CORRELATION AND ITS APPLICATION TO CERTAIN PROBLEMS IN THE SUGAR INDUSTRY

By W. O. CHRISTIANSON.

Many members of our Association who have come across the statistical terms "correlation coefficient" and "regression coefficient" would like to have a clearer insight into what these constants mean. They have usually neither the facilities nor the time necessary to garner this knowledge by the laborious study of textbooks on statistical method. The writer therefore has attempted, in as simple a way as possible, to explain something of these two constants, to show how they are calculated and to illustrate the importance of their application to certain types of problems in our industry.

We often wish to know if two factors are correlated because we have reason to believe that variation in one factor causes, or is associated with, variations in the other. Let us take, for example, a study of the effect of fibre per cent. cane on Java Ratio. For this purpose we would collect data, either factory figures or the results of experiments, writing down the fibre figures with the corresponding Java Ratio figures obtained by analysis of the same lots or samples of cane. Having collected these pairs of figures, we need some system whereby we can judge them. One way is to put them on a graph, plotting Java Ratio against fibre per cent. cane. Another popular way is to classify them into groups of certain fixed ranges of, say, fibre content, as shown in the following table.

	Fibre per cent. cane.	Java Ratio.
	17.4	75.2
	17.3	73.9
	17.3	74.6
	17.0	75.9
Averages	17.25	74.90
	16.9	75.3
	16.6	75.5
	16.5	73.0
	16.1	74.6
Averages	16.52	74.60
	15.8	75.1
	15.5	77.4
	15.4	77.0
	15.3	75.3
	15.1	76.5
Averages	15.42	76.26
	14.2	77.1
Averages	14.20	77.10
Totals	226.4	1056.4
Arithmetic average or mean	16.172	75.457

On examining the above figures, are we justified in contending that these two factors are inversely correlated? In other words, does the variate, Java Ratio, increase as the variate, fibre per cent. cane, decreases?

If you wish to support the contention that there is a correlation, you might say that at an average of 17.25 per cent. fibre the Java Ratio averages 74.9, while at 15.42 per cent. fibre the Java Ratio average was 76.26, or that a drop of 1.8 per cent. in fibre caused a rise of 1.36 in Java Ratio. If you wish to deny the relationship, then you point to the exceptions, and might argue that a fibre content of 15.8 per cent. gives a lower Java Ratio than does one of 17.0 per cent., and in general that the relationship claimed to exist is just a chance one. Plotting the figures on a graph does not seem to help, as is shown by the dots in graph A.

The Total Correlation Coefficient.

Many years ago, mathematicians developed a method whereby we can lift this matter of correlation above the level of mere argument and, perhaps, guesswork. This method is to calculate the total correlation coefficient (universally designated r) from these figures, and test its significance. Both these terms are better explained after showing how the correlation is done.

Suppose, for simplicity, we had the following figures:—

Sucrose per cent. cane.	Purity of Juice.
12	88.0
14	89.5
16	91.0

Let us call the variate, sucrose per cent. cane, x, and the other variate, purity, y, and set out the table on page 2.

Notice how we can simplify the calculations thus:—

$$S(x')^2 = 12^2+14^2+16^2 - \frac{(\text{Grand total})^2}{\text{No. of observations}}$$

$$= 596 - \frac{42^2}{3}$$

$$= 8.0$$

$$\text{and } S(y')^2 = 88.0^2+89.5^2+91.0^2 - \frac{268.5^2}{3}$$

$$= 4.5$$

$$S(x'y') = (12\times 88.0)+(14\times 89.5)+(16\times 91.0) - \frac{42\times 268.5}{3}$$

$$= +6.0$$

Sucrose per cent. or x	Mean or $\bar{x}$	Difference from mean or x′	Square of differences or $(x')^2$	Purity y	Mean $\bar{y}$	Difference y′	Square of difference $(y')^2$	Products of differences (x′y′)
12	14.0	−2.0	4.0	88.0	89.5	−1.5	2.25	+3.0
14	14.0	±0.0	0.0	89.5	—	±0.0	0.0	—
16	14.0	+2.0	4.0	91.0	—	+1.5	2.25	+3.0
Sums S(x)= 42			S(x′)²= 8.0	S(y)= 268.5			S(y′)²= 4.5	S(x′y′)= +6.0

Correlation coefficient

$$r = \frac{S(x'y')}{\sqrt{S(x')^2 \times S(y')^2}}$$

$$= \frac{+6.0}{\sqrt{8.0 \times 4.5}} = \frac{+6.0}{\sqrt{36}}$$

$$= +1.0$$

The sign of our correlation coefficient is positive, showing that increase in one variate is associated with increase in the other and that, similarly, decrease in one is associated with decrease in the other. If the sign had been negative, increase in one variate would be associated with decrease in the other variate.

The value of our correlation coefficient 1.0 shows that the correlation is perfect. If it had been zero there would have been no correlation. Values of 1.0 and zero are almost never obtained in practice.

Having demonstrated the calculation of the total correlation coefficient, let us revert to our problem of Java Ratio and fibre. For this we calculate :—

Sum of squared deviations from mean of fibre content :

$$S(x')^2 = (17.4^2 + 17.3^2 + \ldots + 14.2^2) - \frac{226.4^2}{14}$$

$$= 12.5486.$$

Sum of squares for Java Ratio :

$$S(y')^2 = 20.3143.$$

Sum of products of deviations :

$$S(x'y') = -10.0971.$$

Total correlation coefficient :

$$r = \frac{-10.0971}{\sqrt{12.5486 \times 20.3143}}$$

$$= -0.6324.$$

This correlation coefficient is derived from our 14 pairs of figures. Our 14 pairs of figures are, however, only a random sample of a large number of pairs of such figures obtained during an experiment, and it is but a short step to consider them a sample of an infinite number or "population" of such pairs of figures. What is the possibility of our having obtained by accident a sample from a population which is not correlated ? In other words, what are the odds against our sample correlation coefficient of −0.6324 being obtained from a sample drawn from a population the true correlation coefficient of which is zero?

The statistical test is conveniently provided by R. A. Fisher's table of r, included in his work, "Statistical Methods for Research Workers," published by Oliver & Boyd. It is also published by permission in most other modern statistical textbooks. Referring to Professor Fisher's table, we find that, for 12 "degrees of freedom" or (14−2) pairs of observations, if a coefficient is as high as ±0.5324 there is but one chance in 20, or a probability of 5 in 100 (P=.05), that the correlation coefficient of the population is zero. If a coefficient derived from 14 pairs of observations is as high as ±0.6614, then the chances are 99 to 1 (or P=.01) against the population coefficient being zero.

Our coefficient is −.6324, so that it is above the P=.05 level and we call it "significant." If it had been above the P=.01 level we would call it "highly significant." Conventionally, a correlation coefficient below the level P=.05 is considered "not significant," which means *not proved*. From the table of r, it will be seen that the smaller the size of the sample the higher the value of r required for significance, and similarly the larger the sample the smaller the value of r required for significance.

Points which must be stressed about the correlation coefficient are:—

1. The sign of the coefficient shows whether two variates are directly or inversely associated. A coefficient of +.5 is as good as one of −.5.
2. The coefficient is used to measure the strength of an association, or the closeness with which changes in one variate keep in step with changes in another.
3. Its value lies between ±1.0, which indicate perfect association, and zero, which indicates no association.
4. The test of a correlation coefficient is to estimate the probability of its being obtained from an infinite population of zero correlation. Values below the P=.05 level of Fisher's table are considered non-significant, and the correlation unproved. Values above the P=.05 level are called "significant," and values above P=.01 level, "highly significant."

