Acta Cryst. (1978). B**34**, 2527–2533

Computer Retrieval and Analysis of Molecular Geometry. II. Variance and Its Interpretation

BY PETER MURRAY-RUST

Department of Chemistry, University of Stirling, Stirling, Scotland

AND RICHARD BLAND

Department of Sociology, University of Stirling, Stirling, Scotland

(*Received* 17 *January* 1978; *accepted* 9 *March* 1978)

The causes of variance in the observed geometrical parameters of a molecular fragment are outlined. It is shown how, with the help of multivariate techniques such as factor analysis, the variance can be partitioned between experimental errors and structural variation.

Introduction

In the previous paper (Murray-Rust & Motherwell, 1978*a*; hereafter MM*a*) we outlined how molecular geometry can be retrieved from the Cambridge Crystallographic Data File, how the reliability of individual cases can be estimated and how, in principle, statistical analysis can show up trends and patterns in the data to increase our understanding at an empirical level. In this paper we outline the causes of variation in molecular geometry and how this variation can be analysed. In particular we shall distinguish carefully between variation due to crystallographic errors or artefacts and *structural variation* due to chemical and crystallographic effects.

We assume that a system file of geometry for a given molecular fragment has been created (see MM*a*) and that initial screening and searching is complete. (In the course of the subsequent analysis we may have to revise screens, especially the raw-data screen.) The file contains data for *n cases* (one crystal structure may provide several cases if there is more than one fragment in the asymmetric unit) and *m variables*, p_{ij} (bond lengths, angles and other physical quantities, see Appendix). As explained earlier (MM*a*) we shall use the *SPSS* package (Nie, Hull, Jenkins, Steinbrenner & Bent, 1975) to carry out all operations on the file and its data.

Variance in geometrical parameters

The geometry of a molecular fragment in different crystals usually shows considerable variation due to two causes: *structural variation* and *experimental errors* (Fig. 1). Depending on the chemical nature of the fragment and the experimental techniques used, either of these may predominate. In 211 phosphate groups, tabulated by Baur (1974), the variation in P—O length is large (more than 0·2 Å) and cannot be explained by the reported e.s.d.'s (average value 0·007 Å). It has been shown (Murray-Rust, Bürgi & Dunitz, 1978) that 95% of the variance can be convincingly attributed to chemical properties of the phosphate group and the influence of crystal environments. In a study of 11 alanyl fragments from the file with low e.s.d.'s no convincing explanation could be found for the variance in bond lengths and angles except known experimental errors (Murray-Rust, 1977). (This result is unremarkable since considerable forces are needed to deform saturated carbon skeletons significantly.)

When structural variation is small, *mean values* (μ_j) for the parameters can be found; the larger the number of cases the smaller the *standard error* of these means. Mean values for the geometry of molecular fragments are important to chemists and crystallographers (*e.g.* in testing theoretical calculations, comparison with other experimental methods, and model-building). When structural variation is large, however, determination of the means of the parameters constitutes only part of the process, *analysis of variance* being at least as important. The following paper (Murray-Rust & Motherwell,

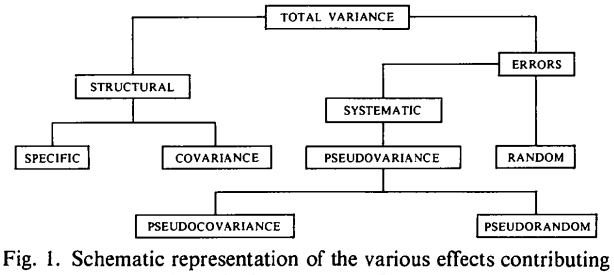

Fig. 1. Schematic representation of the various effects contributing to the total variance in the geometry of a molecular fragment.

