We are grateful to the Science and Engineering Research Council for financial support.

References

ANSTIS, G. R. & O'KEEFE, M. A. (1976) *Proc. 34th Ann. Meet. EMSA*, edited by G. W. BAILEY, pp. 480–481. Baton Rouge: Claitor.

DONALDSON, D. M. & ROBERTSON, J. M. (1954). *Proc. R. Soc. London Ser. A*, **220**, 157–170.

FRYER, J. R. (1979). *Acta Cryst.* A**35**, 327–332.

FRYER, J. R., CAMPS, R. A. & SMITH, D. J. (1982). *Electron Microscopy* 1982, Vol. 2, pp. 449–450. Frankfurt: Deutsche Gesellschaft für Elektronenmikroskopie e.V.

FRYER, J. R. & SMITH, D. J. (1982). *Proc. R. Soc. London Ser. A*, **381**, 225–240.

GOODMAN, P. & MOODIE, A. F. (1974). *Acta Cryst.* A**30**, 280–290.

International Tables for X-ray Crystallography (1974). Vol. IV. Birmingham: Kynoch Press.

JAP, B. & GLAESER, R. M. (1980). *Acta Cryst.* A**36**, 57–67.

KERR, K. A., ASHMORE, J. P. & SPEAKMAN, J. C. (1975). *Proc. R. Soc. London Ser. A*, **344**, 199–215.

KIRKLAND, E. J. (1982). *Ultramicroscopy*, **9**, 45–64.

LYNCH, D. F., MOODIE, A. F. & O'KEEFE, M. A. (1975). *Acta Cryst.* A**31**, 300–307.

MURATA, Y., FRYER, J. R. & BAIRD, T. (1977). *Acta Cryst.* A**33**, 198–200.

O'KEEFE, M. A. (1973). *Acta Cryst.* A**29**, 389–401.

O'KEEFE, M. A. (1979). *Proc. 37th Ann. Meet. EMSA*, edited by G. W. BAILEY, pp. 556–557. Baton Rouge: Claitor.

O'KEEFE, M. A. & BUSECK, P. R. (1979). *Trans. Am. Crystallogr. Assoc.* **15**, 27–46.

O'KEEFE, M. A., BUSECK, P. R. & IIJIMA, S. (1978). *Nature (London)*, **274**, 322–324.

O'KEEFE, M. A., FRYER, J. R. & SMITH, D. J. (1982). *Electron Microscopy and Analysis* 1981, edited by M. J. GORINGE, pp. 337–340. Bristol and London: Institute of Physics.

O'KEEFE, M. A. & PITT, A. J. (1980). *Electron Microscopy* 1980, Vol. 1. *Physics*, edited by P. BREDEROO & G. BOOM, pp. 122–123: Leiden: Seventh European Congress on Electron Microscopy Foundation.

PIROUZ, P. (1981). *Acta Cryst.* A**37**, 465–471.

ROBERTSON, J. M. & TROTTER, J. (1961). *J. Chem. Soc.* pp. 1280–1284.

ROBERTSON, J. M. & WHITE, J. G. (1945). *J. Chem. Soc.* pp. 607–617.

SAIJO, H., KOBAYASHI, T. & UYEDA, N. (1977). *J. Cryst. Growth.* **40**, 118–124.

SCHMIDT, G. M. J. (1971). *Pure Appl. Chem.* **27**, 647–678.

SELF, P. G., O'KEEFE, M. A., BUSECK, P. R. & SPARGO, A. E. C. (1983). *Ultramicroscopy*. In the press.

SMITH, D. J. (1980). *Electron Microscopy* 1980, Vol. 4. *High Voltage*, edited by J. VAN LANDUYT & G. BOOM, pp. 122–129. Leiden: Seventh European Congress on Electron Microscopy Foundation.

SMITH, D. J., CAMPS, R. A., FREEMAN, L. A., HILL, R., NIXON, W. C. & SMITH, K. C. A. (1983). *J. Microsc.* **130**, 127–136.

SMITH, D. J. & FRYER, J. R. (1981). *Nature (London)*, **291**, 481–482.

SMITH, D. J. & O'KEEFE, M. A. (1983). *Acta Cryst.* A**39**, 139–148.

UNWIN, P. N. T. & HENDERSON, R. (1975). *J. Mol. Biol.* **94**, 425–440.

UYEDA, N. & ISHIZUKA, K. (1974). *J. Electron Microsc.* **23**, 79–88.

VEBLEN, D. R. & BUSECK, P. R. (1980). *Am. Mineral.* **65**, 599–632.