5. The correlation coefficient is a pure number, and it does not follow that one coefficient larger than another indicates a stronger association. The coefficient must be judged merely by the value of the probability indicated. Special tests are necessary to find if one coefficient is significantly better than another.
6. A significant correlation coefficient shows merely that two variates are *associated*, not that variation in one *causes* changes in the other. If variation in one causes variation in the other they are, of course, correlated, but correlation is not necessarily evidence of causation. For instance, a significant positive correlation between sucrose per cent. cane and purity of juice would indicate that both these figures are high in mature canes and lower in less mature cane, and not that a high sucrose percentage causes a high purity of juice.

The Regression Coefficient.

The correlation coefficient measures the strength of an association between two variates. The regression coefficient measures the amount of change in one variate associated with unit change in another. It is calculated :—

Linear regression coefficient of y on x,

$$byx = \frac{S(x'y')}{S(x')^2}$$

(There is also the regression of x on y, but we will omit that as its inclusion would only confuse matters.)

Regression coefficient of Java Ratio on fibre per cent. cane, or

$$byx = \frac{-10.0971}{12.5486}$$
$$= -0.8046$$

This means that change of 1 per cent. in fibre content is accompanied by an average change of 0.8046 in Java Ratio, and the regression equation, which enables us to calculate Java Ratio values (Y) from given values of fibre content (x) is as follows :—

$$(Y-\bar{y}) = byx\ (x-\bar{x}).$$

In the case of our example, this reduces to :—

Java Ratio $-75.457 = -.8046$ (Fibre percentage -16.172)

or Java Ratio $= 88.468 - (.8046 \times$ Fibre percentage).

For 14 per cent. fibre we calculate :—

Java Ratio $= 88.468 - (.8046 \times 14)$
$= 77.20.$

Similar calculations give us values which enable us to draw the regression line shown on graph A.

The above equation gives us a straight regression line, hence the term "linear" regression. Such a line passes through the means of both variates and is such that the sum of the deviations from it is zero, while the sum of the squares of the deviations from it is at a minimum. When the association is perfect, all the points lie on the line—there are no deviations from it. The deviations from our line indicate the amount by which variation in fibre per cent. cane fails to account for all the variation in Java Ratio. In other words, variation in fibre per cent. significantly accounts for some of the variation in Java Ratio, but there remains a residual variation not accounted for.

The Partial Correlation Coefficient.

Up to now we have considered the association of only two variates, but it often happens that the effect of one variate on another is seriously upset by the influence of a third variate. It is very often necessary to eliminate this third variate before we can study the association between the two in which we are most interested. The partial correlation coefficient gives us a measure of the association between variates after the effect of other variates is eliminated or "partialled" out.

Let us take as an example the following figures for fibre per cent. cane and extraction. They are the arithmetical averages of the annual results of twelve South African sugar factories during the eleven-year period from 1935 to 1945.

Year.	Fibre per cent. cane.	Extraction.
1935	15.79	90.63
1936	14.85	91.06
1937	14.98	91.54
1938	14.34	92.21
1939	14.87	92.42
1940	15.66	92.02
1941	15.67	92.19
1942	15.33	92.68
1943	15.29	93.04
1944	15.85	93.20
1945	16.02	93.30

If we endeavour to correlate fibre percentage with extraction in these figures, we find a total correlation of +.51, which is not only not significant but actually positive. It is obvious, however, that extraction has improved (due to increased milling efficiency) during the eleven years, so before we can test for any correlation between fibre per cent. cane and extraction, we must eliminate the time effect. This we can do by calculating a partial correlation coefficient associating fibre per cent. cane with extraction, after calculating all possible total correlation coefficients and substituting in the formula :—

$$r_{ef.t} = \frac{r_{ef} - (r_{ft} \times r_{et})}{\sqrt{(1-r_{ft}^2)\ (1-r_{et}^2)}}$$

In this formula :—

$r_{ef.t}$ = Partial correlation coefficient associating extraction with fibre per cent. cane with time eliminated.

r_{ef} = Total correlation coefficient of the association of extraction with fibre per cent. cane.

r_{ft} = Total correlation coefficient of fibre per cent. cane with time.

r_{et} = Total correlation coefficient of extraction with time.

To test for significance we again refer to Fisher's table, but the "degrees of freedom" are now (N−2−1) or 8 degrees of freedom. The above formula is valid only when regressions are linear.

Another way of determining the partial correlation is to fit regression lines of fibre percentage on time and of extraction on time, to measure the deviations in each case from the lines and to correlate these residuals. This method is very convenient when the values of the variate to be eliminated are at equally spaced intervals and there is only one value of a variate for each value of the others. The line best fitted is a polynomial curve, of which the straight line is a special case. The straight line is a first order polynomial curve.

If we fit first order polynomial curves to express the regression of each of our two variates on time, we can conveniently measure time from the mean year 1940, when, for example, 1941 becomes +1 and 1936, −4. The equation for fitting a polynomial curve takes the general form :—

$$Y = a+bt+ct^2+dt^3 \ldots\ldots$$

which Fisher transforms to :

$$Y = A+BT_1+CT_2 \ldots\ldots$$

but for our present problem we require only the portion $Y=A+BT_1$:

$$\text{when } T_1 = (t-\bar{t}) = t \text{ (since } \bar{t}=0)$$

$$A = \frac{1}{N} S(y) = \bar{y}$$

$$B = \frac{12}{N(N^2-1)} S(yt)$$

It is not necessary to estimate the sum of squares of the residuals by fitting the regression line, for the sum of squares is reduced by the amount

$$\frac{N(N^2-1)}{12} B^2,$$

and the residual sum of squares is the difference between the total sum of squares and this amount. Similarly, the residual sum of products is the difference between the total sum of products and the amount

$$\frac{N(N^2-1)}{12} (B_1 \times B_2)$$

Our partial correlation coefficient then becomes :

$$\text{Residual sum of products} \div \sqrt{(\text{Residual sum of squares for extraction}) \times (\text{Residual sum of squares for fibre percentage})},$$

and incidentally the partial regression coefficient of extraction on fibre percentage is :

$$\text{Residual sum of products} \div \text{Residual sum of squares for fibre percentage.}$$

The partial correlation coefficient expressing the association between extraction and fibre per cent. cane for our example, after eliminating the effect of those factors associated with time, is −.8656, which is highly significant. We found the total correlation coefficient to be not significant and actually positive, showing how misleading data collected over a period of time can be, if the time effect is not considered.

Let us now examine some further correlations of importance in which a time effect is involved.

Relation between Juice Purity and Boiling House Recovery, and between Fibre per cent. Cane, Sucrose per cent. Cane and Extraction.

Casual examination of South African sugar factory returns reveals a general improvement year by year in boiling house recovery and in extraction. This is rightly claimed to be due to increased efficiency in the operation of the factories.

The general improvement has been so great that it tends to mask the effect of the so-called "quality factors," fibre percentage of cane and purity of the juice. As a consequence of this masking, doubts have recently been expressed that, within the range shown, purity of mixed juice is any important factor in boiling house recovery, and similarly that fibre per cent. cane has any connection with extraction.

With a view to mathematical examination of the figures reported by the factories, averages have been extracted from the Annual Summaries of Laboratory Returns. These average figures cover the last seventeen years and are derived from all the twelve factories reporting regularly during that period.

The examination of the figures so compiled is necessarily done in two parts. Firstly, boiling house recovery and its connection with purity of mixed juice, and, secondly, extraction and its association with fibre per cent. cane and with sucrose per cent. cane are dealt with. The calculations are shown in the annexure.

PART I.