1978*b*; hereafter MM*b*) shows that nucleosides have a large range of conformations and the idea of mean geometry for the fragment is almost meaningless. Even within one conformation there are variations of up to 20° in torsion angles and the covariance of many of the parameters is important. For sophisticated model-building (*i.e.* an accuracy of 0·1 Å or better) a knowledge of covariance is almost as important as a knowledge of means.

In the nucleoside study most of the variance can be seen to be structural, although large experimental errors are present in some cases. Unfortunately, for many common fragments both structural variation and experimental errors make significant contributions to the overall variance. The main part of this paper is devoted to showing how the results of statistical analysis can be reliably attributed to structural variation. To explain the causes and analysis of structural variation we shall discuss an ideal situation where experimental errors are negligible. Later we discuss the effect of errors and the methods available for treating them.

Analysis of structural variation

We have suggested, then, that the variance in a battery of measures applied to a series of fragments can be conceptualized as coming from two sources: firstly, from experimental error, and, secondly, from real differences in molecular structure. We are not, however, simply interested in partitioning total variance into these two portions – our goal is, rather, to *interpret* structural variance and to see how it can be related to the underlying crystallographic and chemical properties of the fragment being studied. This idea has been briefly discussed in MM*a*, in which we noted a similarity between the analysis of socio-economic census data, for example, and data from a file of molecular geometry. Having noticed this similarity, it is perhaps not surprising to find that one of our principal analytical techniques was first developed in the field of the social sciences.

Factor analysis* is a method of transforming a multivariate set of measures in order to explore the possibility of reducing the set to a smaller number of underlying variables or *factors*. Thus, for example, the answers to batteries of questions in intelligence tests can be reduced to a smaller number of independent variables which are often taken to be components of human ability. The transformation of the original set of measures into the desired factors is a mechanical one, performed by algorithm; the task of the analyst is to see what light is cast by the answers on his original theoretical problem. This he does by considering the possibility of correspondences between the derived factors and properties or processes of the material under study. In so doing he is guided by two things: firstly, the existing body of theory in his discipline and, secondly, the composition of the derived factors in terms of the input measures. Thus, for example, Timms (1975), in a study of the towns of the Central Region of Scotland, found wide variations between them in terms of a large set of socio-economic variables. Factor analysis reduced this set to three main factors, which he conceptualized as representing respectively the 'Social Rank', 'Family Life Cycle' and 'Social Deprivation' aspects of the differences between the towns.

We take up the mathematical properties of the method later in this section, but as an introduction one can say that it can be seen as an analysis of *covariation*. Suppose we have a set of variables which, on inspection, turns out to be made up of two sub-sets, with strong inter-correlations *within* each set, and weak correlations between the sets. In such a situation the algorithm will find two factors, one corresponding to each set. In principle each variate should be normally distributed but in practice the methods are fairly robust and so long as the distribution of cases (in multi-dimensional variable space) is unimodal and deviations from linearity are not great the method works well.

For each variable the mean (μ_j), variance and standard deviation (σ_j) are computed. The deviation from the mean for any case is then expressed as a *z-score*

$$z_{ij} = (p_{ij} - \mu_j)/\sigma_j. \qquad (1)$$

The z_{ij} form the *standardized data matrix*, **Z**, which is used for all subsequent analysis, the merit of which is that most derived quantities (*e.g.* factor scores) will then have zero mean and unit variance. Moreover if all p_j are normally distributed then so will be the scores. It is therefore important to check on the distribution of the *z*-scores for each variable by computing the *skewness* and *kurtosis* which are zero if the distribution is normal.

There is no reason in general why structural variables should be normally distributed, as shown by the multimodal scatter of torsion angles in ribonucleosides (MM*b*). For small deviations from a mode, however, it is not an unreasonable assumption in practice. [Indeed the principle of structural correlation (Murray-Rust, Bürgi & Dunitz, 1975) suggests a Gaussian distribution if the deviations are small enough for the energy of distortion to be represented harmonically (Murray-Rust, Bürgi & Dunitz, 1978).] Variates with severe skewness or kurtosis (but still giving an overall unimodal distribution) can be transformed by logarithmic or similar methods to give new variables with a more nearly normal distribution (see, for example, Rummel, 1970). For multimodal distri-

* The commonly-used generic term. In fact, we use the *Principal Components* variant of the technique.

butions it will be necessary to separate the cases into groups or *clusters** and to analyse each one separately.

