21

THE CORRELATION OF WEEKLY RAINFALL.

By R. A. FISHER, M.A., Fellow of Gonville and Caius College, Cambridge, and Chief Statistician, Rothamsted Experimental Station ;

and WINIFRED A. MACKENZIE, B.Sc. (Econ.), Assistant Statistician, Rothamsted Experimental Station.

(Communicated by R. H. Hooker, M.A., F.S.S.)

1. Introductory.

During the deliberations of the Committee appointed by the Ministry of Agriculture and Fisheries in 1920 to report on the uses being made of meteorological data in connection with the agricultural and fishing industries, the question arose: To what extent can simultaneous observations of weather and crops, carried on in the same season in different parts of the country, replace observations taken at the same station for many consecutive seasons? Apart from the actual differences in climate and the corresponding adaptations of agricultural practice, and variety of plant cultivated, the question turns upon the extent to which different parts of the country experience in any one year similar deviations from their average weather sequence. We require to know, in fact, to what extent the weather deviations are correlated at different stations.

Besides this question, which was the immediate cause of the present inquiry, a knowledge of these correlations is essential if we are to know with what accuracy the rainfall, or any other instrumental observation, may be estimated for any one point, from known observations at neighbouring points. The accuracy of such estimates must evidently depend upon the density with which observing stations are scattered in the area considered, and with a knowledge of the correlations it would be possible to state how thickly the stations ought to be scattered, in order that estimates of the weather at intervening points should have any required accuracy. The same consideration applies when we wish to estimate the total rainfall over any particular area, such as a river basin ; whatever the degree of accuracy required, a certain number of stations will be necessary, and this number will be calculable from the coefficients of correlation between the stations.

In the third place, the study of the correlations between different stations is needed to put upon an experimental and quantitative basis such questions in general meteorological physics as refer to localisation of meteorological effects. It would be conceivable that the greater part of the differences between one season and another were due to very widespread meteorological causes ; in this case high correlations would be found even between distant stations. On the other hand, an important part of these variations may be due to causes which are more or less local in their consequences. The distribution of the variance of any instrumental observation between the widespread and the local fluctuations is closely analogous mathematically to the distribution of the energy of radiation between long and short wave-lengths. This distribution can only be investigated experimentally, but its nature must govern all

calculations of the *accuracy* of meteorological estimates and predictions. That this distribution has a strongly marked annual period is definitely established by the present inquiry.

Since an Agricultural Research Station cannot undertake to devote more than a limited amount of time to purely Meteorological Research, our work is aimed primarily at the testing and development of adequate methods, and the material and treatment have been chosen so as to throw light upon those particular questions which from the point of view of agricultural research most urgently needed answering. Thus our inquiry is confined to rainfall, although the methods used would apply with even greater precision to temperature. The three stations chosen—Rothamsted, York, and Aberdeen—are all on the eastern flank of Great Britain. In choosing the period of one week for each value, we were partly influenced by agricultural considerations, but these were reinforced by the nature of the data. The shorter the period taken the more abundant were the data; a correlation obtained from whole years would have nearly seven times the probable error of the average of correlations obtained from weeks; also we wished to investigate the changes throughout the year, and although the labour of computation was increased, these are shown more clearly and more accurately by weeks than by months. On the other side it should be noted that the shorter the period the less normal becomes the distribution, and the more hazardous the assumption, which our method includes, that the standard error of r is proportional to $1 - r^2$; again, if the weather at one station lagged behind that of a second, this effect would be more serious for the shorter periods, such as single days. It may be noted that the position of the week is more variable with respect to the solar year than is the day or month, though the slowness of the change in the correlation renders correction for this error unnecessary; unlike the months, however, the weeks are of equal and constant duration.

2. The Correlations between Weekly Rainfall Values.

The data used were extracted from the *Weekly Weather Report*[1] published by the Meteorological Office from the sixth week of 1878 to the end of 1920. The Rothamsted and Aberdeen series extend over the whole period, but the York series has a break of nearly seven years from the eighteenth week of 1899 to the third week of 1906 inclusive, observations being resumed at a different station in the same district.

The number of years for which simultaneous observations were obtainable for Rothamsted and Aberdeen was about forty-three, while each of these stations have about thirty-six years in common with York. The numbers of weeks for which each number of pairs was available are shown in Table I.

The variation in the number of pairs available for the different weeks of the year is therefore very small. No allowance will be made for this variation, since we may with ample accuracy consider the size of the sample to be given in these cases by the average values 42·87, 36·15, and 36·15.