Boiling House Recovery and Mixed Juice Purity.
Arithmetical averages of 12 factories :—

Year.	Boiling house recovery.	Mixed juice purity.
1929 ...	84.54 ...	85.96
1930 ...	84.51 ...	85.98
1931 ...	84.18 ...	85.33
1932 ...	84.85 ...	85.33

Year.	Boiling house recovery.	Mixedjuice purity.
1933 ...	85.46 ...	85.25
1934 ...	85.57 ...	84.15
1935 ...	86.95 ...	86.48
1936 ...	88.06 ...	85.48
1937 ...	88.39 ...	85.66
1938 ...	88.82 ...	86.47
1939 ...	89.50 ...	86.58
1940 ...	88.51 ...	85.20
1941 ...	88.68 ...	85.56
1942 ...	89.40 ...	86.00
1943 ...	90.08 ...	86.57
1944 ...	89.40 ...	86.10
1945 ...	89.43 ...	86.28

These figures illustrate clearly the general trend of improvement in boiling house recovery from 1929 to 1945. It is not an absolutely regular improvement, and irregularity appears at times to be associated with purity changes—but not always. Thus, for example, between 1942 and 1943 a rise of 0.57° purity was associated with a rise of 0.68 per cent. in boiling house recovery, but between 1932 and 1933 a small drop in purity of 0.08° was accompanied by a rise of 0.61 per cent. in boiling house recovery.

The figures also show that in the latter part of the seventeen-year period when the boiling house recovery was highest the purity was also highest, as compared with the earlier part when both purity and boiling house recovery were low.

Figures such as these are often graphed, in which case we get the representation in graph B. This method of showing a correlation is often very useful, but it can be quite misleading and might give rise to a statement that while the correlation appears to be quite good since 1936, in the earlier years there is little or no connection between purity of mixed juice and boiling house recovery.

Statistical analysis of the variation in boiling house recovery with time can be shown graphically by a straight line, but is more adequately described by a fourth order polynomial curve. This is proved by the analysis of variance of boiling house recovery (see appendix), and shown graphically in Fig. 1 of Graph C. Now, in order to make comparisons with purity a similar order time curve must be fitted to the purity figures, and we can then measure the residual variation of boiling house recovery with purity. If this partial correlation gives us a coefficient which is significant, we can assert that a correlation between the two factors exists; while if our coefficient is not significant. then we have not *proved* its existence.

In the averages shown above we find a partial correlation coefficient between the factors purity and boiling house recovery of +0.72, which is highly significant. The odds are more than 99 : 1 against this correlation being due to chance. We therefore must conclude that in the yearly figures analysed, an increase in purity was accompanied by an increase in recovery, and a decrease in purity was accompanied by a decrease in boiling house recovery.

This is shown graphically by examining Figs. 1 and 2 of Graph C taken together. In these figures differences are measured from moving means—the mean association of time with boiling house recovery and of time with purity. From these graphs, then, it can be seen that when the purity figure is above the line associating purity with time, the boiling house recovery figure is above the line associating boiling house recovery with time. Similarly, when the purity figure is below the line the boiling house recovery is below the line.

There are a few exceptions, as must be expected for there are other factors (such as the nature of the impurities) which influence boiling house recovery. Notable exceptions are 1931, 1936 and 1937. In 1931 a slight rise in purity of 0.02 is associated with a drop of 0.16 in boiling house recovery. In 1936 a purity drop of 0.20° was accompanied by a rise in recovery of 0.43, while in 1937 a purity drop of 0.16 was accompanied by a rise of 0.15 in boiling house recovery. These values are differences from the regressions.

It is interesting to note how much clearer our correlation is now than the picture provided by the graph B. Figures 1 and 2 taken together would not lead us, for instance, to the mistaken idea that purity and recovery are not associated in the years 1929 to 1936.

PART 2.

Fibre per cent. Cane, Sucrose per cent. Cane and Extraction.

The corresponding figures for extraction are :—

Arithmetical Averages for 12 Factories.

Year.	Fibre per cent. cane.	Extraction.	Sucrose per cent. cane.	Sucrose per cent. bagasse.
1929 ...	15.53	88.73	13.03	—
1930 ...	15.74	88.98	13.82	—
1931 ...	15.71	89.33	14.02	—
1932 ...	15.55	89.86	13.64	—
1933 ...	15.66	90.22	13.99	—
1934 ...	14.95	91.12	11.98	—
1935 ...	15.79	90.63	13.63	3.50
1936 ...	14.85	91.06	13.34	3.43
1937 ...	14.98	91.54	14.04	3.44
1938 ...	14.34	92.21	13.74	3.26
1939 ...	14.87	92.42	13.45	3.04
1940 ...	15.66	92.02	13.15	2.98
1941 ...	15.67	92.19	13.99	3.12
1942 ...	15.33	92.68	13.38	2.88
1943 ...	15.29	93.04	13.16	2.73
1944 ...	15.85	93.20	13.64	2.68
1945 ...	16.02	93.30	14.27	2.76

These averages show a distinct trend of improvement in extraction with the passing of the years. Our problem, then, is to correlate extraction with the other variates, fibre per cent. cane and sucrose per cent. cane, after making sufficient allowance for the time trend.

The statistical analysis of the figures for the whole seventeen-year period shows that expression of the association with time requires the fitting of at least second order regression curves. When these are fitted, and the effect of time thus eliminated, we can correlate the residual effects and calculate partial correlation coefficients which we can test for significance.

In the case of fibre per cent. cane, we find a highly significant partial correlation coefficient expressing association with extraction. This coefficient is −0.7369, showing that extraction varies inversely with fibre per cent. cane, and the odds are greater than 99 : 1 against this association being due to chance. We must therefore conclude that, in the averages for the seventeen-year period, a rise in fibre per cent. cane is associated with a drop in extraction, and a drop in fibre per cent. cane is associated with a rise in extraction. Figs. 3 and 4 of Graph D taken together illustrate this conclusion graphically.

In the case of sucrose per cent. cane and its association with extraction, analysis gave the rather surprising result of a negative partial correlation coefficient. This coefficient was, however, −0.4591, which is not significant.

It is usual to regard the year 1934 as an exceptional one and not to include it in a consideration of annual figures. In 1934, a record low figure for sucrose per cent. cane was associated with a record high extraction not again attained during 1935 and 1936. A further examination was therefore made of the averages for the eleven-year period, 1935 to 1945.

The analysis of the figures for this eleven-year period is a comparatively simple one, for only straight-line correlations with time are found to be required. This analysis yielded the highly significant partial correlation coefficient of −0.8656 for the association of fibre per cent. cane with extraction, while for the association of sucrose per cent. cane and extraction a non-significant coefficient (−0.0856) was again obtained.

In these average figures, then, there is a highly significant correlation between fibre per cent. cane and extraction, but no correlation can be proved to exist between sucrose per cent. cane and extraction.

The surprising lack of correlation between sucrose per cent. cane and extraction was finally examined. This lack of correlation could come about if a rise or fall in sucrose per cent. cane led to a corresponding proportionate rise or fall in sucrose per cent. bagasse. This association was actually found to exist, the correlation coefficient being the highly significant one of +0.870.

A similar correlation between sucrose per cent. cane and sucrose per cent. bagasse is discernable in the monthly averages shown in the Annual Summary of Laboratory Reports for the past season, 1945-46. In these monthly averages, we can calculate the sucrose per cent. bagasse from the sucrose per cent. cane figures, if we take a suitable month as the basis of comparison. Thus, taking the levels reached in September as our basis, we can recalculate sucrose per cent. bagasse figures for July, August, October and December with some degree of accuracy. This is due, of course, to the fact that notwithstanding appreciable variation in sucrose per cent. cane, the extraction figure did not vary to a great extent. The figures are as follows :—

July: Calculated sucrose per cent. bagasse $\left(14.14\times\frac{2.88}{15.02}\right) = 2.70$

Actual figure . . . . . $= 2.78$

August: Calculated sucrose per cent. bagasse $\left(14.73\times\frac{2.88}{15.02}\right) = 2.81$

Actual figure . . . . . $= 2.82$

October: Calculated sucrose per cent. bagasse $\left(14.78\times\frac{2.88}{15.02}\right) = 2.82$

Actual figure . . . . . $= 2.81$

November: Calculated sucrose per cent. bagasse $\left(14.56\times\frac{2.88}{15.02}\right) = 2.78$

Actual figure . . . . . $= 2.70$

Conclusion.

