to distinguish between one-time and continuing losses.

ACKNOWLEDGMENT

I thank Professor T. L. Isenhour, Professor K. N. Berk, and the reviewers for many valuable suggestions.

REFERENCES AND NOTES

(1) "Atlantic Council's Nuclear Fuels Policy Working Group, Nuclear power and Nuclear Weapons Proliferation"; Atlantic Council of the U.S.: Washington, DC, 1978.
(2) Beaton, A. E.; Tukey, J. W. "The Fitting of Power Series Meaning Polynomials, Illustrated on Band-Spectroscopic Data". *Technometrics* **1974**, *16*, 147-186.
(3) Beckman, R. J.; Cook, D. "Outlier...s". *Technometrics* **1983**, *25*, 119-149.
(4) Berenson, M. L.; Levine, D. M.; Goldstein, M. "Intermediate Statistical Methods and Applications"; Prentice-Hall: Englewood Cliffs, NJ, 1983.
(5) Booth, D. E. "Regression Methods and Problem Banks"; COMAP, Inc.: Lexington, MA, 1985; Module No. 626.
(6) Booth, D. E.; Montasser, S. "Robust Discriminant Analysis and the Periods of Modern Egyptian Economic Developments". *Ind. Math.* **1985**, *35* (1), 81-91.
(7) Booth, D. E. "Some Applications of Robust Statistical Methods to Analytical Chemistry". Doctor of Philosophy Dissertation, The University of North Carolina at Chapel Hill, 1984.
(8) Booth, D. E. "A Model for the Early Detection of Loss in Nuclear Material Inventories", Tabor School of Business and Engineering, Millikin University: Decatur, IL, 1983; Faculty Working Paper 83-4.
(9) Booth, D. E. "The Analysis of Outlying Data Points Using Robust Regression: A Multivariate Problem Bank Identification Model". *Decis. Sci.* **1982**, *13*, 71-81.
(10) Booth, D. E. "The Analysis of Outlying Data Points by Robust Regression: I. A Model for the Identification of Problem Banks". *Ind. Math.* **1981**, *31* (2), 85-98.
(11) Booth, D. E.; Isenhour, T. L. "An Application of Robust Time Series Analysis to the Interpretation of Quality Control Charts". Submitted to *J. Quality Technol.*
(12) Box, G. E. P.; Jenkins, G. M. "Time Series Analysis Forecasting and Control"; Holden Day: San Francisco, 1976; revised ed.
(13) Chernick, M. R.; Downing, D. J.; Pike, D. H. "Detecting Outliers in Time Series Data". *J. Am. Stat. Assoc.* **1982**, *77*, 743-747.
(14) Denby, L.; Martin, R. D. "Robust Estimation of the First-Order Autoregressive Parameter". *J. Am. Stat. Assoc.* **1979**, *74*, 140-146.
(15) Downing, D. J.; Pike, D. H.; Morrison, G. W. "Analysis of MUF Data Using ARIMA Models". *Nucl. Mater. Manage.* **1978**, *7* (4), 80-86.
(16) Fox, A. J. "Outliers in Time Series Data". *J. R. Stat. Soc., Ser. B*, **1972**, *34*, 340-363.
(17) Goldman, A. S.; Picard, R. R.; Shipley, J. P. "Statistical Methods for Nuclear Material Safeguards: An Overview". *Technometrics* **1982**, *24*, 267-274.
(18) Hillier, F. S.; Lieberman, G. J. "Operations Research"; Holden-Day: San Francisco, 1974; 2nd ed.
(19) Hogg, R. V. "Statistical Roubustness: One View of Its Use in Applications Today". *Am. Stat.* **1979**, *33*, 108-115.
(20) Hull, C. H.; Nie, N. H. "SPSS Update 7-9"; McGraw-Hill: New York, 1981.
(21) Makridakis, S.; Wheelwright, S. "Forecasting Methods and Applications"; Wiley: New York, 1978.
(22) Mallows, C. "Robust Methods—Some Examples of Their Use". *Am. Stat.* **1979**, *33*, 179-184.
(23) Martin, R. D. "Robust Methods for Time Series". In "Applied Time Series Analysis II"; Findley, D. F., Ed.; Academic Press: New York, 1981.
(24) Martin, R. D. "Robust Estimation for Time Series in Autoregressions". In "Robustness in Statistics"; Launer, R. L.; Wilkinson, G., Eds.; Academic Press: New York, 1979.
(25) Martin, R. D.; Zeh, J. E. "Determining the Character of Time Series Outliers". "Proceedings of the American Statistical Association"; Business and Economics Statistics Section, American Statistical Association: Washington, DC, 1977.
(26) Martin, R. D. "Time Series: Model Estimation, Data Analysis, and Robust Procedrues". In "Modern Statistics: Methods and Applications"; American Mathematical Society: Providence, RI, 1980.
(27) Montgomery, D.; Johnson, L. "Forecasting and Time Series Analysis"; McGraw-Hill: New York, 1976.
(28) Mosteller, F.; Tukey, J. "Data Analysis and Regression"; Addison-Wesley: New York, 1977.
(29) Neter, J.; Wasserman, W. "Applied Linear Statistical Models"; Irvin: Homewood, IL, 1974.
(30) Pankratz, A. "Forecasting with Univariate Box-Jenkins Models"; Wiley: New York, 1983.
(31) Preston, D. B. "Robust Forecasting". In "Applied Time Series Analysis"; Findley, D. F., Ed.; Academic Press: New York, 1981.
(32) "Nuclear Energy and National Security"; Research and Policy Committee, Committee for Economic Development: New York, 1976.
(33) Wagner, H. M. "Principles of Management Science"; Prentice-Hall: Englewood Cliffs, NJ, 1970.

