U. S. Geological Survey (Reports, open file series)
Water Resources Division

Open file report
Sept. 1955
55-91

FITTING CURVES TO CYCLIC DATA

by W. B. Langbein

A common problem in hydrology is to fit a smooth curve to cyclic or periodic data, either to define the most probable values of the data or to test some principle that one wishes to demonstrate. This study treats of those problems where the length or period of the cycle is known beforehand - as a day, year, or meander length for example. Curve-fitting can be made by free-hand drawing, and where the data are closely aligned this method offers the simplest and most direct course. However, there are many problems where the best fit is far from obvious, and analytical methods may be necessary. There are many analytical methods available for smoothing, some of which are listed below:

1. Moving arithmetic average
2. Moving arc
3. Fitting polynomial by least squares
4. Double integration
5. Fitting Fourier (sine and cosine) series
6. Fitting special curves such as skewed cyclic probability functions by method of moments

Except for the use of moving arithmetic averages, the writer knows of no common application of the above methods in hydrologic practice. This article describes the moving arc and double integration methods because they are practical and seem to deserve consideration in hydrologic practice.

Moving arc method. The objective is to smooth out erratic departures without obscuring the significant fluctuations. The ordinary running average applied to data that conform, for example, to a simple sine curve will greatly attenuate and modify the given curve. The method is therefore useless in this connection. However, there are moving arcs, described in the actuarial literature for example, which can follow a curve successfully. Our references are Whittaker and Robinson (1932), and Sasuly (1934) who discuss the problem thoroughly. Fitting moving parabolic arcs seems simplest. The respective weights for parabolic arcs of various lengths are given below:

Table 1 - Weights for n-point least-square parabolas. *

5	7	9	11	13	15	17	19	21
-3	-2	-21	-36	-11	-78	-21	-136	-171
12	3	14	9	0	-13	-6	-51	-76
17	6	39	44	9	42	7	24	9
12	7	54	69	16	87	18	89	84
-3	6	59	84	21	122	27	144	149
35	3	54	89	24	147	34	189	204
	-2	39	84	25	162	39	224	249
	21	14	69	24	167	42	249	284
		-21	44	21	162	43	264	309
		231	9	16	147	42	269	324
			-36	etc	etc	etc	264	329
			429	143	1105	323	etc	324
							2261	etc
								3059

Central value underlined

* Whittaker and Robinson, p. 295

To illustrate how the above weights can follow a curve, consider the following values on a sine curve

n	y
1	0
2	.276
3	.500
4	.707
5	.866
6	.966
7	1.000
8	.966
9	.866

A straight arithmetical average of 5 points centered on the 7th item above for example gives a value of 0.933 compared with 1.00 given. Applying the 5 point parabolic weights gives as a result 1.00.

Table 2 illustrates the application of a 5-point moving parabolic arc to the residual monthly corrections to a correlation between the flow of two streams in Utah. In this problem, the smoothed values of the monthly corrections are believed to be superior to the corrections as originally computed, because logically there should be a uniform variation in these corrections among the months. It is desired, however, not to destroy the intrinsic character of the variations.

The procedure is as follows: The corrections as originally computed are listed under y in table 2 and plotted on figure 1. Each value of y is then multiplied successively by each weight; the value for January, -.07, is multiplied by -3, by 12, and by 17. To obtain the 5-point total for April, for example, the weighted values for February, March, April, May and June are totaled: +.15 -.48 +1.36 +2.52-66 = +2.89. The 5-point average is found by dividing 2.89 by the sum of the weights, 35, to yield +.083.

There are problems where the data indicate a less obvious smoothed curve; i.e., where there is a considerable random component.

