Statistical analysis of X-ray data for olivine

By F. P. Agterberg

Geological Survey of Canada, Ottawa

[Taken as read 30 January 1964]

Summary. A diagram of the relationship between d_{174} and % forsterite in olivines is presented in fig. 1 of the preceding paper by Jambor and Smith. In this note, statistical aspects of Jambor and Smith's data are considered in detail and compared to equivalent data by Yoder and Sahama (1957). This analysis leads to the conclusion that the relationship of mol. % Fo and d_{174} or d_{130} can be expressed as a straight line for Fo $>$ 30 %. Olivines with Fo $<$ 10 % are not related to this straight line.

J. L. JAMBOR and C. H. Smith (this vol., p. 730) have studied the variation of the d_{174} spacing of olivines with their chemical composition; their fig. 1 shows that the d_{174} determinative curve is approximately a straight line with a possible curvature near the fayalite end. The first hypothesis to test is whether the relationship between mol % Fo and d_{174} for the 26 points is continuous and may be represented by a power series in d_{174}, and if one or more higher-order terms of this series are justified.

Neither of the variables may be considered as independent for a linear regression analysis. Therefore, there are two possible linear regression equations, for % Fo on d_{174} and for d_{174} on % Fo; they are, respectively:

$$\% \text{ Fo} = 4446{\cdot}29-4264{\cdot}55\, d_{174} \quad \text{and} \quad \% \text{ Fo} = 4458{\cdot}98-4276{\cdot}89\, d_{174}.$$

The difference between the equations is small and it appears that regardless of which assumption is used for further tests the statistical conclusions would be similar. The conclusions based on the use of either equation may therefore be applied to the true relationship between d_{174} and % Fo, which are both liable to a certain unknown error. The following analysis contains d_{174} as the independent variable.[1]

[1] When mol. per cent. Fo is considered as the independent variable, the estimated residual variances for linear, quadratic, and cubic fit are respectively $2{\cdot}19\times10^{-7}$, $1{\cdot}06\times10^{-7}$, and $0{\cdot}97\times10^{-7}$, and for linear and quadratic fit for data with Fo $>$ 30 %: $0{\cdot}88\times10^{-7}$ and $0{\cdot}92\times10^{-7}$ Å^2. When fitting curves to the data with Fo $>$ 30 %, the estimated residual variance increases with higher-order fits. This is the effect of the decrease of the number of degrees of freedom on the approximately constant unaccountable sum of squares.

The second-order equation is: % Fo = $-25295{\cdot}61+53431{\cdot}07\ d_{174} -27978.62\ d^2{}_{174}$. The standard error of the coefficient of the second-order term amounts to 6034·76 and a t-test shows that the second-order term is significantly different from zero.

The analysis of variance presented in table I also demonstrates the justification of a high-order term. The F-ratio of the estimated variance due to the quadratic effect and the unaccountable variance is significant at the 99 % confidence level of the F-distribution for 1 and 22 degrees of freedom. The second-order fit reduces the unaccountable or residual variance from 3·990 (linear fit) to 2·152. The effect of the third-order fit (reduction to 1·960) is seen to be negligible by the F-ratio test.

Table I. Analysis of variance, range Fo 0 to 100 %
(Cf. Quenouille, 1952, p. 96, table 6. 3*a*)

	% Fo for d_{174} data			% Fo for d_{130} data		
Variation	d.f.	Sum of Squares	Estimated Variance	d.f.	Sum of Squares	Estimated Variance
Linear effect	1	33092·758	33092·758	1	29951·755	29951·755
Quadratic effect	1	47·376	47·376	1	36·685	36·685
Cubic effect	1	5·272	5·272	1	14·883	14·883
Overall effect	3	33145·406	—	3	30003·323	—
Unaccountable	22	43·114	1·960	25	50·698	2·028
Total effect of *d*	25	33188·520	—	28	30054·021	—

The analysis of variance is based on the assumption of continuity for the relationship between % Fo and d_{174}. More satisfactory results are obtained when discontinuity instead of continuity is assumed: the residual variance is further reduced from 2·152 (quadratic fit) or 1·960 (cubic fit) to 1·386 when the data with Fo $>$ 30 % are fitted by a straight line. The analysis of variance for the data with Fo $>$ 30 % is given in table II; it will be seen that the effect of the quadratic fit is negligible.