Cambridge Crystallographic Data Centre. 7. Estimating Average Molecular Dimensions from the Cambridge Structural Database

ROBIN TAYLOR* and OLGA KENNARD

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

Received June 11, 1985

The Cambridge Structural Database contains the atomic coordinates of some 40 000 organocarbon crystal structures. It is therefore likely to be a major source of data in future determinations of average molecular dimensions. From a statistical point of view, there is no single "optimum" method of obtaining such averages. Practical guidelines are suggested here on the basis of computer-simulation results.

INTRODUCTION

The Cambridge Crystallographic Data Centre (CCDC) maintains and distributes the Cambridge Structural Database (CSD), which currently contains the results of about 40 000 organocarbon crystal structure determinations. Previous papers in this series[1-3] document the development of CSD, concentrating mainly on the organization and content of the database and its associated search and retrieval software. This software is still under active development in Cambridge, but increasingly, the CCDC is addressing the problems of database *utilization*.

The value of CSD as a research tool is well-known,[4] and there are currently over one-hundred published papers describing CSD utilization projects. Moreover, it is now recognized as an important facility in molecular graphics.[5,6] One of the most significant areas in which CSD can be of use is in the estimation of average molecular dimensions. This is a major objective of many chemical and crystallographic research projects. It is also of fundamental importance in molecular graphics, e.g., in the construction of "fragment libraries".

The statistical problems involved in estimating average molecular dimensions were examined from a theoretical point of view in two earlier papers.[7,8] In the present paper, we have two objectives. First, we use the theoretical results obtained earlier to devise practical guidelines for estimating average molecular dimensions from CSD. Second, we outline the computer-simulation algorithm used to obtain these guidelines,

0095-2338/86/1626-0028$01.50/0

Table I. Cambridge Structural Database AS Flags

AS	uncorrected ESD range (Å)	corrected ESD range (Å)	no. of entries[a]
0	unassigned	unassigned	4205
1	$0.001 \leq \bar{\sigma}(C-C) \leq 0.005$	$0.0015 \leq \bar{\sigma}(C-C) \leq 0.0075$	3824
2	$0.005 < \bar{\sigma}(C-C) \leq 0.010$	$0.0075 < \bar{\sigma}(C-C) \leq 0.015$	4123
3	$0.010 < \bar{\sigma}(C-C) \leq 0.030$	$0.015 < \bar{\sigma}(C-C) \leq 0.045$	3237
4	$0.030 < \bar{\sigma}(C-C)$	$0.045 < \bar{\sigma}(C-C)$	839
			16228[b]

[a] In late 1983, chemical classes[1] 1–60 inclusive. [b] Total.

so that they may (if necessary) be updated by future users of the database.

THEORETICAL BACKGROUND

Although CSD contains the atomic coordinates of some 40000 organocarbon crystal structures, it does not contain the estimated standard deviations (ESD's) of these coordinates. Some limited information about experimental precision 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 ESD's quoted for carbon–carbon (or other light atom–light atom) distances in the report of the structure. Thus, if the average value of these ESD's falls in the range 0.001–0.005 Å, the entry is given an AS flag of 1. Interexperimental comparisons[9,10] suggest that ESD's should be increased by about 50% in order to allow for systematic errors in the diffraction experiment. Thus, the "corrected" ESD range for structures with AS = 1 is $0.0015 \leq \bar{\sigma}(C-C) \leq 0.0075$ Å. The corresponding ranges for AS = 2–4 are given in Table I. If no bond-length ESD information is available, the entry is assigned an AS flag of zero. The last column of Table I shows the number of entries in CSD with AS = 0–4. These figures were computed in late 1983 and pertain to all organic structures (CSD chemical classes 1–60)[1] for which atomic coordinates are available in the database. The results described in this paper cannot necessarily be extrapolated to organometallic compounds.

