmental effects are negligible, we can always legitimately use model B (*i.e.* calculate the semi-weighted mean) except when the estimate of $\sigma^2(\mu)$ obtained from equation (16) is negative. Although physically meaningless, this will occasionally happen because of random sampling errors.

Table 3. *Calculation of semi-weighted mean: uracil* C(5)—C(4)—O(4) *bond angles* (°)

x_i	$\sigma(x_i)$	$\sigma^2(x_i)$	W_i $\{=1/[\sigma^2(\mu) + \sigma^2(x_i)]\}$	w_i $[=1/\sigma^2(x_i)]$
124·4	0·45	0·2025	0·7491	4·9383
126·3	0·45	0·2025	0·7491	4·9383
126·8	0·9	0·8100	0·5148	1·2346
125·9	1·8	3·2400	0·2287	0·3086
125·6	2·1	4·4100	0·1804	0·2268
129·9	2·25	5·0625	0·1614	0·1975

$\overline{\sigma^2(x_i)} =$ 2·3213	$\sigma^2(sample) =$ 3·4537	$\sigma^2(\mu) =$ 1·1324
$\bar{x}_s =$ 125·99	$\bar{x}_w =$ 125·60	$\bar{x}_u =$ 126·48
$\sigma(\bar{x}_s) =$ 0·62	$\sigma(\bar{x}_w) =$ 0·29	$\sigma(\bar{x}_u) =$ 0·76

The calculation of $\bar{x}_s$ and $\sigma(\bar{x}_s)$ is illustrated in Table 3 for six of the uracil C(5)—C(4)—O(4) angles given in Table 1. For comparison, the weighted and unweighted means are also given, although we emphasize that it would be inappropriate to use the weighted mean for this sample. The semi-weighted and unweighted means are not significantly different, but $\sigma(\bar{x}_s)$ is somewhat smaller than $\sigma(\bar{x}_u)$. The *apparent* standard error of the weighted mean, calculated from (10), is far too small: 0·29°, compared with $\sigma(\bar{x}_s) = 0{\cdot}62°$.

The calculation of $\bar{x}_s$ and $\sigma(\bar{x}_s)$ is further illustrated in Table 4 for the eight (C)O—P—O(H) angles listed in Table 1. This molecular dimension is very sensitive to environmental effects and the incorrect application of equations (6) and (10) therefore leads to serious errors. *In particular, the standard error estimated from equation* (10) *is too small by approximately an order of*

Table 4. *Calculation of semi-weighted mean:* (C)O—P—O(H) *bond angles* (°)

x_i	$\sigma(x_i)$	$\sigma^2(x_i)$	W_i $\{=1/[\sigma^2(\mu) + \sigma^2(x_i)]\}$	w_i $[=1/\sigma^2(x_i)]$
105·5	0·15	0·0225	0·1678	44·4444
106·2	0·3	0·0900	0·1659	11·1111
99·3	0·3	0·0900	0·1659	11·1111
101·5	0·3	0·0900	0·1659	11·1111
105·5	0·3	0·0900	0·1659	11·1111
106·1	0·45	0·2025	0·1629	4·9383
104·3	0·45	0·2025	0·1629	4·9383
103·5	0·45	0·2025	0·1629	4·9383

$\overline{\sigma^2(x_i)} =$ 0·1238	$\sigma^2(sample) =$ 6·0613	$\sigma^2(\mu) =$ 5·9375
$\bar{x}_s =$ 103·99	$\bar{x}_w =$ 104·36	$\bar{x}_u =$ 103·99
$\sigma(\bar{x}_s) =$ 0·87	$\sigma(\bar{x}_w) =$ 0·10	$\sigma(\bar{x}_u) =$ 0·87