Acta Cryst. (1983). A**39**, 847–853

Moments of the Probability Density Function of R_2 Approached *Via* Conditional Probabilities. III. Models Containing Correctly as well as Incorrectly Placed Atoms in the Space Groups $P1$ and $P\bar{1}$

BY W. K. L. VAN HAVERE AND A. T. H. LENSTRA

University of Antwerp (UIA), Department of Chemistry, Universiteitsplein 1, B-2610 Wilrijk, Belgium

(*Received* 3 *June* 1982: *accepted* 13 *May* 1983)

Abstract

First and second moments of $P(R_2)$ are evaluated for models containing correctly as well as incorrectly placed atoms, denoted symbolically by $\{g, f\}$. Formulas are derived, valid for the space groups $P1$ and $P\bar{1}$, using explicitly the set of observed reflections. Extrapolation through the introduction of an *averaged* structure allows some general conclusions to be drawn concerning possible strategies used in automated structure determinations. A select elimination of data points from the R_2 check on the correctness of an atomic position severely limits the usefulness of the R_2 criterion. A check based on R_2^n has no better characteristics than one based on R_2. An R_2 criterion together with a zero-atom strategy has better chances of being successful than a random-atom approach.

0108-7673/83/060847-07$01.50

1. Introduction

In automated crystal-structure determination one needs to discriminate between correct and incorrect models related to the observed structure. The models to be tested can come out of any traditional solution procedure. In order to describe the various situations

involved we introduce the following nomenclature. A tentative model, containing n atoms ($n \leq N$, the number of atoms in the observed structure) of which g atoms are correctly positioned and f atoms are *badly* misplaced, is denoted by $\{g,f\}$. Obviously, the purpose of a structure determination is to produce a model $\{N,0\}$. We take the simple case of adding in $P1$ one new trial atom to a model $\{g,0\}$. We then arrive either at the situation $\{g+1,0\}$ if the trial atom is correct or at the situation $\{g,1\}$ if it is incorrect.

As discriminator function we use R_2, defined as

$$R_2 \equiv \sum_H (E_o^2 - \eta^2 E_c^2)^2 \Big/ \sum_H E_o^4, \tag{1.1}$$

where E_o represents the magnitude of the normalized structure factor of the observed N-atom structure. Similarly, E_c refers to the tentative n-atom model ($g+f=n$), and η^2 describes the fraction of the scattering power of the model *versus* the total structure. For point atoms with equal scattering power,

$$\eta^2 = \eta_c^2/\eta_o^2 = n/N. \tag{1.2}$$

The decision whether the new model is $\{g+1,0\}$ or $\{g,1\}$ can be made if we can decide whether the R_2 value of the new model belongs to the population $R_2\{g+1,0\}$ or to the population $R_2\{g,1\}$. That is to say, we have to have knowledge about the probability density functions $P(R_2)$ for both situations. The $P(R_2)$'s must be defined over the sample space of all possible (partial or complete, correct or incorrect) models of the structure under investigation. Even if the $P(R_2)$'s are simple Gaussian functions one needs to know their first moment (mean value) and their second moment (spread). For Gaussian distributed $P(R_2)$'s a statistically safe criterion to accept the trial atom as correct would be R_2 (new model) $< A$, see Fig. 1.

Because existing formalisms were unable to predict meaningful $\sigma(R_2)$, a new theory had to be developed. In parts I and II of this series (Van Havere & Lenstra, 1983*a*,*b*) we gave the fundamentals and evidence for correctness of the new theory using the extreme models $\{g,0\}$ and $\{0,f\}$ in space groups $P1$ and $P\bar{1}$ as examples. In this paper we will evaluate the first and second moment of $P(R_2)$ for models of the general type $\{g,f\}$, starting from the equations

$$\langle R_2;\mathscr{E}_o\rangle = 1 + \eta^4 \frac{\sum_H \langle E_c^4;E_o\rangle}{\sum_H E_o^4} - 2\eta^2 \frac{\sum_H E_o^2\langle E_c^2;E_o\rangle}{\sum_H E_o^4} \tag{1.3}$$

and

$$\begin{aligned}\sigma^2(R_2;\mathscr{E}_o) = \Bigg\{ &\sum_H \eta^8(\langle E_c^8;E_o\rangle - \langle E_c^4;E_o\rangle^2) \\ &- \sum_H 4\eta^6 E_o^2(\langle E_c^6;E_o\rangle \\ &- \langle E_c^4;E_o\rangle\langle E_c^2;E_o\rangle) \\ &+ \sum_H 4\eta^4 E_o^4(\langle E_c^4;E_o\rangle \\ &- \langle E_c^2;E_o\rangle^2)\Bigg\} \Big/ \left\{\sum_H E_o^4\right\}^2. \end{aligned} \tag{1.4}$$

The notation $\langle R_2;\mathscr{E}_o\rangle$ means the value of R_2 averaged over all models under the constraint of the set $\mathscr{E}_o$, the set of observed E values. The conditional notation of the moments, *e.g.* $\langle E_c^4;E_o\rangle$, shows unambiguously that the available intensity data are taken as a set of fixed parameters representing a particular structure under investigation. The derivation of the basic intensity distribution $P(E_c;E_o)$ necessary to evaluate (1.3) and (1.4) is given in §2. In §§3 and 4 the general expression is actualized for the space groups $P1$ and $P\bar{1}$ to give *a priori* values of R_2 and $\sigma(R_2)$ for model $\{g,f\}$ specific for an actual structure.