The absence of ESD's from CSD may or may not be important when estimating average molecular dimensions. Suppose that we have k observations of a molecular parameter, x_i, $i = 1, 2, ..., k$, with "corrected" ESD's of $\sigma(x_i)$, $i = 1, 2, ..., k$. If the parameter is sensitive to changes in its chemical or crystallographic environment (e.g., a hydrogen-bond distance), it is usually adequate[7] to estimate its average value by the unweighted mean $\bar{x}_u$:

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

The standard error of $\bar{x}_u$ can be estimated from

$$\sigma(\bar{x}_u) = \sigma(\text{sample})/k^{1/2} \tag{2}$$

where σ(sample) is the sample standard deviation of the x_i. Thus

$$\sigma(\bar{x}_u) = (\sum_{i=1}^{k}(x_i - \bar{x}_u)^2/[k(k-1)])^{1/2} \tag{3}$$

Since the $\sigma(x_i)$ values are not used in eq 1 and 3, their absence from CSD does not constitute a problem. However, if the molecular parameter is relatively insensitive to changes in its environment (e.g., a valence-bond distance in a rigid type of molecule), its average value is best estimated by the weighted mean $\bar{x}_w$:

$$\bar{x}_w = [\sum_{i=1}^{k} x_i/\sigma^2(x_i)]/[\sum_{i=1}^{k} 1/\sigma^2(x_i)] \tag{4}$$

The standard error of $\bar{x}_w$ can be estimated from

$$\sigma(\bar{x}_w) = (1/[\sum_{i=1}^{k} 1/\sigma^2(x_i)])^{1/2} \tag{5}$$

(but see reference 8 for some reservations concerning this formula). Since the $\sigma(x_i)$ appear in eq 4 and 5, their absence from CSD is now a source of difficulty.

In an earlier paper, we described an approximation to the weighted mean called the "partially weighted mean".[7] This statistic can be calculated from the information stored in CSD and was shown to be a viable alternative to $\bar{x}_w$. However, its use is open to some objections as it is based on a number of approximations concerning the underlying distribution of $\sigma(x_i)$. A straightforward alternative is to approximate $\bar{x}_w$ by the *unweighted mean of a subset of the most precise observations in the sample* (e.g., observations taken from structures with AS = 1 or 2). This is a procedure that merits very serious study because, in practice, it is by far the most common method of using CSD (e.g., references 11 and 12 and many others). It is therefore considered in detail here.

NOTATION

Throughout the paper, $[\bar{x}_u]_{m,n}$ signifies the unweighted mean of those observations in the sample that are taken from structures with $m \leq$ AS $\leq n$ (e.g., $[\bar{x}_u]_{0,3}$ is the unweighted mean of all observations in the sample except those taken from structures with AS = 4). $[\bar{x}_u]_{m,n}$ and its standard error can be estimated from the formulas

$$[\bar{x}_u]_{m,n} = \sum_{i=1}^{k_{m,n}} x_i/k_{m,n} \tag{6}$$

and

$$\sigma([\bar{x}_u]_{m,n}) = (\sum_{i=1}^{k_{m,n}}(x_i - [\bar{x}_u]_{m,n})^2/[k_{m,n}(k_{m,n}-1)])^{1/2} \tag{7}$$

where the summation is over all observations with $m \leq$ AS $\leq n$ and $k_{m,n}$ is the number of such observations.