The actual correlations obtained for these three stations for the fifty-two weeks of the year are given in Table II.

[1] *Weekly Weather Report of the Meteorological Office*, 1878–1920.

Since these values are obtained from small samples, certain special precautions must be taken in their statistical reduction. It has been

TABLE I.

Number of pairs.	Rothamsted and Aberdeen.	Number of pairs.	Rothamsted and York.	Aberdeen and York.
41	1	35	3	3
42	5	36	38	38
43	46	37	11	11
Total . .	52	Total . .	52	52

pointed out[1] that the curves of distribution in normal samples of the coefficient of correlation (r) are liable to be extremely skew, especially for small samples. This is true even for large samples, the curves

TABLE II.—Weekly Rainfall Correlations.

Week of Year.	Rothamsted and Aberdeen.		Aberdeen and York.		York and Rothamsted.		Week of Year.	Rothamsted and Aberdeen.		Aberdeen and York.		York and Rothamsted.	
	r.	z.	r.	z.	r.	z.		r.	z.	r.	z.	r.	z.
1	·50	·55	·55	·62	·46	·50	27	·31	·32	·21	·21	·68	·82
2	·48	·53	·54	·61	·73	·92	28	·15	·15	·49	·54	·30	·31
3	·36	·38	·67	·82	·89	1·41	29	·13	·13	·19	·19	·50	·54
4	·39	·41	·56	·64	·66	·80	30	·18	·18	·17	·17	·48	·53
5	·48	·52	·35	·36	·62	·72	31	·17	·17	·32	·33	·14	·14
6	·66	·80	·81	1·14	·75	·98	32	·50	·55	·45	·48	·43	·46
7	·61	·70	·73	·94	·50	·55	33	·43	·46	·16	·16	·65	·78
8	·56	·64	·37	·39	·66	·79	34	·36	·37	·56	·63	·63	·74
9	·58	·67	·69	·85	·77	1·03	35	·28	·29	·30	·31	·68	·82
10	·24	·25	·47	·51	·60	·70	36	·44	·47	·40	·42	·57	·65
11	·23	·24	·24	·24	·59	·68	37	·47	·51	·70	·86	·69	·85
12	·45	·49	·48	·52	·59	·67	38	·29	·30	·30	·31	·76	·98
13	·62	·72	·85	1·25	·70	·88	39	·51	·57	·43	·46	·42	·45
14	·66	·79	·51	·57	·79	1·06	40	·29	·29	·40	·42	·57	·65
15	·30	·31	·80	1·10	·27	·28	41	·43	·46	·43	·46	·30	·31
16	·17	·17	·53	·60	·41	·43	42	·36	·38	·63	·73	·66	·79
17	·23	·24	·52	·58	·65	·78	43	·24	·25	·44	·47	·57	·64
18	·59	·68	·71	·88	·73	·93	44	·42	·45	·45	·49	·50	·55
19	·10	·10	·25	·26	·75	·97	45	·30	·31	·57	·64	·58	·66
20	·30	·31	·23	·23	·64	·76	46	·28	·29	·50	·55	·56	·63
21	·32	·34	·43	·46	·62	·72	47	·45	·49	·63	·74	·64	·77
22	·29	·30	·43	·46	·32	·33	48	·44	·47	·42	·45	·73	·93
23	·05	·05	·48	·52	·48	·52	49	·32	·33	·69	·85	·30	·31
24	·04	·04	·58	·66	·58	·66	50	·28	·29	·58	·66	·79	1·07
25	·37	·38	·31	·32	·58	·66	51	·25	·26	·32	·33	·58	·67
26	·41	·44	·41	·44	·14	·14	52	·64	·76	·63	·75	·64	·76

The values are here given to the second decimal place only. The further work was based upon correlations taken to the fourth place of decimals.

changing form very rapidly as r approaches ± 1. It has been further shown[2] that by the transformation $z = \tanh^{-1} r$, sampling curves are

[1] R. A. Fisher (1915), "Frequency distribution of the values of the correlation coefficients in samples from an indefinitely large population." *Biom.* x. pp. 507-521.

[2] R. A. Fisher (1921), "On the 'probable error' of a coefficient of correlation deduced from a small sample." *Metron.* i. pt. 4, pp. 1-32.

obtained which are practically normal and of constant standard deviation. Where processes are employed involving the averaging of a large number of values of r, accuracy is not attained owing to the skewness of the distributions, even when the values of r are properly weighted; also individual errors in r, which with small samples may be considerable, introduce errors into the weights. On the other hand, the weight of a value of z depends, for practical purposes, only on the number in the sample from which it was calculated, and when, as in the present case, all samples are nearly of the same size, weighting may be neglected; in addition, correction may be simply made in the averages for the small bias which is introduced into the product moment correlation by its method of calculation. For these reasons the correlation coefficients were transferred to the z scale, as shown in Table II., and the values of z used for the statistical reductions.