In the last section we will discuss how, through the concept of an *averaged structure*, these results can be further generalized, *i.e.* made independent of a specific structure. Then it becomes possible to evaluate the potential applicability of R_2-based criteria to the screening of a set of *MULTAN* solutions, *e.g.* by estimating the number of correctly placed atoms. Moreover, it allows one to investigate and draw general conclusions about the chances various strategies of structure determination have on being successful. One strategy starts from a complete but possibly incorrect model – *i.e.* the situation $\{0,N\}$ in the most extreme case – and tries by somehow rearranging the atoms *via* situations $\{g,f\}$ to arrive at the wanted situation $\{N,0\}$. Another approach starts from an incomplete but correct model – *i.e.* the situation $\{0,0\}$ in the most extreme case – and tries by somehow finding new

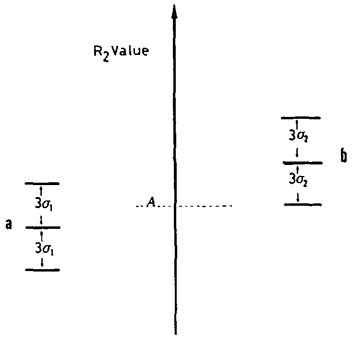

Fig. 1. R_2 ranges for a correct and incorrect trial situation to which the R_2 value of a model should be compared; a represents the first moment of $P(R_2)$ for all models $\{g+1,0\}$ and σ_1 the second moment, while b is the first moment of $P(R_2)$ for all models $\{g,1\}$ and σ_2 the second moment.

atoms *via* situations $\{g,0\}$ also to arrive at $\{N,0\}$. The latter strategy, called the zero-atom or additive approach, together with the R_2 criterion, will be shown to have the better chances. Finally, we look closer into the zero-atom approach by evaluating the chances of finding new atoms as a function of the number of reflections in the data set.

2. General expression for the basic intensity distribution

Let the rigid point-atom structure contain N atoms with scattering power 1. The relation between the sets of atoms indicated by the subscripts o, c, g and f is illustrated in Fig. 2 for the space group $P1$.

The value F_o is composed of F_c and F_q, the latter belonging to some unknown rest structure of size $N-n$. The structure-factor equation is taken as

$$F_o = \sum_{j=1}^{N} \exp(2\pi i \mathbf{H} r_j). \tag{2.1}$$

The step to normalized structure factors can be made by realizing that F deviates from E only by a constant factor $\sqrt{N}$. Equations (1.3) and (1.4) show the need to know the distribution function $P(E_c;E_o)$ and its moments for the situation $\{g,f\}$. The distribution functions we derived for $P1$ and $P\bar{1}$ (parts I and II) for the situations $\{g,0\}$ are of the type $P(E_g;E_o)$. Thus, a more general expression is needed here.

Since E_g, E_c and E_o are interrelated, the basic element in our present set-up is the distribution $P(E_g,E_c,E_o)$. Let $P(E_c;E_o)$ be the marginal of the conditional probability function $P(E_g,E_c;E_o)$, or algebraically

$$P(E_c;E_o) = \int_0^\infty P(E_g,E_c;E_o)\,\mathrm{d}E_g. \tag{2.2}$$

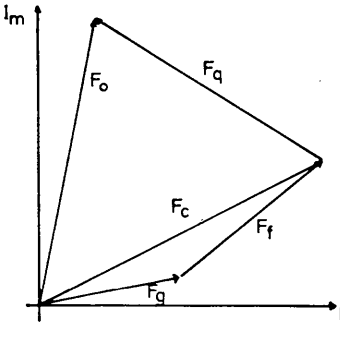

Fig. 2. Relation between structure factors of an N-atom structure (F_o) and an n-atom model (F_c), in which g atoms (F_g) are correctly and f atoms (F_f) are incorrectly placed.

Using the theorem of Bayes (see Appendix *A**) we find

$$P(E_c;E_o) = \int_0^\infty P(E_c;E_g,E_o)P(E_g;E_o)\,\mathrm{d}E_g \tag{2.3}$$

and

$$P(E_g,E_c,E_o) = P(E_c;E_g,E_o)\,P(E_o;E_g)\,P(E_g)$$
$$P(E_g,E_c,E_o) = P(E_o;E_c,E_g)P(E_c;E_g)P(E_g). \tag{2.4}$$

Since E_c differs from E_g only by a set of unrelated incorrect atoms we have

$$P(E_o;E_c,E_g) = P(E_o;E_g) \tag{2.5}$$

and thus

$$P(E_c;E_g) = P(E_c;E_g,E_o). \tag{2.6}$$

Substitution of this result in (2.3) gives

$$P(E_c;E_o) = \int_0^\infty P(E_c;E_g)P(E_g;E_o)\,\mathrm{d}E_g. \tag{2.7}$$

The distributions $P(E_g;E_o)$ are known (parts I and II), while $P(E_c;E_g)$ for the space groups $P1$ and $P\bar{1}$ can be found in Srinivasan & Parthasarathy's (1976) handbook on crystallographic statistics.