Table 2. - Illustration of use of 5 point moving parabolic arc.

	y	Product of y times indicated weight -3	+ 12	+ 17	5 point total	5 point average
Jan	-0.07	+0.21	-0.84	-1.19		
Feb	-.05	+.15	-.60	-.85		
Mar	-.04	+.12	-.48	-.68	-.74	-.021
Apr	+.08	-.24	+.96	+1.36	+2.89	+.083
May	+.21	-.63	+2.52	+3.57	+7.05	+.20
June	+.22	-.66	+2.64	+3.74	+7.16	+.205
July	+.08	-.24	+.96	+1.36	+3.01	+.086
Aug	-.06	+.18	-.72	-1.02	-1.77	-.050
Sept	-.12	+.36	-1.44	-2.04	-4.23	-.12
Oct	-.13	+.39	-1.56	-2.21	-4.55	-.13
Nov	-.11	+.33	-1.32	-1.87	-3.82	-.109
Dec	-.08	+.24	-.96	-1.36	-2.98	-.085
Jan	-.07	+.21	-.84	-1.19	-2.30	-.066
Feb	-.05	+.15	-.60	- .85	-2.17	-.062
Mar	-.04	+.12	-.48	-.68		

Note. - In this example, the cycle is closed, from December to January at the beginning.

The following is an illustration of the efficiency of the moving parabolic arc method in eliminating random variation from a set of cyclic data. An original graph of some variable defined by 52 points shown in Figure 2 was altered by adding quantities randomly selected to the values as read from the given graph. The test is to see how well the original graph can be reconstructed from the randomized points. Using a 21-point parabola, the following results were obtained for selected items:

