RESEARCH PAPERS

Acta Cryst. (1995). A**51**, 95–111

A Systematic Study of Coordinate Precision in X-ray Structure Analyses. I. Descriptive Statistics and Predictive Estimates of E.S.D.'s for C atoms

By Frank H. Allen*

Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, England

and Jason C. Cole and Judith A. K. Howard*

Department of Chemistry, University of Durham, South Road, Durham DH1 3LE, England

(*Received* 10 *March* 1994; *accepted* 29 *June* 1994)

Abstract

This study examines the relationship of structure precision, as expressed by the e.s.d.'s of atomic coordinates, to the R factor and chemical constitution of a given crystal structure. On the basis of the work of Cruickshank [*Acta Cryst.* (1960), **13**, 744–777], it is shown that $\overline{\sigma}$(C–C), the mean e.s.d. of a C–C bond length in a structure, or $\overline{\sigma}$(C), the mean isotropic e.s.d. of a C atom, can be estimated by expressions of the form $\overline{\sigma} = kRN_c^{1/2}$. Here, N_c is taken as $\sum Z_i^2/Z_C^2$, with the atomic numbers Z_i summed over all atoms in the asymmetric unit and $Z_C = 6$. It is also shown that $\overline{\sigma}(E)$, the mean isotropic e.s.d. of a non-C atom, can be estimated by $\overline{\sigma}(E) = kRN_c^{1/2}/Z_E$. Values of k were determined by regression analyses based on subsets of 25984 and 20334 entries in the Cambridge Structural Database (CSD) that contain atomic coordinate e.s.d.'s. 95% of coordinate e.s.d.'s for C atoms can be estimated to within 0.005 Å of their published value and 78% to within 0.0025 Å. These predicted $\overline{\sigma}$ values provide useful estimates of precision for those 39000 structures for which coordinate e.s.d.'s are not available in the CSD. Details of the diffraction experiment, which might provide an improved estimating function in Cruickshank's (1960) treatment, are not available in any CSD entries. However, values of N_r (the number of reflections) and N_p (the number of parameters) used in refinement were added manually for 817 entries, and the variation of $\overline{\sigma}$(C–C) with decreasing N_r/N_p ratios is examined: there is a rapid increase in $\overline{\sigma}$(C–C) as N_r/N_p decreases below *circa* 6.0. A method for approximating $\bar{s}$, the r.m.s. reciprocal radius for the reflections observed, is presented, but it is found that a function of the form $\overline{\sigma}(\text{C–C}) = kRN_c^{1/2}/\bar{s}(N_r - N_p)^{1/2}$ [directly analogous to Cruickshank's (1960) equation] had only slightly improved predictive ability for this data set by comparison with functions based upon R and $N_c^{1/2}$ alone. Possible reasons for this apparent anomaly are discussed.

Introduction

Systematic analyses of the earliest crystallographic results played a vital role in the development of theories of chemical bonding and in the study of hydrogen-bonded and nonbonded interactions (see *e.g.* Pauling, 1940; Pimentel & McClellan, 1960; Sutton, 1958, 1965). As the volume of available results began to grow dramatically during the late 1960's and early 1970's (Fig. 1*a*), the number of systematic studies decreased, perhaps because of the difficulty of locating the appropriate results in the literature and the labour involved in retrieving, organizing and processing the large volume of associated numerical data. Nowadays, the problems have largely been eliminated through the availability of crystallographic databases (see *e.g.* Allen, Bergerhoff & Sievers, 1987) that are fully retrospective and maintained on a current basis. These databases, together with improving software for search, retrieval, analysis and display of the stored information, have provoked a renewed interest in the systematic study of crystal and molecular structures. As a result, the 1980's saw a steady flow of papers reporting systematic applications of the chemical and crystallographic results stored in the Cambridge Structural Database (CSD; Allen *et al.*, 1991). Indeed, a new component of the CSD System is designed to record these references (Allen, Kennard & Watson, 1995) and the preliminary statistics presented in Fig. 1(*b*) show a sharp rise from a plateau of ~20 papers per year until 1988 to the 56 papers published in 1991.

This growing interest in knowledge acquisition from numerical structural data, coupled with the continued growth in crystallographic output sum-

* Authors for correspondence.

marized in Fig. 1(*a*) and Table 1, presents its own problems in terms of the selection of CSD entries to be included in a given analysis. The primary selection will, of course, be dictated by the aims and scope of each research project. Primary criteria are usually defined in terms of the chemical substructure(s), both intramolecular and intermolecular, for which systematic knowledge is required. In most cases, however, we must also consider a variety of secondary (or general) selection criteria through which we may (*a*) restrict the volume of data arising from a particular primary search or, more commonly, (*b*) ensure that we can have statistical confidence in the numerical results of any systematic analysis. In both (*a*) and (*b*), we are using secondary criteria to select the 'best' entries for inclusion in the analysis, where 'best' is essentially defined in terms of the likely precision of the relevant atomic coordinate data.

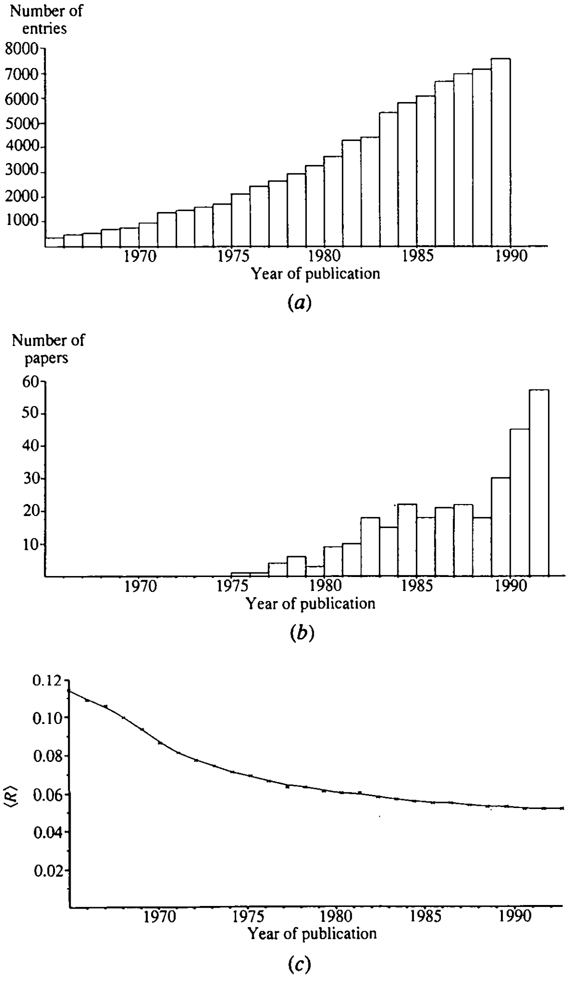

Fig. 1. Retrospective overview of the Cambridge Structural Database: (*a*) growth of the CSD by publication year; (*b*) growth in number of research papers using the CSD as basis; (*c*) improvements in structural precision as measured by the mean *R* factor in each publication year.

Table 1. *Summary statistics for the January* 1992 *release of the CSD*

Number of entries	96731
Number of chemical compounds	86012
Number of entries with three-dimensional coordinates	84598
Number of error-free three-dimensional coordinate sets	82828
Number of atoms with three-dimensional coordinates	4368677
X-ray studies	95966
Neutron studies	765
Absolute configuration by X-ray methods	2499
Low-temperature studies	9672
Number of literature sources	613

Table 2. *Secondary criteria often used in selecting CSD entries for inclusion in systematic analyses*

The percentage of entries that would pass each individual test is indicated where relevant.

Entry is error-free in CSD checks (98%)
No disorder in crystal structure (89%)
Neutron study (0.8%)
Organic structure (CSD classses 1–65 or 70) (57%)
Metallo-organic structure (CSD classes 66–69, 71–86) (43%)
Limitations on:
  Maximum atomic number
  CSD chemical classes
  Number of coordinates in entry
  Year of publication
  Temperature of data collection
  Crystallographic *R* factor
  Mean e.s.d. of a C–C bond [$\bar{\sigma}$(C–C)] flagged as:
    AS = 0 for $\bar{\sigma}$(C–C) not available
    AS = 1 for $0.0 < \bar{\sigma}(\mathrm{C{-}C}) \leq 0.005$
    AS = 2 for $0.005 < \bar{\sigma}(\mathrm{C{-}C}) \leq 0.010$
    AS = 3 for $0.010 < \bar{\sigma}(\mathrm{C{-}C}) \leq 0.030$
    AS = 4 for $0.030 < \bar{\sigma}(\mathrm{C{-}C})$

A number of information items in the CSD that are frequently used as secondary search criteria are listed in Table 2. Thus, we should normally exclude entries containing residual coordinate errors and may choose to treat disordered structures in the same way. In studies involving light-atom (*e.g.* C-, N- or O-atom) geometry, it is common to place some limit on the atomic number(s) of other elements that may also be present in retrieved structures: this limitation is effected in a more general way by restricting the search to organic structures only. If H-atom-coordinate precision is important, then neutron structures only may be used, provided that sufficient data are available (see *e.g.* Taylor & Kennard, 1982). Even the use of some cut-off value for year of publication is related, albeit obliquely, to coordinate precision: Fig. 1(*c*) shows how mean *R* factors have fallen from *circa* 0.09 in 1970 to *circa* 0.05 in recent years. Thus, the use of some of the criteria in Table 2 depends very much upon a qualitative, even intuitive, experiential knowledge of their likely effects on coordinate precision.

Only the crystallographic *R* factor (recorded in the database as the lowest of *R*, R_w *etc*) and the AS flag, defined in Table 2, can be regarded as true indicators of structural precision in the CSD since both arise from the least-squares fit of the structural model to the measured diffraction data. The *R* factor, despite its statistical imperfections, is ubiquitous in the crystallographic literature and less than 1% of CSD entries for which coordinates are available lack an *R*-factor field. The AS flag has been included in the CSD since its inception, but its comprehensive availability is less satisfactory: 16% of entries lack this flag (see Table 5). Until 1985, the AS flag was assigned by CSD editorial staff on the basis of published C–C bond-length e.s.d.'s (occasionally for C–N or C–O values if C–C bonds were not present). Thus, even if atomic coordinate e.s.d.'s were available in the publication, but bond length e.s.d.'s were not, then AS remained unassigned. Further, AS is a discontinuous parameter for which the banding, although appropriate for structures determined in the late 1960's and early 1970's, is no longer suitable for the results of the late 1980's and early 1990's. In particular, the e.s.d. band 0.011–0.030 Å (AS = 3) is too broad and subsumes many organic stuctures with $\overline{\sigma}$(C–C) in the range 0.011–0.015 Å that might be considered sufficiently precise for inclusion in many systematic studies.

A more fundamental objection to the AS flag is that it concentrates entirely on C atoms and, even then, conveys no explicit information concerning the precision of any given C atom or of any geometrical parameters involving that atom. Information of this kind is essential if we are to calculate weighted means in geometrical studies. The best that can be done at present is (*a*) to use the AS values to generate a 'semi-weighted mean', as suggested by Taylor & Kennard (1983), or (*b*) to generate unweighted means using subsets of CSD entries in which geometrical precision is likely to fall within relatively narrow limits, *e.g.* entries with, say, $R \leq 0.070$ and AS = 1 or 2.

For all of these reasons, it was decided to incorporate individual atomic coordinate e.s.d.'s into the CSD, beginning with entries from the 1985 literature. The AS flag has continued but is now assigned automatically using the stored e.s.d.'s. Further, by including the reported e.s.d.'s of invidual bond lengths in raw input to CSD check procedures, it is also possible to perform some numerical consistency checks on the keyboarded e.s.d. data.

The inclusion of coordinate e.s.d.'s is a significant improvement in CSD information content, an improvement that will be made available in CSD System software in the near future. However, this new addition does introduce a discontinuity in the CSD: the addition of any new information field at some point in time automatically creates a backlog of nonupgraded entries equal to the number of entries that existed prior to the upgrade! In this case, the backlog amounts to 50 000 entries and a fully retrospective upgrade is unlikely to be effected within current CSD work schedules. Thus, the CSD still lacks a single indicator of structural precision that can be included for all entries and can be represented as a continuous real-valued numerical variable: for example, the real $\overline{\sigma}$(C–C) or the real mean e.s.d. of a C-atom position [$\overline{\sigma}$(C)]. Further, it is still impossible to obtain reasonable estimates of weighted means for geometrical parameters calculated across a full range of CSD entries.

However, the CSD now contains a large number of entries (45 763, 44%) for which coordinate e.s.d. data *are* available and for which the problems noted above do not apply. In this paper, the available e.s.d. data are used to examine the possibility that we can predict reasonable values of $\overline{\sigma}$(C–C), and hence of $\overline{\sigma}$(C), using some function of variables $(X,Y,Z,...)$ that *do* exist in the vast majority of CSD entries, *i.e.*

$$\overline{\sigma}(\text{C–C})_p \simeq 2^{1/2}\overline{\sigma}(\text{C})_p \simeq f_p(X,Y,Z,...), \qquad (1)$$

where f_p is some predictive function and the factor of $2^{1/2}$ arises from the r.m.s. treatment of pairs of equal and spherical error distributions, $\overline{\sigma}$(C), in any structure. If a suitable function can be found, such that $\overline{\sigma}(\text{C–C})_p$ compares favourably with the observed $\overline{\sigma}(\text{C–C})_o$ calculated directly from the stored σ(C) values, then $\overline{\sigma}(\text{C–C})_p$ can be used as the continuous real-valued indicator of structural precision envisaged earlier. The value of $\overline{\sigma}(\text{C–C})_p$ would supplement the $\overline{\sigma}(\text{C–C})_o$ values embodied in AS flags for pre-1985 entries and even encompass those entries for which an AS flag is not available. A preliminary study of this kind, reported by Allen & Doyle (1987) and based on 4817 entries containing coordinate e.s.d.'s, showed the feasibility of the approach. We now report a more comprehensive study based on 35 747 entries containing coordinate e.s.d.'s.