3. Space group *P*1

In this section we will summarize the results for $\langle R_2 \rangle$ and $\sigma(R_2)$ for models $\{g,f\}$ of a specific structure in the space group $P1$. The actual derivation of the equations is presented in Appendix *B*.* For R_2:

$$\langle R_2;\mathscr{E}_o\rangle = \left\{\sum_H E_o^4\left(\frac{\eta_g^8}{\eta_o^8} - 2\frac{\eta_g^4}{\eta_o^4} + 1\right) + \sum_H E_o^2\left(4\frac{\eta_g^4}{\eta_o^4} - 2\right)\left(\frac{\eta_c^2}{\eta_o^2} - \frac{\eta_g^4}{\eta_o^4}\right) + \sum_H 2\left(\frac{\eta_c^2}{\eta_o^2} - \frac{\eta_g^4}{\eta_o^4}\right)^2\right\} \Big/ \sum_H E_o^4; \tag{3.1}$$

and for $\sigma(R_2)$:

$$\sigma^2(R_2;\mathscr{E}_o) = \left\{\sum_H E_o^6\left(8\frac{\eta_g^{12}}{\eta_o^{12}} - 16\frac{\eta_g^8}{\eta_o^8} + 8\frac{\eta_g^4}{\eta_o^4}\right) \times \left(\frac{\eta_c^2}{\eta_o^2} - \frac{\eta_g^4}{\eta_o^4}\right)\right.$$

* Appendices *A*, *B* and *C* have been deposited with the British Library Lending Division as Supplementary Publication No. SUP 38805 (9 pp.). Copies may be obtained through The Executive Secretary, International Union of Crystallography, 5 Abbey Square, Chester CH1 2HU, England.

$$+\sum_H E_o^4\left(52\frac{\eta_g^8}{\eta_o^8}-48\frac{\eta_g^4}{\eta_o^4}+4\right)\times\left(\frac{\eta_c^2}{\eta_o^2}-\frac{\eta_g^4}{\eta_o^4}\right)^2+\sum_H E_o^2\left(80\frac{\eta_g^4}{\eta_o^4}-16\right)\times\left(\frac{\eta_c^2}{\eta_o^2}-\frac{\eta_g^4}{\eta_o^4}\right)^3+\sum_H 20\left(\frac{\eta_c^2}{\eta_o^2}-\frac{\eta_g^4}{\eta_o^4}\right)^4\Bigg\Bigg/\left\{\sum_H E_o^4\right\}^2. \quad (3.2)$$

For the normalized residual function R_2^n (see part I, §§ 2.3 and 4.3) we find

$$\langle R_2^n;\mathscr{E}_o\rangle=\left\{\sum_H E_o^4\left(\frac{\eta_g^8}{\eta_o^4\eta_c^4}-2\frac{\eta_g^4}{\eta_o^2\eta_c^2}+1\right)+\sum_H E_o^2\left(4\frac{\eta_g^4}{\eta_o^2\eta_c^2}-2\right)\left(1-\frac{\eta_g^4}{\eta_o^2\eta_c^2}\right)+\sum_H 2\left(1-\frac{\eta_g^4}{\eta_o^2\eta_c^2}\right)^2\right\}\Bigg/\sum_H E_o^4 \quad (3.3)$$

and

$$\sigma^2(R_2^n;\mathscr{E}_o)=\left\{\sum_H E_o^6\left(8\frac{\eta_g^{12}}{\eta_o^6\eta_c^6}-16\frac{\eta_g^8}{\eta_o^4\eta_c^4}+8\frac{\eta_g^4}{\eta_o^2\eta_c^2}\right)\times\left(1-\frac{\eta_g^4}{\eta_o^2\eta_c^2}\right)+\sum_H E_o^4\left(52\frac{\eta_g^8}{\eta_o^4\eta_c^4}-48\frac{\eta_g^4}{\eta_o^2\eta_c^2}+4\right)\times\left(1-\frac{\eta_g^4}{\eta_o^2\eta_c^2}\right)^2+\sum_H E_o^2\left(80\frac{\eta_g^4}{\eta_o^2\eta_c^2}-16\right)\left(1-\frac{\eta_g^4}{\eta_o^2\eta_c^2}\right)^3+\sum_H 20\left(1-\frac{\eta_g^4}{\eta_o^2\eta_c^2}\right)^4\Bigg\Bigg/\left\{\sum_H E_o^4\right\}^2 \quad (3.4)$$