The mean values are shown in Table III., in which the correction $-\frac{1}{2}r/(n-1)$ has been inserted; this correction, which is too small to be of importance for single samples, becomes increasingly necessary as larger numbers of correlations are averaged, and in the present case is a perceptible fraction of the probable error.

TABLE III.—Mean Values of the Correlation.

	Aberdeen—Rothamsted.		Rothamsted—York.		York—Aberdeen.	
	r	z	r	z	r	z
Crude mean	·3755	·3948	·5953	·6859	·5330	·5943
Correction		−·0045		−·0085		−·0076
Corrected mean	·3717	·3903 ± ·0141	·5898	·6774 ± ·0221	·5275	−·5867 ± ·0222

The three stations lie very nearly on a great circle, the distance from Rothamsted to York being about 150 statute miles, while from York to Aberdeen is about 225 statute miles. The correlations thus diminish regularly with increasing distance; so regularly indeed that it is easy to construct a formula that will very closely agree with these three results. Thus, if we imagine that z falls off as some inverse power of the distance, we find the formula

$$z = \left(\frac{a}{61{\cdot}96}\right)^{-\frac{3}{5}}$$

where a is the distance between the stations in geographical or sea miles. This formula gives the values ·3918, ·6790, and ·5836, agreeing with the observed results very much more closely than the probable errors require. Without further information than is at present available as to the rainfall correlations between different stations it would be rash to suppose that such a formula were widely true. We should expect the parametric distance, 62 sea miles, to vary from place to place, and, since there is no *a priori* reason to think that such relations are isotropic, to vary with direction of displacement in the same region. All that we wish to emphasize is that, even from a short series of years, such as we have used, it is possible to obtain values of the mean correlation, with

such small errors of sampling as to display clearly those regularities which indicate the general laws of meteorological correlation.

In estimating the standard error of z, three values may be compared. For normal samples the variance [1] is given by $(n-3)^{-1}$ with great accuracy, but the distribution of weekly rainfall is far from normal, and no reliance can be placed on this value, which corresponds to the ordinary formula—

$$\sigma_r^2 = \frac{(1-r^2)^2}{n-1}.$$

Again, if the correlation had been free from seasonal variations we might have found the sampling variance empirically from that of the 52 values obtained from separate weeks; as we shall see, however, marked seasonal variations occur, so that the values of the total variance are larger than the variance due to sampling in any particular week. To obviate this difficulty we have calculated the variance from the differences between the values of z for successive weeks, the differences in the true values being so small that their squares may be neglected. Thus we have the following values.

TABLE IV.—Sampling Variance of z.

	Formula.	Aberdeen—Rothamsted.	Rothamsted—York.	York—Aberdeen.
Normal expectation .	$\frac{1}{n-3}$	·02508	·03017	·03017
Variance of z . .	$\frac{1}{51}S(z-\bar{z})^2$	·03635	·06228	·06210
Squares of differences .	$\frac{1}{104}S(z_{p+1}-z_p)^2$	·02262	·05587	·05635

It is to be noted that in every case the total variance of z is distinctly greater than the sampling variance derived from the squares of differences, thus indicating a genuine seasonal variation. In comparison with the normal expectation it is remarkable that for the Aberdeen—Rothamsted correlations the sampling variance does not appear to differ significantly from the normal expectation, while in the other two cases it is considerably greater.

3. The Annual Oscillation of Rainfall Correlations.

The values given in Table IV. indicate that a genuine cyclic change takes place in the course of the year in the rainfall correlations. It is interesting to determine how much of this change is accounted for by the first few harmonic terms. The first few terms of a Fourier expansion have therefore been fitted to the values of z, on the assumption that the fifty-two weeks are equally spaced round the year; this is sufficiently nearly the case, although the interval between the last and first weeks exceeds the other intervals by nearly thirty hours.

[1] The variance is the square of the standard deviation; its simplicity lies in its additive property. Thus the variance due to two independent causes of variation acting simultaneously is equal to the sum of the values of the variance when each acts separately.