Factors affecting stuctural precision

The theoretical background to this work is provided by Cruickshank (1960), who analysed the precision of X-ray intensity data that is required to yield a mean isotropic coordinate e.s.d., $\overline{\sigma}(A)$, for any element *A* in a structure that may also contain other elements *B*, *C etc.* In particular, he derived a simple approximate formula that related $\overline{\sigma}(A)$ to the residual (*R*), the chemical consitution of the asymmetric unit and limiting values from the data-collection experiment; thus,

$$\overline{\sigma}(A) \simeq R(N_A)^{1/2}/\bar{s}(mp)^{1/2}, \qquad (2)$$

where $\bar{s}$ is the r.m.s. reciprocal radius for the reflections observed and p is the difference between the number of independent reflections (N_r) and the number of parameters determined (N_p). The parameter N_A is the number of atoms similar to A that are required to give a scattering power at $\bar{s}$ that is equal to the scattering power of the N atoms of the asymmetic unit of the structure, *i.e.*

$$\sum_{i=1}^{N} f_i^2 = N_A f_A^2. \qquad (3)$$

The factor m in (2) is 4 for noncentrosymmetric space groups and 8 for centrosymmetric space groups (Cruickshank, 1960).

Because values of $\bar{s}$ are not available in the CSD and this study concentrates primarily on carbon, our best estimate of N_c (the 'equivalent number of C atoms') is given by:

$$N_c = \sum_{i=1}^{N} Z_i^2/Z_C^2. \qquad (4)$$

Here, the Z_i are atomic numbers, the denominator Z_C^2 is 36 and we must assume that all of the diffraction experiments were carried out with similar r.m.s. reciprocal radii for the observed reflections. Equation (2), then, predicts that $\bar{\sigma}$(C) and $\bar{\sigma}$(C–C) will decrease with decreasing R but will increase as the proportion and size of heavier atoms ($Z_i \gg 6$) in the structure increases. Equation (2) also predicts that $\bar{\sigma}$(C) and $\bar{\sigma}$(C–C) will be inversely proportional to $p^{1/2}$ and to $\bar{s}$; we return to this topic later in the paper. We concentrate initially on relationships between $\bar{\sigma}$(C) and $\bar{\sigma}$(C–C), the R factor and functions that involve the atomic numbers of the constituent elements of the structure.

Methodology

The January 1992 release of the CSD (see Table 1) was used throughout this analysis. Three subsets of CSD entries, hereinafter referred to as data sets 1, 2 and 3, were retrieved using both local code and the program *QUEST* (*Cambridge Structural Database User's Manual*, 1992).

Data set 1

This comprised the 83 516 entries for which atomic coordinates were available that have been published since 1965. This data set was used for a preliminary survey of the inter-relationship between AS values, R factors and $Z_{\max}$ (the maximum atomic number of any atom in each crystal structure).

Data set 2

This comprised the 35 747 entries for which atomic coordinate e.s.d.'s were available and that also satisfied the additional criteria: (*a*) the structure was determined by X-ray (not neutron) diffraction; (*b*) intensity data were collected on a diffractometer; (*c*) no residual numerical errors remained after CSD checking and evaluation procedures; (*d*) no disorder or polymeric (catena) bonding was reported; and (*e*) the R factor was less than 0.100. A work file was generated from this data set (see below) and formed the basis for the derivation of predictive functions, f_p, indicated in (1).

Table 3. *Definition of data items included for each entry in the work files generated for data sets* 2 *and* 3

REFCOD	CSD reference code
R	Crystallographic R factor
AS	CSD AS flag defined in Table 2
T	Temperature of data collection (K)
$N_{\rm nh}$	Number of non-H atoms in asymmetric unit
$\rm N_h$	Number of H atoms with coordinates reported
SPGN	Space-group number (*International Tables for X-ray Crystallography*, 1960)
CENT	Noncentrosymmetric = 1, centrosymmetric = 2
V	Unit-cell volume
$Z_{\max}$	Atomic number of heaviest element in structure
$Z_{\rm r.m.s.}$	$\left\{\sum_{i=1}^{N_a} Z_i^2/36N_{\rm nh}\right\}^{1/2}$, Z_i are atomic numbers
$N_c^{1/2}$	$\left\{\sum_{i=1}^{N_a} Z_i^2/36\right\}^{1/2}$, Z_i are atomic numbers
$RZ_{\max}$	The product $RZ_{\max}$
$RZ_{\rm r.m.s.}$	The product $RZ_{\rm r.m.s.}$
$RN_c^{1/2}$	The product $RN_c^{1/2}$
$\bar{\sigma}$(C–C)	Mean calculated e.s.d. of C–C bond lengths
σ(C–C)$_{\min}$	Minimum calculated e.s.d. of a C–C bond length
σ(C–C)$_{\max}$	Maximum calculated e.s.d. of a C–C bond length
$\bar{\sigma}$(C)	Mean isotropic e.s.d. of C atoms
$\bar{\sigma}(E)$	Mean isotropic e.s.d. of heaviest element(s)
N_b	Number of C–C bonds contributing to $\bar{\sigma}$(C–C)$_c$
N_r	Number of independent reflections (data set 3 only)
N_p	Number of parameters refined by least squares (data set 3 only)
N_r/N_p	Ratio of N_r to N_p

Data set 3

This comprised a small subset of 817 entries from data set 2 for which values of N_r (number of independent reflections) and N_p (number of parameters refined in the least-squares process) were abstracted from the original literature and added to the work file. This data set was used in attempts to improve the predictive function obtained from data set 2.

Generation of work file from data sets 2 *and* 3

The extensive binary CSD records for data set 2 were converted to a simple formatted ASCII work file using local software. The work file consisted of a single record for each entry that contained the information items of Table 3. These items were chosen as being related, directly or indirectly, to the precision of the coordinate set. For the small subset referred to as data set 3, values of N_r and N_p were initially abstracted from *Acta Crystallographica Section C*. However, owing to the high proportion of organic structures published there, the selection was balanced by N_r and N_p values abstracted from papers published in *Inorganic Chemistry* and *Organometallics*. The chemical composition of data set 3 was comparable with that of the larger data sets.

Obviously, since the CSD does not record details of the parameter variance matrix arising from least-squares-refinement procedures, analytical approximations were used to compute values for $\overline{\sigma}(\mathrm{C})$, $\overline{\sigma}(E)$, $\overline{\sigma}(\mathrm{C{-}C})$, $\sigma(\mathrm{C{-}C})_{\min}$ and $\sigma(\mathrm{C{-}C})_{\max}$ defined in Table 3. Muir & Mallinson (1993) have recently published a cautionary reminder of the hazards of applying simple analytical equations in oblique coordinate systems. In this work, we have essentially employed their method (2) to make allowance for the lack of variance information. Fractional coordinates $\mathbf{x}$ were first transformed to orthogonal values $\mathbf{X}$ *via* $\mathbf{X} = \beta^T\mathbf{x}$ [Dunitz, 1979, equation (5.30)]. Standard deviations of the orthogonal coordinates $\sigma(\mathbf{X})$ were obtained as the square roots of the diagonal elements of $\mathbf{S} = \beta^T\lambda\beta$, where λ is a symmetrical 3×3 variance matrix for the fractional coordinates having $\lambda_{ii} = \sigma^2(x_i)$, $\lambda_{ij} = \sigma(x_i)\sigma(x_j)r_{ij}$. Here, the correlation coefficients r_{ij} are taken as the cosines of the reciprocal-cell angles ($r_{12} = \cos\gamma^*$, $r_{13} = \cos\beta^*$, $r_{23} = \cos\alpha^*$) following the analysis of Templeton (1959). Variances in the cell parameters are ignored in our procedures.

For any atom A, we then compute an isotropic equivalent standard deviation $\sigma(A)$ as the r.m.s. average of $\sigma(X)$, $\sigma(Y)$, $\sigma(Z)$ arising from the treatment described above. Only the nonzero axial components were included in this averaging for atoms on special positions. The quantities $\overline{\sigma}(\mathrm{C})$ and $\overline{\sigma}(E)$ of Table 3 are then the mean isotropic equivalent standard deviations taken over all instances of carbon (C) and the heaviest element (E) in the structure, respectively. The e.s.d. of a bond length d between atoms A and B was calculated as

$$\begin{aligned}\sigma(A{-}B) = \{&[\sigma^2(X)_A + \sigma^2(X)_B](\Delta X/d)^2\\ &+ [\sigma^2(Y)_A + \sigma^2(Y)_B](\Delta Y/d)^2\\ &+ [\sigma^2(Z)_A + \sigma^2(Z)_B](\Delta Z/d)^2\}^{1/2}. \qquad (5)\end{aligned}$$

For isotropic errors in X, Y, Z, this value is approximated by

$$\sigma(A{-}B) = [\sigma^2(A) + \sigma^2(B)]^{1/2} \qquad (6)$$

and, for $A = B =$ carbon, we have

$$\sigma(\mathrm{C{-}C}) = 2^{1/2}\,\sigma(\mathrm{C}). \qquad (7)$$

Thus, for any structure we might expect that the ratio $\overline{\sigma}(\mathrm{C{-}C})/\overline{\sigma}(\mathrm{C})$ would be close to $2^{1/2}$. In fact, this ratio is 1.377, averaged across the very large variety of structures contained in data set 2 (see Table 6).

A number of additional checks were included in the software for work-file generation in an attempt (*a*) to improve the validity of the statistical analysis of work-file entries and (*b*) to guard against possible numerical errors in the coordinate e.s.d.'s entered in the CSD: these data, particularly in the early years of entry, were subject to limited scrutiny. All entries failing tests (iii) and (iv) below were omitted from the work files and will be further examined by CSD staff. The criteria applied were:

(i) The number of C–C bonds (N_b, Table 3) used to generate $\overline{\sigma}(\mathrm{C{-}C})$ must be ≥ 5.

(ii) The distribution of the individual bond-length e.s.d.'s, $\sigma(\mathrm{C{-}C})_i$ ($i = 1{-}N_b$) was examined for each entry. The sample standard deviation

$$S = \left\{\sum_{i=1}^{N_b}[\sigma(\mathrm{C{-}C})_i - \overline{\sigma}(\mathrm{C{-}C})]^2/N_b - 1\right\}^{1/2} \qquad (8)$$

was calculated and $\sigma(\mathrm{C{-}C})_i$ were eliminated if $|\overline{\sigma}(\mathrm{C{-}C}) - \sigma(\mathrm{C{-}C})_i| > 4S$, N_b was decreased accordingly and, if N_b was still ≥ 5, $\overline{\sigma}(\mathrm{C{-}C})$ was recalculated for the work-file entry.

(iii) Entries with highly skewed distributions of $\sigma(\mathrm{C{-}C})_i$ were eliminated by calculating

$$M = [\sigma(\mathrm{C{-}C})_{\max} + \sigma(\mathrm{C{-}C})_{\min}]/2, \qquad (9)$$

as an approximation of the median value, and then comparing M with the actual $\overline{\sigma}(\mathrm{C{-}C})$ value *via*

$$D_\sigma = [|\overline{\sigma}(\mathrm{C{-}C}) - M|]/\overline{\sigma}(\mathrm{C{-}C}). \qquad (10)$$

Entries with $D_\sigma > 0.25$ were rejected.

(iv) In a few cases, it was found that coordinate e.s.d.'s in the CSD were larger or smaller than reported values by a factor of 10 (integer e.s.d.'s in keyboarded input entered with an erroneous power of 10). These instances are internally consistent and pass tests (ii) and (iii). They were eliminated (*a*) when $\overline{\sigma}(\mathrm{C{-}C}) > 0.04$ Å, or (*b*) when the AS flag calculated from the stored e.s.d.'s did not agree with that recorded in the CSD, within a tolerance of 0.001 Å for AS = 1, 2 and 0.002 Å for AS = 3, 4.

Tests (i)–(iv) are stringent and may have eliminated some valid entries. However, they do guarantee that erroneous e.s.d. data are rejected and could form a basis for improved checking during future database building operations. As a result of these tests, the work file denoted as data set 2 was reduced to 29 362 entries, which were used in subsequent analyses.