4. Space group $P\bar{1}$

Here we summarize the results for $\langle R_2\rangle$ and $\sigma(R_2)$ for models $\{g,f\}$ of a specific structure in the space group $P\bar{1}$. The actual derivation of the formulae is presented in Appendix C.* $\langle R_2\rangle$ is given by

$$\langle R_2;\mathscr{E}_o\rangle=\left\{\sum_H E_o^4\left(\frac{\eta_g^8}{\eta_o^8}-2\frac{\eta_g^4}{\eta_o^4}+1\right)+\sum_H E_o^2\left(6\frac{\eta_g^4}{\eta_o^4}-2\right)\left(\frac{\eta_c^2}{\eta_o^2}-\frac{\eta_g^4}{\eta_o^4}\right)+\sum_H 3\left(\frac{\eta_c^2}{\eta_o^2}-\frac{\eta_g^4}{\eta_o^4}\right)^2\right\}\Bigg/\sum_H E_o^4 \quad (4.1)$$

and $\sigma(R_2)$ by

$$\sigma^2(R_2;\mathscr{E}_o)=\left\{\sum_H E_o^6\left(16\frac{\eta_g^{12}}{\eta_o^{12}}-32\frac{\eta_g^8}{\eta_o^8}+16\frac{\eta_g^4}{\eta_o^4}\right)\times\left(\frac{\eta_c^2}{\eta_o^2}-\frac{\eta_g^4}{\eta_o^4}\right)+\sum_H E_o^4\left(168\frac{\eta_g^8}{\eta_o^8}-144\frac{\eta_g^4}{\eta_o^4}+8\right)\times\left(\frac{\eta_c^2}{\eta_o^2}-\frac{\eta_g^4}{\eta_o^4}\right)^2+\sum_H E_o^2\left(384\frac{\eta_g^4}{\eta_o^4}-48\right)\times\left(\frac{\eta_c^2}{\eta_o^2}-\frac{\eta_g^4}{\eta_o^4}\right)^3+\sum_H 96\left(\frac{\eta_c^2}{\eta_o^2}-\frac{\eta_g^4}{\eta_o^4}\right)^4\Bigg\Bigg/\left\{\sum_H E_o^4\right\}^2. \quad (4.2)$$

5. Discussion and conclusions

In the preceding sections we derived expressions to predict *a priori* values of $\langle R_2\rangle$ and $\sigma(R_2)$ tailored to the structure at hand. However, generally useful insight, *i.e.* independent of a specific structure, into the behaviour of $\langle R_2\rangle$ and $\sigma(R_2)$ requires an extra averaging over all structures, that is knowledge of $\langle\langle R_2\rangle\rangle_{r^o}$ and $\langle\sigma(R_2)\rangle_{r^o}$. As discussed in part I, §§ 1 and 4.1 these quantities can be approximated by replacing the explicit summations over the data set $\mathscr{E}_o$ by distribution averages. Thus $\sum_H E_o^n$ are replaced by $\mathscr{H}\langle E_o^n\rangle_{r^o}$, where $\mathscr{H}$ represents the number of reflections and the $\langle E_o^n\rangle_{r^o}$ are evaluated by averaging the space-group-dependent structure-factor equations with respect to the atomic coordinates. In doing so one introduces an *average* structure of size N. The behaviour of $\langle R_2\rangle$ and $\sigma(R_2)$ for a specific structure compared to $\langle\langle R_2\rangle\rangle_{r^o}$ and $\langle\sigma(R_2)\rangle_{r^o}$ for the averaged structure are sufficiently close to allow generally useful conclusions.

Table 1. *Comparison of theoretical values* $\langle R_2\rangle$ *and* $\sigma(R_2)$ *for an averaged structure with experimental results (Petit & Lenstra, 1982) for a 100-atom structure with 2000 reflections*

Situation	$\langle R_2$(exp)$\rangle$	$\langle R_2$(th)$\rangle$	σ(exp)	σ(th)
Space group $P1$				
{50, 50}	0·750	0·750	0·040	0·031
{30, 30}	0·667	0·670	0·018	0·016
{60, 0}	0·399	0·400	0·011	0·012
{0, 60}	0·755	0·760	0·017	0·016
Space group $P\bar{1}$				
{50, 50}	1·000	1·000	0·060	0·051
{30, 30}	0·840	0·840	0·020	0·025
{60, 0}	0·482	0·480	0·017	0·019
{0, 60}	0·960	0·960	0·020	0·024

* Appendices *A*, *B* and *C* have been deposited. See earlier footnote.

