Acta Cryst. (1983). A**39**, 553–562

Moments of the Probability Density Function of R_2 Approached *Via* Conditional Probabilities. I. Development of the Theory for *P*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* 15 *April* 1982; *accepted* 14 *February* 1983)

Abstract

A critical survey is given of the operations necessary to evaluate the moments of the probability density function $P(R_2)$. The failure of existing theories to give $\sigma(R_2)$ is traced to the fact that averaging over all reflections is only equivalent to averaging over the set of coordinates of a model for an infinite data set. With the help of conditional probabilities the difficulties are overcome and formulas are derived for the first and second moments as a function of the size of the model. The formulas are valid in space group *P*1 for two extreme cases, *viz* a completely incorrect and a completely correct model. Incorporation of observed intensities enables one to obtain accurate *a priori* estimates of $\langle R_2 \rangle$ and $\sigma(R_2)$. The theory agrees very well with simulated experiments. It is demonstrated that R_2 and R_2^n have equal resolving power.

1. Introduction: the need for a new theory

In automated crystal structure analyses one can effectively use residual functions, such as R_2 and R_2^n, as discriminator functions (Lenstra, 1974; Van de Mieroop, 1979) to decide upon the correctness of an atom newly added to a tentative model. The actual decision whether the model at hand (be it partial or complete) is correct or incorrect rests upon a comparison between the actual R_2 value and its expected value. A statistical approach to the latter is needed, because knowledge of coordinates is obviously not available at that stage. The best way to apply R_2 is to implement it into a statistical decision process. In order to do so, either the probability density function $P(R_2)$ itself has to be known, or the moments necessary to reproduce the properties of $P(R_2)$. The latter approach is taken in this paper. For definitions and in-depth discussions of the statistical terminology we refer to standard texts (*e.g.* Lindgren, 1976; Neuts, 1973; Rohatgi, 1976).

Up to now two complementary procedures are followed to evaluate the moments of $P(R_2)$:

(i) In the Patterson approach R_2 is defined over a vector space (Lenstra, 1974; 1979) as

$$R_2 \equiv \int_v (P_o - P_c)^2 \, dv / \int_v P_o^2 \, dv \qquad (1.1)$$

in which P_o and P_c represent the observed and calculated Patterson maps, while v stands for the volume of one unit cell. A Patterson map is then regarded as a weighted set of δ functions representing vectors randomly positioned in the unit cell of the Patterson map.

If all atoms are placed correctly, vectors in P_o and P_c coincide and contribute to the terms in $P_o P_c$, whereas they do not for incorrectly placed atoms. Evaluation of R_2 thus amounts to a counting of vectors. This point of departure, however, is only valid if the data set used to calculate the Patterson maps is infinite, thus enforcing from the start a severe limitation on the theory. We will show that this jeopardizes further developments.

(ii) In the more common statistical approach R_2 is defined in reciprocal space as

$$R_2 \equiv \langle (E_o^2 - \eta^2 E_c^2)^2 \rangle_H / \langle E_o^4 \rangle_H. \qquad (1.2)$$

The angular brackets represent an averaging over the actual set of structure factors, $H \equiv (h,k,l)$. We define η^2 as the fraction of the scattering power of the model relative to the total structure:

$$\eta^2 \equiv \eta_c^2 / \eta_o^2. \qquad (1.3)$$

For an equal-atom structure

$$\eta^2 = n/N, \qquad (1.4)$$

in which n and N represent the number of atoms in model and observed structures, respectively. E_o and E_c are the moduli of the structure factors of, respectively, the total structure and a model. The normalized structure factors are defined as

$$\bar{E}_o \equiv \bar{E}_o(H) = N^{-1/2} \sum_{j=1}^{N} \exp(-2\pi i H r_j). \qquad (1.5)$$

0108-7673/83/040553-10$01.50

An *a priori* value, which is useful as well as feasible, is the mean value of R_2 averaged over all models, $\langle R_2\rangle_{r^c}$, where r^c is defined as

$$r^c \equiv \{\mathbf{r}^c\} = \{\mathbf{r}_j^c, j = 1, \ldots, n\}. \tag{1.6}$$

Should one want to make statements about R_2 with more general validity, *i.e.* irrespective of any structure, one needs to have knowledge about $\langle\langle R_2\rangle_{r^c}\rangle_{r^o}$. In other words, an extra averaging over all structures is necessary. To derive these quantities from (1.2) investigators have always tacitly replaced $\langle\ \rangle_H$ by $\langle\ \rangle_r$. This leads to

$$\langle\langle R_2\rangle_{r^c}\rangle_{r^o} = \langle R_2\rangle_{r^c} \tag{1.7}$$

and

$$\langle R_2\rangle_{r^c} = \{\langle E_o^4\rangle_{r^o} + \eta^4\langle E_c^4\rangle_{r^c} - 2\eta^2\langle E_o^2 E_c^2\rangle_{r^c,r^o}\}/\{\langle E_o^4\rangle_{r^o}\}. \tag{1.8}$$

The necessary intensity moments are then obtained from an intensity distribution (Wilson, 1949; Srinivasan & Parthasarathy, 1976) or evaluated by averaging the space-group-dependent structure-factor equations with respect to r^o and r^c (Wilson, 1950*a*,*b*, 1969, 1978; Shmueli & Kaldor, 1981; Shmueli & Wilson, 1981). However, replacing $\langle\ \rangle_H$ by $\langle\ \rangle_r$ is only permitted if H represents an infinitely large data set. Therefore, the Patterson and statistical approaches have the same limiting condition. The immediate result is that every correct model of n atoms out of an N-atom structure gives the same R_2 value and thus the corresponding $P(R_2)$ is a δ function, *i.e.* $\sigma(R_2) = 0$.

In practical X-ray crystallography one always deals with finite data sets. Replacement of $\langle\ \rangle_H$ by $\langle\ \rangle_r$ is thus no longer an identity operation but an approximation. Consequently, the result of (1.8) is only an approximation of the quantities $\langle R_2\rangle_{r^c}$ and $\langle\langle R_2\rangle_{r^c}\rangle_{r^o}$ we are interested in. Use of them in the decision process will reduce the chances for correct decisions when handling small data sets. More important, however, is that one immediately observes in practice that various correct n-atom models for one N-atom structure give rise to different R_2 values. Thus $\sigma(R_2) = 0$ is a very bad estimate for the true spread in R_2. We conclude that the present formalisms are inadequate to obtain realistic information on $P(R_2)$. In the next sections a new theory is developed which will give better estimates of $\langle R_2\rangle_{r^c}$ and allows the calculation of realistic higher moments of R_2 and the residuals in general.