Analysis of data sets 1, 2 *and* 3

Data sets 1 and 2 were both used to generate the simple descriptive statistics presented below. Data sets 2 and 3 were then further analysed using (*a*) correlation analysis to examine the interdependence of parameters in each work file; (*b*) simple linear regressions of the general form $\overline{\sigma}(\mathrm{C{-}C})_p = a + b(f_p)$, where f_p [(1)] is analogous to the expression of

Cruickshank (1960); (*c*) multiple linear regressions of the general form $\overline{\sigma}(\text{C–C})_p = a + bL + cM + ...$, where *L*, *M etc.* are individual data items held in the respective work files. All calculations were carried out using local software that incorporated calls to appropriate statistical routines from the CAMAL library (Taylor, 1986) or from the NAG library (Numerical Algorithms Group, 1990). Computations were carried out on a stand-alone DEC/VAX 3300 and on a VAX cluster in the Department of Chemistry, University of Durham, and on an IBM 3084Q at the University of Cambridge Computer Laboratory.

Terminology

The mean values of parameters that can vary within a single structure are denoted thoughout by $\overline{\sigma}$(C–C), $\overline{\sigma}$(C), $\overline{\sigma}(E)$ *etc.* Mean values of parameters taken over many structures are denoted by $\langle\overline{\sigma}(X-X)\rangle$, $\langle R\rangle$, $\langle Z_{\max}\rangle$ *etc.*

Results and discussion

Descriptive statistics for data sets 1 *and* 2

Descriptive statistics for the complete database are not only relevant to our overall analysis but are also of intrinsic crystallographic interest. Figs. 2 and 3 show the distributions of *R* (Fig.2) and the AS flag (Fig. 3) for, in each case: (*a*) all structures; (*b*) light-atom structures ($Z_{\max} \le 18$); (*c*) heavy-atom structures ($Z_{\max} > 18$). Fig. 2 shows that the overall mean $\langle R\rangle$ for heavy-atom structures is slightly lower (0.055) than that for light-atom structures (0.060). However, Fig. 3 shows that the AS flag, taken as a rough measure of $\overline{\sigma}$(C) or $\overline{\sigma}$(C–C), is considerably affected by increasing $Z_{\max}$. The overall mean ⟨AS⟩ is 2.27 over all structures (Fig. 3*a*) but falls to 1.90 for the light-atom structures (Fig. 3*b*) and rises to 2.53 for the heavy-atom structures (Fig. 3*c*), despite the lower overall $\langle R\rangle$ exhibited by the latter category. This proportionality between $\overline{\sigma}$(C) and increasing atomic numbers implicit in (2) is further quantified in Table 4, in which ⟨AS⟩ increases systematically over five ranges of $Z_{\max}$.

Table 5 illustrates the expected relationship between AS and *R*, in which the CSD chemical-class system is used to mimic the $Z_{\max}$ divisions of Figs. 2 and 3. Table 5(*a*) shows that ⟨AS⟩ increases smoothly from 1.73 to 3.21 as $\langle R\rangle$ increases from 0.025 to 0.125. This relationship is also observed for the light-atom structures (Table 5*b*) and heavy-atom structures (Table 5*c*), but the general level of ⟨AS⟩ values of Table 5(*b*) is significantly lower than that for Table 5(*c*), as expected from the results of Table 4. Table 5 is given in full, since it allows a rapid visual assessment of typical percentages of CSD entries that will survive secondary search criteria based on various *R* and AS limits.

The data set 2 work file permits us to move from the rather crude AS measure of coordinate precision to the more exact $\overline{\sigma}$(C) and $\overline{\sigma}$(C–C) values calculated from the coordinate e.s.d.'s stored in the CSD. Fig. 4 shows distributions of $\overline{\sigma}$(C–C) for all structures (Fig. 4*a*) and for the light-atom (Fig. 4*b*) and heavy-atom (Fig. 4*c*) subdivisions that show the same functional form as the AS distributions of Fig. 3. The overall mean $\langle\overline{\sigma}(\text{C–C})\rangle$ for 26 529 entries from data set 2 that pass all criteria noted earlier is 0.0103 Å. However, $\langle\overline{\sigma}(\text{C–C})\rangle$ for the light-atom structures ($Z_{\max} \le 18$) is 0.0068 Å, almost exactly half of the $\langle\overline{\sigma}(\text{C–C})\rangle$ of 0.0134 Å exhibited by the heavy-atom ($Z_{\max} > 18$) structures. Again, the effect of increasing $Z_{\max}$ on $\overline{\sigma}$(C–C) is further illustrated by the results of data set 2 that are included in Table 4. The distributions of Fig. 4 are also clear visual proof of the inappropriate

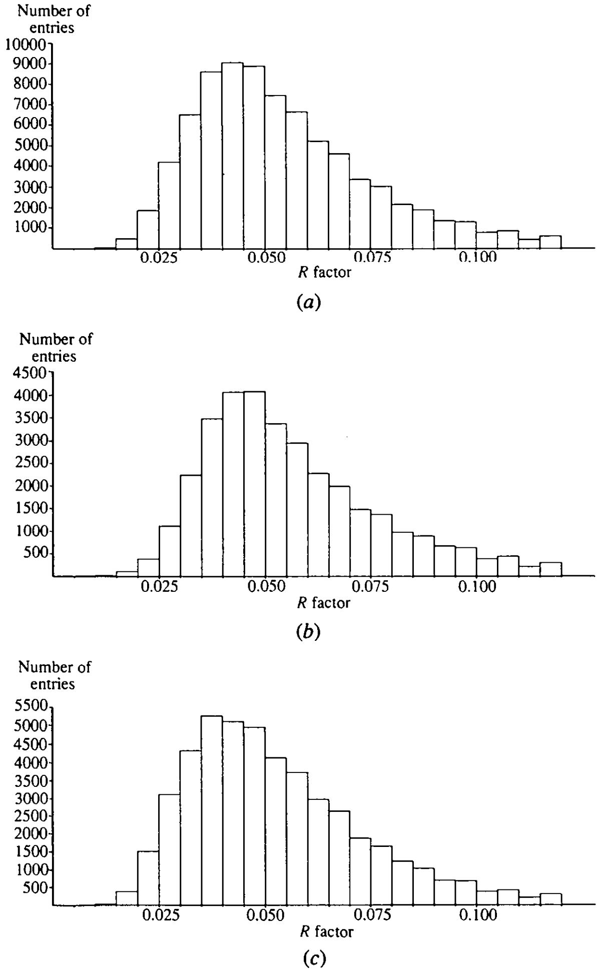

Fig. 2. Distributions of *R* factors in data set 1: (*a*) all structures; (*b*) light-atom structures ($Z_{\max} \le 18$); (*c*) heavy-atom structures ($Z_{\max} > 18$).

banding of the current AS flags for the classification of $\overline{\sigma}$(C–C) values obtained in modern structure determinations; this is especially true for light-atom structures.

Table 4. *Distribution of the AS flag (data set* 1) *and of* $\overline{\sigma}$(C–C) *(data set* 2) *for ranges of* Z_{max}

$N_{\rm ent}$ is the number of entries in a range and ⟨AS⟩, ⟨R⟩, $\langle Z_{\rm max}\rangle$ and $\langle\overline{\sigma}$(C–C)⟩ are mean values for that range. The AS distribution is qualified by values in parentheses that are percentages of the total number of entries in data set 1. For data set 2, results are quoted for entries that pass all criteria described in the text.

	$Z_{\rm max}$ range				
	(1) 6–10	(2) 11–18	(3) 19–36	(4) 37–57	(5) ≥58
Data set 1					
$N_{\rm ent}$	20137	28091	21777	8001	5510
⟨R⟩	0.059	0.056	0.060	0.053	0.051
$\langle Z_{\rm max}\rangle$	7.84	15.99	28.68	46.46	76.50
⟨AS⟩	1.71	2.28	2.47	2.61	3.06
AS = 0	2697 (3)	4608 (6)	3664 (4)	1221 (1)	859 (1)
AS = 1	8212 (10)	5243 (6)	2563 (3)	656 (1)	104 (–)
AS = 2	6335 (8)	8276 (10)	6223 (7)	2143 (3)	708 (1)
AS = 3	2700 (3)	8200 (10)	7614 (9)	3160 (4)	2639 (3)
AS = 4	193 (–)	1764 (2)	1713 (2)	821 (1)	1200 (1)
Data set 2					
$N_{\rm ent}$	7831	4627	6005	4581	3485
⟨R⟩	0.052	0.051	0.050	0.044	0.042
$\langle Z_{\rm max}\rangle$	7.88	15.94	28.35	46.00	76.15
$\langle\overline{\sigma}$(C–C)⟩	0.0064	0.0076	0.0112	0.0127	0.0180

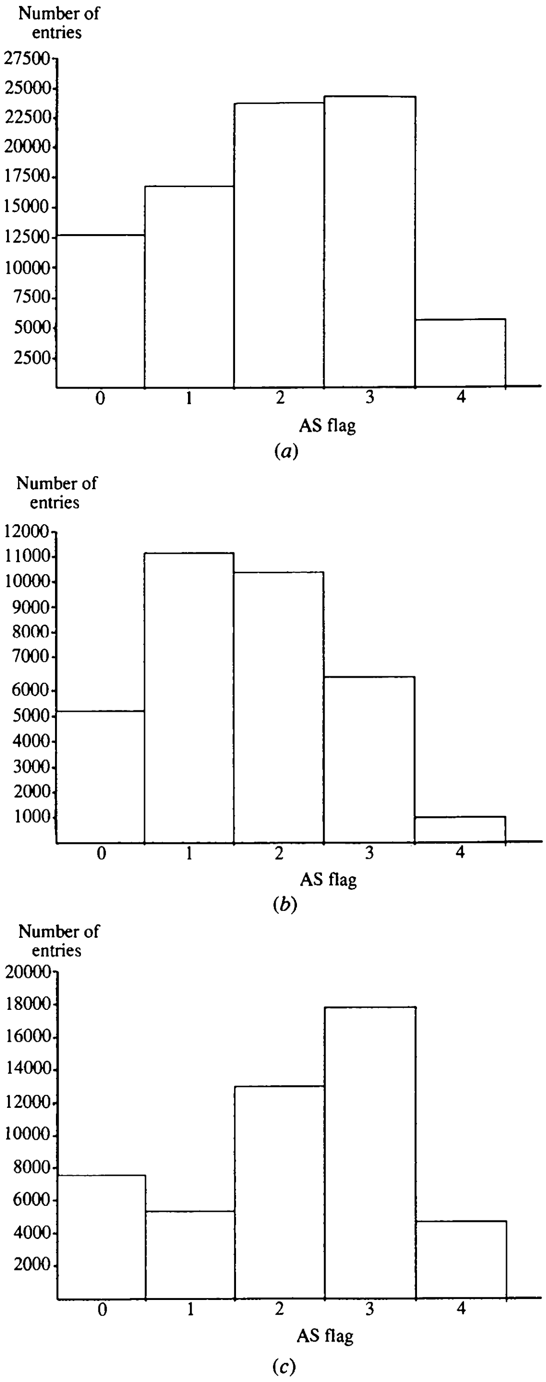

Fig. 3. Distributions of AS flags in data set 1: (*a*) all structures; (*b*) light-atom structures ($Z_{\rm max} \le 18$); (*c*) heavy-atom structures ($Z_{\rm max} > 18$).

The overall relationship between $\overline{\sigma}$(C–C), R and $Z_{\rm max}$ is explored in Fig. 5. To generate Fig. 5(*a*), data set 2 was divided into the five ranges of $Z_{\rm max}$ identified in Table 4. Then, for each $Z_{\rm max}$ range, the mean value $\langle\overline{\sigma}$(C–C)⟩ was determined for each of five R-factor ranges: 0.001–0.035, 0.036–0.045, 0.046–0.055, 0.056–0.070 and 0.071–0.100. The $\langle\overline{\sigma}$(C–C)⟩ values were then plotted against ⟨R⟩ for each of the $Z_{\rm max}$ ranges to generate Fig. 5(*a*). The 25 'bins' used in this procedure ranged in size from 433 to 2251 entries except for the highest-R-factor bins of $Z_{\rm max}$ ranges (4) and (5), which contained only 300 and 141 entries, respectively. Within any $Z_{\rm max}$ range, the mean, $\langle Z_{\rm max}\rangle$, is virtually identical for each of its five R-factor bins. Thus, Fig. 5(*a*) shows a clear linear relationship between the $\langle\overline{\sigma}$(C–C)⟩ and increasing ⟨R⟩ in each (effectively constant) $Z_{\rm max}$ range (full lines) and a similar linear increase in $\langle\overline{\sigma}$(C–C)⟩ with increasing $\langle Z_{\rm max}\rangle$ at an effectively constant R factor (dotted lines). Despite some disparities in the bin sizes, the results of Fig. 5(*a*) are in clear agreement with the predictions of Cruickshank (1960).

Finally, in Fig. 5(*b*), we explore composite relationships of the form $\overline{\sigma}$(C–C) = $kRf(Z)$ as suggested by (2), in which $f(Z)$ is taken as (see Table 3): (i) $Z_{\rm max}$; (ii) $Z_{\rm r.m.s.}$; (iii) $N_c^{1/2}$ evaluated with (4). Since it is impractical to plot ~30 000 points for each relationship, a simple binning process has been applied so as to visualize the subset of data set 2 denoted S140 in Table 6. The 25 984 entries were divided into eight equal 3248-entry bins for each of $RZ_{\rm max}$, $RZ_{\rm r.m.s.}$ and $RN_c^{1/2}$ by the use of sorted lists. Mean values of $\langle Rf(Z)\rangle$ and $\langle\overline{\sigma}$(C–C)⟩ were calculated for each bin to generate the three eight-point plots that are superimposed as Fig. 5(*b*). It can be seen that all

Table 5. *Distributions of the AS flag* (*Table* 2) *versus R factor for CSD entries having coordinates that have been published since* 1965 (*data set* 1)

N_{ent} is the number of entries in a given R-factor range, $\langle R \rangle$ is the mean of the available ($R > 0.0$) R factors, $\langle \text{AS} \rangle$ is the mean (see text) of the available nonzero AS flags. Integers (≥ 1) in parentheses are percentages of the total number of entries in each subdivision.

