LIMITATIONS IN THE USE OF DOUBLE-MASS CURVES

(NOTE)

M. Wigbout*

ABSTRACT

The double-mass curve is a well-known tool for checking and adjusting inconsistencies in hydrological data caused by changes in the methods of observation or data processing. Some limitations in the use of this technique are mentioned, and an alternative method for investigating the data is suggested and applied to an example from the literature.

INTRODUCTION

The consistency of the observations in time at a given hydrological station can be checked by comparing the data with the pattern of data from one or more other stations in the area. For the sake of convenience only a reliable pattern is considered here. The same comments will also apply to other cases.

It is assumed that there are annual observations of p years, $y_1, \ldots, y_p$ for station A (which are to be checked) and $x_1, \ldots x_p$ for the reliable pattern. From these two records the cumulative data

$$Y_k = \Sigma_1^k y_t \text{ and } X_k = \Sigma_1^k x_t \qquad (k = 1, \ldots p)$$

are obtained, giving two derived records. The curve through the points (X_k, Y_k) in a graph is called a double-mass curve.

Usually the relationship between the two records is a fixed ratio and the double-mass curve a straight line. A break in this curve indicates a change in the relationship between the two records and warns the investigator that there may be something wrong with the data of station A. The break may be caused by errors in the data, by changes in the method of data observation or by the processing of the data.

* Water and Soil Division, Ministry of Works, Wellington.

The following discussion will be concentrated on an example from Searcy and Hardison (1960). The data are given in Table 1 and the corresponding double-mass curve has been drawn in Fig. 1.

According to Searcy and Hardison there was a break between the observations of 1938 and 1939. To test the change for significance they divided the observational results into two groups and carried out an analysis of covariance between these two groups. According to their calculations they used the following model:

$$\text{group 1: } y_t = a_1 + bx_t + e_t \quad (t = 1, \ldots, 18); \tag{1}$$

$$\text{group 2: } y_t = a_2 + bx_t + e_t \quad (t = 19, \ldots, 25); \tag{2}$$

with the e_t uncorrelated and from the same normal distribution with expectation 0 and some variance σ^2. With this model the means of the groups of the observations (both of stream A) were significantly different ($P = 0.02$).

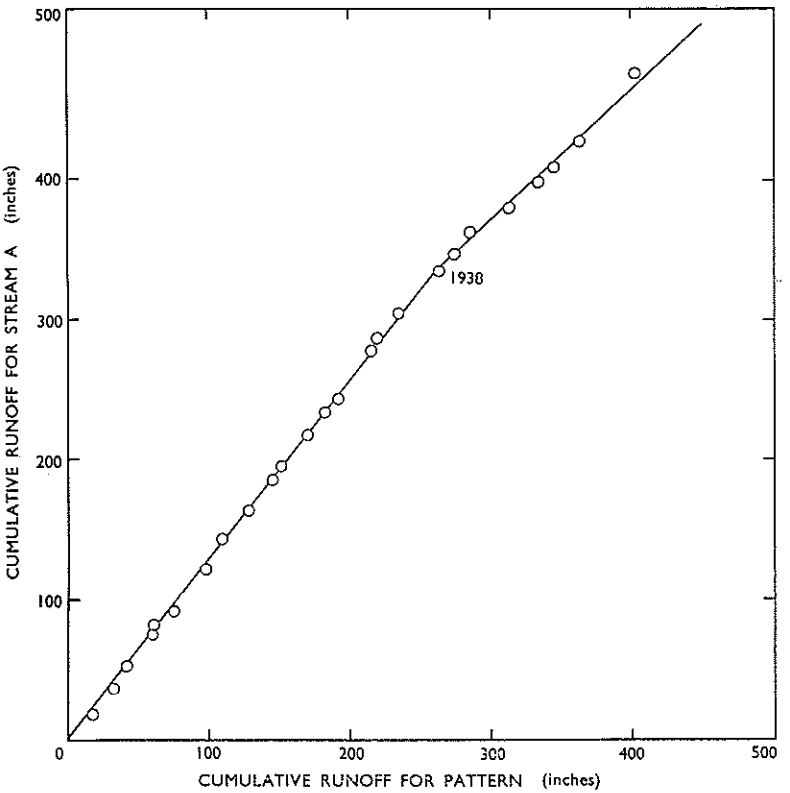

FIG. 1 — Double-mass curve of streamflow data, according to Fig. 2 of Searcy and Hardison (1960) and the data in Table 1.