Table 5. *Lower limits of $\sigma(\bar{x}_s)/\sigma(\bar{x}_u)$*

$r = \dfrac{[\sigma(x_i)]_{\text{max.}}}{[\sigma(x_i)]_{\text{min.}}}$	$I = \sigma(\mu)/[\overline{\sigma^2(x_i)}]^{1/2}$						
	0	1	2	3	4	5	10
2	0·80	0·95	0·99	1·00	1·00	1·00	1·00
3	0·60	0·89	0·98	0·99	1·00	1·00	1·00
4	0·47	0·86	0·97	0·99	1·00	1·00	1·00
5	0·38	0·83	0·96	0·99	0·99	1·00	1·00
10	0·20	0·77	0·93	0·97	0·99	0·99	1·00
20	0·10	0·74	0·91	0·96	0·98	0·99	1·00
30	0·07	0·73	0·91	0·96	0·98	0·99	1·00

magnitude. Because environmental effects are very important, the semi-weighted and unweighted means are virtually identical, as are their standard errors. This suggests that there is little advantage in using $\bar{x}_s$ rather than $\bar{x}_u$ when $\sigma^2(\mu)$ is relatively large. Cochran (1954) gives an approximate method for calculating the lower limit of the relative precision of the semi-weighted and unweighted means [*i.e.* the lowest possible value of $\sigma(\bar{x}_s)/\sigma(\bar{x}_u)$] for given values of I and r, where:

$$I = \sigma(\mu)/[\overline{\sigma^2(x_i)}]^{1/2} \qquad (22)$$

$$r = [\sigma(x_i)]_{\text{max.}}/[\sigma(x_i)]_{\text{min.}}. \qquad (23)$$

Representative results are given in Table 5. For example, the table shows that when $\sigma^2(\mu)$ is equal to $\overline{\sigma^2(x_i)}$ (*i.e.* $I = 1$) and the largest observational e.s.d. is five times the smallest one (*i.e.* $r = 5$), the lowest possible value of $\sigma(\bar{x}_s)/\sigma(\bar{x}_u)$ is approximately 0·83. Clearly, the unweighted mean is a good alternative to the semi-weighted mean except when environmental effects are relatively small ($I < \sim 1$) or the range spanned by the $\sigma(x_i)$ is large ($r > \sim 10$).

VI. Estimation using the Cambridge Structural Database (CSD)

VI. 1. '*AS flags*'

CSD contains the atomic coordinates from over 30 000 organo-carbon crystal structure determinations. It is therefore likely to be a major source of crystallographic data in future determinations of average molecular dimensions. Unfortunately, the e.s.d.'s of the atomic coordinates are not stored in the database. Thus, calculation of weighted and semi-weighted means is impossible without reference to the original literature. A limited amount of information about the e.s.d.'s is incorporated in the form of '*AS* flags'. Whenever possible, each new entry to the database is assigned an *AS* flag of 1, 2, 3 or 4, depending on the e.s.d.'s quoted for carbon–carbon distances in the report of the structure (*Cambridge Crystallographic Data Centre User Manual*, 1978). Thus, if the average value of these e.s.d.'s falls in the range 0·001–0·005 Å, the entry is given an *AS* flag of 1. Multiplication by the 'correction factor' of 1·5 (see § II.3) produces a 'true' e.s.d. range of 0·0015–0·0075 Å for $AS = 1$. The corresponding ranges for $AS = 2, 3, 4$, are summarized in Table 6.

It is possible to make use of the *AS* flags in estimating average molecular dimensions. This is illustrated here for the simplest possible case: estimation of the average value of a bond length involving two first-row atoms [*e.g.* the N(7)—C(8) distance in adenine derivatives]. We assume that: (i) each observation in the sample has assigned to it an *AS* flag of 1, 2, 3 or 4, (ii) the sample is moderately large ($k > \sim 20$), (iii) the e.s.d. of the *i*th observation, $\sigma(x_i)$, can fall with equal probability anywhere within the range corresponding to the *AS* flag of the observation, but it cannot fall outside this range. Assumption (ii) is reasonable, since if the sample were small it would be practicable to retrieve the $\sigma(x_i)$ from the original literature. Assumption (iii) is an approximation, one aspect of which is considered later.