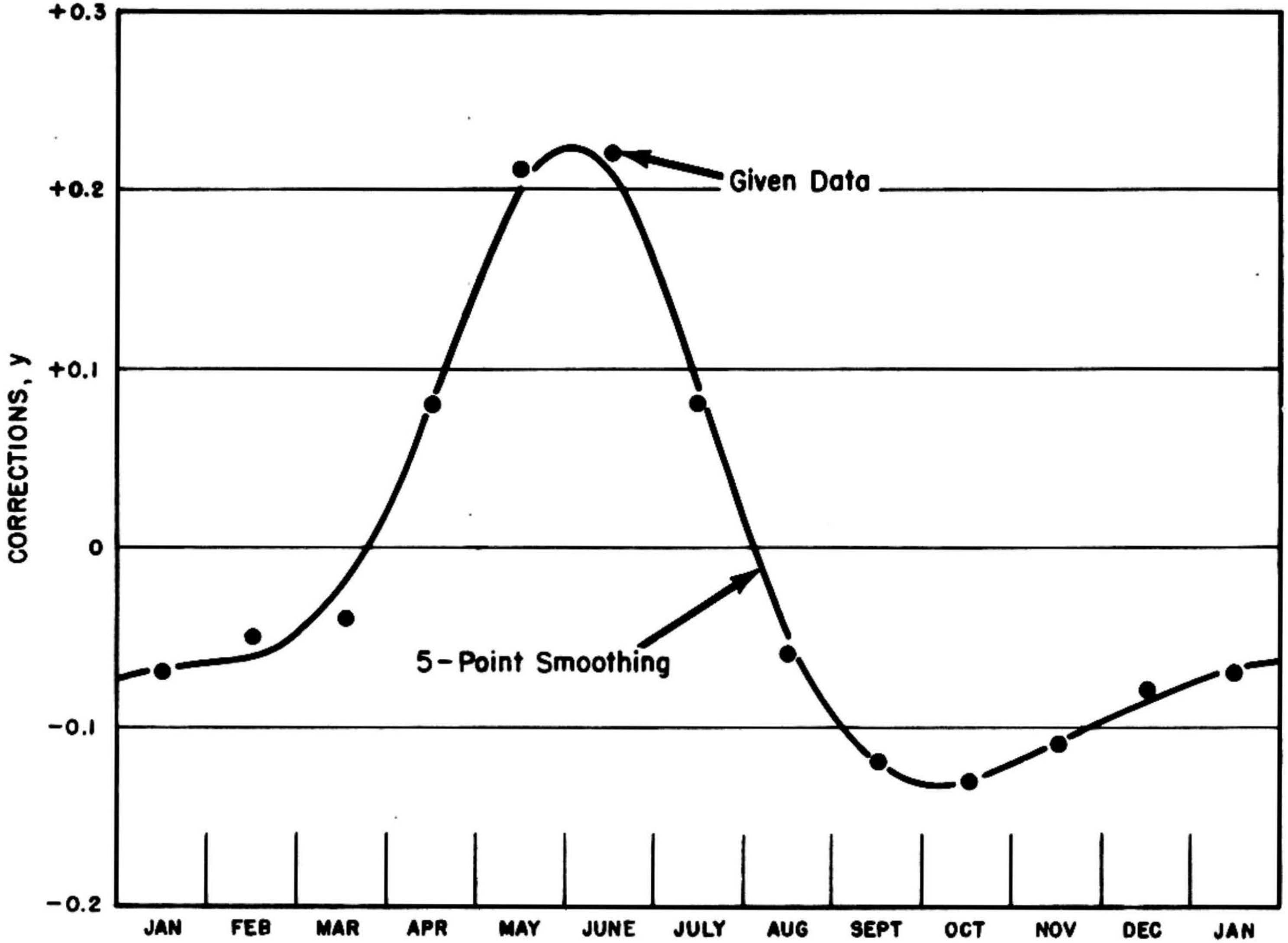

Figure 1.-- Illustration of smoothing by 5-point moving arc.