R range	N_{ent}	$\langle R \rangle$	AS = 0	AS = 1	AS = 2	AS = 3	AS = 4	$\langle \text{AS} \rangle$
(*a*) All entries (83 516)								
Unknown	581 (1)		261	73	95	117	35	2.36
0.001–0.030	5832 (7)	0.025	667 (1)	2228 (3)	2123 (3)	773 (1)	41	1.73
0.031–0.040	14725 (17)	0.035	1602 (2)	5028 (6)	4869 (6)	3016 (4)	210	1.88
0.041–0.050	18268 (22)	0.045	2249 (3)	5191 (6)	5806 (7)	4475 (5)	547 (1)	2.02
0.051–0.060	14650 (17)	0.054	2034 (2)	2556 (3)	4695 (6)	4592 (6)	773 (1)	2.28
0.061–0.070	10267 (12)	0.064	1599 (2)	1023 (1)	2937 (4)	3866 (5)	842 (1)	2.52
0.071–0.080	6721 (8)	0.074	1232 (2)	392	1572 (2)	2720 (3)	805 (1)	2.71
0.081–0.100	7174 (8)	0.088	1548 (2)	213	1208 (1)	3028 (4)	1177 (1)	2.91
≥ 0.100	5298 (6)	0.125	1857 (2)	74	380	1726 (2)	1261 (2)	3.21
Totals	82935 (99)	0.057	13049 (16)	16778 (20)	23685 (28)	24313 (29)	5691 (7)	2.27
(*b*) ‘Organic’ structures: CSD classes 1–59 (34 992)								
Unknown	301 (1)		119	60	59	54	9	2.07
0.001–0.030	1334 (4)	0.025	199 (1)	751 (2)	286 (1)	90	8	1.43
0.031–0.040	5317 (15)	0.035	571 (2)	3167 (9)	1259 (4)	284 (1)	36	1.41
0.041–0.050	8110 (23)	0.045	898 (3)	3986 (11)	2501 (7)	670 (2)	55	1.55
0.051–0.060	6474 (19)	0.054	795 (2)	2081 (6)	2549 (7)	967 (3)	82	1.83
0.061–0.070	4484 (13)	0.064	620 (2)	832 (2)	1814 (5)	1129 (3)	89	2.12
0.071–0.080	2946 (8)	0.074	502 (1)	305 (1)	1062 (3)	986 (3)	91	2.35
0.081–0.100	3338 (10)	0.088	682 (2)	173 (–)	869 (3)	1389 (4)	225 (1)	2.63
≥ 0.101	2688 (8)	0.125	999 (3)	50 (–)	285 (1)	946 (3)	408 (1)	3.01
Totals	34691 (99)	0.060	5385 (15)	11405 (33)	10684 (31)	6515 (19)	1003 (2)	1.90
(*c*) ‘Metallo-organic’ structures: CSD classes 60–86 (48 524)								
Unknown	280 (1)		142	13	36	63	26	2.73
0.001–0.030	4498 (9)	0.025	468 (1)	1477 (3)	1837 (4)	683 (1)	33	1.82
0.031–0.040	9408 (19)	0.035	1031 (2)	1861 (4)	3610 (7)	2732 (6)	174	2.14
0.041–0.050	10158 (21)	0.045	1351 (3)	1205 (3)	3305 (7)	3805 (8)	492 (1)	2.40
0.051–0.060	8176 (17)	0.054	1239 (3)	475 (1)	2146 (4)	3625 (7)	691 (1)	2.65
0.061–0.070	5783 (12)	0.064	979 (2)	191	1123 (2)	2737 (6)	753 (2)	2.84
0.071–0.080	3775 (8)	0.074	730 (2)	87	510 (1)	1734 (4)	714 (2)	3.00
0.081–0.100	3836 (8)	0.088	866 (2)	40	339 (1)	1639 (3)	952 (2)	3.18
≥ 0.101	2610 (5)	0.124	858 (2)	24	95	780 (2)	853 (2)	3.40
Totals	48244 (99)	0.055	7664 (16)	5273 (11)	13001 (27)	17798 (37)	4688 (10)	2.53

three composite functions are reasonably linear in $\overline{\sigma}$(C–C). However, the $RN_c^{1/2}$ function of Cruickshank (1960) appears to generate the ‘best’ straight-line fit, and one that passes close to the origin, in agreement with (2).

Correlation analysis of data set 2

The full symmetric correlation matrix for all of the numerical data items of Table 3 was calculated from the data set 2 work file. The calculation was carried out for two subsets (denoted *S*140 and *S*220) of the complete data set using the $\overline{\sigma}$(C–C) and *R*-factor limits given in Table 6(*a*). Correlation coefficients C_{ij}, linking R and $\overline{\sigma}$(C–C) with the most relevant data items, are given in Table 6(*b*).

Many of the C_{ij}'s simply provide a numerical confirmation of the observations made in the previous section. Thus, R is positively correlated with AS and with $\overline{\sigma}(E)$, $\overline{\sigma}$(C) and $\overline{\sigma}$(C–C) and is negatively correlated with functions derived from the atomic numbers of the constituent elements. The correlations of R and of $\overline{\sigma}$(C–C) with T, the temperature of data collection, are close to zero, a reflection of the total domination of the data set by room-temperature studies. $\overline{\sigma}$(C–C) is positively correlated with N_{nh}, the number of independent non-H atoms, and with V, the unit-cell volume. However, R and Z_{max} are also positively correlated with these quantities. It is not possible to locate further independent parameters from the wide selection in Table 3 that can validly be included in regression experiments. What is important in Table 6 is the high level of correlation between $\overline{\sigma}$(C–C) and the Cruickshank-like functions RZ_{max}, $RZ_{\text{r.m.s.}}$ and $RN_c^{1/2}$.

Preliminary regression analysis of data set 2

The 11 simple and multiple linear regressions carried out on subsets *S*140 and *S*220 of data set 2 (see Table 6*a*) are enumerated and defined in Table 7(*a*). The simple linear regressions of type 1 have the functional form [(2)] proposed by Cruickshank

(1960), with an assumed constancy of his denominator across all structures. If we define $\overline{\sigma}$(C–C), calculated from published coordinate e.s.d.'s, as the 'observed' value ($\overline{\sigma}_o$) and that 'predicted' by any regression equation as $\overline{\sigma}_p$, then there are many ways in which the two distributions may be compared. The quantities chosen for use in this study are defined in Table 7(*b*) and include calculations of the ability of each regression to predict the current AS flag stored in each CSD entry.

The main numerical results of the regression analyses are presented in Table 8(*a*) for the 25 984 entries of subset *S*140 and in Table 8(*b*) for the 20 334 entries of subset *S*220; at this stage, no distinction is made between centrosymmetric and non-centrosymmetric structures. Both sets of results show that regressions 1.1 and 2.1 based only on the *R* factor are significantly inferior to those that involve some function of the atomic numbers Z_i. It remains, then, to make a choice for that function between $Z_{\max}$, $Z_{\rm r.m.s.}$ and the $N_c^{1/2}$ suggested by Cruickshank (1960), and also between the various forms of the regression equations [(1), (2) or (3): Table 7(*b*)].

In all cases, the introduction of $Z_{\max}$ (the simplest function of the Z_i) results in a significant improvement over the use of *R* alone. However, there is no doubt that the assessment criteria are all further improved by the use of $Z_{\rm r.m.s.}$ or $N_c^{1/2}$ functions that take account of all of the Z_i rather than the simple maximum. The choice between $Z_{\rm r.m.s.}$ and $N_c^{1/2}$ is, perhaps, a little more subjective. However, for the larger subset *S*140, 20 of the 30 assessment criteria from the three types of equations are improved by the use of $N_c^{1/2}$, while, for subset *S*220, 18 are improved. There appears to be no valid statistical reason to choose $Z_{\rm r.m.s.}$ in preference to the $N_c^{1/2}$ values.

The original Cruickshank (1960) expression [(2)] has the form of a straight line passing through the origin: a type 1 regression in our analysis. The results of Table 8 for both subsets show that the assessment criteria are not significantly improved (and are sometimes marginally worsened) by the introduction of additional degrees of freedom in regressions of types

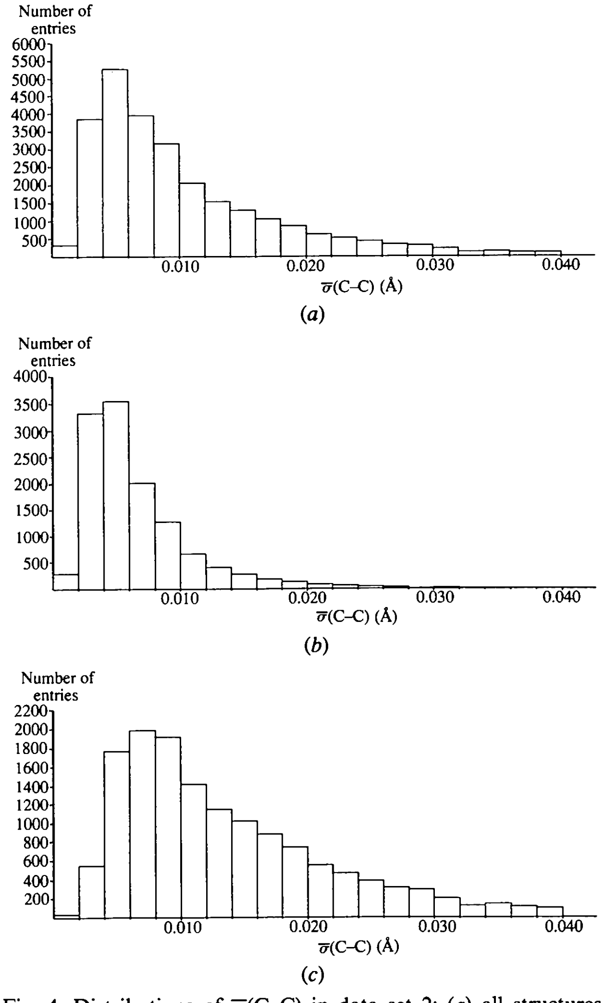

Fig. 4. Distributions of $\overline{\sigma}$(C–C) in data set 2: (*a*) all structures; (*b*) light-atom structures ($Z_{\max} \leq 18$); (*c*) heavy-atom structures ($Z_{\max} > 18$).

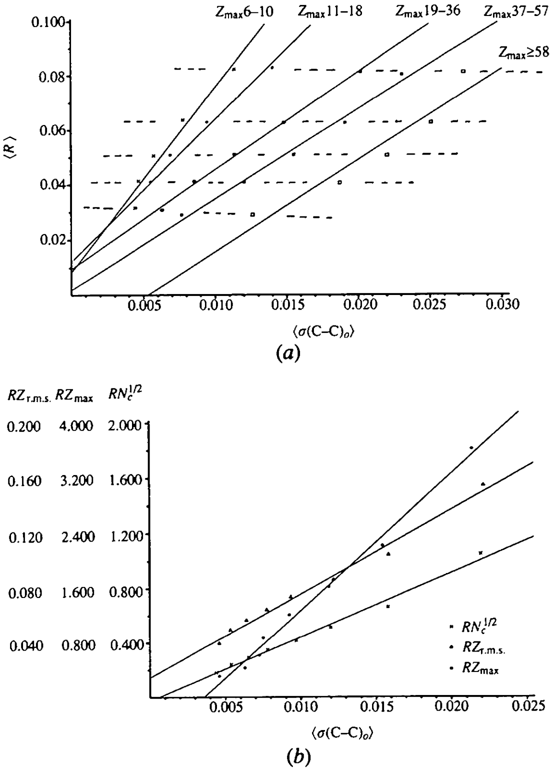

Fig. 5. Relationships between $\overline{\sigma}$(C–C), *R* and functions of atomic numbers, Z_i, in the structure: (*a*) plot of binned means ⟨$\overline{\sigma}$(C–C)⟩ *versus* ⟨*R*⟩ for five ranges of $Z_{\max}$; (*b*) plots of binned means ⟨$\overline{\sigma}$(C–C)⟩ *versus* ⟨$RZ_{\max}$⟩, ⟨$RZ_{\rm r.m.s.}$⟩ and ⟨$RN_c^{1/2}$⟩. Binning procedures are described in the text.

Table 6. *Correlation analysis of data set 2*

Data items and their names are as defined in Table 3. Two subsets of entries, denoted *S*140 and *S*220, were chosen for analysis on the basis of $\bar{\sigma}$(C–C) and *R*-factor limits. The C_{ij} are correlation coefficients selected from the complete matrix for each subset.