2. New theory

Starting from the general definition of R_2

$$R_2 \equiv \sum_H (E_o^2 - \eta^2 E_c^2)^2 / \sum_H E_o^4, \tag{2.1}$$

where H may be any subset of points in reciprocal space, we are faced with the problem to construct a probability space which probability measure allows an intelligent guess of R_2. Defining a probability space $[\mathscr{R}, P(R_2)]$, we can take as sample space $\mathscr{R}$ the set of all real numbers. Recognizing that R_2 is a function of structure factors, which themselves are functions of the fractional coordinates, it is easily seen that in our context the natural primary variables for R_2 are coordinates and that the logical choice is to define R_2 *over* an underlying probability space $[\Omega, P(r)]$ based on these coordinates. The concrete choice of this space depends of course on whether the model is correct or incorrect. Instead of altering the probability space for each particular structure and for each change in the set of structure factors used, we prefer to use the concept of conditional probabilities. In doing so the original probability space can be kept. The sample space Ω can be taken as an N-fold Cartesian product of the unit cube, $[0,1]^3$, that is we consider structures containing N-point atoms with fractional coordinates between 0 and 1. The probability measure $P(r)$ connected with the set of coordinates is easily defined when we consider the atoms randomly positioned, meaning $P(r)$ associates an equal probability with all points in the direct unit cell. Using this framework, we may calculate $P(R_2)$ with standard statistical procedures. It will, however, be convenient to work with the sets of E values as intermediates between the coordinate space and the probability space of R_2, since this will allow us to use published results on intensity statistics. The observed intensities, which form a set of *a priori* fixed parameters, will form the set of conditions under which the appropriate set of stochastic variables, the set of structure factors belonging to the model, must be handled. The way in which these conditional probabilities are made explicit will then allow us to express the correctness or incorrectness of a model. This brings us back to a point where standard statistical procedures allow the calculation of $P(R_2)$.

For all practical purposes we take the probability associated with R_2 as Gaussian (Fig. 1), that is, we

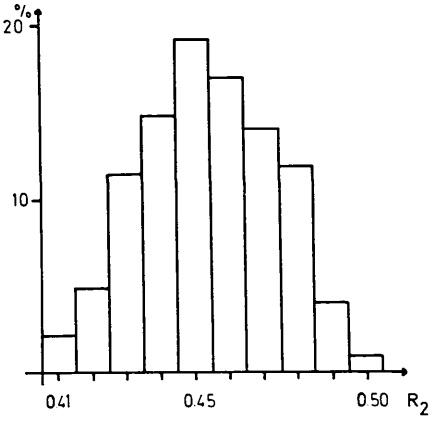

Fig. 1. $P(R_2)$ for a model size of 6 out of 11 atoms. For details of the calculation see § 3.3.

confine ourselves to the first two moments of R_2, although the evaluation of higher moments presents no special problems.

The first moment is obtained by simply rephrasing the definition (2.1) into

$$\langle R_2;\mathscr{E}_o\rangle = 1 + \eta^4 \frac{\sum_H \langle E_c^4;\mathscr{E}_o\rangle}{\sum_H E_o^4} - 2\eta^2 \frac{\sum_H E_o^2\langle E_c^2;\mathscr{E}_o\rangle}{\sum_H E_o^4}, \tag{2.2}$$

where $\mathscr{E}_o$ is the subset of all possible reflections actually used in the calculations. The averaging is done over the coordinates of the model unless stated otherwise. The second moment about the mean is given by

$$\sigma^2(R_2;\mathscr{E}_o) = \langle R_2^2;\mathscr{E}_o\rangle - \langle R_2;\mathscr{E}_o\rangle^2. \tag{2.3}$$

After substitution and sorting we obtain

$$\begin{aligned}\sigma^2(R_2;\mathscr{E}_o) = \Bigg\{ & \sum_H \eta^8(\langle E_c^8;\mathscr{E}_o\rangle - \langle E_c^4;\mathscr{E}_o\rangle^2) \\ & + \sum_{H\neq K}\sum \eta^8[\langle E_c^4(H)\,E_c^4(K);\mathscr{E}_o\rangle \\ & - \langle E_c^4(H);\mathscr{E}_o\rangle\langle E_c^4(K);\mathscr{E}_o\rangle] \\ & - \sum_H 4\eta^6 E_o^2[\langle E_c^6;\mathscr{E}_o\rangle \\ & - \langle E_c^4;\mathscr{E}_o\rangle\langle E_c^2;\mathscr{E}_o\rangle] - \sum_{H\neq K}\sum 4\eta^6 E_o^2(H) \\ & \times [\langle E_c^4(H)\,E_c^2(K);\mathscr{E}_o\rangle \\ & - \langle E_c^4(H);\mathscr{E}_o\rangle\langle E_c^2(K);\mathscr{E}_o\rangle] \\ & + \sum_H 4\eta^4 E_o^4[\langle E_c^4;\mathscr{E}_o\rangle - \langle E_c^2;\mathscr{E}_o\rangle^2] \\ & + \sum_{H\neq K}\sum 4\eta^4 E_o^4(H)[\langle E_c^2(H)\,E_c^2(K);\mathscr{E}_o\rangle \\ & - \langle E_c^2(H);\mathscr{E}_o\rangle\langle E_c^2(K);\mathscr{E}_o\rangle]\Bigg\} \Big/ \Big\{\sum_H E_o^4\Big\}^2. \end{aligned} \tag{2.4}$$

Note that $\langle a;b\rangle$ means the average of a under the condition that b has occurred. Equations (2.2) and (2.4) are generally valid for all space groups and all models including completely correct as well as completely incorrect models. The only difference between correct and incorrect models is the way in which the conditional relation towards $\mathscr{E}_o$ is formulated. This relation and this relation alone modifies $P(R_2)$ and its moments. With (2.2) and (2.4) the problem of the moments of R_2 is reduced to finding the moments of the intensity distributions $P(E_c;\mathscr{E}_o)$ and $P[E_c(H)\ E_c(K);\ \mathscr{E}_o]$. For clarity and brevity we confine ourselves in this paper to two extreme cases in space group $P1$ only: (i) completely incorrect models and (ii) completely correct models. The extreme cases in $P\bar{1}$, as well as cases in $P1$ and $P1$ containing a mixture of correct and incorrect atoms, will be dealt with in subsequent papers in this series. Also, small positional errors in otherwise correctly placed atoms and random errors in the measured E values will be ignored for the time being.