VI. 2. *Choice of mathematical model*

Suppose that the sample contains k_1 observations with $AS = 1$, k_2 observations with $AS = 2$, *etc*. The expected mean square value of the (corrected) e.s.d.'s of the observations with $AS = 1$ is given by:

$$E[\overline{\sigma^2(x_i)}]_1 = \int_{0\cdot 0015}^{0\cdot 0075} \sigma^2 \, Pr(\sigma) \mathrm{d}\sigma$$

$$= \int_{0\cdot 0015}^{0\cdot 0075} [\sigma^2/(0\cdot 0075 – 0\cdot 0015)] \mathrm{d}\sigma \qquad (24)$$

(Hamilton, 1964). Hence:

$$E[\overline{\sigma^2(x_i)}]_1 \simeq 0\cdot 0000233 \text{ Å}^2. \qquad (25)$$

The corresponding quantities $E[\overline{\sigma^2(x_i)}]_2$ and $E[\overline{\sigma^2(x_i)}]_3$ can be calculated in an analogous fashion (Table 6). The expected mean square value of the e.s.d.'s of observations with $AS = 4$, $E[\overline{\sigma^2(x_i)}]_4$, cannot be estimated, because the upper limit of the e.s.d. range is undefined. Therefore, these observations must be rejected.

The expected mean square value of the e.s.d.'s of *all* observations in the sample (with $AS = 4$ entries eliminated) can now be estimated as:

$$E[\overline{\sigma^2(x_i)}] \simeq 10^{-7} \left(\frac{233k_1 + 1313k_2 + 9750k_3}{k_1 + k_2 + k_3} \right) \text{Å}^2. \qquad (26)$$

Table 6. *Cambridge Structural Database 'AS' flags*

AS	Uncorrected e.s.d. range (Å)	Corrected e.s.d. range (Å)	$E[\overline{\sigma^2(x_i)}]$	Partial weight $\{=1/E[\overline{\sigma^2(x_i)}]\}$
1	$0\cdot 001 \le \sigma(x_i) \le 0\cdot 005$	$0\cdot 0015 \le \sigma(x_i) \le 0\cdot 0075$	0·0000233	43011
2	$0\cdot 005 < \sigma(x_i) \le 0\cdot 010$	$0\cdot 0075 < \sigma(x_i) \le 0\cdot 015$	0·0001313	7619
3	$0\cdot 010 < \sigma(x_i) \le 0\cdot 030$	$0\cdot 015 < \sigma(x_i) \le 0\cdot 045$	0·0009750	1026
4	$0\cdot 030 < \sigma(x_i)$	$0\cdot 045 < \sigma(x_i)$	—	—

For example, application of equation (26) to the sample of adenine N(7)—C(8) distances in Table 1 (for which $k_1 = 2, k_2 = 2, k_3 = 8$) gives:

$$E[\overline{\sigma^2(x_l)}] \simeq 0{\cdot}00068\ \text{Å}^2. \qquad (27)$$

The correct value of $\overline{\sigma^2(x_l)}$, obtained from equation (7), is 0·00074 Å^2. Thus, in this example the value estimated from equation (26) is a good approximation of the true value, even though the sample size is small. The approximate value of $\overline{\sigma^2(x_l)}$ may now be used to assess the importance of environmental effects. We suggest that environmental effects should be regarded as being significant unless $\sigma^2(sample)$ is less than $E[\overline{\sigma^2(x_l)}]$, and there are no strong chemical grounds for believing that environmental effects are important. In view of the approximations involved in deriving equation (26), it is doubtful whether any more sophisticated procedure is justified.