QUALITATIVE DISCUSSION OF $[\bar{X}_U]_{M,N}$

A simple example will illustrate the advantages and disadvantages of different $[\bar{x}_u]_{m,n}$. Suppose that we have four observations of a valence-bond distance that is not significantly affected by changes in its environment. Let the observations have corrected ESD's and AS flags as follows: 0.002 Å, 1; 0.008 Å, 2; 0.009 Å, 2; 0.010 Å, 2. Since environmental effects have been assumed negligible, differences between the observations must be entirely due to experimental errors in their measurement. If these errors are approximately normally distributed, there is a 99% chance that the *true* value of the dimension will lie within three standard deviations of each observation, i.e., within ±0.006 Å of the first, ±0.024 Å of the second, etc. These confidence ranges suggest that the first observation gives more information about the true value of the dimension than the other three observations put together. This is reflected in the weights that would be used in calculating the weighted mean ($1/0.002^2 = 250000$; $1/0.008^2 = 15625$; $1/0.009^2 = 12346$; $1/0.010^2 = 10000$); these are such that the value of $\bar{x}_w$ is almost entirely determined by the first observation. If we wished to approximate $\bar{x}_w$ by one of the $[\bar{x}_u]_{m,n}$, it is evident that the best choice would be $[\bar{x}_u]_{1,1}$. This would give the first observation a weight of one while the other

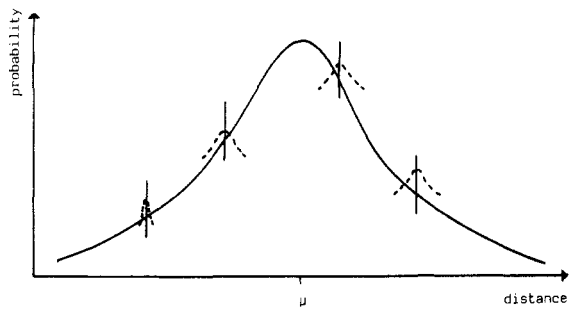

Figure 1. Schematic drawing of environmental (solid curve) and experimental (broken curves) distributions for the hypothetical example discussed in the text.

three would be zero weighted.

Suppose, now, that the four observations are of a hydrogen-bond distance. Differences between the observations are now likely to be due to environmental effects rather than experimental errors. This is shown schematically in Figure 1: the solid curve represents the probability distribution of the hydrogen-bond distance over all possible environments, the four vertical lines represent the individual observations, and the broken curves show the experimental uncertainties associated with the observations. Effectively, we now seek to estimate the average value of the hydrogen-bond distance over all possible crystal-field environments (i.e., μ in Figure 1). Since the broken curves are much narrower than the solid curve, the differences in experimental precisions are unimportant: i.e., the "information content" of the first observation is not significantly greater than that of the last. If we now use $\bar{x}_w$ or $[\bar{x}_u]_{1,1}$, we are simply wasting three-fourths of our data. The best estimate is $[\bar{x}_u]_{0,4}$ ($=[\bar{x}_u]_{1,2}$ in this example), since this uses all of the observations and wastes no information.

In choosing a suitable $[\bar{x}_u]_{m,n}$, we must therefore compromise between two opposing factors. If environmental effects appear to be small, we should exclude all but the most precise observations from our calculations; if they seem to be large, we should avoid any needless waste of data. Methods of assessing the importance of environmental effects are discussed in an earlier paper.[7] We now consider the consequences of using various $[\bar{x}_u]_{m,n}$ in the limiting situations: *environmental effects* $\ll$ *experimental errors* and *environmental effects* $\gg$ *experimental errors*.

ENVIRONMENTAL EFFECTS $\ll$ EXPERIMENTAL ERRORS

Computer simulations were used to investigate the relative precision of $[\bar{x}_u]_{m,n}$ and $\bar{x}_w$ when environmental effects are negligible. Several simulations were performed, but they differed only in the size of the samples generated. The procedure in each simulation was as follows. Six-thousand artificial samples of observations were constructed, using pseudo random number generators. Each observation, x_i, had a "corrected" ESD, $\sigma(x_i)$, and an AS flag, AS_i. The procedure used to generate these quantities is outlined in the appendix to this paper. The quantities $\bar{x}_w$, $\sigma(\bar{x}_w)$, and $[\bar{x}_u]_{m,n}$ (for m,n = 0,4; 0,3; 0,2; 0,1; 1,3; 1,2; and 1,1) were estimated for each generated sample from eq 4–6, respectively. The $\sigma([\bar{x}_u]_{m,n})$ could have been estimated from eq 7 but were instead calculated *precisely* from[13]

$$\sigma([\bar{x}_u]_{m,n}) = [\sum_{i=1}^{k_{m,n}} \sigma^2(x_i)]^{1/2}/k_{m,n} \qquad (8)$$

This expression is not usually applicable to crystallographic data but can be used here because the $\sigma(x_i)$ are known exactly and environmental effects are completely nonexistent. The precision of each $[\bar{x}_u]_{m,n}$ relative to the weighted mean was then calculated as

$$R_{m,n} = \sigma(x_w)/\sigma([x_u]_{m,n}) \qquad (9)$$

At the end of the simulation, the average values of the $R_{m,n}$ were computed and histograms of the $R_{m,n}$ distributions printed out.