Fig. 3 shows the R_2 surface and Fig. 4 the $\sigma(R_2)$ surface for such a generalized average structure in the space group $P1$, as functions of the fractions η_c^2/η_o^2 and η_g^2/η_c^2. For example, the section parallel to the η_c^2/η_o^2 axis at $\eta_g^2/\eta_c^2 = 0$ gives the generalized paths of R_2 or $\sigma(R_2)$ for situations $\{0,n\}$ with n varying from zero to N. A similar section at $\eta_g^2/\eta_c^2 = 1{\cdot}0$ gives these paths for situations $\{n,0\}$ (see Figs. 3 and 4 of part I for comparison). Finally, sections parallel to the η_g^2/η_c^2 axis at $\eta_c^2/\eta_o^2 = 1{\cdot}0$ give the variation of R_2 and $\sigma(R_2)$ going from situations $\{0,N\}$ to $\{N,0\}$ *via* $\{g,f\}$ while $g+f=N$.

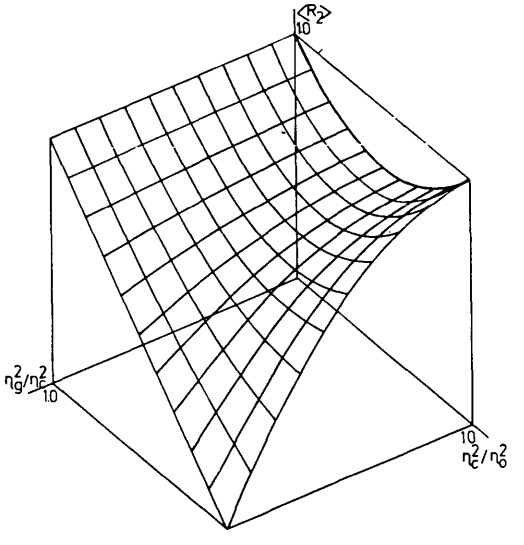

Fig. 3. $\langle R_2 \rangle$ for models of the type $\{g,f\}$ for averaged structures in the space group $P1$.

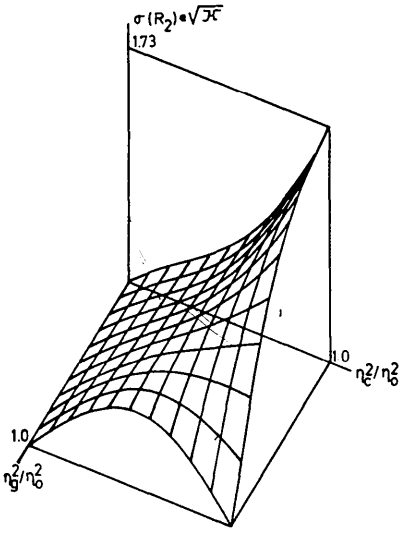

Fig. 4. $\sigma(R_2)$ for models of the type $\{g,f\}$ for averaged structures in the space group $P1$. $\mathscr{N}$ is the number of reflections in the data set.

Table 1 shows that the theoretical values generated in this way fit rather well to experimental values obtained by Petit & Lenstra (1982). The agreement is very satisfactory if one takes into account that the theoretical values are for an average structure, not using the data set for the actual structure at hand.

Turning to the applications of the residuals, one could possibly see their use in the screening of a set of *MULTAN* solutions. This would require the ability to determine η_g^2 and η_f^2 for a particular set of peaks. Fig. 5 gives $\langle R_2 \rangle$ values and their 3σ ranges as a function of η_g^2 and η_f^2 for complete models:

$$\eta_g^2 + \eta_f^2 = N. \tag{5.1}$$

When about 2000 reflections are taken in the R_2 calculations, reasonably good estimates of the percentage of false atomic positions can be obtained, provided the solution already contains >75% correctly placed atoms.

To transfer this result to structures of different sizes it is essential to realize that an increase in the size of the structure, *e.g.* from ten to 100 atoms, will reduce the change in $\langle R_2 \rangle$ per added atom ten times, while the σ values stay nearly constant, because they depend only on the fraction of the atoms placed and not on the absolute size of the structure. Thus to get near *equivalent* data sets one must have, because $\sigma(R_2)$ is inversely proportional to $\sqrt{\mathscr{N}}$, a 100-fold increase in the size of the data set. Note that the quality of the estimates cannot be improved by replacing R_2 by R_2^n, because in the limit of a complete model these indicators are identical.

The consequences of our present knowledge can further be pursued to investigate the chances two extreme procedures have to bring a structure determination to a successful end.

The first strategy, advocated by Lenstra (1974), starts from a zero-atom model. Atoms are added to the asymmetric unit, one at a time, such that *via* the situation $\{g,0\}$ one finally arrives at the complete,

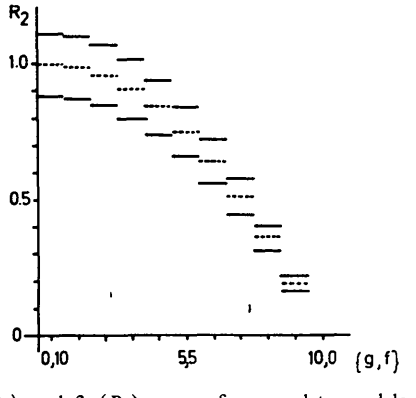

Fig. 5. $\langle R_2 \rangle$ and $3\sigma(R_2)$ ranges for complete models (see text); Solid lines give the 3σ ranges while the dashed lines give the average R_2 values. The space group is $P1$.