To test the effect of so-called "quality factors" of cane on boiling house recovery and extraction, arithmetical averages of figures reported by twelve South African factories during the past seventeen years have been compiled. Statistical analysis shows that, in these averages, there is a highly significant association between mixed juice purity and boiling house recovery, there is a highly significant association between fibre per cent. cane and extraction, and that no association can be proved to exist between sucrose per cent. cane and extraction.

Experiment Station,
South African Sugar Association,
Mount Edgecombe.
March, 1946.

ADDENDUM TO PART 1.

Arithmetical Averages for 12 Factories over the 17-year period 1929 to 1945.

Year.		Purity of mixed juice.		Recovery on mixed juice.
1929	...	85.96	...	84.54
1930	...	85.98	...	84.51
1931	...	85.33	...	84.18
1932	...	85.33	...	84.85
1933	...	85.25	...	85.46

Year.		Purity of mixed juice.		Recovery on mixed juice.
1934	...	84.15	...	85.57
1935	...	86.48	...	86.95
1936	...	85.48	...	88.06
1937	...	85.66	...	88.39
1938	...	86.47	...	88.82
1939	...	86.58	...	89.50
1940	...	85.20	...	88.51
1941	...	85.56	...	88.68
1942	...	86.00	...	89.40
1943	...	86.57	...	90.08
1944	...	86.10	...	89.40
1945	...	86.28	...	89.43
Totals	...	1458.38		1486.33
Means	...	85.78706		87.43118

Sum of squared deviations from means—
for *Purity*

$$=(85.96^2+85.98^2+\ldots 86.28^2)-\frac{1458.38^2}{17}$$
$$=125{,}116.6654-125{,}110.13085$$
$$=6.53455$$

for *Recovery*

$$=(85.54^2+84.51^2+\ldots 89.43^2)-\frac{1486.33^2}{17}$$
$$=130{,}022.2975-129{,}951.58052$$
$$=70.71698.$$

Sum of products: Purity × Recovery

$$=[(85.96\times84.54)+(85.98\times84.51)+\ldots (86.28\times89.43)]-\frac{1458.38\times1486.33}{17}$$
$$=127{,}518.7050-127{,}507.87914$$
$$=+10.82586.$$

Correlation with Time.

Time (t) is conveniently measured in periods from the middle year, 1937, when, e.g. for 1935, t = −2, and for 1944, t = +7. Correlation with time can often best be fitted by a curved regression line and the equation be of the polynomial form :—

Boiling house recovery $=A+BT_1+CT_2$, etc., for which we calculate up to the 5th order polynomial for our seventeen-year period (N = 17):—

$$T_1=t$$
$$T_2=t^2-\frac{N^2-1}{12}=t^2-24.$$
$$T_3=t^3-\left(\frac{3N^2-7}{20}\right)t=t^3-43t.$$
$$T_4=t^4-\left(\frac{3N^2-13}{14}\right)t^2+\frac{3(N^2-1)(N^2-9)}{560}$$
$$=t^4-61t^2+432.$$
$$T_5=t^5-\frac{5(N^2-7)}{18}t^3-\left(\frac{15N^4-230N^2+407}{1008}\right)t$$
$$=t^5-\frac{235}{3}t^3+\frac{3532}{3}t.$$

And :—

A = mean of A (Boiling house recovery in this case)
$$=\frac{1}{N}S(b).$$
$$B=\frac{12}{N(N^2-1)}S(bT_1).$$
$$C=\frac{180}{N(N^2-1)(N^2-4)}(S\,bT_2).$$
$$D=\frac{2800}{N(N^2-1)(N^2-4)(N^2-9)}S(bT_3).$$
$$E=\frac{44100}{N(N^2-1)(N^2-4)(N^2-9)(N^2-16)}S(bT_4)$$
$$F=\frac{698{,}544}{N(N^2-1)(N^2-4)(N^2-9)(N^2-16)(N^2-25)}S(bT_5)$$

We can conveniently calculate the sums of products, boiling house recovery × time, S (bt), S (bt^3) and S (bt^5), from the differences, e.g. 1945-1929, 1944-1928, and S (bt^2), S (bt^4), from the sums, as shown below :—

Boiling house recovery.	Differences.	t	t^3	t^5	Sums.	t^2	t^4
89.43−84.54	4.89	8	512	32768	173.97	64	4096
89.40−84.51	4.89	7	343	16807	173.91	49	2401
90.08−84.18	5.90	6	216	7776	174.26	36	1296
89.40−84.85	4.55	5	125	3125	174.25	25	625
88.68−85.46	3.22	4	64	1024	174.14	16	256
88.51−85.57	2.94	3	27	243	174.08	9	81
89.50−86.95	2.55	2	8	32	176.45	4	16
88.82−88.06	0.76	1	1	1	176.88	1	1
88.39	29.70	36	1296	61776	1397.94	204	8772

Sum of recovery figures or S(b) = 1486.33.
Sum of products B.H.R. × time or S(bt)
$$=(4.89\times8)+(4.89\times7)+\ldots(0.76\times1)$$
$$=+159.06.$$
Sum of products B.H.R. × time² or S(bt^2)
$$=+35520.92,$$
and similarly S(bt^3) = +6330.72,
S(bt^4) = +1,526,566.64,
S(bt^5) = +306,612.96.

We then calculate :—

$$A=\frac{S(b)}{N}=\frac{1486.33}{17}=87.43118.$$
$$B=\frac{12}{N(N^2-1)}S(bT_1)=\frac{+159.06}{408}$$
$$=+0.3898531.$$
$$C=\frac{180}{N(N^2-1)(N^2-4)}S(bT^2)$$
$$=\frac{1}{7752}[S(bt^2)-24\,S(b)].$$
$$=\frac{-151.00}{7752}=-0.0194788.$$
$$D=\frac{1}{139536}[S(bt^3)-43\,S(bt)].$$
$$=\frac{1}{139536}[6330.72-43\,(159.06)].$$
$$=\frac{-508.86}{139536}=-0.0036468.$$

$$E=\frac{1}{2418624}[S(bt^4)-61\,S(bt^2)+432\,S(b)].$$

$$=\frac{+1885.08}{2418624}=+0.000779402.$$

$$F=\frac{1}{40310400}\left[S(bt^5)-\frac{235}{3}S(bt^3)+\frac{3532}{3}S(bt)\right]$$

$$=\frac{-2026.80}{40310400}=-0.00005028.$$

To construct our analysis of variance we calculate the reduction in sum of squares associated with the different order regressions :—

Total sum of squares for boiler house recovery
=70.71698 (see page 7).

Reduction in sum of squares due to fitting 1st order regression

$$=\frac{N(N^2-1)}{12}B^2$$

$$=408\times\left(\frac{159.06}{408}\right)^2$$

$$=\frac{159.06^2}{408}$$

$$=62.01001.$$

Reduction in sum of squares due to 2nd order regression

$$=\frac{N(N^2-1)(N^2-4)}{180}C^2$$

$$=\frac{-151.00^2}{7752}$$

$$=2.94131.$$

Reduction in sum of squares due to 3rd order regression

$$=\frac{-508.86^2}{139536}=1.85571.$$

Reduction in sum of squares due to 4th order regression
=1.46923.

Reduction in sum of squares due to 5th order regression
=0.10191.

Analysis of Variance of Boiling House Recovery.