VI. 3. *Model A: environmental effects negligible*

If environmental effects are negligible, the best estimate of the mean is the weighted mean. However, since this cannot be calculated without recourse to the original literature, we are obliged to use an alternative. One possibility is to calculate the unweighted mean of the most precise observations in the sample, *e.g.* all those with $AS < 3$. The example in Table 2 suggests that this may give excellent results and our own general experience supports this belief. However, if the sample contains relatively few precise observations, we may be reluctant to use this method. An alternative is to weight each observation according to its AS flag. For example, an observation with $AS = 1$ is weighted by a factor of $1/E[\overline{\sigma^2(x_l)}]_1$, *i.e.* 43 011. Weights for $AS = 2$, 3 are calculated in an analogous fashion (Table 6). Thus, we define a *partially-weighted mean*, $\bar{x}_p$, as:

$$\bar{x}_p = \frac{\left(43\,011 \sum_{j=1}^{k_1} x_j + 7619 \sum_{l=1}^{k_2} x_l + 1026 \sum_{m=1}^{k_3} x_m\right)}{(43\,011k_1 + 7619k_2 + 1026k_3)} \qquad (28)$$

where $\sum_{j=1}^{k_1}$ represents summation over all observations with $AS = 1$, *etc.* Equation (28) can also be written as:

$$\bar{x}_p = \frac{(43\,011k_1[\bar{x}_u]_1 + 7619k_2[\bar{x}_u]_2 + 1026k_3[\bar{x}_u]_3)}{(43\,011k_1 + 7619k_2 + 1026k_3)} \qquad (29)$$

where $[\bar{x}_u]_1$ is the unweighted mean of all observations with $AS = 1$, *etc.* The standard error of $\bar{x}_p$ can be estimated approximately as:

$$\sigma(\bar{x}_p) = [1/(43\,011k_1 + 7619k_2 + 1026k_3)]^{1/2}. \qquad (30)$$

The calculation of $\bar{x}_p$ and $\sigma(\bar{x}_p)$ is illustrated in Table 7 for the adenine N(7)—C(8) bond lengths listed in Table 1. For comparison, the weighted and unweighted means are also given. The results show that, in this example, $\bar{x}_p$ and $\sigma(\bar{x}_p)$ are very good approximations of $\bar{x}_w$ and $\sigma(\bar{x}_w)$, respectively. In contrast, the unweighted mean and standard error are relatively poor approximations.

Table 7. *Calculation of partially-weighted mean: adenine* N(7)—C(8) *bond lengths* (Å)

x_l	$\sigma(x_l)$	$\sigma^2(x_l)$	AS	Partial weight $\{=1/E[\overline{\sigma^2(x_l)}]\}$	w_l $[=1/\sigma^2(x_l)]$
1·315	0·003	0·000009	1	43011	111111
1·311	0·003	0·000009	1	43011	111111
1·322	0·012	0·000144	2	7619	6944
1·329	0·012	0·000144	2	7619	6944
1·347	0·021	0·000441	3	1026	2268
1·301	0·0225	0·000506	3	1026	1975
1·378	0·0285	0·000812	3	1026	1231
1·325	0·030	0·000900	3	1026	1111
1·314	0·030	0·000900	3	1026	1111
1·333	0·0315	0·000992	3	1026	1008
1·294	0·045	0·002025	3	1026	494
1·315	0·045	0·002025	3	1026	494

$\bar{x}_p = 1{\cdot}3157$ $\quad \bar{x}_w = 1{\cdot}3144$ $\quad \bar{x}_u = 1{\cdot}3237$

$\sigma(\bar{x}_p) = 0{\cdot}0030$ $\quad \sigma(\bar{x}_w) = 0{\cdot}0020$ $\quad \sigma(\bar{x}_u) = 0{\cdot}0064$