It is concluded that if we assume continuity a second-order fit for all the data is satisfactory, or if we assume discontinuity consisting of a break in the fayalite side of the curve a first-order fit is satisfactory and the degree of fit is improved considerably.[1]

By analysis of variance it has been shown that a straight line is the best fit for the data for Fo $>$ 30 %. Further analysis is necessary

[1] Another explanation would be that the residual variance is very large in the fayalite portion of the curve. This hypothesis is rejected because the residual variance for linear fit to the data with Fo $<$ 10 % amounts to 2·736, and the F-ratio for this value and the residual variance for 20 points is 1·97, which is less than 2·93, the 95 % confidence level for the F-distribution with 4 and 18 degrees of freedom.

because the errors of the variables need consideration, and because the above test by analysis of variance lacks a significance level. Two significance tests will be given to demonstrate that the data for Fo $<$ 30 % are not related to the straight line of the first 20 points for Fo $>$ 30 %.

Both variables d_{174} and % Fo are subject to error. An estimate of the error of d_{174} can be made from the differences in 2θ, which has been measured by two different observers (Jambor and Delabio) in $m = 18$ out of 26 cases (see the preceding paper, table II). The variance of 2θ is estimated[1] by $S^2_{2\theta} = \sum_{i=1}^{m} \frac{1}{2}\Delta_i^2 \big/ m = 0{\cdot}002056$, where Δ_i is the difference between the two measurements of 2θ. It follows that $S_{2\theta} = 0{\cdot}0453° = 7{\cdot}912 . 10^{-4}$ radian. The corresponding standard error in d_{174} is found by application of the theory of the propagation of error (Deming, 1943, pp. 37–48) to the equation $2\, d_{174} \sin \theta = 1{\cdot}93579$, giving $S_d \approx 0{\cdot}1875 . S_{2\theta} = 1{\cdot}48 . 10^{-4}$. The d_{174} data used for the determinative curve are, for the larger part, values for two observations, so that S_d must be corrected by a factor somewhat smaller than $1/\sqrt{2}$. The ultimate estimate of S_d is therefore slightly larger than $1 . 10^{-4}$; this is a minimum estimate, since other sources of error may be present above the one detected by comparing measurements by two observers. A maximum estimate is obtained by assuming that % Fo is free of error, giving after regression analysis S_d $2{\cdot}96 . 10^{-4}$. The true standard error is thus between 1 and $3 . 10^{-4}$.

TABLE II. Analysis of variance, range Fo 30 to 100 %

	% Fo for d_{174} data			% Fo for d_{130} data		
Variation	d.f.	Sum of Squares	Estimated Variance	d.f.	Sum of Squares	Estimated Variance
Linear effect	1	7847·224	7847·224	1	6105·646	6105·646
Quadratic effect	1	0·005	0·005	1	0·082	0·082
Overall effect	2	7847·229	—	2	6105·728	—
Unaccountable	17	24·940	1·467	19	16·045	0·844
Total effect of d	19	7872·170	—	21	6121·773	—

If, on the other hand, d_{174} is assumed to be free of error, a maximum estimate of the error in the mol. % Fo data is obtained, amounting to 1·18 %. In this case, the true standard error is probably $\frac{1}{2}$–1 %.

For calculating the least square fit of associated points with both variables subject to error, the ratio of the variances of these variables

[1] Each combination of the two values provides an independent estimate of $S^2_{2\theta}$ equal to $\{(\frac{1}{2}\Delta_i)^2+(\frac{1}{2}\Delta_i)^2\}/(2-1)$; the overall estimate is the average of m of these individual values.

should be known (Kummell's equation, see Deming, p. 184). Although, in the present case, their order of magnitude is known, the values cannot be established with enough precision for further analysis along these lines. Fortunately, the true regression line must lie between the estimates of regression of % Fo on d_{174} and regression of d_{174} on % Fo, and these equations differ very little, being:

$$\% \text{ Fo} = 4144{\cdot}99 - 3970{\cdot}14\, d_{174} \quad \text{and} \quad \% \text{ Fo} = 4157{\cdot}92 - 3982{\cdot}76\, d_{174}.$$

TABLE III. Values of d_{174} calculated from: *a*, the regression of % Fo on d_{174}; *b*, the regression of d_{174} on % Fo; *c*, the mean of these regressions. Lt, ±95 % confidence limits for % Fo from the mean of the regressions

	d_{174}			
a	*b*	*c*	% Fo	Lt
1·01885	1·01887	1·01886	100	0·87
1·02389	1·02389	1·02389	80	0·57
1·02893	1·02892	1·02892	60	0·72
1·03397	1·03394	1·03395	40	1·16

It is reasonable to consider the average of these lines as a satisfactory estimate[1] of the regression line:

$$\% \text{ Fo} = 4151{\cdot}46 - 3976{\cdot}45\, d_{174}. \tag{1}$$

Values of d_{174} calculated from these equations for selected % Fo values are given in table III.