If all the variates now have a nearly normal distribution the correlation matrix, **R**, is calculated where

$$\mathbf{R}_{mm} = \frac{1}{n} \mathbf{Z}^T_{mn} \mathbf{Z}_{nm}. \quad (2)$$

(The individual r_{jk} are the Pearson correlation coefficients of p_j with p_k.) If there are no missing values in the data matrix **Z** (as will be the case for output from *GEOM*) then **R** is Gramian (or semi-positive definite) with no negative eigenvalues.

Factor analysis of molecular geometry may involve group-theoretical considerations (which will be discussed elsewhere) but when the fragment cannot show any symmetry the simple treatment here is appropriate. The m factors **F** are formed from the eigenvalues λ and eigenvectors **E** of **R** by

$$\mathbf{F}_{mm} = \lambda^{1/2} \mathbf{E}_{mm} \quad (3)$$

where λ is the diagonal matrix of the eigenvalues. The factors satisfy the relation:

$$\mathbf{F}\mathbf{F}^{\mathrm{T}} = \mathbf{R}_{mm} \quad (4)$$

and are linear combinations of the original variables. Factor analysis thus corresponds to an orthogonal rotation of axes so that the factors (in decreasing order) explain as much variance as possible.

Ideally some of the eigenvalues will be zero, in which case only p ($<m$) factors will be significant, resulting in a reduction in the dimensionality of the problem. This has been achieved in studies of visible spectra (Bulmer & Sturvell, 1975) but is unlikely to apply to molecular geometry [except where there are mathematical constraints, *e.g.* if angles are linearly related (MM*b*)]. Typically, however, some of the eigenvalues may be small enough to be regarded as insignificant, effectively reducing the dimensionality. In applications of factor analysis in other fields, a wide variety of rules have been proposed to assist in the decision as to whether a factor is important enough to warrant its retention. Among these are *Kaiser's criterion*, which retains factors whose eigenvalues are greater than unity, and the *scree test*. This test involves an inspection of the plot of factor number against eigenvalue. Such a plot typically shows eigenvalues falling rapidly in magnitude as one progresses through the first few factors, followed by a lower rate of decrease for the remainder, and has been likened to a cliff with scree at its base. Factors corresponding to the scree area are rejected. Plots which do not show this feature, but instead resemble a steady slope, are often taken as being indicative of a situation where dimensional reduction is not possible. In crystallographic applications, however, the e.s.d.'s are of great help in making this decision, and we consider this question later in this paper.

We can then compute the matrix of factor scores, **S**, from

$$\mathbf{S}_{np} = \mathbf{Z}_{nm} \mathbf{F}_{mp}$$

where **F** contains only the significant factors.

Since factor analysis merely represents a transformation of the data its value lies in the reification of the factors, and this is as good a guide as any to determining their significance. The mathematical basis of factor analysis parallels closely that of normal coordinate analysis (especially in the symmetry aspects and redundant coordinates) and it is to be expected that the most important factors will be closely related to soft normal coordinates. Factor analysis is only strictly appropriate for linear combinations of the parameters but if the distortions are not too large (*i.e.* up to 0·2 Å) this approximation works well. Second-order effects (corresponding to cubic force constants) can be revealed if two-dimensional scattergrams of the factor scores show slightly non-linear plots (Murray-Rust, Bürgi & Dunitz, 1978).