correct solution $\{N,0\}$. Suppose we can start from the situation $\{g,0\}$, add one atom and want to monitor the process by the R_2 criterion. We must be able to discriminate a wrong move ending in $\{g,1\}$ from a correct move ending in $\{g+1,0\}$. As a measure of the resolving power of the discriminator function we take the quantity S, defined as

$$S(g) = \frac{\langle R_2\{g,1\}\rangle - \langle R_2\{g+1,0\}\rangle}{3[\sigma(R_2\{g,1\}) + \sigma(R_2\{g+1,0\})]}. \quad (5.2)$$

The second strategy, proposed by Booth (1947, 1949), starts from a random but complete model, *i.e.* in general from the situation $\{0,N\}$. By somehow rearranging the atoms one hopes *via* intermediate situations $\{g,f\}$ to arrive at $\{N,0\}$. Suppose we have a way of moving only false atoms, leaving correct atoms at their positions as soon as they are found. If we can now start from the situation $\{g,f\}$, move one false atom and monitor the process by the R_2 criterion, we must be able to discriminate a wrong move ending in $\{g,f\}$ from a correct move ending in $\{g+1,f-1\}$. The above definition of S can be applied to both strategies and shows that for $S > 1$ a perfect discrimination is possible between wrong and correct ending moves. Obviously the lower the g value for which $S(g) > 1$, the better solution strategy and R_2 criterion fit together and the better chances we have that the strategy will bring an automated structure determination to a successful end.

Table 2 summarizes the $S(g)$ values for the two strategies applied to an *average* structure in $P1$.

Table 2. *$S(g)$ values for an average structure in $P1$ with ten equal atoms and 2000 reflections*

Zero-atom strategy

Start	End: Wrong	End: Correct	$S(R_2)$	$S(R_2^n)$
{1,0}	{1,1}	{2,0}	0·805	0·712
{2,0}	{2,1}	{3,0}	0·912	0·865
{3,0}	{3,1}	{4,0}	1·017	1·013
{4,0}	{4,1}	{5,0}	1·161	1·194
{5,0}	{5,1}	{6,0}	1·376	1·439
{6,0}	{6,1}	{7,0}	1·724	1·803
{7,0}	{7,1}	{8,0}	2·341	2·413
{8,0}	{8,1}	{9,0}	3·634	3·649
{9,0}	{9,1}	{10,0}	7·487	7·487

Random-model strategy

Start	End: Wrong	End: Correct	$S(R_2)$
{1,9}	{1,9}	{2,8}	0·131
{2,8}	{2,8}	{3,7}	0·226
{3,7}	{3,7}	{4,6}	0·332
{4,6}	{4,6}	{5,5}	0·460
{5,5}	{5,5}	{6,4}	0·629
{6,4}	{6,4}	{7,3}	0·881
{7,3}	{7,3}	{8,2}	1·319
{8,2}	{8,2}	{9,1}	2·332
{9,1}	{9,1}	{10,0}	7·487

Contrary to Wilson's (1977) opinion, the zero-atom strategy in combination with R_2 seems to have the better chance of being successful. Provided sufficient data points are available the zero-atom strategy may automatically lead to the correct structure if about 25% of the atoms are already properly placed. On the other hand, chances are dim that the random-model strategy will automatically give the correct structure unless more than 65% of the atoms in the starting model are correct. It is of interest to note, Table 2, that when we base the S values on the R_2^n criterion the turning points at which $S(g) > 1$ do not change.

Low $S(g)$ values (<1) indicate that the route to the end of the determination is endangered, but do not necessarily predict a fatal outcome. In the zero-atom strategy the introduction of a false atom at any stage is fatal, but the rejection of a correct one is merely unfortunate, since it may become acceptable at some later stage. The process continues in the correct direction as long as we can find one more atom that is correct. It would be of value to have information about the possibilities of finding such an atom amongst the $N-g$ atoms at every stage $\{g,0\}$. The situation may be analyzed from Fig. 6. Curve F represents the distribution $P(R_2)$ for the fatal ending $\{g,1\}$ and curve G the distribution $P(R_2)$ for the correct situation $\{g+1,0\}$. The area under F to the left of $R_2(C)$ gives the chance of a type I error, the addition of a fatal incorrect atom. The area under G to the right of $R_2(C)$ gives the chance of a type II error, the rejection of a correct atom. We take F and G as Gaussian (part I) and put $R_2(C)$ at a distance $3\sigma(R_2\{g,1\})$ away from the average $R_2\{g,1\}$. Thus the chance of a fatal error becomes negligible ($<0{\cdot}3\%$). We can now calculate the area under G to the left of $R_2(C)$ as the chance to find a correct atom amongst the $N-g$ candidates.