The coefficients of these equations may not be rounded off further if d_{174} is to remain accurate to the third decimal place. However, the number of digits does not represent their accuracy; the standard error of the coefficient of d_{174}, b, may be estimated by:

$$S_b = \surd\{(\tfrac{1}{2}\delta b)^2 + b^2(1-r^2)/(n-2)r^2\} = 53{\cdot}22,$$

where δb is the difference of the coefficients of the two regressions and r is the correlation coefficient of d_{174} and % Fo. The standard error of the constant term of equation (1) amounts to 54·55, and the corrected value for the residual variance is 1·420 (as against 1·386 for the linear regression of % Fo on d_{174}, see table III).

The ±95 % confidence limits for % Fo of table III have been calculated by the formula $\text{Lt}_{\text{true \% Fo}} = \pm S.t\surd(R+1/n)$, where t is taken

[1] The 'reduced major axis' (Imbrie, 1956), which is simply the geometric mean of the two regressions, gives 3976·44±50·10 for the coefficient of d_{174}.

from a table of double-sided confidence limits, for $n-2$ degrees of freedom, and $R = (d-\bar{d})^2 \big/ \sum_{i=1}^{n} (d_i-\bar{d})^2$. This[1] confidence limit defines a band within which we expect to find 95 % of the true points, on error-laden measurements of which equation (1) was based.

Two procedures may be applied for testing whether the 6 points at the fayalite end of the determinative curve are related to equation 1 or not.

The average slope of the regression lines for all 26 points is

$$b' \pm s_{b'} = -4270{\cdot}72 \pm 47{\cdot}30.$$

Let us assume that the slope of this line should be the same as that of equation (1), the line for the first 20 points only, which has slope $b = -3976{\cdot}45$; the quotient $|b-b'|/s_{b'} = 6{\cdot}221$ is a measure of the probability of this assumption, a large value suggesting that the difference of slopes is real; at the 95 % and 99 % significance levels of Student's t-distribution for 25 degrees of freedom the quotient should be 2·093 and 2·861 respectively, and we conclude that the slopes b' and b are significantly different.

Again, in fig. 1A, the 95 % confidence belt for testing whether further observations are related to the line of equation 1 has been plotted for points with Fo $<$ 30 %. This belt defines a band within which 95 % of any new data may be expected to fall: $\mathrm{Lt}_{\mathrm{new}} = \pm S.t\surd(1+R+1/n)$, where t and R are as defined above. The six values at the fayalite end of the curve fall below this belt when they are tested individually, while combinations of further observations such as the three values for pure fayalite should be tested in relation to a confidence belt narrower than that of fig. 1A (Quenouille, p. 65).

For testing whether the first 20 points individually are related to the fitted straight line, a third confidence limit must be used:

$$\mathrm{Lt}_{\mathrm{obs}} = \pm S.t\surd(1-R-1/n);$$

this limit defines the band within which 95 % of the observations used in deriving the equation should lie. It is plotted in fig. 1A, and it will be seen that one point (5 % of 20) falls just outside the belt; this exact agreement is, of course, fortuitous.

It is concluded that statistical analysis suggests that there is a linear fit for the data with Fo $>$ 30 %, but in the fayalite portion a break probably occurs that cannot be described more precisely for lack of observations in the 7 to 30 % Fo part of the curve.

[1] For this and the other confidence limits defined below see Quenouille, 1952, pp. 64–66.