Since the factors are orthogonal, their scores in particular cases can be independently examined to help in reification. The cases can be sorted by scores (SORT CASES) and the chemical formulae of those with scores outside the 5- and 95-percentiles (*i.e.* $|S_{ij}| > 2$) plotted graphically (*PLOD*). This may show chemical features responsible for the exceptional scores which can then be included in a revised screen (either to eliminate compounds likely to show this factor, or, conversely, to include them specifically). Factor analysis on the rescreened data should result in an even smaller number of factors and hence a clearer separation of the chemical causes of molecular variability. Alternatively the crystal environment of the fragments with high scores can be examined* and a quantitative measure of crystal packing forces obtained. At present, however, rescreening on the basis of the environment of the fragment is not possible.

This analysis of one-dimensional scores applies, of course, to other quantities (*e.g.* z-scores, regression residuals). A danger is that the statistics can be seriously affected by *outliers*, isolated points not conforming to the normal distribution. This is a likely occurrence since crystal structures are not analysed at random but undertaken because of their interest, and unusual geometries are therefore quite common! Outliers due to structural effects are of enormous importance in formulating theories of molecular variability. For example, there are tens of thousands of C—N lengths on file but only about 10 are in the range

* If these clusters cannot be seen from one- or two-dimensional statistics, multi-dimensional cluster analysis is in principle appropriate. It has not yet been used for molecular geometry.

* It is planned that *GEOM* will be able to plot this information on the line-printer.

1·6–2·8 Å. Consideration of these 10 cases, however, led Bürgi, Dunitz & Shefter (1973) to develop the idea of reaction pathways. Unfortunately, outliers are also commonly caused by severe experimental error and we tackle this problem in the next section.

Errors

Much of the variance in the molecular geometry of the compounds on file is due to experimental errors of every sort known to crystallography. An obvious example is the variation in C—H lengths which is almost entirely due to the difficulty in locating H atoms precisely by X-rays. These errors make the identification and analysis of structural variance much more difficult than has so far been outlined. We believe, however, that so long as the importance of errors is realized before undertaking an analysis of molecular geometry on file worthwhile results can be obtained in a large number of cases. We outline the most important types of error and show their effect on the statistical procedures already described.

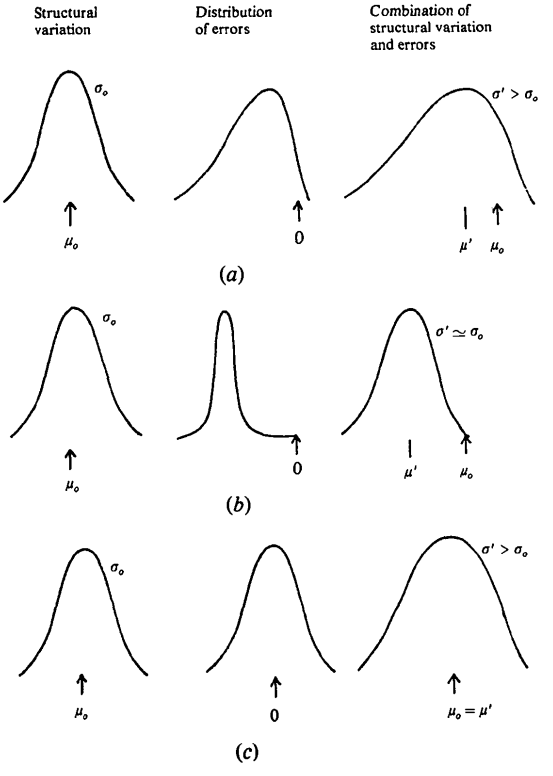

Fig. 2. The effect of some crystallographic errors on the mean and variance of a bond length. Structural variation is normally distributed with mean μ_o and standard deviation σ_o. (*a*) Thermal motion: the error distribution is left-skew and results in a slightly skew observed distribution with lower mean (μ') and increased variance (σ'). (*b*) Non-coincidence of electronic and nuclear density: the error distribution is narrow and only the mean of the observed distribution is affected. (*c*) Random errors (absorption, disorder, *etc.*): the mean is unaffected but the variance is increased.