2.1. *Incorrect models*

The intensity moments for incorrect models can be derived realizing that no correlation exists between observed and calculated structure amplitudes. Thus,

$$\langle E_c^n;\mathscr{E}_o\rangle = \langle E_c^n\rangle \tag{2.1.1}$$

and

$$\langle E_c^n(H)\,E_c^m(K);\mathscr{E}_o\rangle = \langle E_c^n(H)\,E_c^m(K)\rangle. \tag{2.1.2}$$

This condition, in combination with our original sample space Ω, allows one to evaluate $\langle E_c^n\rangle$, (2.1.1), using one of the intensity distributions available in the literature. For example, one can use the asymptotic distribution of Wilson (1949), provided the model contains a sufficiently large number of atoms:

$$P(E_c) = 2E_c\exp(-E_c^2). \tag{2.1.3}$$

Although Wilson's original derivation employs an averaging over reflections and not over coordinates, Karle & Hauptman, 1953; Hauptman & Karle, 1953, confirmed that (2.1.3) conforms with the consequences of our choice of sample space. For most practical purposes Wilson's distribution gives sufficiently accurate results also for models containing a finite number of atoms (see § 3 *Experimental verification*). The moments of (2.1.3) are given by (Shmueli, 1982)

$$\langle E_c^{2n}\rangle = n! \quad \text{for } n = 0, 1, 2, 3, \ldots. \tag{2.1.4}$$

Next, we turn to the evaluation of the moments given in (2.1.2). This is a more complicated problem since there exists correlation between some reciprocal-lattice points. Direct methods (*e.g.* Klug, 1958; Giacovazzo, 1980) provide us with arguments that this correlation is of minor importance to our results. Firstly, a non-zero correlation is only to be expected in $P1$ for reflections with indices that are multiples of one another. Secondly, the influence of correlation on the intensity moments is inversely proportional to some power of n (number of atoms in the model) and, hence, will decrease as the size of the model increases. For the majority of non-correlated reflections in the set,

$$\langle E_c^n(H)\,E_c^m(K)\rangle = \langle E_c^n(H)\rangle\langle E_c^m(K)\rangle \tag{2.1.5}$$

holds, and the terms containing summations running over two sets of reciprocal vectors cancel out two by two. For the relatively few correlated reflections the

contributions of the double summations can be shown to be negligible (see § 3 *Experimental verification*). Thus the results of $\sigma^2(R_2)$ are hardly affected by ignoring the existence of any correlation between lattice points in the set used.

Substitution of the intensity moments (2.1.4) into (2.2) gives

$$\langle R_2;\mathscr{E}_o\rangle = \left\{\sum_H (E_o^4 - 2\eta^2 E_o^2 + 2\eta^4)\right\} / \left\{\sum_H E_o^4\right\} \quad (2.1.6)$$

and into (2.4) yields

$$\sigma^2(R_2;\mathscr{E}_o) = \left\{\sum_H (4\eta^4 E_o^4 - 16\eta^6 E_o^2 + 20\eta^8)\right\} / \left\{\sum_H E_o^4\right\}^2. \quad (2.1.7)$$

2.2. *Correct models*

To derive the intensity moments for correct models the correlation between E_c and the reflections in $\mathscr{E}_o$ has to be taken explicitly into account. Unfortunately, the required moments $\langle E_c^n;\mathscr{E}_o\rangle$ are too complicated to handle and will be replaced by $\langle E_c^n;E_o\rangle$. That is, *under the constraint of a set of E_o values* is replaced by *under the constraint of one E_o value*. Of course, the latter is a necessary component of the former, but the question is: will it be a sufficient condition? The consequences of the approximation can be far reaching, because the correctness of a model is now defined at the level of a single reflection. A careful consideration of our conception of correctness is thus required. Suppose we have a correct model containing n atoms. Evidently, many sets of $3(N-n)$ coordinates, which complete the $3n$ coordinates, may be found in the sample space Ω, which, together with the coordinates of the model, will yield E_o. These sets definitely include the ultimately correct combination. However, all others which ought to be regarded as incorrect as far as the real structure is concerned are also incorporated in the final results. This might leave us empty handed, and in fact we are convinced that this problem is lethal if we stop at the level of a single reflection. However, the effect on R_2 will be limited, because in the enumeration one adds up, and many reflections contribute to the final result. At the level of a single reflection many models must be called acceptable, including the subset we would normally call correct. In the summation over a large number of reflections the latter subset becomes dominant, because it is always present in each group of the so-called acceptable models offered to us by each single reflection. Comparison of the theoretical results with the outcome of experiments has to demonstrate the correctness of our logic at this point. As we did with incorrect models we disregard correlation between reciprocal-lattice points. That is, we suppress again the summations running over two indices in (2.4).

From here on the technical derivation of $P(E_c;E_o)$ is straight-forward. For incomplete models with sufficiently large rest structures (in our case of size $N-n$), Srinivasan & Parthasarathy (1976) have given the conditional probability function $P(E_o;E_c)$ as

$$P(E_o;E_c) = \frac{2E_o\eta_o^2}{\eta_o^2-\eta_c^2}\exp\left\{-\frac{\eta_o^2E_o^2+\eta_c^2E_c^2}{\eta_o^2-\eta_c^2}\right\} \times I_0\left\{\frac{2\eta_o\eta_cE_oE_c}{\eta_o^2-\eta_c^2}\right\}, \quad (2.2.1)$$

where $I_0(x)$ is a modified Bessel function of the first kind and order zero. In order to obtain the required distribution we need to interchange E_o and E_c in (2.2.1) using the theorem of Bayes. Thus,

$$P(E_c;E_o) = P(E_o;E_c)\,P(E_c)/P(E_o). \quad (2.2.2)$$

As marginal distributions $P(E_o)$ and $P(E_c)$ we use asymptotical distributions of Wilson, (2.1.3). Consequently, the rest structure, $(N-n)$, the original model n, and thus by definition the total structure N, all have to contain a sufficiently large number of atoms. We obtain now

$$P(E_c;E_o) = \frac{2E_c\eta_o^2}{\eta_o^2-\eta_c^2}\exp\left\{-\frac{\eta_o^2E_c^2+\eta_c^2E_o^2}{\eta_o^2-\eta_c^2}\right\} \times I_0\left\{\frac{2\eta_o\eta_cE_oE_c}{\eta_o^2-\eta_c^2}\right\}. \quad (2.2.3)$$

The moments of this distribution are

$$\langle E_c^{2n};E_o\rangle = n!\left(\frac{\eta_o^2-\eta_c^2}{\eta_o^2}\right)^n {}_1F_1\left(-n;1;-\frac{\eta_c^2E_o^2}{\eta_o^2-\eta_c^2}\right), \quad (2.2.4)$$

where ${}_1F_1(a;c;x)$ represents a confluent hypergeometric function. The derivation of (2.2.4) is given in Appendix *A*. Introduction of these moments into (2.2) gives

$$\langle R_2;\mathscr{E}_o\rangle = \left\{\sum_H E_o^4(\eta^8 - 2\eta^4 + 1) + \sum_H E_o^2(4\eta^6 - 2\eta^2)\,(1-\eta^2) + \sum_H 2\eta^4(1-\eta^2)^2\right\} / \sum_H E_o^4 \quad (2.2.5)$$

and substitution into (2.4) yields

$$\sigma^2(R_2;\mathscr{E}_o) = \left\{\sum_H E_o^6(8\eta^{14} - 16\eta^{10} + 8\eta^6)(1-\eta^2) + \sum_H E_o^4(52\eta^{12} - 48\eta^8 + 4\eta^4)(1-\eta^2)^2 + \sum_H E_o^2(80\eta^{10} - 16\eta^6)(1-\eta^2)^3 + \sum_H 20\eta^8(1-\eta^2)^4\right\} / \left\{\sum_H E_o^4\right\}^2. \quad (2.2.6)$$