The form of expansion is,

$$z = \bar{z} + A_1 \sin\theta + B_1 \cos\theta + A_2 \sin 2\theta + B_2 \cos 2\theta + \ldots,$$

where $\theta = 0$, at the mean date of the first week, January 4, at 9 p.m. Since

$$A_s = \tfrac{1}{26} S(z \sin s\theta), \text{ and } S(\sin^2 s\theta) = 26,$$

it follows that the sampling variance of each of these coefficients is one twenty-sixth of that of z. The variance contributed by any pair of terms is

$$V_s = \tfrac{1}{2}(A_s^2 + B_s^2),$$

and in the absence of any real oscillation in this period, the mean value of this quantity must be the same as the variance of the coefficients. If V_s bears to this mean value a ratio, k, then it is easy to see that the probability that such a value should be exceeded by chance is

$$P = e^{-k}.$$

The coefficients of the first two periods are given in Table V.

TABLE V.—Terms of Yearly and Half-yearly Periods.

	Aberdeen—Rothamsted.	Rothamsted—York.	York—Aberdeen.
A_1	+·030,37 ± ·0199	+·051,66 ± ·03127	+·082,52 ± ·03140
B_1	+·116,54 ± ·0199	+·122,63 ± ·03127	+·145,39 ± ·03140
A_2	+·080,97 ± ·0199	+·032,62 ± ·03127	−·007,98 ± ·03140
B_2	−·035,42 ± ·0199	−·025,88 ± ·03127	−·055,60 ± ·03140
V_1	·007,246	·008,854	·013,973
V_2	·003,905	·000,867	·001,578
Mean chance value	·000,870	·002,149	·002,167
P_1	·000,24	·016	·001,6
* P_2	·015	·67	·48

In all cases the first harmonic is definitely significant; for the second harmonic, this is only the case with the Aberdeen—Rothamsted series. The second harmonics agree, however, in their general tendency; as will be seen in the diagram, the tendency is in all cases to make the winter maximum fall later and the summer minimum earlier, thus causing a rapid fall of correlations in the spring, while lengthening the rise in the autumn period.

TABLE VI.—Residual Variance of z.

	Aberdeen—Rothamsted.	Rothamsted—York.	York—Aberdeen.
Total variance .	·036,35	·062,28	·062,10
Residue after 1st harmonic	·029,97	·055,58	·050,29
Residue after 2nd harmonic	·026,94	·056,86	·050,88
Sampling variance .	·022,62	·055,87	·056,35

If we inquire to what extent these harmonic terms suffice to account for the excess of the total variance over that due to sampling, we may successively subtract from the former the amounts contributed by the

* For .015, read .011.

harmonic terms, adding on each occasion the mean values attributable to chance, which addition serves to allow for the reduction by fitting of the