Errors (Fig. 2) affect parameters in two main ways: by altering the mean and increasing the variance. (The skewness and kurtosis may also be affected but in general this is less important.) Changes in the mean will not affect our analysis of molecular variability, but increased variance can be serious. In particular if most of the variance is due to errors it will be very hard to estimate the structural variance. Typically, errors are classified as *random*, which affect the variance, and *systematic*, which affect the mean, but the latter term can be misleading in the present context. Bond lengths derived from the file are affected by thermal motion, a serious systematic error, but present in different amounts in each structure. The mean bond length taken over all cases is lowered but the variance is also increased, a phenomenon we shall call *pseudorandom* error. Because the error can sometimes be very large the skewness may also be somewhat altered (Fig. 2*a*). An even more serious consequence of variable amounts of systematic error we call *pseudocovariance*. If an error affects more than one parameter (as is often the case with thermal motion or disorder) then varying amounts of the error will suggest that the parameters are covarying. There is no simple statistical way of eliminating pseudocovariance and it must be kept constantly in mind when factors are being interpreted. Another case of pseudocovariance is the uncertainty of position of light atoms in a framework of heavy atoms, causing covariance among several bonds and angles. In many cases this can be detected because it produces precise geometrical relationships; for example, it was shown for phosphates that Baur's (1974) model of P rattling in an O_4 tetrahedron did not yield the observed covariance of bond lengths and angles (Murray-Rust, Bürgi & Dunitz, 1978).

It must be stressed that the Cambridge Data Centre makes no attempt to analyse crystallographic errors in any structure and reports only the authors' estimate of random errors (e.s.d.'s). The term *error-free set* refers only to the fact that the bond lengths calculated from the atomic coordinates are not inconsistent with the published values or with normally accepted values. It is merely a guarantee that the data set is free from most gross typographical errors.

Random errors

Ideally if all random and pseudorandom errors can be accurately estimated then the structural variance can be calculated. For accurate structures this is possible if there are few gross errors.

Gross random errors

Catastrophic errors can occur in structure solution, refinement and publication and in general will cause

large random errors in parameters. Typical examples are: incorrect structures, *e.g.* molecules in the wrong position in the cell; mis-publication of cell dimensions; refinement with an inadequate amount of data. (We have encountered an alanyl fragment on file with all angles at C_α less than 70°!) Some of these errors may lead to bond lengths which result in Cambridge flagging the structure as an error set. Most of the other cases will have large e.s.d.'s (category 4 for the AS flag) and will probably show up as widely separated outliers. A policy of manual inspection of outliers will mean most of these can be safely identified and rejected.

Other random errors

In crystal structure analysis it is fortunate that most systematic and random errors in the reflexion data are transmitted to the positional parameters as normally distributed errors, which can be estimated from the inverse normal matrix (Cruickshank, 1967). It is, of course, these estimates which give rise to the standard deviations of bond lengths coded on file. Though some refinement methods tend to underestimate the e.s.d.'s, there are now several half-normal plots (Hamilton, 1974) in the literature which suggest that for well-refined structures with good weighting schemes and data the e.s.d.'s are a fair estimate of the effect of experimental error on the variability of the positional parameters.

Apart from gross errors we shall assume that the e.s.d.'s on file give a reasonable estimate of the random variability of positional parameters. These e.s.d.'s, although only crudely coded on file, can be used to help decide how much of the variance is structural, and hence how many factors are significant. Bearing in mind that e.s.d.'s may be underestimated we shall be cautious about giving credibility to too many factors unless there is good crystallographic justification.

Systematic errors

Known systematic crystallographic errors can be roughly divided into those which have no effect on the variance and those which cause pseudovariance (Fig. 2).