(*a*) Subset definition

Subset name	*S*140	*S*220
$\bar{\sigma}$(C–C) minimum (Å)	0.001	0.002
$\bar{\sigma}$(C–C) maximum (Å)	0.040	0.020
R minimum	0.001	0.001
R maximum	0.100	0.070
No. of entries	25984	20334
$\langle\bar{\sigma}(\text{C–C})/\bar{\sigma}(\text{C})\rangle$	1.377	1.377

(*b*) Correlation coefficients

Item *i*	Item *j*	C_{ij}	C_{ij}
R	AS	0.359	0.214
R	*T*	0.051	0.050
R	N_{nh}	0.166	0.092
R	N_{h}	0.022	0.038
R	*V*	0.132	0.063
R	$Z_{\max}$	−0.227	−0.337
R	$Z_{\text{r.m.s.}}$	−0.229	−0.343
R	$N_c^{1/2}$	−0.114	−0.232
R	$\bar{\sigma}(E)$	0.113	0.361
R	$\bar{\sigma}$(C)	0.369	0.221
$\bar{\sigma}$(C–C)	*R*	0.363	0.214
$\bar{\sigma}$(C–C)	AS	0.851	0.888
$\bar{\sigma}$(C–C)	*T*	0.044	0.045
$\bar{\sigma}$(C–C)	N_{nh}	0.338	0.303
$\bar{\sigma}$(C–C)	N_{h}	−0.030	0.026
$\bar{\sigma}$(C–C)	*V*	0.351	0.303
$\bar{\sigma}$(C–C)	$Z_{\max}$	0.525	0.520
$\bar{\sigma}$(C–C)	$Z_{\text{r.m.s.}}$	0.529	0.505
$\bar{\sigma}$(C–C)	$N_c^{1/2}$	0.584	0.558
$\bar{\sigma}$(C–C)	$RZ_{\max}$	0.700	0.646
$\bar{\sigma}$(C–C)	$RZ_{\text{r.m.s.}}$	0.737	0.657
$\bar{\sigma}$(C–C)	$RN_c^{1/2}$	0.729	0.657
$\bar{\sigma}$(C–C)	$\bar{\sigma}(E)$	0.031	0.035
$\bar{\sigma}$(C–C)	$\bar{\sigma}$(C)	0.985	0.978

Table 7. *Regression analysis summary for data set 2*

(*a*) Enumeration and definition of the simple linear and multiple linear regressions performed on data set 2

Regression no. (*RN*)	Functional form
1. Simple linear, no constant	
1.1	$\bar{\sigma}(\text{C–C})_p = kR$
1.2	$\bar{\sigma}(\text{C–C})_p = kRZ_{\max}$
1.3	$\bar{\sigma}(\text{C–C})_p = kRZ_{\text{r.m.s.}}$
1.4	$\bar{\sigma}(\text{C–C})_p = kRN_c^{1/2}$
2. Simple linear with constant	
2.1	$\bar{\sigma}(\text{C–C})_p = a + kR$
2.2	$\bar{\sigma}(\text{C–C})_p = a + kRZ_{\max}$
2.3	$\bar{\sigma}(\text{C–C})_p = a + kRZ_{\text{r.m.s.}}$
2.4	$\bar{\sigma}(\text{C–C})_p = a + kRN_c^{1/2}$
3. Multiple linear	
3.2	$\bar{\sigma}(\text{C–C})_p = a + bR + cZ_{\max}$
3.3	$\bar{\sigma}(\text{C–C})_p = a + bR + cZ_{\text{r.m.s.}}$
3.4	$\bar{\sigma}(\text{C–C})_p = a + bR + cN_c^{1/2}$

(*b*) Parameters used to assess regression results [$\bar{\sigma}_o$ is $\bar{\sigma}$(C–C) as observed in crystal structures: $\bar{\sigma}_p$ is the value of $\bar{\sigma}$(C–C) predicted by the regression equation]

Item	Description
R_σ	A pseudo *R* factor measuring the discrepancy between the $\bar{\sigma}_o$ and $\bar{\sigma}_p$ distributions where $R_\sigma = \sum\lvert\bar{\sigma}_o - \bar{\sigma}_p\rvert/\sum\bar{\sigma}_o$
r.m.s. (σ)	The r.m.s. error, *i.e.* r.m.s.(σ) = $[\sum(\bar{\sigma}_o - \bar{\sigma}_p)^2/n]^{1/2}$ for *n* observations in the subset
N_{50}	Percentage of entries with $\lvert\bar{\sigma}_o - \bar{\sigma}_p\rvert \le 0.0050$ Å
N_{25}	Percentage of entries with $\lvert\bar{\sigma}_o - \bar{\sigma}_p\rvert \le 0.0025$ Å
N_{10}	Percentage of entries with $\lvert\bar{\sigma}_o - \bar{\sigma}_p\rvert \le 0.0010$ Å
n_{50}	Percentage of entries for which $\bar{\sigma}_p$ is within 50% of $\bar{\sigma}_o$
n_{25}	Percentage of entries for which $\bar{\sigma}_p$ is within 25% of $\bar{\sigma}_o$
n_{10}	Percentage of entries for which $\bar{\sigma}_p$ is within 10% of $\bar{\sigma}_o$
AS_0	Percentage of AS flags that are predicted exactly by the regression equation
AS_2	Percentage of AS flags that are predicted exactly or within 0.002 Å of the relevant AS flag limits of Table 2

2 or 3. In particular, the regression intercept *a* is inconsistently predicted and adopts negative values of appreciable magnitude in type 3 equations. Again, there are no statistical reasons to choose the more complex regression forms over the orignal type 1 formulation given by Cruickshank (1960), despite our assumption of a constant denominator in (2).

Centrosymmetric and noncentrosymmetric structures

We subsequently performed type 1 regressions on subdivisions of *S*220 that contained only centrosymmetric (*S*220*c*: Table 8*c*) or noncentrosymmetric (*S*220nc; Table 8*d*) structures. Assessment criteria for both subdivisions show small but consistent improvements by comparison with the overall results of Table 8(*b*). Regressions of type 1.3, based on $RZ_{\text{r.m.s.}}$, and type 1.4, based on $RN_c^{1/2}$, give the best results for both *S*220*c* and *S*220nc and the regressions of type 1.4 appear marginally preferable on the basis of improvements in the assessment criteria.

Cruickshank's (1960) equation (2) indicates that the proportionality constant *k* should differ by a factor of $2^{1/2}$ for centrosymmetric and noncentrosymmetric structures, *i.e.* $k_{\text{nc}} = 2^{1/2}k_c$. The ratio of k_{nc}/k_c from our regressions of type 1.4 in Tables 8(*c*) and (*d*) is 1.399, remarkably close to the expected value of 1.414. Values of k_{nc} (Table 8*d*) are consistently larger than values of k_c (Table 8*c*) for regressions of types 1.1, 1.2 and 1.3 but their ratios (1.071, 1.568, 1.307) are appreciably different from $2^{1/2}$.

Thus, we would recommend the use of regression equation (1.4*c*) or (1.4*d*), *viz*:

$$\bar{\sigma}(\text{C–C})_p = 0.01814RN_c^{1/2} = \bar{\sigma}_p \qquad (11a)$$

or

$$\bar{\sigma}(\text{C–C})_p = 0.02537RN_c^{1/2} = \bar{\sigma}_p, \qquad (11b)$$

as the best predictive equations for centrosymmetric (11*a*) and noncentrosymmetric (11*b*) structures. In Fig. 6, we show (*a*) the composite numerical error distribution ($\bar{\sigma}_o - \bar{\sigma}_p$) and (*b*) the composite percentage error distribution $[(\bar{\sigma}_o - \bar{\sigma}_p)|\bar{\sigma}_o]$ calculated for subset *S*220 by use of (11). Each bar on the histograms is annotated with the mean value $\langle\bar{\sigma}_o\rangle$ for the structures represented by that bar. The distribution of numerical errors is approximately normal with a

Table 8. *Results of simple linear and multiple linear regressions for data set* 2

The regression equations are those of Table 7(*a*) and the parameters used to assess the results are described in Table 7(*b*).

RN	*a*	*b* or *k*	*c*	R_σ (%)	r.m.s. (σ) (Å)	N_{50} (%)	N_{25} (%)	N_{10} (%)	n_{50} (%)	n_{25} (%)	n_{10} (%)	AS_0 (%)	AS_2 (%)
(*a*) Overall results for the 25 984 entries of subset *S*140 defined in Table 6(*a*)													
1.1		0.2070 (8)		50.8	0.0069	59.1	30.2	12.1	52.4	26.7	10.5	39.6	59.7
1.2		0.00651 (2)		39.1	0.0058	73.2	47.5	21.4	67.0	35.3	14.0	59.0	76.9
1.3		0.1350 (3)		35.5	0.0051	78.8	45.4	18.0	67.2	39.2	16.4	54.2	77.4
1.4		0.02157 (6)		33.5	0.0051	79.3	54.8	24.4	75.1	44.0	18.4	61.1	83.2
2.1	0.00195 (13)	0.1707 (27)		51.1	0.0069	58.0	28.7	11.3	51.4	26.1	10.4	39.5	59.0
2.2	0.00376 (5)	0.00481 (3)		36.9	0.0053	76.9	45.5	17.6	65.6	37.1	15.0	54.1	77.2
2.3	−0.00125 (6)	0.1482 (8)		34.2	0.0050	79.3	50.0	21.1	70.8	41.6	17.6	57.7	80.4
2.4	0.00138 (6)	0.01937 (11)		34.4	0.0051	79.6	50.0	20.2	70.0	40.9	17.4	57.0	80.7
3.2	−0.00755	0.23906	0.00021	35.9	0.0051	75.6	47.4	20.5	68.9	39.2	16.4	58.6	77.0
3.3	−0.01284	0.24010	0.00692	35.4	0.0051	76.4	48.7	21.2	70.1	40.3	16.9	58.5	78.4
3.4	−0.00855	0.20464	0.00092	35.1	0.0051	77.0	49.0	21.5	70.6	40.5	17.3	59.1	78.7
(*b*) Overall results for the 20 334 entries of subset *S*220 defined in Table 6(*a*)													
1.1		0.1755 (7)		42.4	0.0044	77.2	43.9	17.9	62.6	32.0	13.1	40.1	68.9
1.2		0.00564 (2)		38.5	0.0042	80.6	52.3	21.8	66.4	34.2	14.1	58.2	77.3
1.3		0.1169 (3)		29.9	0.0032	89.1	63.2	27.6	78.1	47.6	20.3	57.1	85.6
1.4		0.01928 (5)		30.2	0.0034	87.3	64.7	31.7	82.5	49.4	20.5	60.8	86.6
2.1	0.00465 (11)	0.0777 (2)		41.7	0.0042	79.9	42.2	16.3	61.6	32.3	13.1	40.3	61.2
2.2	0.00437 (4)	0.00322 (2)		31.1	0.0033	89.3	59.6	23.8	74.4	43.6	18.1	53.9	84.8
2.3	0.00079 (6)	0.1068 (9)		30.3	0.0032	89.4	61.7	26.0	76.2	46.3	19.6	55.1	84.9
2.4	0.00262 (5)	0.01399 (11)		30.4	0.0032	89.3	61.8	25.1	75.8	45.2	19.0	54.6	85.3
3.2	−0.00276	0.15959	0.00013	30.3	0.0032	89.4	60.7	26.1	76.4	45.9	19.3	56.8	84.2
3.3	−0.00627	0.15933	0.00457	30.7	0.0032	88.7	60.5	26.1	76.2	45.7	19.3	55.9	84.1
3.4	−0.00336	0.13170	0.00062	30.3	0.0032	89.0	61.0	26.3	76.7	46.0	19.5	56.4	84.7
(*c*) Results for the 15 170 centrosymmetric structures of subset *S*220 defined in Table 6(*a*)													
1.1		0.1725 (7)		44.0	0.0044	76.1	42.0	17.6	60.9	30.5	12.2	39.9	67.9
1.2		0.00528 (2)		34.2	0.0037	84.9	58.9	25.7	73.8	39.6	16.6	59.1	82.4
1.3		0.1109 (3)		29.0	0.0031	90.2	66.0	28.1	78.6	48.0	20.4	61.2	87.5
1.4		0.01814 (5)		28.4	0.0032	89.0	67.9	34.0	84.8	52.4	21.8	62.3	88.7
(*d*) Results for the 5164 noncentrosymmetric structures of subset *S*220 defined in Table 6(*a*)													
1.1		0.1847 (12)		38.1	0.0041	81.6	46.8	18.4	66.6	35.3	14.0	41.2	71.7
1.2		0.00828 (6)		40.7	0.0045	77.0	47.3	18.1	65.5	29.7	11.5	41.8	77.4
1.3		0.1449 (7)		28.3	0.0031	90.1	62.9	25.9	80.0	48.7	20.5	57.4	85.8
1.4		0.02537 (13)		29.0	0.0033	88.2	64.0	29.9	83.4	50.9	20.4	60.1	86.2
(*e*) Results for 371 centrosymmetric structures from subset *S*220 (Table 6*a*) for which N_r and N_p values are available in data set 3													
1.4		0.0194 (4)		27.4	0.0028	92.2	74.4	38.5	84.9	55.0	20.0	69.3	90.6
4.4		0.0539 (9)		25.0	0.0025	94.1	75.2	41.0	89.0	59.6	24.0	70.3	93.0
5.4		0.923 (16)		27.2	0.0027	92.5	73.9	35.6	85.4	54.7	20.2	69.8	92.7
6.4		0.854 (15)		26.9	0.0027	92.5	75.2	36.1	86.0	55.5	18.9	70.1	92.7
7.4		0.0138 (3)		26.0	0.0028	92.5	74.1	46.1	90.8	58.0	27.5	70.1	92.7
8.4		0.616 (13)		27.5	0.0032	90.6	75.2	43.9	89.8	58.2	26.2	70.2	92.5
9.4		0.567 (13)		27.6	0.0032	90.0	75.5	43.7	89.0	57.7	28.3	69.8	92.7
(*f*) Results for 184 noncentrosymmetric structures from subset *S*220 (Table 6*a*) for which N_r ands N_p values are available in data set 3													
1.4		0.0253 (7)		28.3	0.0030	92.4	64.7	31.0	83.7	48.9	20.1	67.5	91.9
4.4		0.0592 (16)		27.8	0.0030	92.4	70.1	35.3	87.0	55.4	25.0	69.7	92.4
5.4		1.083 (20)		20.8	0.0021	96.2	79.9	43.5	89.1	66.9	26.6	72.1	96.2
6.4		0.823 (28)		24.9	0.0036	94.6	76.1	46.2	94.0	64.7	34.2	71.7	94.0
7.4		0.0220 (6)		26.0	0.0028	94.0	71.2	37.0	88.0	53.8	23.9	69.9	92.9
8.4		0.897 (19)		21.2	0.0024	95.6	80.4	47.3	92.4	65.8	32.6	70.8	94.5
9.4		0.692 (24)		26.6	0.0036	91.9	73.9	45.1	94.6	59.2	27.2	70.5	94.0