Table 3 gives the results for a ten-atom structure when 1000 or 100 reflections are used in the data set from which the criterion is calculated.

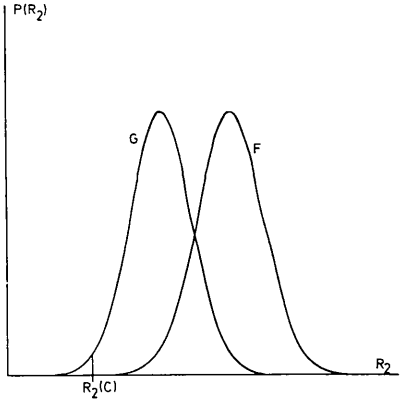

Fig. 6. Distribution $P(R_2)$ for the situation $\{5,0\}$, curve G, and of $P(R_2)$ for the situation $\{4,1\}$, curve F. The structure contains ten atoms, space group $P1$. 100 reflections were used.

Table 3. *Probability of finding a correct atom at various stages of the determination in the zero-atom strategy (see text), with two data sets*

Start situation	P (%) 1000 reflections	P (%) 100 reflections
{1, 0}	72·2	5·6
{2, 0}	82·8	5·4
{3, 0}	90·6	5·1
{4, 0}	97·0	5·2
{5, 0}	99·7	6·2
{6, 0}	99·9	9·2
{7, 0}	~100·0	21·2
{8, 0}	~100·0	83·7
{9, 0}	~100·0	~100·0

For large data sets we have a relatively good (here about 70%) chance of finding an additional correct atom right from the start. Under the unfavourable conditions of a small data set, say 100 reflections for ten atoms, automation becomes almost impossible.

Since in an automated structure determination many R_2 checks will be made and the computing time per check increases with the number of reflections involved in the R_2 calculations, one is tempted to limit the number of data points per test. Unfortunately, the results of Table 3 confirm and extend a previous conclusion (Petit & Lenstra, 1982) that in this way one can only save computer time if one is willing to decrease one's chance to find new atoms.

At this point it seems appropriate to make some remarks against an over-optimistic transfer of our conclusions to experimental situations. An incorrectly placed atom in our analysis is completely and randomly misplaced. Sometimes, however, tentative atomic positions are generated (*e.g.* by *MULTAN*) which exhibit systematic errors, for instance, a geometrically correct fragment at an incorrect location. This makes the magnitudes of E_o and E_c interrelated, which means an invalidation of our premise for handling incorrect atomic sites. Also, the influence of measuring errors in the data set and the influence of small misplacements in otherwise correctly placed atoms is left out of the present theory. This will be the subject of further investigations.

As a final conclusion one can state that a successful application of R_2 to the analysis of *MULTAN* maps seems improbable, particularly if translation problems are present. However, an iterative automated procedure, taking as input peaks from a heavy-atom Fourier or from a *DIRDIF* Fourier (Beurskens & Noordik, 1972), seems well within the possibilities of the discriminating power of the residual functions.

WVH thanks the Belgian organisation IWONL for financial support. The help of Professor H. J. Geise in preparation of this manuscript is gratefully acknowledged. We also wish to thank G. H. Petit and J. F. Van Loock for stimulating discussions.

References

BATEMAN, H. (1953). *Higher Transcendental Functions*. Bateman Manuscript Project, Vols. I and II. New York: McGraw-Hill.

BEURSKENS, P. T. & NOORDIK, J. H. (1972). *Acta Cryst.* A**27**, 187–188.

BOOTH, A. D. (1947). *Nature (London)*, **160**, 196.

BOOTH, A. D. (1949). *Proc. R. Soc. London Ser. A*, **197**, 336–355.

LENSTRA, A. T. H. (1974). *Acta Cryst.* A**30**, 363–369.

LINDGREN, B. W. (1976). *Statistical Theory*, 3rd ed. New York: Macmillan.

NEUTS, M. F. (1973). *Probability*. Boston: Allyn & Bacon.

PETIT, G. H. & LENSTRA, A. T. H. (1982). *Acta Cryst.* A**38**, 67–70.

ROHATGI, V. K. (1976). *An Introduction to Probability Theory and Mathematical Statistics*. New York: John Wiley.

SRINIVASAN, R. & PARTHASARATHY, S. (1976). *Some Statistical Applications in X-ray Crystallography*. Oxford: Pergamon.

VAN HAVERE, W. K. L. & LENSTRA, A. T. H. (1983*a*). *Acta Cryst.* A**39**, 553–562.

VAN HAVERE, W. K. L. & LENSTRA, A. T. H. (1983*b*). *Acta Cryst.* A**39**, 562–565.

WATSON, G. N. (1966). *A Treatise on the Theory of Bessel Functions*. Cambridge Univ. Press.

WILSON, A. J. C. (1977). *Acta Cryst.* A**33**, 523–524.