Variance unaffected

Few systematic errors are present in the same amount in each structure determination. The non-coincidence of the centroid of electron density with nuclear position is probably the most important common effect, especially for H atoms, where the systematic error in C—H lengths (~0·1 Å) is much greater than the variance caused by this particular error.

Thermal motion

This is probably the most serious cause of pseudo-random error (and to a lesser extent of pseudo-covariance). Although a few structures on file have corrected bond lengths these depend on the model used and there is no real alternative to using uncorrected values. For certain types of compound (*e.g.* perchlorates) thermal motion makes any attempt at analysis of structural variance totally impossible, but these situations are well known and can be avoided. Searches can usually be designed so that thermal-motion errors are most unlikely to be >0·05 Å. (In many organic structures the thermal-motion correction will be quite small, 0·002–0·010 Å; *i.e.* the standard deviation from this effect is only about 0·003 Å.) Even if the distribution is somewhat skew, the main effect will be to add a small amount to the pseudovariance, much of which will be specific and will not affect the factors. Severe effects will almost certainly produce outliers and if these are automatically inspected it is easy to refer to the original literature in the few cases that are involved. (Since outliers caused by large structural variance are so valuable in formulating theories it is important to verify that they are not caused by thermal motion.)

Thermal motion can also produce pseudocovariance in some cases. Suppose a benzene ring librates about the C(1)—C(4) axis, resulting in an apparent shortening of C(1)—C(2) but not C(2)—C(3), then the angle C(1)—C(2)—C(3) will be apparently increased, causing pseudocovariance with C(1)—C(2).

Disorder

This is a serious cause of pseudorandom errors and pseudocovariance, and can be very difficult to detect even in the original experiment. Consider the case of the dimerization of carboxylic acids through hydrogen bonding, where there is considerable negative covariance between the C=O and C—O lengths. This could easily be ascribed to structural variance since it makes good chemical sense. The dimer is almost symmetrical, however, and disorder is frequently present, probably in varying amounts from structure to structure. Much (if not all) of the covariance must therefore be attributed to disorder rather than structural variation.

Pseudosymmetry

If this occurs in a structure, the refinement necessarily involves highly covariant atomic parameters, and hence pseudocovariance in the internal molecular coordinates. We optimistically expect this effect to be fairly uncommon (or at least unimportant) in organic structures, since few crystallographic publications make any mention of it!

Artefacts

Constrained refinement of, for instance, rigid benzene rings may totally invalidate analysis of structural variance. Unfortunately, the data file does not flag this technique but in general it is used only in relatively few systems and normally only when accuracy is limited or there is disorder. It is possible to construct derived-data screens to detect some occurrences but it is best to avoid cases where it might be used.

Statistical analysis

It is clear from the nature of crystallographic errors that even if they were accurately recorded on file it would be impossible to calculate the structural variance of a fragment with any precision. A rough estimate is, however, extremely valuable particularly in deciding how many factors, if any, are significant. Unfortunately, the file only gives e.s.d.'s for C—C bonds, but other lengths and angles normally have e.s.d.'s in rough proportion [*e.g.* σ(CCC) (in degrees) is usually about 50–100 × σ(C—C) Å]. When the standard deviations of the variables are computed they can be compared with the e.s.d.'s and if they are of comparable magnitude the structural variance can be assumed to be small.

Standard deviations larger than e.s.d.'s in at least some of the parameters can be due to the following causes: (*a*) underestimation of e.s.d.'s, (*b*) gross random errors, (*c*) pseudovariance from systematic errors, (*d*) structural variance.

To determine whether the variance is really due to structural effects we suggest the following procedure:

(i) Compute *z*-scores for all variables and see if there are any outliers. These should be examined manually (they will only constitute 1–2% of the data) to see if there are unusual structural effects or serious errors. Outliers with genuine structural variance may be of great help in outlining the ways in which molecular geometry can vary.