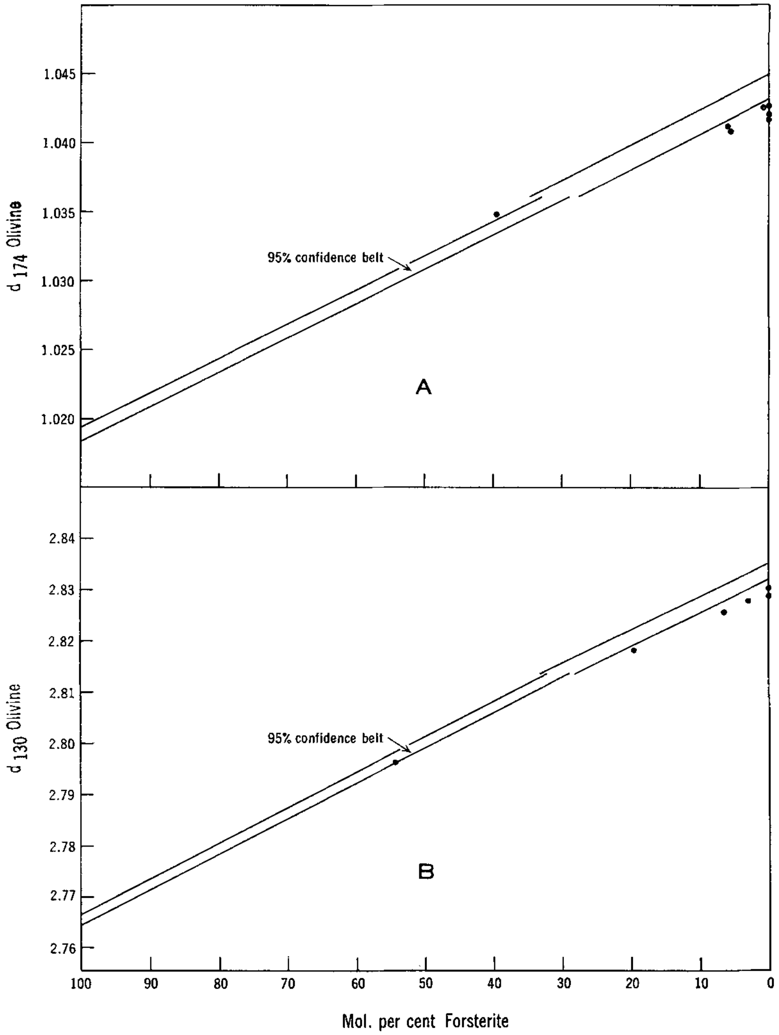

FIG. 1. 95 % confidence belts for d_{174} and d_{130} olivine data. Formulae for calculating the belts are given in text, and only samples falling outside the belts are indicated. Note that samples with less than 10 % Fo fall below belts in both cases. Samples over Fo_{30} represent the normally expected scatter for 95 % confidence limits.

Analysis of the d_{130} data

The preceding analysis for the d_{174} data may be repeated for the d_{130} data of Yoder and Sahama (1957). The analysis of variance[1] leads to similar conclusions (tables I and II). The residual variance amounts to 3·79 for linear fit to all 29 points. It is reduced to respectively 2·52 and 2·03 for quadratic and cubic fits. If the 22 points for Fo $>$ 30 % are analysed separately, the residual variance becomes 0·81 for linear fit, which is a considerable improvement.

The average regression equation, equivalent to equation (1), for olivines with Fo $>$ 30 % is:

$$\% \text{Fo} = 4088{\cdot}89 - 1442{\cdot}44\, d_{130}. \qquad (2)$$

The corrected standard errors of constant term and coefficient of d_{130} are respectively 46·41 and 16·69. The 95 % confidence limits for Fo equal to 100, 80, 60, and 40 % are respectively 0·69, 0·41, 0·57, and 0·97 %. The corrected residual variance of 0·819 is less than the value of 1·420 for the d_{174} data, indicating that the diffractometer method is more precise than the powder camera method.

Extreme observations are tested in fig. 1B, which is comparable to fig. 1A. The d_{130} values for synthetic fayalite and forsterite belong to the point clusters for natural olivines. There is no reason to assume that there are different determinative curves for synthetic and natural olivine.

Acknowledgement. Charles H. Smith and John L. Jambor are heartily thanked for discussion and critical reading of the manuscript.

References.

DEMING (W. E.), 1943. Statistical adjustment of data. John Wiley & Sons, New York.

IMBRIE (J.), 1956. Bull. Amer. Mus. Nat. Hist., vol. 108, art. 2.

QUENOUILLE (M. H.), 1952. Associated measurements. Butterworth's Scientific Publications, London.

SIMONEN (S.), 1961. Bull. Comm. géol. Finlande, vol. 196, p. 371.

YODER (H. S.) and SAHAMA (Th. G.), 1957. Amer. Min., vol. 42, p. 475.

[1] The 29 data have been obtained by adding Yoder and Sahama's two values for synthetic fayalite and forsterite and a new determination by Simonen (1961) to the original 26 values of Yoder and Sahama. For 26 points only, the discussed residual variances are respectively 2·87, 2·02, 1·79, and 0·78.