TABLE 1 — Annual runoff (in inches), from Table 2 of Searcy and Hardison (1960).

Water year	Stream A		Pattern	
	Yearly (y_t)	Cumulative (Y_t)	Yearly (x_t)	Cumulative (X_t)
1921	19.73	19.73	19.61	19.61
1922	15.80	35.53	12.29	31.90
1923	17.52	53.05	8.12	40.02
1924	16.58	69.63	14.39	54.41
1925	5.33	74.96	3.53	57.94
1926	16.45	91.41	13.80	71.74
1927	30.67	122.08	24.03	95.77
1928	21.22	143.30	12.40	108.17
1929	21.96	165.26	19.70	127.87
1930	19.34	184.60	18.10	145.97
1931	9.87	194.47	5.13	151.10
1932	24.81	219.28	18.30	169.40
1933	15.53	234.81	12.20	181.60
1934	9.35	244.16	7.94	189.54
1935	32.75	276.91	25.58	215.12
1936	7.57	284.66	4.06	219.18
1937	19.72	304.38	13.76	232.94
1938	28.33	332.71	28.64	261.58
1939	15.04	347.75	10.41	271.99
1940	13.65	361.40	10.68	282.67
1941	17.42	378.82	30.15	312.82
1942	17.82	396.64	21.60	334.42
1943	9.41	406.05	8.96	343.38
1944	21.13	427.18	20.01	363.39
1945	37.85	465.03	40.25	403.64

The use of the same slope b in both regression equations (1) and (2) is curious, as the test was carried out because there was a different slope (according to the double-mass curve). In general it would make more sense to test whether separate regression lines for the two separate groups would give a better fit than one regression line with all observations together. To put it more exactly, the question is whether the fitting of the model

$$y_t = a_1 + b_1 x_t + e_t \quad (t = 1, \ldots, 18) \tag{3}$$

and

$$y_t = a_2 + b_2 x_t + e_t \quad (t = 19, \ldots, 25) \tag{4}$$

is significantly better than the fitting of one equation for all 25 observation pairs together:

$$y_t = a + b x_t + e_t \quad (t = 1, \ldots, 25) \tag{5}$$

Whatever statistical test is applied, there are some drawbacks to the use of a double-mass curve approach, mainly due to the fact

that the occurrence of certain phenomena (e.g. an outlier) will also affect the cumulative data in subsequent time periods. The more obvious difficulties are:

1. It is not easy to detect an outlier.
2. It is not always easy to determine a breakpoint (see Fig. 1).
3. It is not easy to obtain a good idea of the structure of the data – for example, an indication of change of variation.

The definitions of an outlier, a breakpoint and a change in variation are not given.* They depend mainly on the judgment of the investigator and on the data in any special case. It is important that physical explanations are given for the phenomena observed, such as a break in the curve being caused by a change in equipment.

As pointed out previously, the double-mass curve forms a straight line if the relationship between the record of station A and the reliable pattern is a fixed ratio, that is $y=bx$. With the models used, for example equation (5) ($y_t=a+bx_t+e_t$), the cumulative points are of the form

$$Y_k=ka+bX_k+\Sigma_1^k e_t \qquad (k=1,\ldots p)$$

and these points do not necessarily result in a straight line. The deviation from a straight line depends on the relationship between k and X_k. If for example $k=cX_k$, the equation can be written

$$Y_k=(ac+b)X_k+\Sigma_1^k e_t$$

which plots as a straight line. In practice the departure from a straight line will not be important because a is nearly always relatively small and because at least a part of the term ka (linearly increasing with k) can be put into the term bX_k (increasing with k).