(ii) Carry out factor analysis on the standardized data matrix. The eigenvalues of the correlation matrix, in decreasing order, are proportional to the amount of variance explained by each factor. In general there will be no zero eigenvalues but if the proportion of variance due to random errors is roughly known the number of factors describing structural variance can be estimated, *i.e.* choose p so that

$$\frac{1}{m}\sum_{1}^{p}\lambda_i \simeq 1 - [(\text{e.s.d.})/(\text{observed s.d.})]^2.$$

(iii) Calculate scores for the p factors, and examine them manually to see if any are due to outliers produced by errors. If not, attempt to reify the factors; in general, the larger the factor the less its coefficients will be affected by error, and the easier it should be to reify.

(iv) If the factors cannot be convincingly reified then it must be assumed that the covariance is not structural and is due to crystallographic errors, either systematic or random (underestimated). Before reaching this conclusion, however, the molecular formulae of the cases with high factor scores should be examined for common chemical features. If the factors can be reified (as a chemical or crystallographic effect) then the scores give a quantitative measure of this effect in each case. These scores may then be related to the molecular or crystal environment of the fragment.

Conclusion

Automatic analysis of molecular geometry from the data file is clearly enormously faster than searching and transcribing the original literature. The ability to bypass papers in journals is, however, a mixed blessing and highlights the danger of using data files uncritically. It is unlikely that for any particular analysis the results can be confidently published without looking at some, at least, of the original experimental data. However, the statistical procedures will pinpoint the few cases it is important to check (in MM*b* it was only necessary to refer to two papers out of nearly a hundred). Thus, without sacrificing critical judgement, the crystallographer is able to carry out analyses of molecular geometry very rapidly and accurately. An example of such a study, carried out in a short time using the data file, is given in MM*b*.

APPENDIX

The use of *SPSS* for screening and statistical analysis

The *SPSS* system is so widely available that it is almost certainly implemented on any computer which is large enough for the data file. The package has a complex housekeeping system for files and variables and can be directed to carry out simple arithmetic and logical operations in a subset of Fortran. The raw-data input consists of a table of an indefinite number of cases composed of up to 500* variables (which can be real numbers or alphabetic characters), and is totally compatible with the tabulated output from *GEOM*. Arithmetic and logical operations can be carried out on any variable (including alphabetic ones such as atom names), and trigonometric and similar functions are available. New variables can be created and their values computed for all cases or a subset by a Fortran-like IF statement. The cases are held in a *system file* (an $n \times m$ matrix) to which cases (n') or variables (m') can be

* In some versions, 1000.