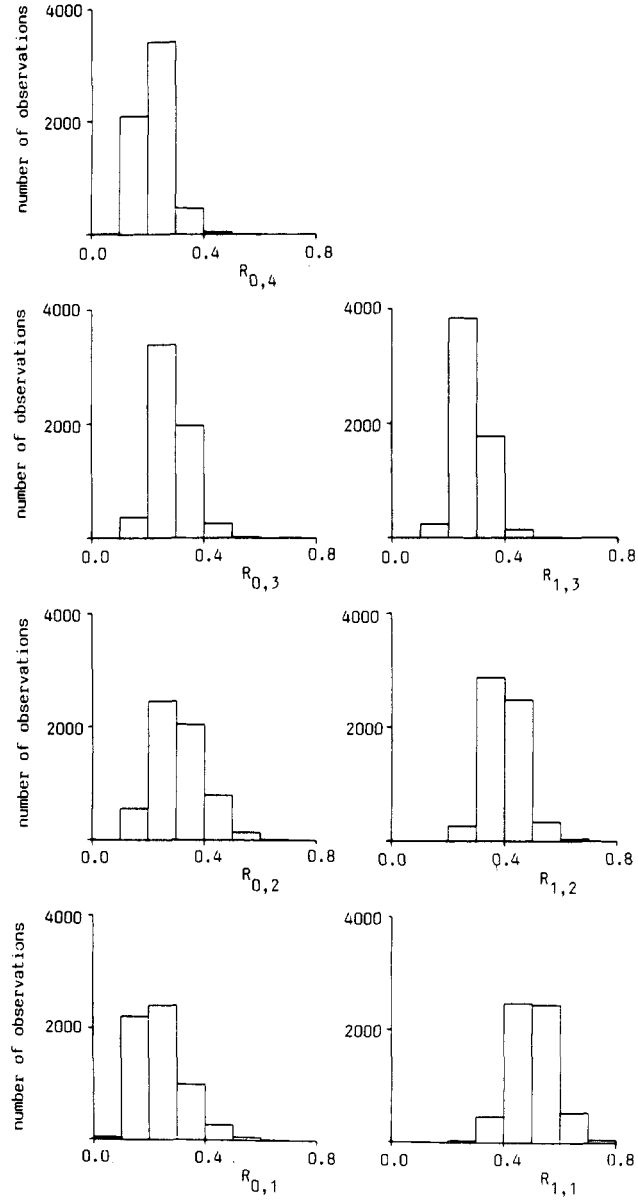

Figure 2. Histograms of $R_{m,n} = \sigma(\bar{x}_w)/\sigma([\bar{x}_u]_{m,n})$ for m, n = 0, 4; 0, 3; 0, 2; 0, 1; 1, 3; 1, 2; and 1, 1. Histograms are based on simulation 1 of Table IIA.

Representative results are given in part A of Table II and Figure 2. As expected, the optimum values of m and n are $m = n = 1$. On average, $[\bar{x}_u]_{1,1}$ is about half as precise as $\bar{x}_w$ (i.e., the mean value of $R_{1,1}$ is about 0.5), and Figure 2 shows that $R_{1,1}$ very rarely falls below 0.3. Inclusion of observations from structures with AS = 2 lowers the mean relative precision to about 0.4. The average value of $R_{0,3}$ is marginally higher than that of $R_{1,3}$, suggesting that $[\bar{x}_u]_{0,3}$ is usually preferable to $[\bar{x}_u]_{1,3}$. Finally, we note that the unweighted mean of the complete sample, $[\bar{x}_u]_{0,4}$, performs rather poorly: $R_{0,4}$ is rarely higher than 0.3, and its average value is about 0.2–0.25.

Figure 3 gives some additional results from simulation 2 of Table IIA. The histograms show the observed distributions of $\bar{x}_w - \mu$, $[\bar{x}_u]_{1,1} - \mu$, and $[\bar{x}_u]_{0,4} - \mu$, where μ is the true mean. Despite the small sample size, the distributions show that the estimated mean values are usually very accurate. For example,