We further examined the performance of the partially-weighted mean by a simulation. The procedure was as follows. A pseudo-random number generator was used to generate a hypothetical sample of 28 observations, four with $AS = 1$, eight with $AS = 2$, and sixteen with $AS = 3$. The e.s.d. of each observation was chosen at random from a uniform distribution in the appropriate range (*e.g.* 0·0075–0·015 Å for an observation with $AS = 2$). The value of the observation itself was chosen at random from a normal distribution with a mean of 1·54 Å and a standard deviation equal to the e.s.d. already assigned to the observation. When all of the observations and e.s.d.'s were chosen, the quantities $\bar{x}_p$, $\sigma(\bar{x}_p)$, $\bar{x}_w$, $\sigma(\bar{x}_w)$, $\bar{x}_u$ and $\sigma(\bar{x}_u)$ were calculated. The complete procedure was then repeated until a total of 3000 hypothetical samples had been generated. At this point, the results were as summarized in Table 8. The r.m.s. values of $(\bar{x}_w - \bar{x}_p)$ and $[\sigma(\bar{x}_w) - \sigma(\bar{x}_p)]$ were 0·0012 and 0·0005 Å, respectively. In contrast, the r.m.s. values of $(\bar{x}_w - \bar{x}_u)$ and $[\sigma(\bar{x}_w) - \sigma(\bar{x}_u)]$ were 0·0043 and 0·0031 Å, respectively. Thus, the performance of the partially-weighted mean was much better than that of the unweighted mean. The maximum and minimum values of $(\bar{x}_w - \bar{x}_p)$ and $[\sigma(\bar{x}_w) - \sigma(\bar{x}_p)]$ give some

Table 8. *Results of simulation comparing partially-weighted, unweighted and weighted means* (Å)

	$(\bar{x}_w - \bar{x}_p)$	$(\bar{x}_w - \bar{x}_u)$	$[\sigma(\bar{x}_w) - \sigma(\bar{x}_p)]$	$[\sigma(\bar{x}_w) - \sigma(\bar{x}_u)]$
Mean	0·0000	0·0000	−0·0004	−0·0030
R.m.s.	0·0012	0·0043	0·0005	0·0031
Min.	−0·0049	−0·0148	−0·0011	−0·0065
Max.	0·0050	0·0137	0·0004	−0·0002

Table 9. *Results of simulations comparing precisions of semi-weighted and unweighted means* (Å)

	$\sigma(\bar{x}_s)/\sigma(\bar{x}_u)$		
$\sigma(sample)$	Mean	Min.	Max.
0·030	0·76	0·28	0·92
0·035	0·89	0·79	0·95
0·040	0·94	0·89	0·97
0·045	0·96	0·94	0·98
0·050	0·97	0·96	0·99
0·055	0·98	0·97	0·99
0·060	0·99	0·98	0·99

indication of the reliability of the partially-weighted mean when assumption (iii) of § VI. 1 breaks down (*i.e.* when the distribution of e.s.d.'s for observations with a given *AS* flag is not very uniform). The results are encouraging. Thus the maximum absolute discrepancy between $\bar{x}_w$ and $\bar{x}_p$ was about 0·005 Å; the largest difference between $\sigma(\bar{x}_w)$ and $\sigma(\bar{x}_p)$ was only 0·001 Å.

VI. 4. *Model B: environmental effects not negligible*

When environmental effects are not negligible, the best estimate of the mean is $\bar{x}_s$ [equations (17), (19)]. However, this quantity cannot be calculated from the information in CSD. Fortunately, the results in Tables 3–5 suggest that the unweighted mean, $\bar{x}_u$, is usually a good alternative. We confirmed this by means of a series of simulations. The procedure in each simulation was as follows. A pseudo-random number generator was used to generate 3000 hypothetical samples, each containing 50 observations. The e.s.d.'s of the hypothetical observations were chosen at random from a uniform distribution in the range 0·0015–0·045 Å (corresponding to observations with $AS \leq 3$). All of the samples were assumed to be of equal variance, this variance being set to a predetermined value [for example, in the first simulation we chose $\sigma^2(sample) =$ 0·0009 Å^2, *i.e.* $\sigma(sample) = 0{\cdot}03$ Å]. The quantities $\sigma(\bar{x}_u)$ and $\sigma(\bar{x}_s)$ were calculated for each sample from equations (11) and (19), respectively. Thus, the relative precision of the semi-weighted and unweighted means, $\sigma(\bar{x}_s)/\sigma(\bar{x}_u)$, was estimated. The average, minimum and maximum values of this quantity over the 3000 samples were then printed out.