2.3. *Moments of* R_2^n

A completely analogous analysis can be given for the normalized residual function R_2^n, being defined as

$$R_2^n \equiv \sum_H (E_o^2 - E_c^2)^2 / \sum_H E_o^4. \tag{2.3.1}$$

For incorrect models we obtain

$$\langle R_2^n; \mathscr{E}_o \rangle = \sum_H (E_o^4 - 2E_o^2 + 2) / \sum_H E_o^4 \tag{2.3.2}$$

$$\sigma^2(R_2^n; \mathscr{E}_o) = \sum_H (4E_o^4 - 16E_o^2 + 20) / \left\{\sum_H E_o^4\right\}^2. \tag{2.3.3}$$

For correct models one obtains

$$\langle R_2^n; \mathscr{E}_o \rangle = \left\{\sum_H E_o^4(\eta^4 - 2\eta^2 + 1) + \sum_H E_o^2(4\eta^2 - 2)\,(1 - \eta^2) + \sum_H 2(1 - \eta^2)^2\right\} / \sum_H E_o^4 \tag{2.3.4}$$

and

$$\sigma^2(R_2^n; \mathscr{E}_o) = \left\{\sum_H E_o^6(8\eta^6 - 16\eta^4 + 8\eta^2)(1 - \eta^2) + \sum_H E_o^4(52\eta^4 - 48\eta^2 + 4)(1 - \eta^2)^2 + \sum_H E_o^2(80\eta^2 - 16)(1 - \eta^2)^3 + \sum_H 20(1 - \eta^2)^4\right\} / \left\{\sum_H E_o^4\right\}^2. \tag{2.3.5}$$

3. Experimental verification

In the development of the theory we introduced a number of approximations. So a careful verification is in order. An ideal way of judging how the approximations affect the final formulas is to compare the numbers they produce with numbers obtained from a Monte Carlo simulation of the same problem. A Monte Carlo procedure is by its nature an *ab initio* method, since only the definition of R_2, the basic ideas of the theory and the definitions of correct and incorrect models are used. That is to say that we will use as *observed* structure a simulated structure with randomly placed atoms. Two test structures were constructed in $P1$ with a unit cell of 5 × 5 × 10 Å. The first contains ten equal atoms and is represented by the E_o values of 70 reflections in the range $0 \leq \theta \leq 10°$, denoted as data set (10,70). The second contains 100 atoms and is represented similarly by data set (100,70). The choice of the θ range and the size of the unit cell may seem peculiar at first sight. One has to remember, however, that the averages are taken over $\{\mathbf{r}_j^c\}$ [(1.6)], distributed evenly in the cell. The averaging is not performed over **H**. Consequently, in this homogeneous field of fractional coordinates, the quantity $2\pi\mathbf{H}\mathbf{r}_i$ covers the circle completely and evenly. Therefore, the actual θ range and cell size are immaterial, the data are given for future reference.

3.1. *Incorrect models of a simulated structure*

We selected data set (10,70) for the observed structure because if the theory fits to this data set it will certainly fit to sets containing larger numbers of atoms and reflections. The distribution $P(E_o)$ of the set showed a non-centric character. Then 12 000 independent, unrelated (incorrect) coordinate sets were generated for each model containing n ($n \leq 10$) atoms. The corresponding R_2 values were calculated from (2.1), leading to

$$\langle R_2(\text{exp}) \rangle = \sum_{12000} R_2(\text{exp}) / 12\,000 \tag{3.1.1}$$

and to

$$\sigma^2[R_2(\text{exp})] = \langle R_2^2(\text{exp}) \rangle - \langle R_2(\text{exp}) \rangle^2. \tag{3.1.2}$$

Averages and spread of R_2 are given in Table 1, column 1. They are compared (Table 1, column 2) with values calculated from (2.1.6) and (2.1.7). From the point of view of an experimental crystallographer the agreement between theory and experiment is very satisfactory, as it will serve his needs in practice.

For the theoretician, however, it is worthwhile to pinpoint the reasons for the small discrepancies. The most dominant factor in these differences was traced to

Table 1. *Comparison for incorrect models between* $\langle R_2(exp) \rangle$, $\sigma^2(R_2)$ *and theoretical values, obtained at various levels of approximation*

n	$\langle R_2(\text{exp}) \rangle$	$\langle R_2(t1) \rangle$	$\langle R_2(t2) \rangle$	$\langle \langle R_2 \rangle_{r^c} \rangle_{r^o}$
0	1·0000	1·0000	1·0000	1·0000
1	0·9062	0·9106	0·9062	0·910
2	0·8301	0·8389	0·8300	0·840
3	0·7716	0·7849	0·7716	0·790
4	0·7308	0·7486	0·7309	0·760
5	0·7077	0·7300	0·7079	0·750
6	0·7023	0·7292	0·7026	0·760
7	0·7146	0·7460	0·7150	0·790
8	0·7448	0·7806	0·7452	0·840
9	0·7922	0·8329	0·7931	0·910
10	0·8580	0·9029	0·8586	1·000
	$\sigma^2(\text{exp})$	$\sigma^2(t1)$	$\sigma^2(t2)$	$\langle \sigma^2 \rangle_{r^o}$
0	0·0000	0·0000	0·0000	0·0000
1	0·0000	0·0002	0·0000	0·0002
2	0·0004	0·0007	0·0004	0·0008
3	0·0010	0·0014	0·0010	0·0016
4	0·0018	0·0023	0·0018	0·0027
5	0·0027	0·0036	0·0027	0·0045
6	0·0040	0·0056	0·0040	0·0072
7	0·0060	0·0088	0·0061	0·0116
8	0·0093	0·0137	0·0092	0·0183
9	0·0143	0·0210	0·0142	0·0284
10	0·0218	0·0316	0·0216	0·0429

Table 2. *Values of* $\langle E_c^n(H)E_c^m(K)\rangle_{r^c}$ *and* $\langle E_c^n\rangle_{r^o}$ *for ten atoms,* $H(1,1,1)$ *and* $K(2,2,2)$