median at *circa* 0.001 Å. Numerical errors are larger for larger $\overline{\sigma}_o$'s. The distribution of percentage errors is, however, heavily skewed towards negative errors and shows a steady increase of $\langle\overline{\sigma}_o\rangle$ from ~0.005 to ~0.015 Å and a median value of *circa* 20%. This is to be expected: a small numerical error in a small $\overline{\sigma}$ value will, of course, generate a relatively large percentage error. However, the negative skewing towards lower $\overline{\sigma}_o$ values implies that $\overline{\sigma}_p > \overline{\sigma}_o$ in these cases and could be taken as an indication that some of the lower $\overline{\sigma}_o$ values may be underestimated in the last-squares process, a problem discussed elsewhere by Taylor & Kennard (1986).

Mean isotropic e.s.d.'s of C *atoms*

The regression analysis has so far concentrated on the prediction of $\overline{\sigma}$(C–C) as the mean e.s.d. of a C–C bond in a structure of known elemental constitution and known *R* factor. This permits a direct link between the regression results and the existing AS flags in the CSD. Nevertheless, (11) may also be used to predict a mean isotropic e.s.d. for a C atom $\overline{\sigma}$(C) by use of (7). Thus,

$$\overline{\sigma}(\mathrm{C}) = 0.01814RN_c^{1/2}/2^{1/2} = 0.01283RN_c^{1/3}$$

for centrosymmetric structures (12*a*)

or

$$\overline{\sigma}(\mathrm{C}) = 0.02537RN_c^{1/2}/2^{1/2} = 0.01794RN_c^{1/2}$$

for noncentrosymmetric structures. (12*b*)

However, since this is an empirical study, it may be more appropriate to use

$$\overline{\sigma}(\mathrm{C}) = 0.01814RN_c^{1/2}/1.377 = 0.01317RN_c^{1/2}$$

for centrosymmetric structures (13*a*)

or

$$\overline{\sigma}(\mathrm{C}) = 0.02537RN_c^{1/2}/1.377 = 0.01842RN_c^{1/2}$$

for noncentrosymmetric structures, (13*b*)

where the denominator is the mean value of $\overline{\sigma}$(C–C)/$\overline{\sigma}$(C) determined from our calculations for both subsets *S*140 and *S*220 (see Table 6). The proximity of this ratio (1.377) to the value of 1.414 expected for perfectly spherical error distributions, $\overline{\sigma}$(C), confirms that errors in C-atom positions are nearly isotropic over a very large sample of structures, a fact noted by Taylor & Kennard (1986) in an analysis of atomic e.s.d.'s for a much smaller sample of 200 structures.

The validity of (13) was checked by a regression analysis of the form $\overline{\sigma}(\mathrm{C}) = kRN_c^{1/2}$ for the centrosymmetric and noncentrosymmetric subdivisions of *S*220. This yielded $k = 0.01320$ (4) and values of R_σ and r.m.s. (σ) of 27.8% and 0.023 Å for centrosymmetric structures. The *k* value for noncentrosymmetric structures was 0.01871 (10) with R_σ and r.m.s. (σ) being 28.7% and 0.0026 Å, respectively. Values of N_{50}, N_{25} and N_{10} (Table 7) are 95.4, 79.9 and 45.8% (94.6, 76.5 and 40.0%), respectively, for centrosymmetric (noncentrosymmetric) structures. The composite error distribution (Fig. 7) is near normal and the composite percentage error distribu-

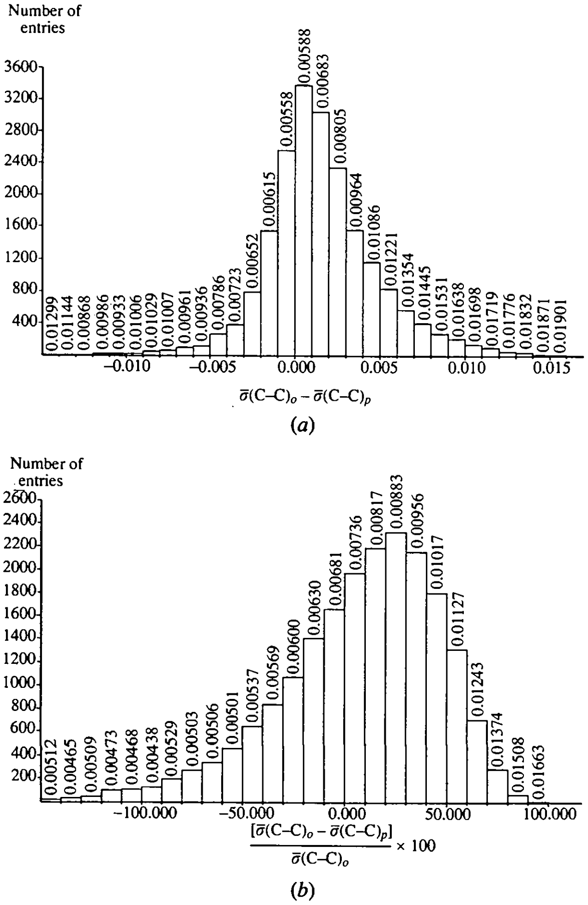

Fig. 6. Distributions of (*a*) $\overline{\sigma}(\mathrm{C{-}C})_o - \overline{\sigma}(\mathrm{C{-}C})_p$, the real numerical difference between observed and predicted $\overline{\sigma}$(C–C) values, and (*b*) the percentage difference $100[\overline{\sigma}(\mathrm{C{-}C})_o - \overline{\sigma}(\mathrm{C{-}C})_p]/\overline{\sigma}(\mathrm{C{-}C})_o$. Predicted values were calculated using (11) for subset *S*220 of Table 6. Values of the mean $\langle\overline{\sigma}(\mathrm{C{-}C})_o\rangle$ are shown for each bar of each distribution.

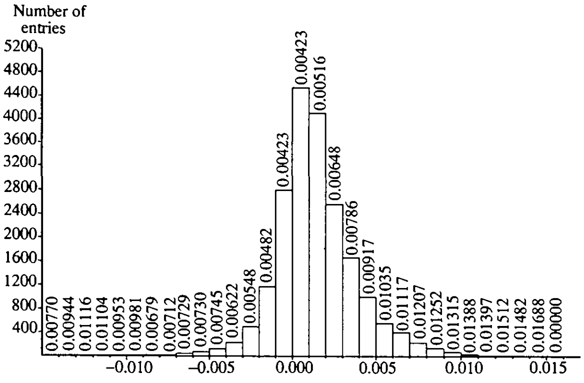

Fig. 7. The distribution of real numerical differences $\overline{\sigma}(\mathrm{C})_o - \overline{\sigma}(\mathrm{C})_p$ derived using (13) for subset *S*220. Values of the mean $\langle\overline{\sigma}(\mathrm{C})_o\rangle$ are shown for each bar of the distribution.

tion (not shown) is almost identical in form to Fig. 6(*b*). The ratio of the regression constants k_{nc}/k_c is 1.417, almost identical to the expected value of 1.414 (Cruickshank, 1960).

*E.s.d.'s of non-*C *atoms*

For any particular structure determination, the quantities R, $\bar{s}$ and p in Cruickshank's (1960) original expression [(2)] are constants. Hence, within that structure, the mean isotropic e.s.d.'s $\sigma(A)$ and $\sigma(B)$ of two different elements A and B would be related according to

$$\sigma^2(A)/\sigma^2(B) = N_A/N_B. \qquad (14)$$

Simple manipulation using (3), in which the $\sum Z_i^2$ term is also a constant for a specific structure, then yields.

$$\sigma(A)/\sigma(B) = Z_B/Z_A, \qquad (15)$$

i.e. an inverse ratio of atomic numbers. Thus, we may express $\overline{\sigma}(E)$, the isotropic e.s.d. of the heaviest atom(s) E, in terms of the isotropic e.s.d. of a C atom, $\overline{\sigma}(\mathrm{C})$,

$$\overline{\sigma}(E) = 6\overline{\sigma}(\mathrm{C})/Z_E, \qquad (16)$$

where Z_E is the atomic number of E.

The validity of (16) is explored in Table 9. Here, we have used values of $\overline{\sigma}(E)$, the mean e.s.d. of the heaviest (non-carbon) element (E) in any structure, and the corresponding $\overline{\sigma}(\mathrm{C})$ value, as calculated for each CSD entry in subsets *S*140 and *S*220 of Table 6. These 'observed' values were then averaged over ranges of Z_E, the atomic number of the heaviest element, and the mean atomic number, $\langle Z_E\rangle$, was calculated for each range. This binning procedure generates seven 'representative structures', each with a unique $\langle\overline{\sigma}(\mathrm{C})\rangle$ and $\langle Z_E\rangle$ value that can be used in (16) to obtain $\langle\overline{\sigma}(E)_p\rangle$, a predicted value for the e.s.d. of an element of atomic number Z_E. Ideally, $\langle\overline{\sigma}(E)_p\rangle$ should equal $\langle\overline{\sigma}(E)_o\rangle$ and their graph should be a straight line passing through the origin and having a gradient K [$=\langle\overline{\sigma}(E)_o\rangle/\langle\overline{\sigma}(E)_p\rangle$] of unity.

The results in Table 9 show values of K that are close to unity for the lowest Z_E ranges but that tend towards lower values (*circa* 0.75–0.80) when heavier atoms are present. This is entirely reasonable since our estimate of a unitary value for K involves the term Z_E/Z_{C} arising from our estimate [(4)] of the N_A term in Cruickshank's (1960) original equation (2). However, the correct expression for N_A [(3)] involves scattering factors ($\bar{f}_A$) at the $\bar{s}$ appropriate for each structure. It has been pointed out to us (Cruickshank, 1993) that, since scattering factors for heavy atoms drop off more slowly than for carbon (and heavy atoms usually have smaller displacement parameters than C atoms in the same structure), then our estimate of N_c [from (4)] will be somewhat smaller than that given by (3). Further, it is the ratio $\bar{f}_e/\bar{f}_c$ that should be involved in generating a unitary value for K in Table 9 and this, then, is likely to be greater than the ratio Z_E/Z_{C} used by us. This in turn leads to K values that are systematically less than unity for the higher Z_E ranges.

Table 9. *Values of* $\langle\overline{\sigma}(E)\rangle_o$ *and* $\langle\overline{\sigma}(\mathrm{C})\rangle_o$, *the observed isotropic e.s.d.'s of a non-*C *atom* (E) *and a* C *atom* (C) *in* Å, *averaged over ranges of* Z_E, *the atomic number of* E

N_{ent} is the number of entries in each range and $\langle Z_E\rangle$ is the mean value for the range. The quantity $\overline{\sigma}(E)_p$ is calculated using (16) with the appropriate $\langle Z_E\rangle$ as denominator. K is the ratio $\langle\overline{\sigma}(E)\rangle_o/\overline{\sigma}(E)_p$. Calculations were carried out for subsets *S*140 and *S*220 of Table 6.