Table II. Precision of $[\bar{x}_u]_{m,n}$ under Two Extreme Conditions

(A) Environmental Effects ≪ Experimental Errors

simulation no.	minimum sample size, k_{min}	maximum sample size, k_{max}	average value of $\sigma(\bar{x}_w)/\sigma([\bar{x}_u]_{m,n})$ for m,n of 0,4	0,3	0,2	0,1	1,3	1,2	1,1
1	25	125	0.22	0.29	0.31	0.24	0.28	0.40	0.50
2	25	25	0.25	0.31	0.35	0.30	0.30	0.42	0.53
3	50	50	0.23	0.29	0.32	0.25	0.29	0.40	0.51
4	75	75	0.22	0.28	0.30	0.23	0.28	0.40	0.50
5	100	100	0.22	0.28	0.29	0.22	0.27	0.39	0.49
6	125	125	0.21	0.28	0.29	0.22	0.27	0.39	0.49

(B) Environmental Effects ≫ Experimental Errors

expected value of $\sigma([\bar{x}_u]_{0,4})/\sigma([\bar{x}_u]_{m,n})$ for m,n of 0,4	0,3	0,2	0,1	1,3	1,2	1,1
1.00	0.97	0.87	0.70	0.83	0.70	0.49

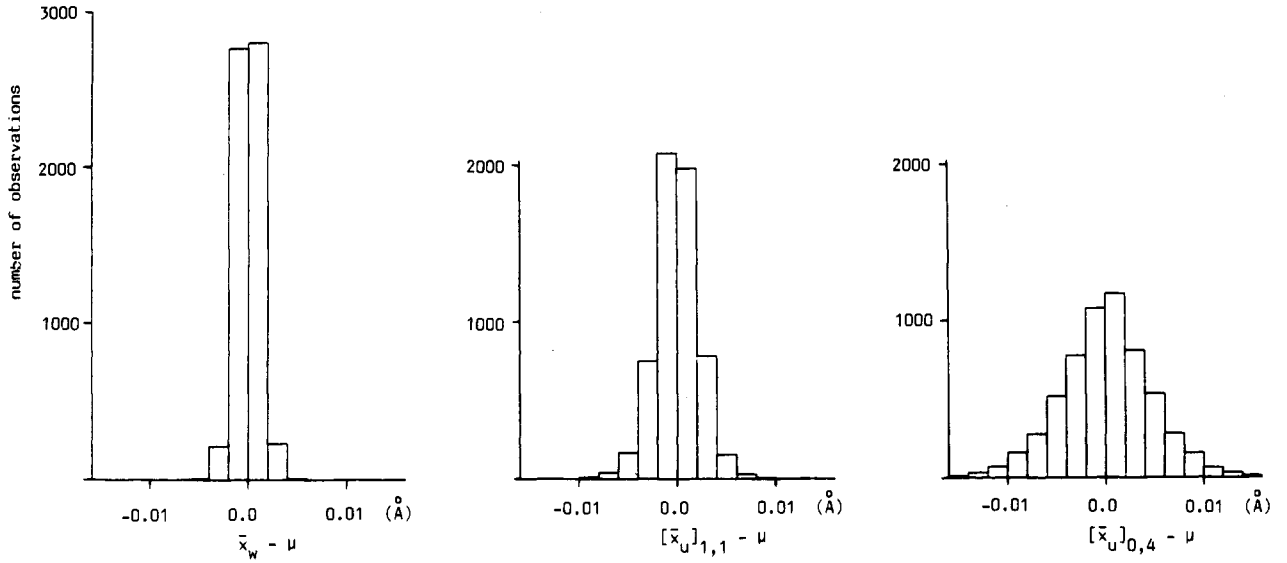

Figure 3. Histograms of $\bar{x}_w - \mu$, $[\bar{x}_u]_{1,1} - \mu$, and $[\bar{x}_u]_{0,4} - \mu$, based on simulation 2 of Table IIA.

74% of the $[\bar{x}_u]_{0,4}$ values lie within ±0.005 Å of the true mean. This may explain why early compilations of average molecular dimensions[14] are quite reliable, even though they were based on very limited data.

ENVIRONMENTAL EFFECTS ≫ EXPERIMENTAL ERRORS

When experimental errors are negligible compared with environmental effects, the optimum estimate of the mean is $[\bar{x}_u]_{0,4}$; this uses all the available data. If other $[\bar{x}_u]_{m,n}$ are chosen, the effect will be to reduce the size of the sample without reducing the expectation value of the sample standard deviation. Equation 2 then indicates that a loss of precision may be expected, since the denominator of the right-hand side will be reduced without a concomitant reduction in the numerator. The figures in the last column of Table I enable us to predict the likely loss in precision; for example, since there are only 3824 entries with AS = 1, we may expect that

$$\sigma([x_u]_{0,4})/\sigma([x_u]_{1,1}) \simeq 3824^{1/2}/16\,228^{1/2} = 0.49 \qquad (10)$$

Corresponding values for the other $[\bar{x}_u]_{m,n}$ are given in Table IIB. The results show that the likely loss in precision is relatively small for most values of m, n.