$\langle E_c^2(H)E_c^2(K)\rangle$	1·00225	$\langle E_c^2(H)\rangle$	1·00094	$\langle E_c^2(K)\rangle$	1·00059
$\langle E_c^2(H)E_c^4(K)\rangle$	1·90563	$\langle E_c^4(H)\rangle$	1·90852	$\langle E_c^4(K)\rangle$	1·90020
$\langle E_c^4(H)E_c^2(K)\rangle$	2·00385	$\langle E_c^6(H)\rangle$	5·18945	$\langle E_c^6(K)\rangle$	5·13349
$\langle E_c^4(H)E_c^4(K)\rangle$	3·97452	$\langle E_c^8(H)\rangle$	17·8715	$\langle E_c^8(K)\rangle$	17·5190
$\langle E_c^2(H)E_c^6(K)\rangle$	5·14338				
$\langle E_c^6(H)E_c^2(K)\rangle$	5·93261				

the form used for the intensity distributions. A more accurate intensity distribution, in which the actual number of atoms is explicitly taken into account is, for instance, given by Srinivasan & Parthasarathy (1976). See Appendix *B* for the moments. When they are incorporated into the theory the values for the averaged value and spread are obtained as given in Table 1, column 3. The discrepancies are decreased in absolute value to less than 10^{-3} in $\langle R_2\rangle$ and to about 10^{-4} in $\sigma^2(R_2)$. The next important reason for the remaining discrepancies is the convergence of the simulations. If values obtained after 2 000 000 trials are used instead of 12 000, one finds a perfect fit (differences less than 10^{-5}) for the average R_2 value. This is to be expected, since in the calculations of $\langle R_2\rangle$ for incorrect models no other approximations than the ones stated above were introduced. With respect to σ^2, the differences are then less than 10^{-4}. These last remaining differences must be caused by the oppression of the summations running over two indices (= neglect of correlation between lattice points). Indeed, when we incorporated the contributions of the latter summations we found that the last entry of Table 1, column 3 had changed from $\sigma^2(R_2)$ is 0·02164 into 0·02170, that is the difference from experiment had become about 10^{-5}.

For future developments it can be of interest to have some notions about the behaviour of the moments $\langle E_c^n(H)\,E_c^m(K)\rangle$. We calculated such values for a series of reflection pairs, each pair averaged over 2 000 000 independent ten-atom models. Some results are listed in Table 2, taking the pair $H(1,1,1)$ and $K(2,2,2)$ as an example. The calculations indicate that (i) the effect of correlation increases with increasing powers of n and m, (ii) if $K = 2H$, significant correlation effects show only if $n \geq m$, whereas if $H = 2K$ they show only if $m \geq n$. Furthermore, the effect of correlation diminishes with increasing size of the model or structure. Note that even for the small ten-atom models the correlation effects prove to be small.

3.2. *Correct models of a simulated structure*

The observed structure is represented by the E_o values of the data set (100,70). A number of atoms larger than in the incorrect case is chosen in order to be able to construct a sufficently large number of incomplete but correct models. This has the additional advantage that (2.2.3) can be more safely applied. In our opinion 100 atoms is a reasonable compromise between the need for asymptotically large models and the wish to keep the computations within bounds. We restricted ourselves to 70 reflections to save computer time, and, more importantly, because if the theory holds sufficiently for this small data set it will certainly hold for larger sets. Thus, 60 000 different modes were generated each containing n ($n \leq 100$) correct atomic positions, randomly chosen out of the original 100 atoms. It was secured that a particular atom occurred only once in a model. For each model an R_2 value was calculated from (2.1) leading as before to $\langle R_2(\text{exp})\rangle$ and $\sigma^2[R_2(\text{exp})]$, listed in Table 3, column 1. We needed 60 000 trials to converge these numbers to the fourth decimal place. Theoretical values were calculated from (2.2.5) and (2.2.6), see Table 3, column 2. The discrepancies between theory and experiment are small, less than 10^{-3} in $\langle R_2\rangle$ and less than 2×10^{-4} in $\sigma^2(R_2)$, in spite of the unusual small data set.

Table 3. *Comparison for correct models between experimental and theoretical values for* $\langle R_2\rangle$ *and* $\sigma^2(R_2)$

n	$\langle R_2(\text{exp})\rangle$	$\langle R_2(\text{th})\rangle$	$\langle\langle R_2\rangle_{r^c}\rangle_{r^o}$
0	1·0000	1·0000	1·00
25	0·7396	0·7394	0·75
50	0·4999	0·5006	0·50
75	0·2597	0·2613	0·25
100	0·0000	0·0000	0·00
n	$\sigma^2(\text{exp})$	$\sigma^2(\text{th})$	$\langle\sigma^2\rangle_{r^o}$
0	0·0000	0·0000	0·0000
25	0·0021	0·0023	0·0052
50	0·0052	0·0054	0·0079
75	0·0031	0·0032	0·0061
100	0·0000	0·0000	0·0000

This clearly shows that the step in the theoretical development in which we replace $\langle E_c^n;\mathscr{E}_o\rangle$ by $\langle E_c^n;E_o\rangle$ does not lead to serious errors, even for relatively small numbers of reflections. Obviously, in practical situations, with 1000 or more reflections in the data set, the influence of this approximation will undoubtedly be less important than the use of the asymptotic intensity distribution, (2.3.3).

3.3. *Correct models of an actual molecule*

It may be argued that simulated experiments using virtually ideal random models are not a fair test of the basic ideas underlying the practicability of the theory. A real structure may exhibit non-random characteristics, for example owing to relations between atoms caused by a near equality of bond lengths and angles, or to packing effects. None of these special features is so far modelled into the theory. Comparison against a real structure will show how serious the omission of such features is in practice.

As an example we took the structure of (1*S*,4*S*)-5-acetyl-3-oxo-2-oxa-5-azabicyclo[2.2.1]heptane (Lenstra, Petit & Geise, 1979). The compound,

$C_7H_9NO_3$,crystallizes in space group $P2_12_12_1$ with $a = 7{\cdot}10$, $b = 9{\cdot}20$ and $c = 11{\cdot}17$ Å. The asymmetric part was taken as an equal-atom structure containing 11 atoms in $P1$. From the atomic coordinates we calculated 1047 E_o values. We confined ourselves to testing only correct models because such models are the most interesting ones in practice as well as the most difficult ones to handle in theory. Each correct model of n atoms was constructed in its C_{11}^n different settings. The corresponding minimum, maximum and average R_2 values are indicated in Fig. 2 as crosses. Theoretical values for the moments of R_2 were calculated as before from (2.2.5) and (2.2.6). By drawing lines at $3\sigma[R_2(\text{th})]$ on either side of $\langle R_2(\text{th})\rangle$ and assuming $P(R_2)$ to be Gaussian, we have mapped out a 99·73% confidence area. The crosses fit perfectly well into this area. They do so despite the fact that in some instances the number of models is far from statistically large and therefore R_2(average), R_2(minimum) and R_2(maximum) are rather crude estimators for $\langle R_2(\text{exp})\rangle$ and $\sigma[R_2(\text{exp})]$, and despite the fact that the size of the models is small to use safely the asymptotical intensity distribution given by (2.2.3).