Z_E range	N_{ent}	$\langle\overline{\sigma}(E)_o\rangle$	$\langle\overline{\sigma}(\mathrm{C})\rangle_o$	$\langle Z_E\rangle$	$\langle\overline{\sigma}(E)_p\rangle$	K
(*a*) Subset *S*140						
7–9	7692	0.00332	0.00464	7.88	0.00353	0.941
10–19	4586	0.00190	0.00542	16.00	0.00203	0.936
20–29	3795	0.00117	0.00785	26.34	0.00179	0.654
30–39	1622	0.00126	0.00948	34.04	0.00167	0.755
40–49	2958	0.00088	0.00897	43.67	0.00123	0.715
50–69	1350	0.00088	0.01050	53.49	0.00118	0.746
≥ 70	2979	0.00077	0.01338	77.04	0.00104	0.740
(*b*) Subset *S*220						
7–9	6051	0.00306	0.00427	7.89	0.00325	0.942
10–19	3786	0.00143	0.00472	16.02	0.00177	0.808
20–29	3088	0.00098	0.00623	26.33	0.00142	0.690
30–39	1243	0.00099	0.00727	34.01	0.00128	0.773
40–49	2425	0.00076	0.00710	43.63	0.00098	0.776
50–69	1019	0.00072	0.00787	53.20	0.00089	0.809
≥70	1856	0.00063	0.00946	76.79	0.00074	0.851

In the event, the plots of $\langle\overline{\sigma}(E)_o\rangle$ *versus* $\langle\overline{\sigma}(E)_p\rangle$ for data sets *S*140 and *S*220 (Fig. 8) are almost collinear and pass very close to the origin. In view of the results of Table 9 and Fig. 8, we performed two further type 1 regressions using data set *S*220 with

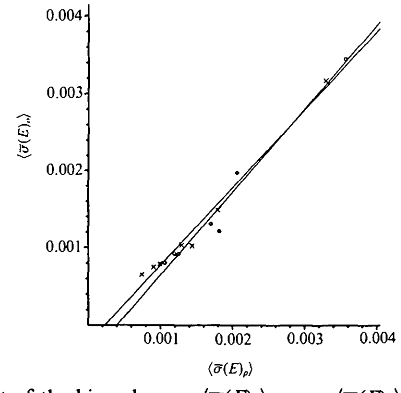

Fig. 8. Plot of the binned mean $\langle\overline{\sigma}(E)_o\rangle$ *versus* $\langle\overline{\sigma}(E)_p\rangle$ for subsets *S*140 (denoted by circles) and *S*220 (denoted by crosses). Binning procedures are described in the text.

$\overline{\sigma}(E)$ as the dependent variable and with the independent variable as (i) $\overline{\sigma}(\mathrm{C})/Z_E$ [from (16)] and (ii) $RN_c^{1/2}/Z_E$ [on the basis of (13)]. For (i), we do not need to form centrosymmetric and noncentrosymmetric divisions of the data set and, in the light of the discussion above, we would expect the slope of the simple regression line to be less than its 'ideal' value of 6.0. For (ii), subdivision is essential and we might expect [(13)] the slopes of the regression lines to approach (but be less than) $6 \times 0.01317 = 0.079$ for centrosymmetric structures and $6 \times 0.01842 = 0.1105$ for noncentrosymmetric structures. Regression results for *S*220 with an imposed zero intercept gave

$$\text{(i)} \qquad \overline{\sigma}(E) = 5.203\overline{\sigma}(\mathrm{C})/Z_E \qquad (17)$$

$$\text{(ii)} \qquad \overline{\sigma}(E) = 0.0678RN_c^{1/2}/Z_E$$

for centrosymmetric structures (18*a*)

or

$$\overline{\sigma}(E) = 0.1006RN_c^{1/2}/Z_E$$

for noncentrosymmetric structures. (18*b*)

Values of R_σ and r.m.s. (σ) from (17) were 28.7% and 0.00088 Å and from (18*a*) [(18*b*)] they were 40.8% and 0.0010 Å [34.9% and 0.0014 Å]. The k_{nc}/k_c ratio here is 1.484.

These regressions confirm the validity of (16) and provide useful predictions of $\overline{\sigma}(E)$ when $\overline{\sigma}(\mathrm{C})$ is already known [(17)] or, more valuably with respect to CSD analyses, when only the *R* factor and atomic constitution of a structure are known [(18)]. We note, however, that the $\overline{\sigma}(E)$ values used here refer only to the heaviest element in any structure. Nevertheless, an expression having the form of (18) is general and we explore this generality for all non-H atoms in subsets of the CSD in paper II (Allen, Cole & Howard, 1995).

Analysis of data set 3

The 817 entries of data set 3 represent a small subset of the CSD for which some information items concerning the structure-refinement process, in the form of N_r (number of reflections) and N_p (number of refined parameters), have been manually edited into the work file for this project. Despite the fact that the quantity $p = N_r - N_p$ occurs in the denominator of the expression used to calculate parameter e.s.d.'s following least-squares refinement and also occurs in the denominator of (2) (Cruickshank, 1960), crystallographers habitually use the ratio N_r/N_p as a rapid guide to the likely 'quality' of a given structure. Hence, for crystallographic interest, we begin this section with a brief comparative analysis of structures in terms of this N_r/N_p ratio. In Table 10, we have divided the full N_r/N_p range into eight unitary bins, together with two extra bins covering values <4 and ≥ 12, respectively. For each bin, we tabulate the mean values $\langle N_r/N_p\rangle$, $\langle\overline{\sigma}(\mathrm{C–C})\rangle$, $\langle R\rangle$, $\langle Z_{\max}\rangle$ and $\langle RN_c^{1/2}\rangle$.

The overall results for all 817 entries (Table 10*a*) show that $\langle\overline{\sigma}(\mathrm{C–C})\rangle$ decreases quite rapidly over the first five bins, despite the fact that $\langle Z_{\max}\rangle$ (and hence $\langle RN_c^{1/2}\rangle$) increases steadily as $\langle N_r/N_p\rangle$ increases, but $\langle R\rangle$ is essentially constant. Obviously, those structures that contain a heavier element tend to diffract well and exhibit higher values of N_r/N_p. Correlation coefficients linking N_r/N_p with $Z_{\max}$ are 0.346 for subset *S*140 and 0.374 for subset *S*220. Nevertheless, the increasing N_r/N_p ratio, particularly in the lower half of the complete range, more than compensates for the increasing $\langle Z_{\max}\rangle$ so that $\langle\overline{\sigma}(\mathrm{C–C})\rangle$ actually falls rather than rising with $\langle Z_{\max}\rangle$.

In order to examine the real effect of increasing N_r/N_p, we have normalized the mean $\langle\overline{\sigma}(\mathrm{C–C})\rangle$ values to the $\langle RN_c^{1/2}\rangle$ value of the first bin, using the already established proportionality between $\overline{\sigma}(\mathrm{C–C})$ and $RN_c^{1/2}$; thus,

$$\langle\overline{\sigma}(\mathrm{C–C})\rangle_n = \langle\overline{\sigma}(\mathrm{C–C})\rangle\langle RN_c^{1/2}\rangle_n/\langle RN_c^{1/2}\rangle, \qquad (19)$$

where $\langle\overline{\sigma}(\mathrm{C–C})\rangle_n$ is the normalized mean value of a $\langle\overline{\sigma}(\mathrm{C–C})\rangle$ associated with a specific $\langle RN_c^{1/2}\rangle$ and $\langle RN_c^{1/2}\rangle_n$ is taken here as 0.31955. The downward trend in $\langle\overline{\sigma}(\mathrm{C–C})\rangle_n$ (Table 10*a*, Fig. 9) now extends to $N_r/N_p \simeq 9–10$ with the most dramatic section of the curve occurring for $N_r/N_p < 6–7$. These results confirm the typical experiential knowledge of small-molecule crystallographers.

Tables 10(*b*) and (*c*) divide the overall Table 10(*a*) into two subgroups: (*b*) 'organic' structures having $Z_{\max} \leq 18$ and (*c*) 'metallo-organic' structure having $Z_{\max} > 18$. In Table 10(*b*), the trends in $\langle\overline{\sigma}(\mathrm{C–C})\rangle$ and $\langle\overline{\sigma}(\mathrm{C–C})\rangle_n$ are similar and are directly comparable with those for the normalized $\langle\overline{\sigma}(\mathrm{C–C})\rangle_n$ for the

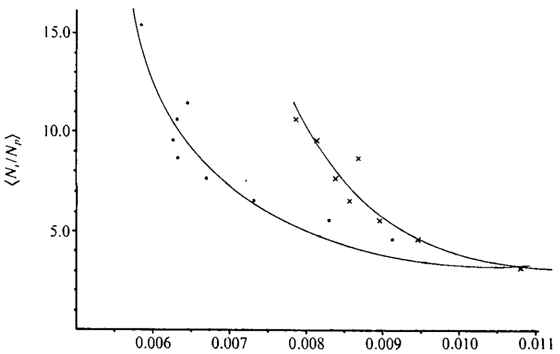

Fig. 9. Plot of binned values of $\langle\overline{\sigma}(\mathrm{C–C})\rangle$ (denoted by crosses) and of $\langle\overline{\sigma}(\mathrm{C–C})\rangle_n$ (denoted by filled circles) *versus* $\langle N_r/N_p\rangle$ for data set 3. Binning and normalization procedures are described in the text.

Table 10. *Analysis of structural precision for ranges of N_r/N_p (see Table 3 for definitions)*

For each range, mean values are cited for $\langle N_r/N_p\rangle$, $\langle\bar{\sigma}\text{C–C)}\rangle$, $\langle R\rangle$, $\langle Z_{\max}\rangle$ and $\langle RN_c^{1/2}\rangle$, where parameters are defined in Table 3. N_{ent} is the number of CSD entries in each range and the normalized quantity $\langle\bar{\sigma}(\text{C–C})\rangle_n$ is described in the text.

N_r/N_p range	$\langle N_r/N_p\rangle$	$\langle\bar{\sigma}(\text{C–C})\rangle$	$\langle R\rangle$	$\langle Z_{\max}\rangle$	$\langle RN_c^{1/2}\rangle$	N_{ent}	$\langle\bar{\sigma}(\text{C–C})\rangle_n$
(*a*) For all entries in data set 3							
0–4	3.16	0.01076	0.0488	10.29	0.319	31	0.01076
4–5	4.58	0.00943	0.0497	14.15	0.331	77	0.00910
5–6	5.52	0.00893	0.0487	16.13	0.345	105	0.00827
6–7	6.50	0.00854	0.0480	20.30	0.373	103	0.00730
7–8	7.53	0.00836	0.0488	22.50	0.399	103	0.00669
8–9	8.56	0.00866	0.0477	28.48	0.438	98	0.00631
9–10	9.47	0.00811	0.0464	28.72	0.414	76	0.00625
10–11	10.52	0.00784	0.0430	32.71	0.397	63	0.00630
11–12	11.34	0.00925	0.0440	36.30	0.459	53	0.00644
≥ 12	15.28	0.00989	0.0438	45.68	0.541	108	0.00584
(*b*) For entries with $Z_{\max} \le 18$							
0–4	3.15	0.01066	0.0487	8.89	0.294	29	0.01155
4–5	4.56	0.00829	0.0499	10.12	0.291	64	0.00910
5–6	5.49	0.00691	0.0489	9.94	0.287	79	0.00768
6–7	6.48	0.00602	0.0492	11.01	0.300	69	0.00641
7–8	7.56	0.00516	0.0510	10.68	0.295	66	0.00557
8–9	8.59	0.00479	0.0488	11.94	0.291	50	0.00526
9–10	9.42	0.00440	0.0481	12.57	0.284	35	0.00495
10–11	10.48	0.00407	0.0463	13.15	0.267	26	0.00486
11–12	11.37	0.00435	0.0484	11.17	0.277	17	0.00501
≥ 12	14.22	0.00501	0.0509	12.64	0.300	25	0.00533
(*c*) For entries with $Z_{\max} > 18$							
0–4	3.32	0.01216	0.0500	30.50	0.675	2	0.00575
4–5	4.68	0.01504	0.0486	34.00	0.528	13	0.00910
5–6	5.61	0.01507	0.0481	34.92	0.520	26	0.00925
6–7	6.54	0.01364	0.0457	39.14	0.522	34	0.00834
7–8	7.49	0.01408	0.0448	43.59	0.583	37	0.00771
8–9	8.53	0.01268	0.0465	45.72	0.592	48	0.00684
9–10	9.51	0.01129	0.0450	42.51	0.525	41	0.00686
10–11	10.55	0.01048	0.0406	46.45	0.488	37	0.00686
11–12	11.33	0.01157	0.0419	48.16	0.545	36	0.00678
≥ 12	15.60	0.01135	0.0417	55.63	0.613	83	0.00591

complete data set (Table 10*a*), since $\langle Z_{\max}\rangle$ and $\langle RN_c^{1/2}\rangle$ now show a very small range. By contrast, the heavy-atom structures of Table 10(*c*) show an approximate, and rather misleading, constancy of $\langle\bar{\sigma}(\text{C–C})\rangle$ over the N_r/N_p range. The true effects cf increasing this ratio are only revealed in the normalized $\langle\bar{\sigma}(\text{C–C})\rangle_n$ values. A comparison of the N_{ent} values of Tables 10(*a*) and (*c*) show that only 26.7% of entries having $N_r/N_p \le 7$ are metallo-organic, but 62.8% of entries with $N_r/N_p > 7$ are of this class, despite the fact that less than half (43.7%) of entries in the total data set have $Z_{\max} > 18$. The imbalance is clear, for reasons already stated.