The results are summarized in Table 9. They show that when $\sigma(sample) \geq 0{\cdot}035$ Å, the unweighted mean is a very satisfactory alternative to the semi-weighted mean. When $\sigma(sample) = 0{\cdot}030$ Å, the precision of the unweighted mean is sometimes very poor relative to that of the semi-weighted mean. However, the *average* relative precision ($\simeq$0·76) is still quite satisfactory.

VII. Summary

The estimation of average molecular dimensions is a statistical problem of surprising complexity. When the observational e.s.d.'s are available, the procedure may be summarized as follows: (1) Determine whether environmental effects are important by comparing $\overline{\sigma^2(x_i)}$ with $\sigma^2(sample)$ [equations (7), (8)], and by calculating the χ^2 statistic, equation (4). (2) If environmental effects are small, calculate the weighted mean, $\bar{x}_w$ [equations (6), (10)]. (3) If environmental effects are moderate, calculate the semi-weighted mean, $\bar{x}_s$ [equations (17), (19)]. (4) If environmental effects are large, calculate the unweighted mean, $\bar{x}_u$ [equations (9), (12)].

When the observational e.s.d.'s are not available (*i.e.* we are using CSD), the procedure must be modified as follows: (1) Determine whether environmental effects are important by comparing the quantities $E[\overline{\sigma^2(x_i)}]$ and $\sigma^2(sample)$ [equations (26), (8)], and by considering the nature of the molecular dimension (*e.g.* is it easily distorted from its equilibrium value?). (2) If environmental effects are small, *either* calculate the unweighted mean of the most precise observations in the sample, *or* calculate the partially-weighted mean, $\bar{x}_p$ [equations (28), (30)]. (3) If environmental effects are moderate or large, calculate the unweighted mean.

Many of the formulae given in this paper are based on the approximation that the $\sigma(x_i)$ are exact estimates of the experimental standard deviations of the corresponding x_i. This is equivalent to assuming that sampling errors in the quantities w_i and W_i [equations (5), (18)] are negligible. The assumption is made in order to simplify the statistical analysis and will probably result in a slight underestimation of the standard errors of weighted and semi-weighted means. Our analysis shows that $\bar{x}_u$ is often a good alternative to $\bar{x}_w$ and $\bar{x}_s$, even assuming that the $\sigma(x_i)$ are exact. Thus, the unweighted mean will probably be satisfactory for most samples of crystallographic data. The weighted, partially-weighted and semi-weighted means will be useful when environmental effects are small and the range spanned by the $\sigma(x_i)$ is large.

Olga Kennard is a member of the external staff of the Medical Research Council. The authors wish to thank Mr Jim Hunt of the Westminster Hospital Medical Computer Centre, London, for referring them to the paper by Cochran. We also thank a referee for pointing out an error in the original version of the manuscript.

References

Allen, F. H., Bellard, S., Brice, M. D., Cartwright, B. A., Doubleday, A., Higgs, H., Hummelink, T., Hummelink-Peters, B. G., Kennard, O., Motherwell, W. D. S., Rodgers, J. R. & Watson, D. G. (1979). *Acta Cryst.* **B35**, 2331–2339.

Arnott, S. & Scott, W. E. (1972). *J. Chem. Soc. Perkin Trans.* 2, pp. 324–335.

Cambridge Crystallographic Data Centre User Manual (1978). 2nd ed. Cambridge Univ.

Cochran, W. G. (1954). *Biometrics*, **10**, 101–129.