Source of variation.	Degrees of freedom.	Sum of squares.	Mean square or variance.	"F" value.
1st order regression ...	1	62.01001	62.01001	106.83*
Residual	15	8.70697	0.58046	
2nd order regression ...	1	2.94131	2.94131	7.14
Residual	14	5.76566	0.41183	
3rd order regression ...	1	1.85571	1.85571	6.17
Residual	13	3.90995	0.30077	
4th order regression ...	1	1.46923	1.46923	7.22
Residual	12	2.44072	0.20339	
5th order regression ...	1	0.10191	0.10191	0.48
Residual	11	2.33881	0.21262	
Total sum of squares	16	70.71698		

$$*\ 106.83=\frac{62.01001}{0.58046}$$

Significant "F" values at the level P=.05 when degrees freedom for reduction in sum of squares $(n_1)=1$.

Degrees of freedom (n_2)	15	14	13	12	11
Significant "F" value	4.54	4.60	4.67	4.75	4.84

From the above analysis of variance the 5th order curve describing the correlation of recovery with time fits the data, but not significantly better than the 4th order. The 4th order curve is a significantly better fit than the 3rd order, the 3rd better than the 2nd, and the 2nd better than the 1st order (straight line). The 4th order polynomial curve is therefore required to adequately describe the association.

Our regression equation then becomes :—

$$Y=A+BT_1+CT_2+DT_3+ET_4,$$

in which, by substituting the factors calculated, we can obtain the values which give us the 4th order polynomial curve describing the association of boiling house recovery with time. These values are calculated on page 9 and shown in Fig. 1 of Graph C.

Correlation of Purity with Time.

Analysis of Variance.

Source of variation.	Degrees of freedom.	Sum of squares.	Variance.	"F" value.
1st order regression ...	1	0.95797	0.95797	2.58
Residual	15	5.57658	0.37177	
2nd order regression ...	1	0.28629	0.28629	0.75
Residual	14	5.29029	0.37788	
3rd order regression ...	1	0.42506	0.42506	1.14
Residual	13	4.86523	0.37425	
4th order regression ...	1	0.40239	0.40239	1.08
Residual	12	4.46284	0.37190	
5th order regression ...	1	0.00824	0.00824	0.02
Residual	11	4.45460	0.40496	
Total sum of squares	16	6.53455		

No "F" values are significant.

This analysis of variance for purity with time reveals no simple time trend which is significant. The variations appear to be random ones. But now, for our analysis of co-variance of purity with boiling house recovery, it is necessary for us to fit a 4th order polynomial curve, as that order significantly describes the association between boiling house recovery with time.

So we again use a regression equation of the form

$$Y=A+BT_1+CT_2+DT_3+ET_4,$$

which equation gives us the values calculated on page 9 and the line shown on Fig. 2 of Graph C.

We could, from these values, measure the residuals (actual value, minus value calculated from regression equation) for each year in the case of both recovery and purity. These residuals could then be correlated and a partial correlation coefficient calculated. Such a correlation is, however, more conveniently calculated from the analysis of co-variance.

Analysis of Co-Variance.

Total sum of products (purity × boiling house recovery $= +10.82586$.

1st order sum of products $= \frac{(+159.06) \times (+19.77)}{408}$

$= +7.70739$.

2nd order sum of products $= \frac{(-151.00)\ (+47.11)}{7,752}$

$= -0.91765$.

Products for the higher orders are calculated similarly.

Analysis of Co-Variance.

Source of variation.	Degrees of freedom.	Sum of products.
1st order regression	1	+7.70739
2nd order regression	1	−0.91765
3rd order regression	1	+0.88814
4th order regression	1	+0.76890
Residual	12	+2.37908
Total	16	+10.82586

Partial correlation coefficient (after eliminating time effect)

$$= \frac{\text{Residual sum of products}}{\sqrt{\text{Residual sum of squares for recovery} \times \text{Residual sum of squares for purity.}}}$$

$$= \frac{+2.37908}{\sqrt{2.44072 \times 4.46284}}$$

$= +0.7208$.

Significant for 10 degrees of freedom at P=.01. (Minimum value required for significance at P=.01 is 0.7079.)

Regression Values—Purity × Time.

4th Order Regression Values.

Year.	$A+BT_1+CT_2+DT_3+ET_4$	Year.	$A+BT_1+CT_2+DT_3+ET_4$
1929	86.19	1937	85.82
1930	85.61	1938	85.92
1931	85.31	1939	85.98
1932	85.20	1940	86.00
1933	85.24	1941	86.00
1934	85.36	1942	86.00
1935	85.52	1943	86.03
1936	85.68	1944	86.14
		1945	86.38

Calculation of Regression Values Recovery × Time.

$A = 87.43118$ $\quad D = -0.0036468$ $\quad T_1 = t$ $\quad T_3 = t^3 - 43t$

$B = +0.3898531$ $\quad E = +0.0007794$ $\quad T_2 = t^2 - 24$ $\quad T_4 = t^4 - 61t^2 + 432$

$C = -0.0194788$

Year.	T_1	$A+BT_1$	T_2	CT_2	T_3	DT_3	T_4	ET_4	4th Order Regression Values. $A+BT_1+CT_2+DT_3+ET_4$
1929 ...	−8	84.31236	+40	−0.77915	−168	+0.61266	+624	+0.48635	84.63
1930 ...	−7	84.70221	+25	−0.48697	− 42	+0.15317	−156	−0.12159	84.25
1931 ...	−6	85.09206	+12	−0.23375	+ 42	−0.15317	−468	−0.36476	84.34
1932 ...	−5	85.48191	+ 1	−0.01948	+ 90	−0.32821	−468	−0.36476	84.77
1933 ...	−4	85.87177	− 8	+0.15583	+108	−0.39385	−288	−0.22447	85.41
1934 ...	−3	86.26162	−15	+0.29218	+102	−0.37197	− 36	−0.02806	86.15
1935 ...	−2	86.65147	−20	+0.38958	+ 78	−0.28445	+204	+0.15900	86.92
1936 ...	−1	87.04133	−23	+0.44801	+ 42	−0.15317	+372	+0.28994	87.63
1937 ...	±0	87.43118	−24	+0.46749	± 0	±0.00000	+432	+0.33670	88.24
1938 ...	+1	87.82103	−23	+0.44801	− 42	+0.15317	+372	+0.28994	88.71
1939 ...	+2	88.21089	−20	+0.38958	− 78	+0.28445	+204	+0.15900	89.04
1940 ...	+3	88.60074	−15	+0.29218	−102	+0.37197	− 36	−0.02806	89.24
1941 ...	+4	88.99059	− 8	+0.15583	−108	+0.39385	−288	−0.22447	89.32
1942 ...	+5	89.38045	+ 1	−0.01948	− 90	+0.32821	−468	−0.36476	89.32
1943 ...	+6	89.77030	+12	−0.23375	− 42	+0.15317	−468	−0.36476	89.32
1944 ...	+7	90.16015	+25	−0.48697	+ 42	−0.15317	−156	−0.12159	89.40
1945 ...	+8	90.55000	+40	−0.77915	+168	−0.61266	+624	+0.48635	89.64

ADDENDUM TO PART 2.

Analysis of Variance. Extraction × Time.

Source of variation.	Degrees of freedom.	Sum of squares.	Mean variance.	"F" value.
1st order regression ...	1	33.99454	33.99454	312.0
Residual	15	1.63428	0.10895	
2nd order regression ...	1	0.68498	0.68498	10.10
Residual	14	0.94930	0.06781	
3rd order regression ...	1	0.01319	0.01319	0.18
Residual	13	0.93611	0.07201	
4th order regression ...	1	0.00438	0.00438	0.06
Residual	12	0.93173	0.07764	
5th order regression ...	1	0.01345	0.01345	0.16
Residual	11	0.91828	0.08348	
Total sum of squares...	16	35.62882		

1st and 2nd order significant fit.

Analysis of Variance. Fibre × Time.