For the regression experiments on data set 3, we employ the general form of (2) and test its predictive ability when $(N_r/N_p)^{1/2}$, $N_r^{1/2}$ or $(N_r - N_p)^{1/2}$ is included as a denominator. Further, a method has been suggested to us (Cruickshank, 1993) for obtaining a simple estimate of $\bar{s}$ (the r.m.s. reciprocal radius) from the unit-cell volume, the space group and N_r. This derivation is included as an Appendix to this paper. All regressions were based on 555 entries from subset *S*220 (Table 6*a*) for which N_r and N_p values were available in data set 3. Centrosymmetric structures (371) and noncentrosymmetric structures (184) were treated separately. Seven different variations of (2) were used, denoted as regression types 1.4, 4.4–9.4 in Tables 8(*e*) and (*f*), *viz*:

$$\text{type 1.4} \quad \bar{\sigma}(\text{C–C}) = kRN_c^{1/2} \tag{20}$$

$$\text{type 4.4} \quad \bar{\sigma}(\text{C–C}) = kRN_c^{1/2}/(N_r/N_p)^{1/2} \tag{21}$$

$$\text{type 5.4} \quad \bar{\sigma}(\text{C–C}) = kRN_c^{1/2}/(N_r)^{1/2} \tag{22}$$

$$\text{type 6.4} \quad \bar{\sigma}(\text{C–C}) = kRN_c^{1/2}/(N_r - N_p)^{1/2} \tag{23}$$

$$\text{type 7.4} \quad \bar{\sigma}(\text{C–C}) = kRN_c^{1/2}/\bar{s} \tag{24}$$

$$\text{type 8.4} \quad \bar{\sigma}(\text{C–C}) = kRN_c^{1/2}/\bar{s}(N_r)^{1/2} \tag{25}$$

$$\text{type 9.4} \quad \bar{\sigma}(\text{C–C}) = kRN_c^{1/2}/\bar{s}(N_r - N_p)^{1/2}. \tag{26}$$

The type 1.4 regressions [(20)] that were used to analyse the larger data set 2 are included here to provide a benchmark against which to compare the predictive abilities of (21)–(26). Equations (23), (24) and (26) represent progressive stages of the Cruickshank (1960) formulation.

Regression results for centrosymmetric and noncentrosymmetric subsets are given in Tables 8(*e*) and (*f*), respectively. The 'benchmark' results from the type 1.4 regressions yield similar values of k, R_σ and r.m.s. (σ) to those given by the large data set 2 and a ratio of k_{nc}/k_c of 1.304 (1.399 for data set 2). Results from the other six regressions are inconsistent although, in general, inclusion of some function involving N_r appears to yield small improvements in the assessment criteria. Surprisingly, (21) with $(N_r/N_p)^{1/2}$ as denominator generates the best overall set of criteria for centrosymmetric structures, although the best overall values of n_{50}, n_{25} and n_{10} are generated by (26), which has the functional form of (2). For the noncentrosymmetric structures, the best overall criteria are given by (22) and (25), which both involve $(N_r)^{1/2}$ in the denominator. We also note that the ratio k_{nc}/k_c for (21), (22), (23), (24), (25) and (26) at 1.098, 1.173, 0.964, 1.594, 1.456 and 1.220 shows wide variations one with another and from the expected value of 1.414.

We are left, then, with an apparent anomaly: why does our best available approximation to the complete Cruickshank (1960) equation [(26) above], involving four variable parameters, give such a small improvement in predictive power by comparison with the much simpler equation [(21) above] which involves only two parameters? Firstly, we accept that the numbers of structures used in the regressions of Tables 8(*e*) and (*f*) are rather small. Secondly, $\bar{s}$ as given in the Appendix is only an estimate and, for this study, we have not modified N_c so that it conforms to (3).

Despite these caveats, we have briefly explored the statistics of the $\bar{s}$ and $(N_r - N_p)^{1/2}$ distributions and also the intercorrelations between $\bar{s}$ and functions involving N_r and N_p with $\overline{\sigma}$(C–C), $N_c^{1/2}$ and R. The $\bar{s}$ distribution is almost normal and ranges from 0.48 to 1.37 Å^{-1}. However, variation in $\bar{s}$ is small, with 88% of values in the range 0.6–1.0 Å^{-1} and 58% between 0.7 and 0.9 Å^{-1}. The $N_r - N_p$ range is much broader: from 241 to 9896, corresponding to $(N_r - N_p)^{1/2}$ in the range 15.5–99.5, although the effective range is rather smaller at *circa* 26.0–75.0.

It is the correlation results of Table 11 that are, perhaps, the most revealing. The $C_{i,j}$ values for $i = R$ are all very low, indicating that R is essentially independent of $\bar{s}$ and $f(N_r,N_p)$. However, for $i = N_c^{1/2}$, the $C_{i,j}$ are appreciable with values for $j = N_r^{1/2}$ and $(N_r - N_p)^{1/2}$ being as high as 0.650. This, of course, is a confirmation of the deductions made from Table 10: that structures having increasing proportions of heavier atoms are likely to generate higher numbers of observed reflections. The expected appreciable inverse correlations of $\overline{\sigma}$(C–C) with $f(N_r - N_p)$ are not observed and, indeed, three of these $C_{i,j}$ show small positive values. Only $C[\overline{\sigma}\text{(C–C)},\bar{s}]$ approaches the behaviour that might have been expected. It would appear that the very small range of $\bar{s}$ and the nonindependence of $(N_r - N_p)^{1/2}$ for this data set are contributory factors to the apparently anomalous results of Tables 8(*e*) and (*f*).

Table 11. *Correlation coefficients $C_{i,j}$ derived from the 555 entries of subset S220 for which N_r and N_p values were available in data set* 3

Names of data items are defined in the text or in Table 3.

		Item *i*	
Item *j*	$\overline{\sigma}$(C–C)	R	$N_c^{1/2}$
N_r/N_p	−0.141	−0.059	0.273
$(N_r/N_p)^{1/2}$	−0.147	−0.055	0.270
N_r	0.102	−0.031	0.364
$N_r^{1/2}$	0.108	−0.054	0.650
$(N_r - N_p)^{1/2}$	0.086	−0.060	0.644
$\bar{s}$	−0.334	−0.114	−0.219
$\bar{s}(N_r - N_p)^{1/2}$	−0.079	−0.098	0.437

Concluding remarks

This study of structural precision has, of necessity, been restricted by the information content of the CSD, particularly in relation to experimental details of the structure-determination process. Nevertheless, it has proved possible to generate the empirical predictive equations that quantify $\overline{\sigma}$(C–C), $\overline{\sigma}$(C) and $\overline{\sigma}(E)$ in terms of the crystallographic R factor and the atomic constitution of the structure under study, using the work of Cruickshank (1960) as the theoretical basis. Specifically, we recommend the use of (11*a*) and (11*b*) [for $\overline{\sigma}\text{(C–C)}_p$], (13*a*) and (13*b*) [for $\overline{\sigma}\text{(C)}_p$] and (18*a*) and (18*b*) [for $\overline{\sigma}(E)_p$], where the *a* and *b* forms should be used for centrosymmetric and noncentrosymmetric structures, respectively. These predictions are shown to provide an acceptable estimate of the AS flag for ~86% of the 13000 structures that lack that information in the CSD. It also appears possible to predict the real $\overline{\sigma}$(C–C) and $\overline{\sigma}$(C) values to within ±0.005 Å in a similar proportion (87 and 95%) of those *circa* 40000 structures that currently lack e.s.d. information in the CSD, and to know that these estimates are within ±0.0025 Å in 65% and 78% of cases.

There remains, of course, the problem of those 13–14% of structures for which the predicted values are less valid. In the absence of further experimental details (through which we might, in any case, have generated better predictive equations), it remains difficult to identify exactly which structures might be so affected. Here, we would have to rely on the currently encoded AS flags, where available, to aid this identification. We would hope that, through a careful analysis of AS flags and predicted $\bar{\sigma}$(C–C) values, it would be possible to incorporate improved precision indicators within the CSD that have known reliability for some 75–80% of database content. Such indicators should, for example, form a suitable basis for the generation of semi-weighted mean values of geometric parameters in the manner suggested by Taylor & Kennard (1983).

This study has also confirmed a number of pieces of experiential crystallographic knowledge, *e.g.* (i) that, within a given structure, the e.s.d.'s of different elements are approximately related by the inverse ratio of their atomic numbers, (ii) that N_r/N_p ratios below *circa* 6.0 lead to rapid increases in atomic e.s.d.'s and (iii) that structures containing heavier elements are more likely to generate diffraction data for which $N_r/N_p \geq 6.0$.

Finally, we note that the predictive equations derived in this work have positive implications for CSD data-processing activities. Specifically, the predictions can be compared with values calculated from input information so as to detect gross errors in the initial keyboarding operations and/or in the published e.s.d. data contained in journals or in deposition documents.

We are greatly indebted to Professor Durward Cruickshank FRS for his detailed comments and discussion of early versions of this manuscript. We thank Drs Olga Kennard FRS, Robin Taylor and Michael Doyle for their interest in this work and for helpful suggestions. The Science and Engineering Research Council (UK) is thanked for financial support to JCC.

APPENDIX

We are indebted to Cruickshank (1993) for suggesting this simple method of estimating $\bar{s}$ from quantities available in the work file for data set 3. In this derivation, V_c is the unit-cell volume, V_c^* $(=1/V_c)$ is the reciprocal-cell volume, $s_{\max}$ $(=2\sin\theta/\lambda)$ is the radius of the sphere of observations in reciprocal space, V_s^* is the volume of this sphere, N_r is the number of reflections and m^* is multiplicity of a general reflection. If we suppose that there are no unobserved reflections within the limiting sphere and we ignore the effect of principal zones where reflection multiplicities may be less than m^*, then we can estimate $s_{\max}$ *via*

$$N_r = V_s^*/m^* V_c^* = (V_c/m^*)(4\pi s^3{}_{\max}/3), \qquad (27)$$

$$s_{\max} = (3m^* N_r/4\pi V_c)^{1/3}. \qquad (28)$$

For a solid sphere, we may calculate the mean square radius $\overline{s^2}$ *via*

$$\bar{s}^2 = \int_0^{s_{\max}} s^2 4\pi s^2 \mathrm{d}s \Big/ \int_0^{s_{\max}} 4\pi s^2 \mathrm{d}s, \qquad (29)$$

$$\bar{s}^2 = 1/5(4\pi s^5_{\max})/[1/3(4\pi s^3_{\max})] = 3/5(s^2_{\max}). \qquad (30)$$

The r.m.s. reciprocal radius $\bar{s}$ is then given by

$$\bar{s} = (\bar{s}^2)^{1/2} = (3/5)^{1/2} s_{\max}. \qquad (31)$$

Hence, we can estimate $\bar{s}$ as

$$\bar{s} = (3/5)^{1/2}(3m^* N_r/4\pi V_c)^{1/3}. \qquad (32)$$

References

Allen, F. H., Bergerhoff, G. & Sievers, R. (1987). Editors. *Crystallographic Databases*, IUCr/Cambridge Univ. Press.

Allen, F. H., Cole, J. C. & Howard, J. A. K. (1995). *Acta Cryst.* **A51**, 112–121.

Allen, F. H., Davies, J. E., Galloy, J. J., Johnson, O., Kennard, O., Macrae, C. F., Mitchell, E. M., Mitchell, G. F., Smith, J. M. & Watson, D. G. (1991). *J. Chem. Inf. Comput. Sci.* **31**, 187–204.

Allen, F. H. & Doyle, M. J. (1987). *Acta Cryst.* **A43**, C291.

Allen, F. H., Kennard, O. & Watson, D. G. (1995). In preparation.

Cambridge Structural Database User's Manual (1992). Version 5.1. Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, England.

Cruickshank, D. W. J. (1960). *Acta Cryst.* **13**, 774–777.

Cruickshank, D. W. J. (1993). Private communication.

Dunitz, J. D. (1979). *X-ray Analysis and the Structure of Organic Molecules.* Ithaca, NY: Cornell Univ. Press.

International Tables for X-ray Crystallography (1960). Vol. I. Birmingham: Kynoch Press.

Muir, K. W. & Mallinson, P. R. (1993). *J. Appl. Cryst.* **26**, 142–143.

Numerical Algorithms Group (1990). Subroutine Library. Numerical Algortithms Group Ltd, Wilkinson House, Jordan Hill Road, Oxford OX2 5DR, England.

Pauling L. (1940). *The Nature of the Chemical Bond.* Ithaca, NY: Cornell Univ. Press.

Pimentel, G. C. & McClellan, A. L. (1960). *The Hydrogen Bond.* San Francisco: W. H. Freeman.

Sutton, L. E. (1958). Editor. *Tables for Interatomic Distances and Configuration in Molecules and Ions*, Special Publication No. 11. London: The Chemical Society.

Sutton, L. E. (1965). Editor. *Tables for Interatomic Distances and Configuration in Molecules and Ions* (*Supplement*), Special Publication No. 18. London: The Chemical Society.

Taylor, R. (1986). *J. Appl. Cryst.* **19**, 90–91.

Taylor, R. & Kennard, O. (1982). *J. Am. Chem. Soc.* **104**, 5063–5070.

Taylor, R. & Kennard, O. (1983). *Acta Cryst.* **B39**, 517–525.

Taylor, R. & Kennard, O. (1986). *Acta Cryst.* **B42**, 112–120.

Templeton, D. H. (1959). *Acta Cryst.* **12**, 771–773.