Cochran, W. G. & Carroll, S. P. (1953). *Biometrics*, **9**, 447–459.

Dyke, T. R., Mack, K. M. & Muenter, J. S. (1977). *J. Chem. Phys.* **66**, 498–510.

Hamilton, W. C. (1964). *Statistics in Physical Science*, pp. 15, 24. New York: Ronald Press.

Hamilton, W. C. & Abrahams, S. C. (1970). *Acta Cryst.* A**26**, 18–24.

James, M. N. G. & Williams, G. J. B. (1973). *Acta Cryst.* B**29**, 1172–1174.

Jeffrey, G. A., Ruble, J. R., McMullan, R. K., DeFrees, D. J. & Pople, J. A. (1981). *Acta Cryst.* B**37**, 1885–1890.

Johnson, C. K. (1970). *Thermal Neutron Diffraction*, edited by B. T. M. Willis, pp. 132–159. Oxford Univ. Press.

Kamb, B. (1968). *Structural Chemistry and Molecular Biology*, edited by A. Rich & N. Davidson, p. 509. San Francisco, London: Freeman.

Robiette, A. G. (1973). *Molecular Structure by Diffraction Methods*, edited by G. A. Sim & L. E. Sutton, Ch. 4. London: Chemical Society.

Sutton, L. E. (1958, 1965). *Tables of Interatomic Distances and Configuration in Molecules and Ions.* Chemical Society Special Publications 11, 18. London: Chemical Society.

Taylor, R. & Kennard, O. (1982*a*). *J. Am. Chem. Soc.* **104**, 3209–3212.

Taylor, R. & Kennard, O. (1982*b*). *J. Mol. Struct.* **78**, 1–28.

Taylor, R. & Kennard, O. (1983). *Acta Cryst.* B**39**, 133–138.

Taylor, R., Kennard, O. & Versichel, W. (1983). To be published.

Verbist, J. J., Lehmann, M. S., Koetzle, T. F. & Hamilton, W. C. (1972). *Acta Cryst.* B**28**, 3006–3013.

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*

Acta Cryst. (1983). B**39**, 525

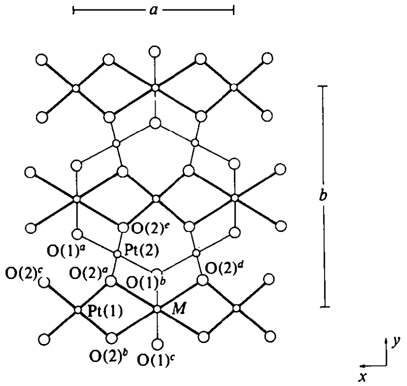
Structure and crystal chemistry of mixed-valence ternary platinum oxides: $MnPt_3O_6$, $CoPt_3O_6$, $ZnPt_3O_6$, $MgPt_3O_6$, and $NiPt_3O_6$: Erratum.

By K. B. Schwartz,* J. B. Parise and C. T. Prewitt, *Department of Earth and Space Sciences, State University of New York at Stony Brook, Stony Brook, New York* 11794, *USA*, and R. D. Shannon, *Central Research and Development Department, E.I. du Pont de Nemours and Company, Wilmington, Delaware* 19898, *USA*

(*Received* 24 *May* 1983)

Abstract

An error in printing is corrected. In the paper by Schwartz, Parise, Prewitt & Shannon [*Acta Cryst.* (1983). B**39**, 217–226] Fig. 3(*a*) is shown incorrectly. The correct version of the figure showing the *C*-centered MPt_3O_6 structure projected down [001] is presented.



(*a*)

Fig. 3. (*a*) *ORTEP* drawing of the *C*-centered MPt_3O_6 structure projected down [001].

* Current address: Lawrence Livermore National Laboratory, PO Box 808, L-396, Livermore, CA94550, USA.

0108-7681/83/040525-01$01.50