CONCLUSIONS

There is no "right" or "wrong" way of using the Cambridge Structural Database when estimating the average dimensions of organic molecules. Each choice of $[\bar{x}_u]_{m,n}$ is a compromise, with advantages and disadvantages dependent on the nature of the molecular dimension being studied. Table II gives data for the limiting situations environmental effects ≪ experimental errors and environmental effects ≫ experimental errors. Any real case will lie somewhere between these two extremes. Methods for assessing the importance of environmental effects have been discussed previously,[7] and these, together with the data in Table II, should enable the individual investigator to select values of m and n suitable for his or her experiment.

As a general rule, we suggest $[\bar{x}_u]_{1,1}$ or $[\bar{x}_u]_{1,2}$ for organic bond lengths. The former is preferable if the investigator suspects that environmental efects are very small (e.g., if the bond is part of a rigid ring system); the latter is recommended when the investigator is unsure about the magnitude of environmental effects, particularly if the sample size is very small. $[\bar{x}_u]_{0,3}$ and $[\bar{x}_u]_{1,3}$ are suitable for "softer" dimensions such as hydrogen-bond distances and, probably, the majority of valence angles. Of the two, $[\bar{x}_u]_{0,3}$ is likely to be preferable under most circumstances because it is slightly more precise than $[\bar{x}_u]_{1,3}$ in Table II. However, observations with AS = 0 may be very imprecise, so any calculation of $[\bar{x}_u]_{0,3}$ should be preceded by a visual inspection for outlying observations; this is a useful routine precaution in any case.

APPENDIX

Simulation Algorithm. The results given in Table IIA and Figure 2 should remain relevant for several years; indeed, they will only cease to be accurate if the distribution of AS flags within CSD changes appreciably. In case this happens, we outline below the algorithm used to generate artificial samples in the simulations on which Table IIA is based.

(1) Select sample size (k) at random from a uniform distribution in the range $k_{min} - k_{max}$ (k_{min} and k_{max} are set by the

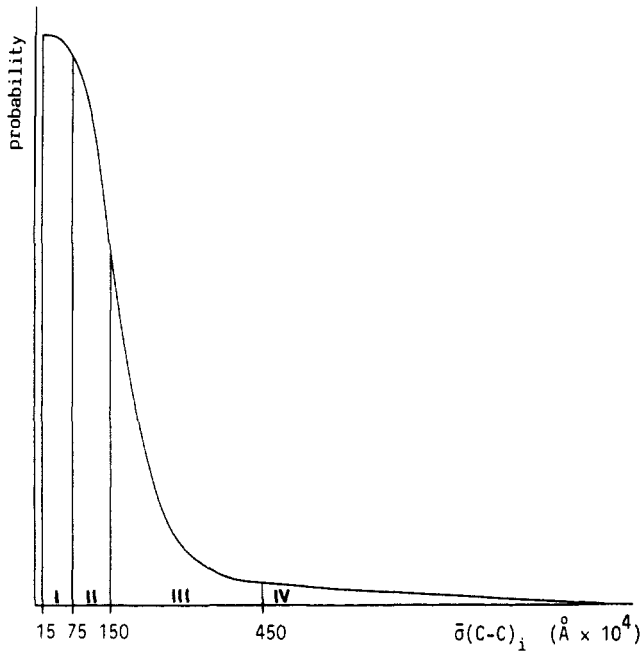

Figure 4. Empirical probability curve used for selection of $\bar{\sigma}(C-C)_i$ values in computer simulations.

user at the beginning of the simulation).

(2) Select the true mean (μ) at random from a uniform distribution in the range 1.1–1.7 Å.

(3) Set $i = 1$.

(4) Select $\bar{\sigma}(C-C)_i$ at random from the empirical probability curve shown in Figure 4. The quantity $\bar{\sigma}(C-C)_i$ is the average ESD of C–C bond lengths in the hypothetical structure from which the ith observation is assumed to be taken. The curve in Figure 4 is continuous in the range 0.0015–0.112 Å, and areas I–IV are in the approximate relative proportions 3824:4123:3237:839 (cf. Table I, final column). The results of the simulations are insensitive to small changes in this curve.

(5) Select a random number from a uniform distribution in the range 1–16 228. If the random number is less than or equal to 4205 (see final column of Table I), set $AS_i = 0$ and go to step 7. Otherwise, go to step 6.