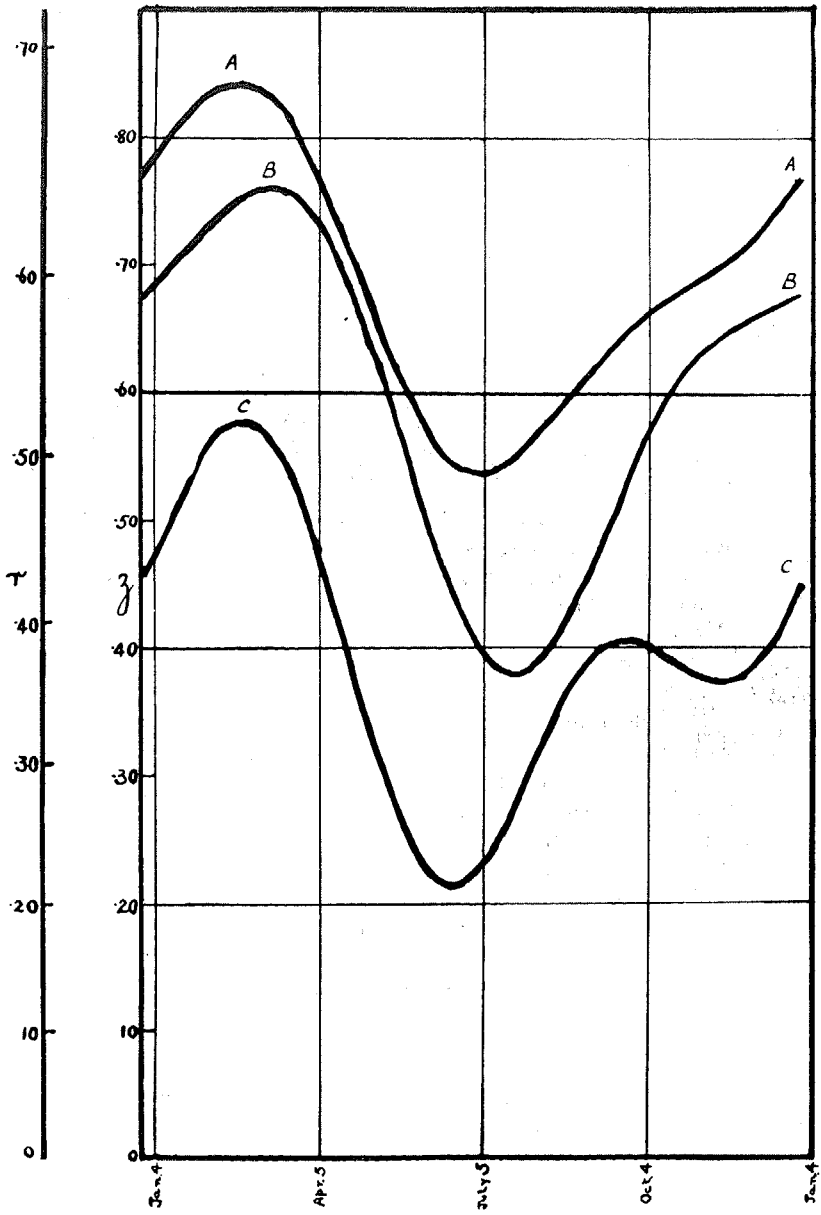

FIG. 1.—Harmonic curves, fitted to the values in Table II., showing average rainfall correlations at different seasons of the year.
A, Rothamsted—York; B, York—Aberdeen; C, Aberdeen—Rothamsted.

number of degrees of freedom, which carries with it a proportional reduction of the variance. Table VI. shows the approach of the residual variance to the sampling value.

The second and third series thus show no evidence of any seasonal variation beyond the first harmonic; this explains why the second harmonic, though probably improving the form of the curve, has been found to be statistically insignificant. In the Aberdeen—Rothamsted series, on the other hand, there would seem to be some progressive change beyond the second harmonic. The third harmonic was fitted, and was insignificant; an examination of the actual figures suggesting that the residue is to be found in periods of one seventh or one eighth of a year. Such a change can scarcely be regarded as seasonal, and as the total variance unaccounted for is small, we have not attempted to investigate it further.

As regards phase, that of the first harmonic is nearly constant; for the three series its maximum is at 14·°61, 22·°85, and 29°·52 from the origin, or at January 19, January 28, and February 3. These agree within the probable errors, and suggest agreement with the principal harmonic of temperature. The shortening of the spring decline is most marked in the Rothamsted—Aberdeen series, where it is reduced to about sixteen weeks (Feb. 20–June 15). For York—Aberdeen it is about eighteen weeks (Mar. 10–July 17), and for Rothamsted—York about nineteen weeks (Feb. 19–July 2). The actual maxima, comprised in a range of twenty days, thus show a closer agreement than the minima in a range of thirty-three days.

In the form of their curves, as in their other statistical characters, the Rothamsted—York and the York—Aberdeen series resemble each other, the visible differences in the curves being chiefly due to the correlation between the more northern pair occurring some three weeks later than that of Rothamsted—York. The Rothamsted—Aberdeen series, which covers the whole range in distance, is principally characterized by the strength of the second harmonic; this produces not only the shortening of the spring decline, noted above, but the autumn rise is broken by an almost stationary period of about three months. Whether there is a true maximum and minimum in this period may be disputed; it is at least certain that from the middle of September to the middle of December no marked change takes place in the correlation.

Finally, it will be noted that, plotted on the z scale, the amplitude of the seasonal variation in correlation displays, in these three series, no distinct change due to the varying distances between the stations.

4. Summary.

Forty years' rainfall data, divided into weekly periods, supply about 2000 values; this quantity is sufficient, with proper treatment, to yield information of some accuracy in respect of rainfall correlation.

The existing data for York, Aberdeen, and Rothamsted have been examined with a view to establishing adequate methods and exploring the main features of weather localization.

The mean correlation, measured on the z scale, falls off with increasing distance in a regular manner. It is not improbable that simple laws may connect these quantities over considerable areas, which will afford sufficiently exact knowledge of the accuracy of meteorological estimates, based on a limited number of stations.

A well-marked annual periodicity in the rainfall correlations exists in our region. The rainfall is most highly localized about July, and least so about the end of February. The correlation rises relatively slowly in the autumn, and for our longest distance (375 miles) the autumn values remain for about three months close to the mean value for the year.

COMMENT

The original paper is followed by a discussion which includes a reply by Fisher.