Source of variation.	Degrees of freedom.	Sum of squares.	Mean variance.	"F" value
1st order regression ...	1	0.00194	0.00194	0.01
Residual	15	3.29915	0.21994	
2nd order regression ...	1	1.30941	1.30941	9.21
Residual	14	1.98974	0.14212	
3rd order regression ...	1	0.14873	0.14873	0.11
Residual	13	1.84101	0.14161	
4th order regression ...	1	0.22702	0.22702	1.69
Residual	12	1.61399	0.13450	
5th order regression ...	1	0.18768	0.18768	1.45
Residual	11	1.42631	0.12966	
Total sum of squares...	16	3.30109		

Only 2nd order significant.

Equation of the form $Y=A+BT_1+CT_2$ adequate for both factors.

Analysis of Co-Variance.

Source of variation.	Degrees of freedom.	Sum of products.
1st order regression	1	+0.25690
2nd order regression	1	−0.94707
Residual	14	−1.01278
Total sum of products ...	16	−1.70295

Partial correlation coefficient (after eliminating time)

$$= \frac{-1.01278}{\sqrt{0.94930 \times 1.98974}}$$

$$= -0.7369$$

Significant at P=.01.

Regression Values.

Extraction × Time. Year.	2nd Order Value. $A+BT_1+CT_2$	Fibre per cent. Cane × Time. Year.	2nd Order Value. $A+BT_1+CT_2$
1929	88.64	1929	15.90
1930	89.07	1930	15.71
1931	89.48	1931	15.54
1932	89.87	1932	15.38
1933	90.25	1933	15.29
1934	90.60	1934	15.20
1935	90.94	1935	15.14
1936	91.25	1936	15.10
1937	91.55	1937	15.09
1938	91.83	1938	15.10
1939	92.09	1939	15.14
1940	92.33	1940	15.21
1941	92.56	1941	15.30
1942	92.76	1942	15.42
1943	92.94	1943	15.57
1944	93.11	1944	15.74
1945	93.26	1945	15.94

Correlation with Time—11 years, 1935-1945. Extraction.

Analysis of Variance.

Source of variation.	Degrees of freedom.	Sum of squares.	Mean variance.	"F" value.
1st order regression ...	1	6.68631	6.68631	81.3
Residual	9	0.74005	0.08223	Significant.
Total	10	7.42636		

Up to and including 5th order, only 1st order significant.

Fibre per cent. Cane.

Anaylsis of Variance.

Source of variation.	Degrees of freedom.	Sum of squares.	Mean variance.	"F" value.
1st order regression ...	1	0.71363	0.71363	Not significant
Residual	9	2.01514	0.22390	
Total	10	2.72877		

No "F" values up to and including 5th order significant.

Sucrose per cent. Cane.

Analysis of Variance.

Source of variation.	Degrees of freedom.	Sum of squares.	Mean variance.
1st order regression ...	1	0.02269	0.02269
Residual	9	1.32534	0.14726
Total	10	1.34802	

No "F" values up to and including 5th order significant.

Analysis of Co-Variance.

Extraction × fibre per cent. cane and sucrose per cent. cane with 1st order regressions on time fitted.

Sums of products :—

Source of variation.	Degrees of freedom.	Extraction × Fibre per cent. cane.	Extraction sucrose per cent. cane.
1st order regression	1	+2.18439	+0.38954
Sum of products ...	9	−1.05705	−0.08489
Totals	10	+1.12734	+0.30465

Partial correlation coefficients.

Correlation between *fibre per cent. cane and extraction* with time linearly eliminated, or

$$r\,fe.T_1 = \frac{-1.05705}{\sqrt{0.74005 \times 2.01514}} = -0.8656$$

Significant at P=.01.

Correlation between *sucrose per cent. cane and extraction* with time linearly eliminated, or

$$r\,se.T_1 = \frac{-0.08489}{\sqrt{0.74005 \times 1.32534}} = -0.0856.$$

Not significant.

Analysis of Variance (Sucrose per cent. Bagasse with Time).

Source of variation.	Degrees of freedom.	Sum of squares.	Mean square.
1st order regression ...	1	0.82218	0.82218
Residual	9	0.07649	0.00850
Total	10	0.89867	

Analysis of Co-Variance of Sucrose per cent. Bagasse and Sucrose per cent. Cane.

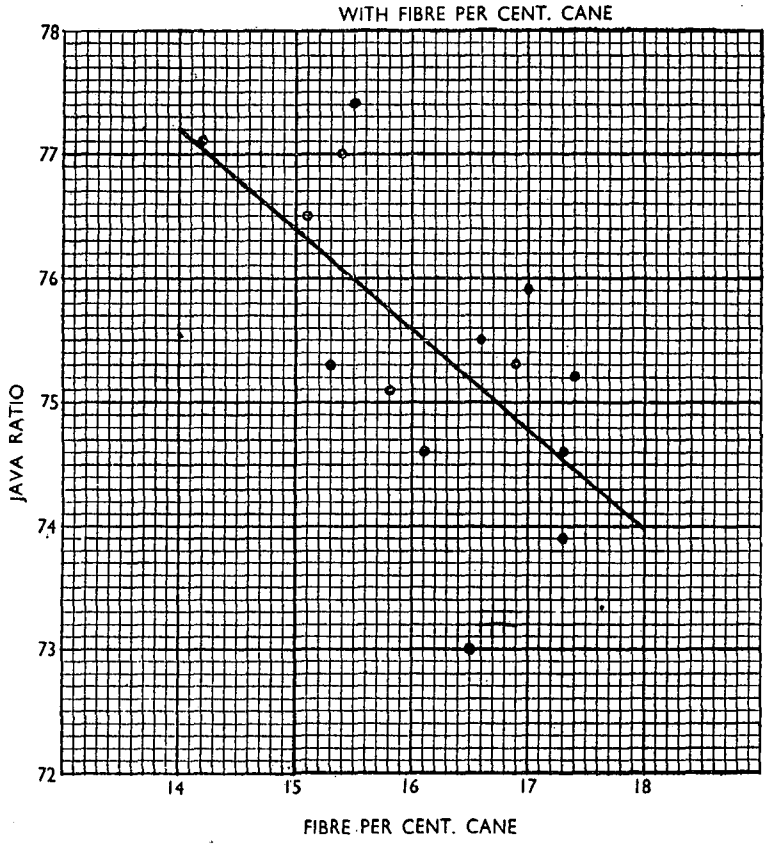
Source of variation.	Degrees of freedom.	Sum of products.
1st order regression...	1	−0.13660
Residual	9	+0.27704
Total	10	+0.14044

Partial correlation coefficient, sucrose per cent. bagasse × sucrose per cent. cane with time linearly eliminated, or