USE OF DIFFERENCES

The problems connected with using the cumulative data suggest that it is better to use noncumulative data. To obtain an indication that there is something wrong with the data of station A it is necessary to eliminate the effect of the natural changes, which are represented in both records. For this purpose it seems to be useful to plot the differences y_t-x_t against time t. However, if the regression coefficient b in the model $y_t=a+bx_t+e_t$ is not equal to unity, the difference $y_t-x_t=a+(b-1)x_t+e_t$ still changes with the natural changes. So in general, y_t-bx_t should be used, because $y_t-bx_t=a+e_t$ is independent of x_t and y_t. A sudden change in this

* In a formal way these phenomena can all be seen as a change in the distribution of the variable x_t.

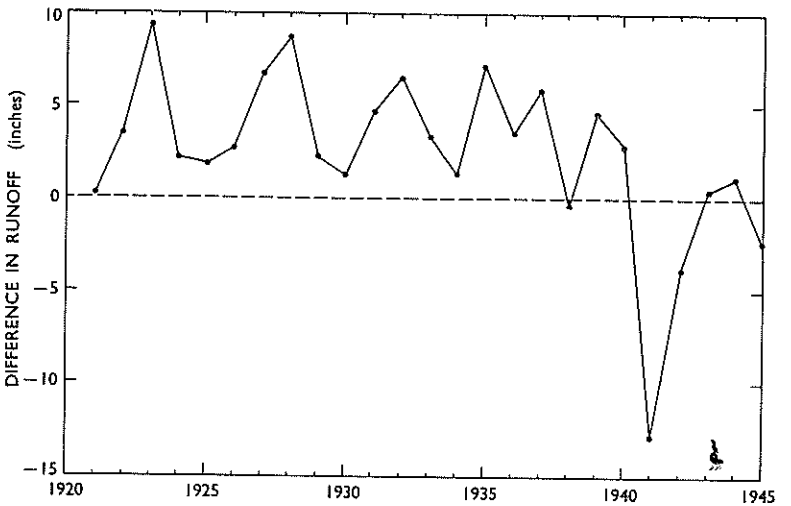

FIG. 2 — Differences in annual runoff of stream A and the reliable pattern ($y_t - x_t$: t=1921, 1922, . . . , 1945; in text t=1, 2, . . . , 25).

series of differences is an indication that there is something wrong with the data of station A.

To check the consistency of the record of station A, it appears reasonable to use the difference $y_t - x_t$ if the value of b is within the range approximately 0.8–1.2 (see Cox, 1957). This will be the case in many hydrological situations where x_t and y_t are of similar size (if they represent the same variable). In order to decide whether to use $y_t - x_t$ or $y_t - bx_t$, it is necessary to relate x_t and y_t. A rough estimate of the value of b may be obtained from other calculations or from a regression analysis carried out with all the data under consideration. The contradiction in using one value of b to determine a break, which is to be associated with a change in the value of b, is only theoretical. In practical applications the method will not be very sensitive to a small change of b.

As an illustration, the data from Searcy and Hardison (1960) are used again (Table 1). Because the y values and the x values for each year are of the same order of magnitude, the equation for the best-fitted regression line will be about $y=x$.* In Fig. 2 the differences $y_t - x_t$ are set out as a function of the years. It seems that there is a slow decrease in the differences with increasing time,

* In fact b=1.07 in the best-fitted line $y=bx$ (by the least-squares method) for all 25 observation pairs.

with an outlier in 1941. A check was made that the trend is not due to the fact that $y_t - x_t$ has been used instead of $y_t - bx_t$. With more information about the stations and the method of observation it would perhaps be possible to explain the structure in Fig. 2. Still, it is not clear that there is a break between the observations of 1938 and 1939, as was concluded by Searcy and Hardison. But for comparison with their results the same breakpoint is used here.

The mean of the 18 differences until 1938 (group 1) is $m_1 = 3.95$, the mean of the 7 differences from 1938 (group 2) is $m_2 = -1.39$ (in inches). With the 'modified' Student test – described, for example, by Pearson and Hartley (1966) – it was found that the means m_1 and m_2 were significantly different ($P = 0.02$).*

It should be noted that the use of the ordinary Student test here is doubtful, because the variations of the values of $y_t - x_t$ for the two groups considered $s_1^2 = 8.34$; $s_2^2 = 33.37$ are significantly different (F test: $P = 0.01$). The cause of this difference seems to be the observation value of 1941. Without this value the variance of the later data is $s_2^2 = 10.06$. The importance of the 'outlier' of 1941 can be seen also in Fig. 3. The dangers of using the ordinary Student test with different variances to test differences in means are discussed, for example, by Cochran and Cox (1966).