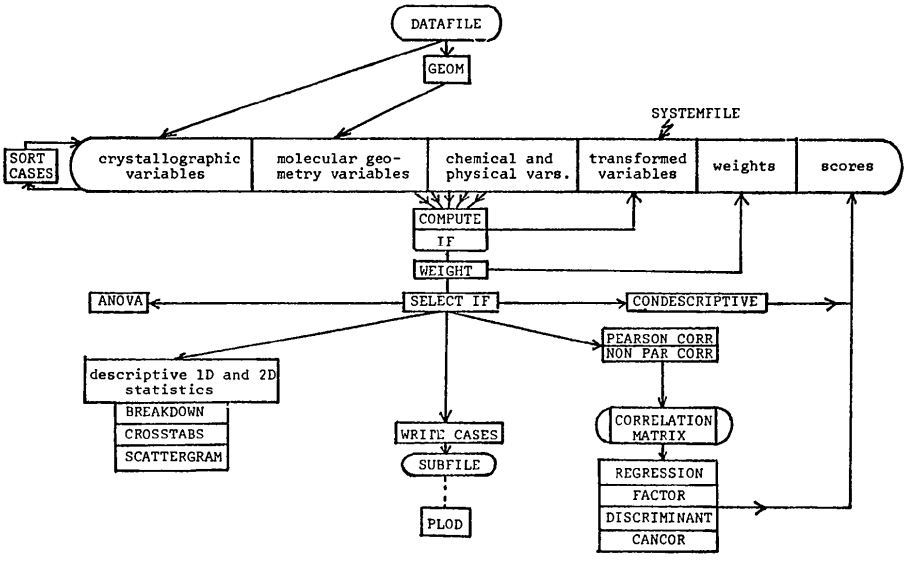

Fig. 3. Relationship of system file to *SPSS* procedures, including data transformation, selection and statistical analysis. (Raw-data, stereochemical and derived-data screening can occur at SELECT IF, and a new system file could be created by WRITE CASES if required.) Note that the system file can be repeatedly enlarged by output from *SPSS* procedures. *PLOD* and *GEOM* are CCDC programs.

added at any time. The cases can be sorted according to the value of any variable. If values for variables in a particular case are missing default values can be entered. (The treatment of data with missing values is carefully controlled in the statistical procedures.) Weights for any case can be read in or generated. For large files a random sample can be taken for analysis. Input and output options are flexible.

Any number of the statistical procedures can be called and scores can be written to file for further analysis. The SELECT IF statement allows selection of a subset of the data fulfilling a certain condition. (We use this statement as the raw-data and stereochemical screens, sometimes in conjunction with WRITE CASES.) The main statistical procedures (Fig. 3) are:

CONDESCRIPTIVE, FREQUENCIES	One-dimensional statistics, including histograms and subsets.
CROSSTABS	Contingency tables for variables which take discrete values.
T-TEST	Test of significance.
PEARSON CORR, NONPAR CORR	Bivariate correlation for interval and ordinal variates.
SCATTERGRAM	Line-printer scatter diagram. This is enormously useful for finding relationships. It can be used to draw sections of a three-dimensional scattermap.
PARTIAL CORR	Partial correlation.
REGRESSION	Multiple regression including stepwise addition of variables.
ANOVA, ONEWAY	Analysis of variance and covariance.
DISCRIMINANT	Discriminant analysis.
FACTOR	Factor analysis including principal components.
CANCORR	Canonical correlation.

References

BAUR, W. H. (1974). *Acta Cryst.* B**30**, 1195–1215.

BULMER, J. T. & STURVELL, H. F. (1975). *Can. J. Chem.* **53**, 1251–1254.

BÜRGI, H. B., DUNITZ, J. D. & SHEFTER, E. (1973). *J. Am. Chem. Soc.* **95**, 5065–5066.

CRUICKSHANK, D. W. J. (1967). In *International Tables for X-ray Crystallography*, Vol. II, 2nd edition. Birmingham: Kynoch Press.

HAMILTON, W. C. (1974). In *International Tables for X-ray Crystallography*, Vol. IV. Birmingham: Kynoch Press.

MURRAY-RUST, P. (1977). Unpublished results.

MURRAY-RUST, P., BÜRGI, H. B. & DUNITZ, J. D. (1975). *J. Am. Chem. Soc.* **97**, 921–922.

MURRAY-RUST, P., BÜRGI, H. B. & DUNITZ, J. D. (1978). *Acta Cryst.* B**34**, 1793–1803.

MURRAY-RUST, P. & MOTHERWELL, S. (1978*a*). *Acta Cryst.* B**34**, 2518–2526.

MURRAY-RUST, P. & MOTHERWELL, S. (1978*b*). *Acta Cryst.* B**34**, 2534–2546.

NIE, N. H., HULL, C. H., JENKINS, J. G., STEINBRENNER, K. & BENT, D. H. (1975). *Statistical Package for the Social Sciences*, 2nd edition. New York, London: McGraw-Hill.

RUMMEL, R. J. (1970). *Applied Factor Analysis*. Evanston: Northwestern Univ. Press.

TIMMS, D. W. G. (1975). *The Stirling Region*, pp. 253–283. The British Association.