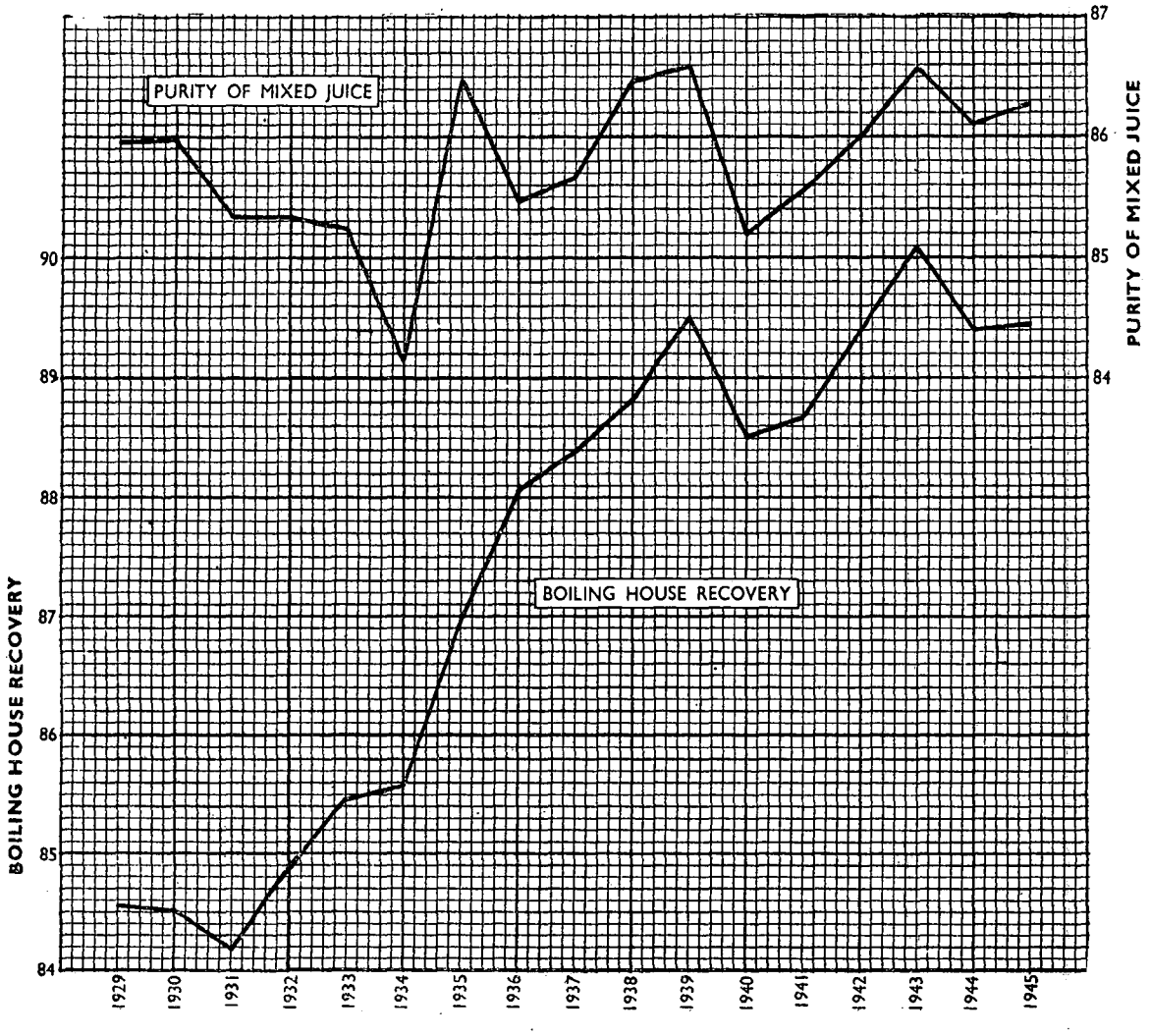
$$r\,bs.T_1 = \frac{+0.27704}{\sqrt{0.07649 \times 1.32534}} = +0.870.$$

Significant at P=.01.