(6) If $0.0015 \leq \bar{\sigma}(C-C)_i \leq 0.0075$, set $AS_i = 1$. If $0.0075 < \bar{\sigma}(C-C)_i \leq 0.015$, set $AS_i = 2$. If $0.015 < \bar{\sigma}(C-C)_i \leq 0.045$, set $AS_i = 3$. If $0.045 < \bar{\sigma}(C-C)_i$ set $AS_i = 4$.

(7) Select a random number (n_i) from a normal distribution with mean = 1, SD = 0.25.

(8) Set $\sigma(x_i) = n_i\bar{\sigma}(C-C)_i$. This permits $\sigma(x_i)$ to differ from $\bar{\sigma}(C-C)_i$, since, in a real structure, the ESD's of individual bond lengths would show some variation about their mean. The standard deviation of the normal distribution from which n_i is chosen (=0.25) was selected by trial and error to give a reasonable spread of $\sigma(x_i)$ values for any given $\bar{\sigma}(C-C)_i$. Results of the simulation are relatively insensitive to changes in this parameter.

(9) If $\sigma(x_i)$ is less than 0.00075, set $\sigma(x_i) = 0.00075$.

(10) Select x_i at random from a normal distribution with mean = μ, SD = $\sigma(x_i)$.

(11) Increment i by 1. If i is less than or equal to k, go to step 4. Otherwise, the sample is generated.

ACKNOWLEDGMENT

Olga Kennard is a member of the external staff of the Medical Research Council.

REFERENCES AND NOTES

(1) Kennard, O.; Watson, D. G.; Town, W. G. "Cambridge Crystallographic Data Centre. I. Bibliographic File". *J. Chem. Doc.* **1972**, *12*, 14–19.
(2) Allen, F. H.; Kennard, O.; Motherwell, W. D. S.; Town, W. G.; Watson, D. G. "Cambridge Crystallographic Data Centre. II. Structural Data File". *J. Chem. Doc.* **1973**, *13*, 119–123.
(3) 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. "The Cambridge Crystallographic Data Centre: Computer-Based Search, Retrieval, Analysis and Display of Information". *Acta Crystallogr., Sect. B: Struct. Crystallogr. Cryst. Chem.* **1979**, *B35*, 2331–2339.
(4) Allen, F. H.; Kennard, O.; Taylor, R. "Systematic Analysis of Structural Data as a Research Technique in Organic Chemistry". *Acc. Chem. Res.* **1983**, *16*, 146–153.
(5) Vinter, J. G. "Molecular Graphics for the Medicinal Chemist". *Chem. Br.* **1985**, *21*, 32–38.
(6) Elder, M.; Machin, P.; Hull, S. E. "CDA: An Interactive Program for the Comparative Analysis of Crystal Structure Coordinate Data". *J. Mol. Graphics* **1984**, *2*, 70–78.
(7) Taylor, R.; Kennard, O. "The Estimation of Average Molecular Dimensions from Crystallographic Data". *Acta Crystallogr., Sect. B: Struct. Sci.* **1983**, *B39*, 517–525.
(8) Taylor, R.; Kennard, O. "The Estimation of Average Molecular Dimensions. 2. Hypothesis Testing with Weighted and Unweighted Means". *Acta Crystallogr., Sect. A: Found Crystallogr.* **1985**, *A41*, 85–89.
(9) Hamilton, W. C.; Abrahams, S. C. "Least-Squares Refinement of Structural Parameters". *Acta Crystallogr., Sect. A: Cryst. Phys., Diffr., Theor. Gen. Crystallogr.* **1970**, *A26*, 18–24.
(10) Taylor, R.; Kennard, O. "Accuracy of Crystal Structure Error Estimates". *Acta Crystallogr., Sect. B: Struct. Sci.*, in press.
(11) Allen, F. H. "The Geometry of Small Rings. III. The Effect of Small-Ring Fusion on the Geometry of Benzene". *Acta Crystallogr., Sect. B: Struct. Sci.* **1981**, *B37*, 900–906.
(12) Schweizer, W. B.; Dunitz, J. D. "Structural Characteristics of the Carboxylic Ester Group". *Helv. Chim. Acta* **1982**, *65*, 1547–1554.
(13) Cochran, W. G. "The Combination of Estimates from Different Experiments". *Biometrics* **1954**, *10*, 101–129.
(14) Sutton, L. E. "Tables of Interatomic Distances and Configuration in Molecules and Ions"; Chemical Society: London, 1958.