We recall once more that the major difference between the present theory and previous approaches to predict estimates of R_2 is the implementation of the set of observed data as information about the structure looked for. In previous theoretical investigations, *e.g.* Srinivasan & Parthasarathy (1976), an *a priori* estimate of R_2 could only be calculated if one assumed an infinite data set. This means, for instance, that there is only one average R_2 value independent of the actual data set, a phenomenon clearly contradicted by practical experience. This drawback is overcome in our present approach, in which the study of the behaviour of R_2 is based on the knowledge of the actual data set of the structure. We demonstrate this in Table 4, where

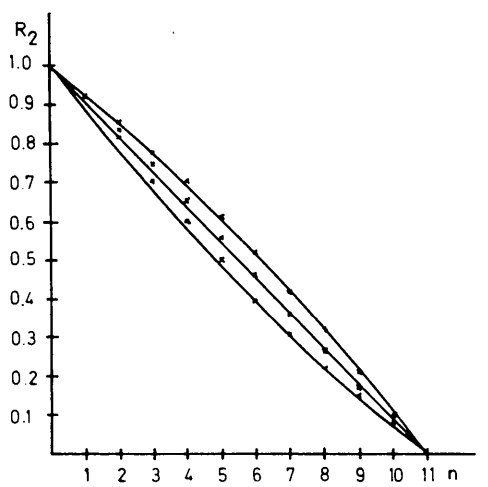

Fig. 2. Experimental minimum, average and maximum R_2 values for an actual structure of 11 atoms. The theoretical values are obtained from equations (2.2.5) and (2.2.6) assuming a Wilson distribution for the E_o values.

Table 4. *Comparison of observed and theoretical* $\langle R_2\rangle$ *values, showing the influence of the size of the data set*

$\langle R_2(\text{th})\rangle$ is calculated from equation (2.2.5). The values for an ∞ data set are obtained from equation (4.1.5).

	$0 \le \theta \le 10°$ 47 reflections		$0 \le \theta \le 30°$ 1045 reflections		∞ reflections
n	$\langle R_2(\text{exp})\rangle$	$\langle R_2(\text{th})\rangle$	$\langle R_2(\text{exp})\rangle$	$\langle R_2(\text{th})\rangle$	$\langle R_2\rangle$
1	0·9514	0·9413	0·9177	0·9099	0·9091
2	0·8830	0·8666	0·8322	0·8199	0·8181
3	0·7972	0·7782	0·7433	0·7298	0·7273
4	0·6967	0·6789	0·6513	0·6393	0·6364
5	0·5853	0·5719	0·5566	0·5485	0·5455
6	0·4676	0·4608	0·4600	0·4572	0·4545
7	0·3489	0·3495	0·3628	0·3656	0·3636
8	0·2354	0·2427	0·2662	0·2739	0·2727
9	0·1341	0·1452	0·1722	0·1821	0·1818
10	0·0527	0·0624	0·0826	0·0907	0·0909
11	0·0000	0·0000	0·0000	0·0000	0·0000

average values of R_2 are listed using different numbers of reflections in the data set. The size of the data set is governed by arbitrarily chosen maximum θ values. As an example, we took the azabicyclo[2.2.1]heptane derivatives in the treatment given above.

The numbers (Table 4) clearly show that variation in the average R_2 value as a function of the size of the data set is correctly followed by the variation in the theoretical results and that the latter values are superior, particularly for small data sets, to estimates of R_2 based on the assumption of an infinite data set. This also suffices to show that the non-randomness of a structure can be taken into account in our new approach. For it is this non-randomness which, when the θ limit is altered, induces changes in the moments $\langle E_o^n(H)\rangle_H$ and thus in the moments of R_2. Although the θ limit does not appear explicitly as a parameter in the present theory, its effects are taken care of simply by using the observed intensities. In this context it is of interest to note that (2.2.3) is strictly speaking a function of the point in reciprocal space chosen, even though this does not show in the actual form of the formula.

4. Discussion

4.1. *Generalization of the moments of* R_2

Because (2.1.6–7) and (2.2.5–6) contain explicitly the summations over the set of observed reflections, they give $\langle R_2\rangle$ and $\sigma(R_2)$ specific for an actual structure. It is, however, sometimes of interest to generalize the present results and to have formulas at one's disposal independent of the structure at hand. As stated before this requires knowledge of $\langle\langle R_2\rangle_{r^c}\rangle_{r^o}$ and $\langle\sigma^2(R_2)\rangle_{r^o}$. An exact evaluation of these quantities is unfortunately too complex, because it involves averages of the type

$$\langle E_o^n(H)/\sum_H E_o^m(H)\rangle_{r^o}. \qquad (4.1.1)$$

If, however, we are satisfied with an approach in which the end justifies the means, (2.1.6–7) and (2.2.5–6) can easily be rewritten to give approximations for the quantities stated above. This is done by replacing

$$\sum_H E_o^n = \mathscr{H}\langle E_o^n\rangle_H \quad \text{by} \quad \sum_H E_o^n = \mathscr{H}\langle E_o^n\rangle_{r^o}, \qquad (4.1.2)$$

in which $\mathscr{H}$ is the number of reflections. The operation is strictly speaking only admissable in the limit of an infinite data set. Substitution of the Wilson distribution (2.1.3) results in the following formulas, for incorrect models:

$$\langle\langle R_2\rangle_{r^c}\rangle_{r^o} \simeq \eta^4 - \eta^2 + 1 \qquad (4.1.3)$$

$$\langle\sigma^2(R_2)\rangle_{r^o} \simeq (5\eta^8 - 4\eta^6 + 2\eta^4)/\mathscr{H}, \qquad (4.1.4)$$

while for correct models:

$$\langle\langle R_2\rangle_{r^c}\rangle_{r^o} \simeq 1 - \eta^2 \qquad (4.1.5)$$

$$\begin{aligned}\langle\sigma^2(R_2)\rangle_{r^o} \simeq \{&(48\eta^{14} - 96\eta^{10} + 48\eta^6)(1-\eta^2)\\ &+ (104\eta^{12} - 96\eta^8 + 8\eta^4)(1-\eta^2)^2\\ &+ (80\eta^{10} - 16\eta^6)(1-\eta^2)^3\\ &+ 20\eta^8(1-\eta^2)^4\}/4\mathscr{H}. \qquad (4.1.6)\end{aligned}$$

It is of importance to note that (4.1.3) and (4.1.5) give the first moments of R_2 identical to those given, for instance, by Petit, Lenstra & Van Loock (1981). The approach of these authors thus leads to a limiting case of the present theory. The last columns of Tables 3 and 4 give the results of (4.1.3–6). They demonstrate that conclusions about the path of $\langle R_2\rangle$ and $\sigma(R_2)$ drawn from the three particular structure examples can be qualitatively transferred to the average structure. On the other hand, the average structure qualitatively predicts the correct order of magnitude of $\langle R_2\rangle$ and $\sigma(R_2)$ for a particular structure. Obviously one should not expect such estimates to be highly accurate.