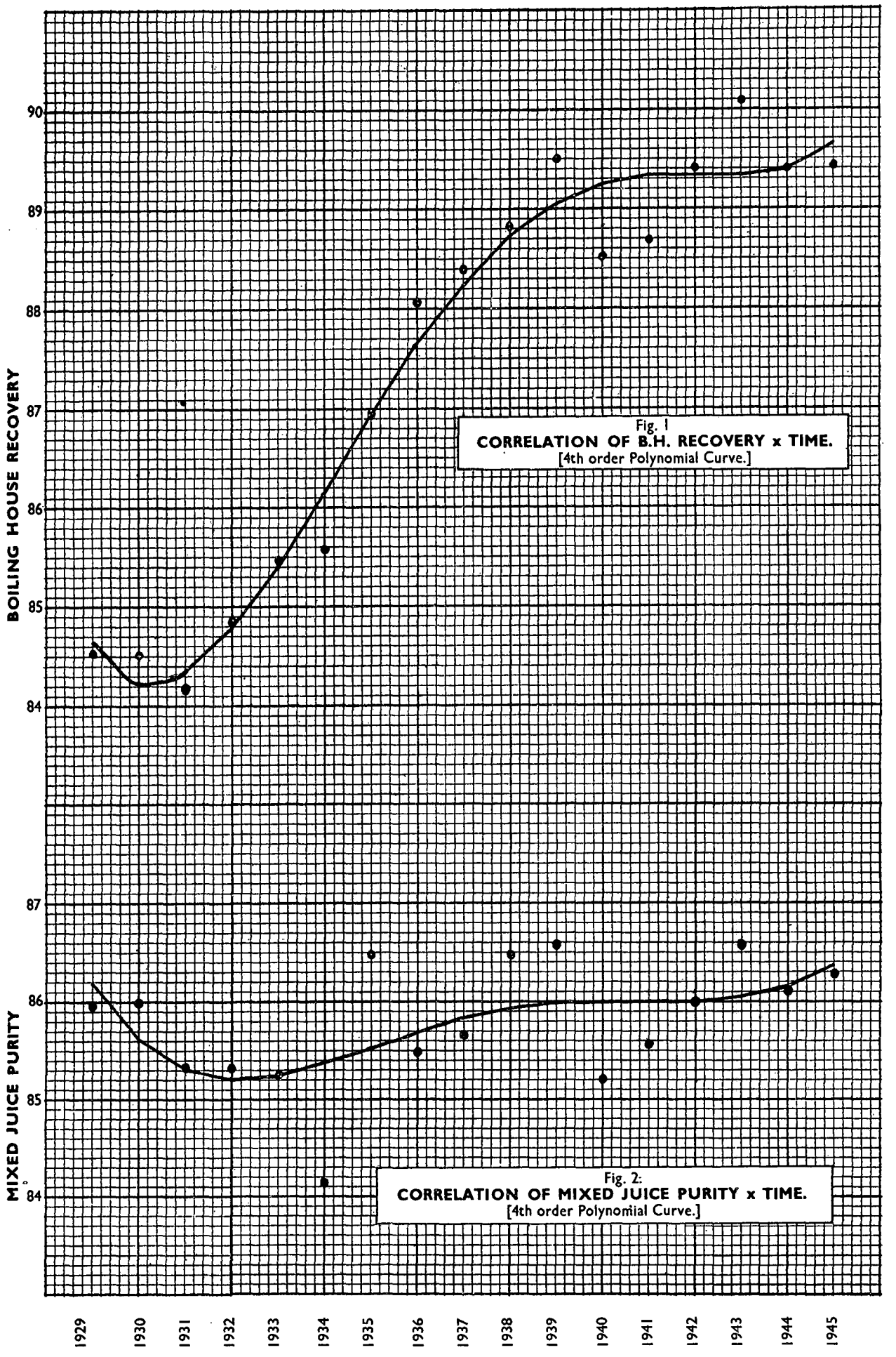
GRAPH A — CORRELATION OF JAVA RATIO WITH FIBRE PER CENT. CANE



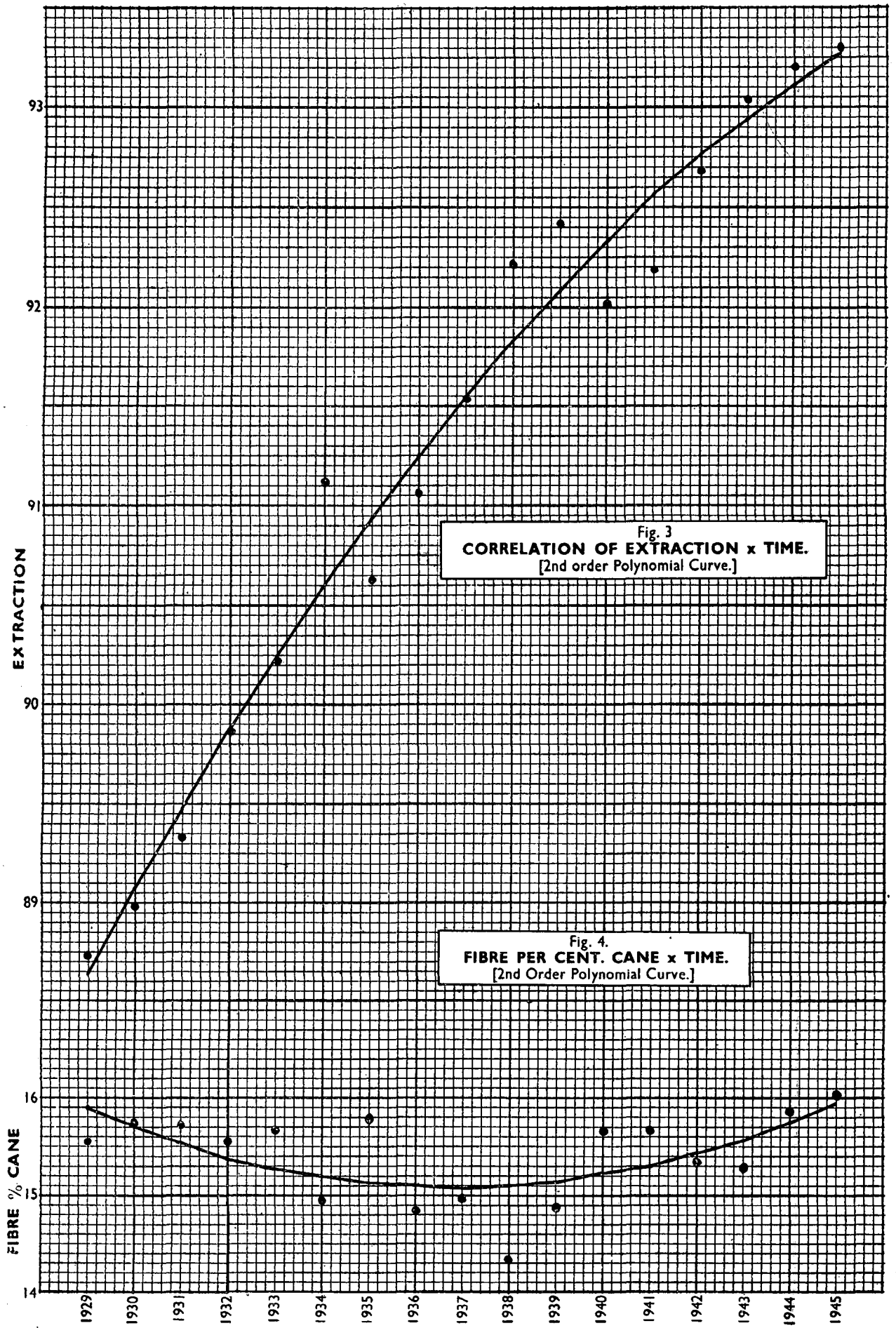
GRAPH B — BOILING HOUSE RECOVERY AND PURITY OF MIXED JUICE
ARITHMETICAL AVERAGES FOR 12 FACTORIES, 1929/45



GRAPH C.


BOILING HOUSE RECOVERY
90
89
88
87
86
85
84
Fig. 1
CORRELATION OF B.H. RECOVERY x TIME.
[4th order Polynomial Curve.]
MIXED JUICE PURITY
87
86
85
84
Fig. 2:
CORRELATION OF MIXED JUICE PURITY x TIME.
[4th order Polynomial Curve.]
1929
1930
1931
1932
1933
1934
1935
1936
1937
1938
1939
1940
1941
1942
1943
1944
1945

GRAPH D.


93
92
91
90
89
16
15
14
EXTRACTION
FIBRE % CANE
1929
1930
1931
1932
1933
1934
1935
1936
1937
1938
1939
1940
1941
1942
1943
1944
1945
Fig. 3
CORRELATION OF EXTRACTION x TIME.
[2nd order Polynomial Curve.]
Fig. 4.
FIBRE PER CENT. CANE x TIME.
[2nd Order Polynomial Curve.]

The PRESIDENT said he did not know a great deal about statistical analyses, but he found the paper clear, and felt that the author had given a lucid exposition of a difficult subject.

It was extremely interesting to see that, as a result of statistical analyses, relationships could be established between fibre per cent. cane and extraction, and between mixed juice purity and boiling-house recovery. No relationship was proved to exist between sucrose per cent. cane and extraction.

Mr. DU TOIT said that for some time now doubt had been expressed about the usefulness of the reduced formulæ and their validity to South African conditions, and even whether fibre had any effect on extraction or purity of juice on recovery. It would, however, be very difficult to attack the conclusions arrived at in this paper. Although not proved, the evidence in this paper seemed to indicate that sucrose per cent. cane had no effect on extraction, and that the sucrose in bagasse varied according to the original sucrose in cane. Thus in 1934 we had a very low sucrose, but the extraction reached a new high level, and one that was not surpassed for some years after.

The paper was also of value as it gave the statistical methods used in such a form that it was easy to follow by those not familiar with statistical analyses, and it could serve as a model for working out similar problems.

Mr. MOBERLY said that he had been studying statistical methods recently, and he found the paper most useful as it explained certain points that had puzzled him.

The method used for finding correlations between certain sets of data was a most powerful one and should be used more often.

It was rather a curious coincidence that the calculated sucrose per cent. bagasse for July was the same as the actual figure for November and vice versa. The same applied to the months of August and September.

Mr. CHRISTIANSON, in reply, stated that the identity was only a coincidence, but the calculated figures were, of course, derived from the mean, and the variations above the mean must equal the variations below the mean.

Mr. RAULT stated that it was wrong to state that a certain rise in fibre would result in a proportionate drop in extraction. The relationship might apply in general when no other factors varied, but the effect of fibre on extraction could not be predicted with any degree of certainty, the high-fibred Co.281 canes being known as favouring extraction as well as capacity. The same applied to the purity and recovery relationship. He was surprised that no correlation could be established between sucrose per cent. cane and extraction. It had been his experience that high sucrose was associated with better extraction, as the sucrose content of the final bagasse did not rise proportionately with richer cane.

Dr. MCMARTIN said that the statistical technique, if intelligently used, would prevent the dissemination of loose or unsound information. For years it had been used very successfully by the agricultural and biological sciences, and he thought its introduction in the milling and manufacturing sides of the industry was overdue.

Mr. MOBERLY, in reply to Mr. Rault, stated that one of the biggest changes in science in recent years was that the idea of absolute truth had been replaced by the concept of probability. In physics, for example, many laws were based on probabilities of millions and millions to one or almost certainty; but in our work, e.g. the effect of fibre on extraction, these probabilities were reduced to about 20 or 100 to 1. Thus no one can predict with absolute certainty that a high fibre will give a lower extraction; the chances were nevertheless in favour of it.

Mr. CHRISTIANSON said that there was hardly any doubt that fibre and extraction could be correlated, and that in general a rise in fibre per cent. cane would give a lower extraction, but that did not mean that you could predict the extraction resulting from fibre per cent. cane. It did not even mean that a higher fibre would always give a lower extraction. The same arguments applied to purity and recovery relationships.

Mr. DUCHENNE asked whether it was possible to get a formula that was applicable to any one factory. At his particular factory purity of mixed juice was graphed against boiling-house recovery for cane payment purposes. Roughly speaking, the recovery was about one and a half per cent. higher than the purity of the average crusher juice. His table was calculated from the S.J.M. formula, but the method outlined here might prove useful.

Mr. CHRISTIANSON, in reply, stated that the regression formulæ worked out in this paper was based on the results of a number of factories over a number of years, and were not applicable to any particular factory under present conditions. If that were desired, he advised basing the formula on weekly averages obtained at that factory.