Although the 1941 observation seems to be an outlier, the data of the next years give an indication of a slow change. As sufficient details of the origin of the data are not available, it is impossible to explain the apparent trend.

CONCLUSIONS

Although the proposed method of plotting $y_t - bx_t$ against time introduces the difficulty of choosing the value of b, this method can give a clearer idea about the structure of the data than the double-mass curve method. This is obvious from the example given. As shown in Fig. 2, the break between the observations of 1938 and 1939 is not as clear as Searcy and Hardison (without mentioning a physical cause) suggested. Another advantage of the proposed method is that it is much easier to detect outliers. Or, more generally, with the proposed method a better idea can usually be obtained about the underlying structure of the data.

---

* Strictly speaking, the results of statistical tests are of limited value, because the data are first studied by eye to find what should be tested. Nevertheless, they can give a guide.

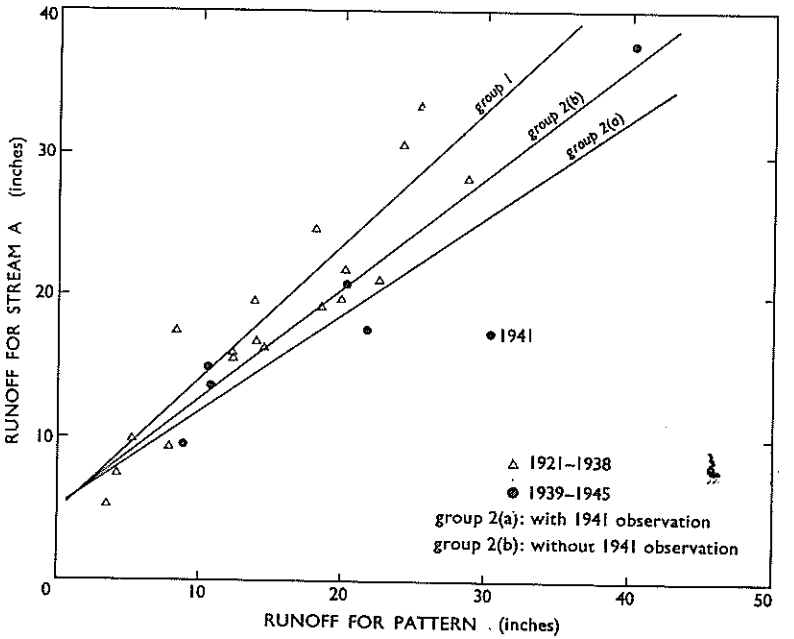

FIG. 3 — Relationships between the annual runoffs of stream A and the reliable pattern.

The discussion with the example also provides a warning against a superficial application of statistical techniques. Because of the 1941 outlier it was not permissible to test with an analysis of covariance as Searcy and Hardison did. With the proposed method it is easy to test with the modified Student test to get around the problem. In general, however, statistical tests should not be overemphasized. The most important thing is that obvious breaks and other phenomena should be physically explained.

ACKNOWLEDGMENT

Permission to publish this note was given by the Commissioner of Works.

REFERENCES

Cochran, W. G.; Cox, G. M. 1966: *Experimental Designs*. 2nd ed. Wiley, New York. (see section 3.9.)

Cox, D. R. 1957: The use of a concomitant variable in selecting an experimental design. *Biometrika 44*: 150–158.

Pearson, E. S.; Hartley, H. O. 1966: *Biometrika Tables for Statisticians*. vol. 1. 3rd ed. Cambridge University Press. (see p. 27 and Table 11.)

Searcy, T. K.; Hardison, C. H. 1960: *Double-mass Curves*. U.S. Geological Survey Water-Supply Paper 1541-B. USGS, Washington. 66 p.