4.2. *Behaviour in the limit of an infinite data set*

In the limit of an infinite data set our results must converge to those of previous investigations: (i) for a chosen size of the model $\langle R_2\rangle$ is a constant and (ii) $\sigma^2(R_2) = 0$. This is easily seen for $\langle R_2\rangle$, (4.1.3, 5). Regarding $\sigma^2(R_2)$, (2.4), one distinguishes two types of summations, those over one set of indices and those over two such sets. Terms of the first type individually converge to zero, since the nominators contain $\mathscr{H}$ and the denominators $\mathscr{H}^2$, see (4.1.4, 6). We have seen that summations of the second type cancel two by two, unless there exists a relation between the indices of the reflection pairs. In other words, the running indices of the summations H and K are no longer independent and the remnants of the double summations collapse to single summations. Their number of terms is now of the same order as the number of terms in the summations running over one index. Thus the contributions of correlated reflections will in absolute value go to zero at the same rate as the other terms do.

4.3. *Comparison of R_2 with R_2^n*

Parthasarathi & Parthasarathy (1975) have stated that R_2^n might be a better indicator function for small models than R_2, mainly because the difference R_2^n(incorrect) − R_2^n(correct) is larger than the corresponding

Table 5. *Comparison of the estimated resolving power between correct and incorrect models for R_2 and R_2^n*

n/N denotes the fraction of the known part of the model and $\mathscr{H}$ the number of reflections.

n/N	$S(R_2)\sqrt{\mathscr{H}}$	$S(R_2^n)\sqrt{\mathscr{H}}$
0·1	0·012	0·010
0·2	0·025	0·021
0·3	0·040	0·033
0·4	0·056	0·046
0·5	0·075	0·062
0·6	0·096	0·080
0·7	0·119	0·100
0·8	0·143	0·125
0·9	0·168	0·155
1·0	0·192	0·192

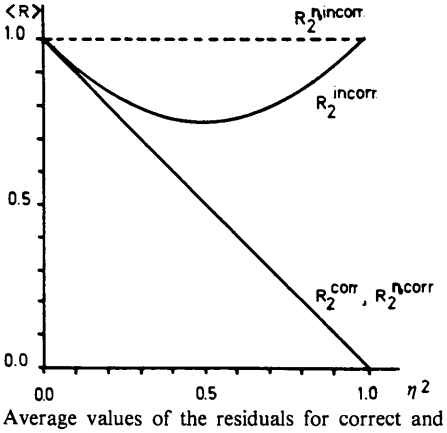

Fig. 3. Average values of the residuals for correct and incorrect models.

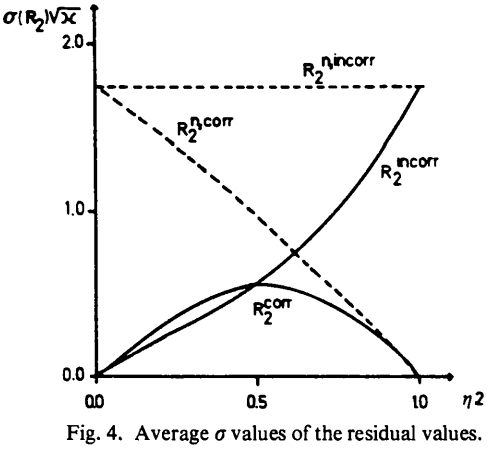

Fig. 4. Average σ values of the residual values.

difference in R_2 values. It is, however, evident that these quantities are bad estimators of the power of the residuals for automation procedures, because no information about the higher moments is used. A better estimator of the resolving power is given by S defined as

$$S \equiv \frac{\langle R_2(\text{incorrect})\rangle - \langle R_2(\text{correct})\rangle}{3\{\sigma[R_2(\text{incorrect})] + \sigma[R_2(\text{correct})]\}}. \quad (4.3.1)$$

To check their statement in a way independent of a particular structure one must use the moments obtained from (4.1.3–6) as parameters in $S(R_2)$ and the corresponding parameters in $S(R_2^n)$. The latter are easily obtained from (2.2.3–5) by replacing, as in the previous paragraph, $\sum_H E_o^n$ by $\mathscr{H}\langle E_o^n\rangle$. Noting that S is proportional to the square root of the number of reflections, we have tabulated (Table 5) the values of $S\sqrt{\mathscr{H}}$ for the two indices in space group $P1$. The numbers show that there is no significant difference in resolving power between R_2 and R_2^n. This is not an unexpected result since the definition of R_2^n differs only in the omission of the trivial weighting factor η^2 from the definition of R_2. Since in our opinion such a scale factor cannot influence the resolving power, it cannot be concluded from the numbers listed in Table 5 that R_2 is a better discriminator function than R_2^n.

As a final remark, one can conclude that a faithful indicator of the resolving power should somehow contain third and higher moments of the residual functions.

Figs. 3 and 4 show the path of $\langle\langle R_2\rangle_{r^c}\rangle_{r^o}$, $\langle\langle R_2^n\rangle_{r^c}\rangle_{r^o}$, $\langle\sigma(R_2)\rangle_{r^o}$ and $\langle\sigma(R_2^n)\rangle_{r^o}$, respectively.

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

APPENDIX *A*

The moments of (2.2.3) can be obtained as

$$\langle E_c^\mu; E_o\rangle = \frac{2\eta_o^2}{\eta_o^2 - \eta_c^2}\exp\left(-\frac{\eta_c^2 E_o^2}{\eta_o^2 - \eta_c^2}\right) \times \int_0^\infty E_c^{\mu+1}\exp\left(-\frac{\eta_o^2 E_c^2}{\eta_o^2 - \eta_c^2}\right) \times I_0\left(\frac{2\eta_o\eta_c E_o E_c}{\eta_o^2 - \eta_c^2}\right) dE_c. \quad (A.1)$$

Using a generalization of Weber's first exponential integral (Bateman, 1953, II):

$$\int_0^\infty J_\nu(at)\exp(-p^2t^2)\,t^{\mu-1}\,dt = \frac{\Gamma(\frac{1}{2}\nu + \frac{1}{2}\mu)\left(\frac{1}{2}\frac{a}{p}\right)^\nu}{2p^\mu\,\Gamma(\nu+1)}\,{}_1F_1\left(\tfrac{1}{2}\nu + \tfrac{1}{2}\mu;\ \nu+1;\ -\frac{a^2}{4p^2}\right)$$
$$\mathrm{Re}(\nu+\mu) > 0, \quad a \in C, \quad \mathrm{Re}(p^2) > 0, \quad (A.2)$$

together with the identity (Bateman, 1953, II)

$$I_\nu(z) = \exp(-i\tfrac{1}{2}\nu\pi)J_\nu[z\exp(i\pi/2)] \quad -\pi < \arg(z) \le \pi/2 \quad (A.3)$$

and Kummer's first transformation (Bateman, 1953, I):

$${}_1F_1(a;b;x) = e^x\,{}_1F_1(b-a;\,b;\,-x), \quad (A.4)$$

we can write equation (*A*.1) as

$$\langle E_c^\mu; E_o\rangle = \Gamma\left(\frac{\mu}{2}+1\right)\left(\frac{\eta_o^2 - \eta_c^2}{\eta_o^2}\right)^{\mu/2} \times {}_1F_1\left(-\frac{\mu}{2};\,1;\,-\frac{\eta_c^2 E_o^2}{\eta_o^2 - \eta_c^2}\right). \quad (A.5)$$

For the even moments $\mu = 2n$ holds:

$$\langle E_c^{2n}; E_o\rangle = n!\left(\frac{\eta_o^2 - \eta_c^2}{\eta_o^2}\right)^n {}_1F_1\left(-n;\,1;\,-\frac{\eta_c^2 E_o^2}{\eta_o^2 - \eta_c^2}\right). \quad (A.6)$$

As can be seen from the definition of ${}_1F_1$ (Bateman, 1953, I),

$${}_1F_1(a;b;x) = 1 + \frac{a}{b}\frac{x}{1!} + \frac{a(a+1)}{b(b+1)}\frac{x^2}{2!} + \ldots, \quad (A.7)$$

the moments reduce in our case to an nth-degree polynomial in x, because a is a negative integer. For instance, the second moment can be written as

$$\langle E_c^2; E_o\rangle = \eta^2 E_o^2 + (1 - \eta^2). \quad (A.8)$$

APPENDIX *B*

A number of authors have derived intensity distributions and intensity moments (Srinivasan & Parthasarathy, 1976; Shmueli & Kaldor, 1981; Shmueli & Wilson, 1981; Karle & Hauptman, 1953; Hauptman & Karle, 1953) whose domain of validity stretches beyond the one obtained by Wilson (1949). The latter distribution is based on the assumption of an asymptotically large number of atoms. For space group $P1$ the moments valid for a small number of atoms in the model take the form:

$$\langle E_c^2\rangle = 1$$
$$\langle E_c^4\rangle = 2 - 1/n$$
$$\langle E_c^6\rangle = 6 - 9/n + 4/n^2$$
$$\langle E_c^8\rangle = 24 - 72/n + 82/n^2 - 33/n^3. \quad (B.1)$$

References

BATEMAN, A. (1953). *Higher Transcendental Functions*, Bateman Manuscript Project, parts I and II. New York: McGraw Hill.
GIACOVAZZO, C. (1980). *Direct Methods in Crystallography*. London: Academic Press.
HAUPTMAN, H. & KARLE, J. (1953). *Acta Cryst.* **6**, 136–141.
KARLE, J. & HAUPTMAN, H. (1953). *Acta Cryst.* **6**, 131–135.
KLUG, A. (1958). *Acta Cryst.* **11**, 515–543.
LENSTRA, A. T. H. (1974). *Acta Cryst.* A**30**, 363–369.
LENSTRA, A. T. H. (1979). *Bull. Soc. Chim. Belg.* **88**, 359–368.
LENSTRA, A. T. H., PETIT, G. H. & GEISE, H. J. (1979). *Cryst. Struct. Commun.* **8**, 1023–1029.
LINDGREN, B. W. (1976). *Statistical Theory*, 3rd ed. New York: Macmillan.
NEUTS, M. F. (1973). *Probability*. Boston: Allyn & Bacon.
PARTHASARATHI, V. & PARTHASARATHY, S. (1975). *Acta Cryst.* A**31**, 38–41.
PETIT, G. H., LENSTRA, A. T. H. & VAN LOOCK, J. F. (1981). *Acta Cryst.* A**37**, 353–360.
ROHATGI, V. K. (1976). *An Introduction to Probability Theory and Mathematical Statistics*. New York: Wiley.
SHMUELI, U (1982). *Acta Cryst.* A**38**, 362–371.
SHMUELI, U. & KALDOR, U. (1981). *Acta Cryst.* A**37**, 76–80.
SHMUELI, U. & WILSON, A. J. C. (1981). *Acta Cryst.* A**37**, 342–353.
SRINIVASAN, R. & PARTHASARATHY, S. (1976). *Some Statistical Applications in X-ray Crystallography*. Oxford: Pergamon.
VAN DE MIEROOP, W. (1979). PhD thesis (in Dutch), Univ. of Antwerp.
WILSON, A. J. C. (1949). *Acta Cryst.* **2**, 318–321.
WILSON, A. J. C. (1950*a*). *Acta Cryst.* **3**, 397–398.
WILSON, A. J. C. (1950*b*). *Research*, **3**, 48.
WILSON, A. J. C. (1969). *Acta Cryst.* B**25**, 1288–1293.
WILSON, A. J. C. (1978). *Acta Cryst.* A**34**, 986–994.

Acta Cryst. (1983). A**39**, 562–565

Moments of the Probability Density Function of R_2 Approached *Via* Conditional Probabilities. II. Completely Correct and Completely Incorrect Models in Space Group $P\bar{1}$

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* 15 *April* 1982; *accepted* 14 *February* 1983)

Abstract

With the help of conditional probabilities formulas are derived for the first and second moment of R_2 as a function of the size of the model. The formulas are valid in the space group $P\bar{1}$ for two extreme cases, *viz* completely correct and completely incorrect models. Incorporation of the observed intensities enables one to obtain accurate *a priori* estimates of $\langle R_2 \rangle$ and $\sigma(R_2)$. The theory agrees very well with simulated experiments.

1. Introduction

In automated structure determinations of single crystals, one may use the mathematical residual function R_2 to discriminate between correct and incorrect models. The applicability of R_2 as a discriminator function increases sharply if one has at one's disposal an *a priori* evaluation of its average value and spread. That is to say, in order to be able to use statistical decision methods in an automated analysis one needs to know for the crystallographic situation at hand either the probability distribution of the residual R_2 or the moments of this distribution.

Until recently, the assumption of an infinite data set allowed only the prediction of the first moment (mean value) but precluded the evaluation of the higher moments. The break-through came with the introduction of the calculus of conditional probability. In part I (Van Havere & Lenstra, 1983) we laid down the general principles of the new theory and derived expressions for the first and second moments of the probability density function of the residual R_2 for completely correct and completely incorrect structure models in space group $P1$. The results for $P1$ may serve as a model for all primitive non-centrosymmetric space groups. In this paper we will derive similar expressions for space group $P1$, which may serve as a parent for all primitive centrosymmetric space groups.

2. Moments of R_2

Throughout this work E_o will refer to the observed magnitude of the normalized structure factor belonging to a structure containing N atoms in the asymmetric unit. Likewise E_c will refer to the calculated magnitude of an E value of a model containing n atoms in the asymmetric unit. The definition of R_2 is

$$R_2 \equiv \sum_H (E_o^2 - \eta^2 E_c^2)^2 / \sum_H E_o^4 \qquad (2.1)$$

0108-7673/83/040562-04$01.50