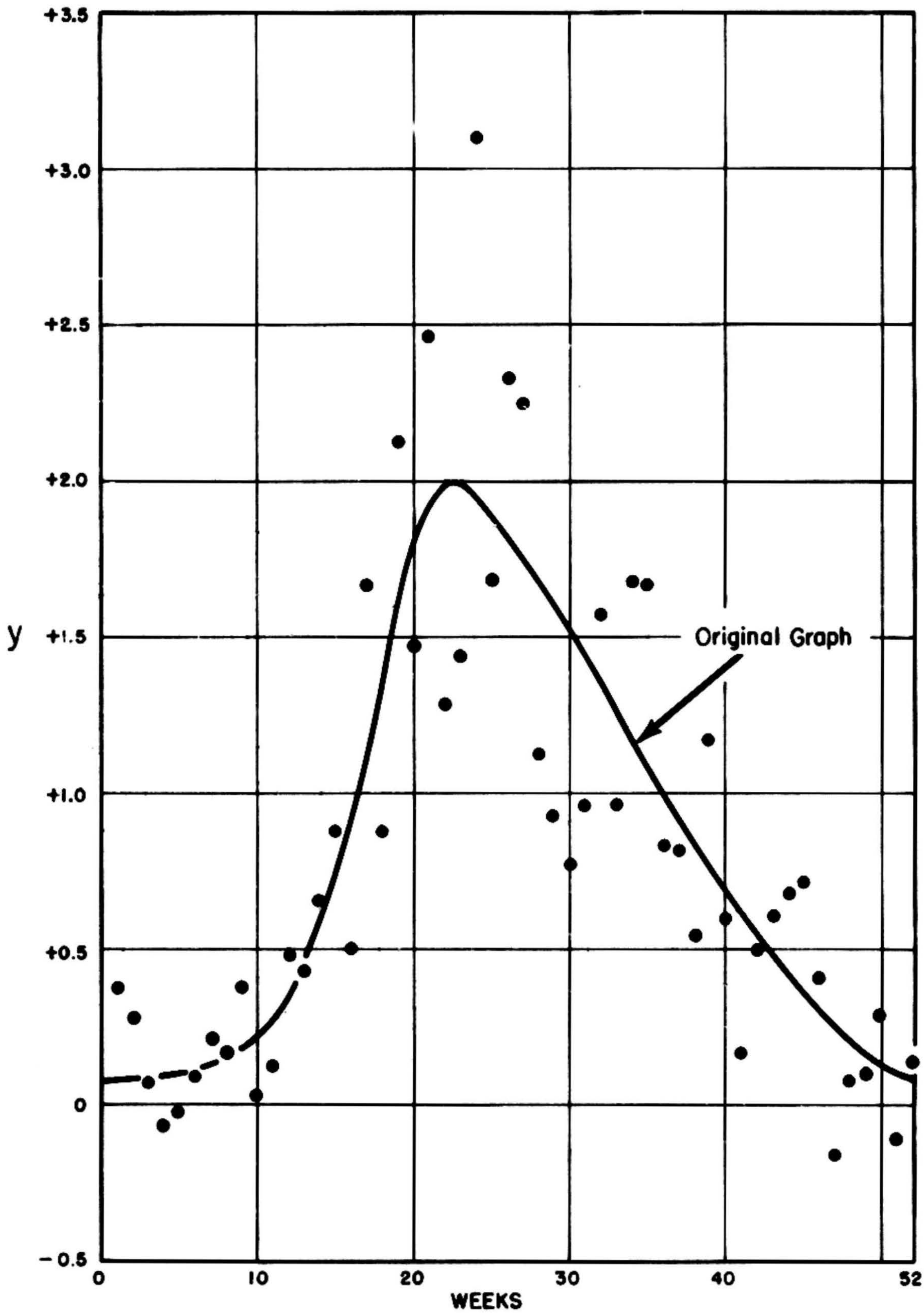

Figure 2.-- Test example; random quantities added to given graph.

Item number	Computed value	Given value
0 or 52	0.10	0.08
10	.22	.22
22 (peak)	1.87	2.00
25	1.79	1.88
30	1.50	1.49
40	.77	.72

The standard deviation of the random points about the mean curve is 0.35 and therefore a derived curve will have an irreducible standard error of $0.35/\sqrt{52} = 0.05$ regardless of the method used.

The moving arc has also the advantage that it makes the choice of number of points less significant upon the final result. The short arc will of course provide close local fits, but will provide little averaging. On the other hand long arcs may tend to alter the intrinsic character of the cycle. A useful compromise would be to use about one-half the number of points available for defining the cycle. An odd number of items in the arc is advised in any case because it provides a definite central value. To provide sufficient averaging of random components at least 5 points should be used, which requires at least 10 points in the cycle for use of a parabolic arc. Where there are fewer than 10 points in the cycle, the 5 point parabola will not make a satisfactory fit. If there are fewer points available in the cycle, a 7 or 9 point quartic arc should be used with weights given below. In this case, some of the points will be used twice in both tails of the weighting procedure.

Table 3. - Weights for 7 and 9 point least square quartic arcs

7 point quartic	9 point quartic
5	+ 15
-30	-55
75	+30
131	+135
75	+179
-30	+135
5	+30
231	-55
	+15
	429

Double integration. This is an adaptation of a process used by Powell (1930) for "cycle seeking." Let us suppose that the data define a graph of the simple form

$$y = a + b \cos \frac{2\pi x}{p} \pm \epsilon$$

in which, a represents the mean, b the amplitude, p the length of the cycle, x the position in the cycle, and ϵ a random variable. If we substract a, the mean from each of these values, and take the progressive totals or first integral of these differences, we get a series of the form

$$\frac{bp}{2\pi} \sin \frac{2\pi x}{p} + \sum \epsilon$$

A second integration will give a series of the form

$$-b\left(\frac{p}{2\pi}\right)^2 \cos \frac{2\pi x}{p} \pm \sum\sum\epsilon$$

The term $\sum\sum\epsilon$ will in general be small, since the positive chance variations in ϵ will tend to balance the negative terms.

This series is in phase though opposed in sign with the original data and substantially all random variations will have been averaged out.

The result of the second integration is not necessarily a simple sine curve - it may be asymmetric depending on the nature of the basic data but in general it tends strongly towards the sine curve in form. Further integration will tend toward further smoothing until the pure sine curve is produced after which further integration has no modifying effect. The second integration provides adequate smoothing and so the integrating process need go no further. The double integration method is peculiarly adapted to those problems where a sine curve may be considered a close approximation of the result sought.

The process is illustrated by the following computations. The data shown on figure 3 and in table 4 are for monthly evapotranspiration from ground water as determined from a hydrologic budget for Beaverdam Creek in eastern Maryland. Although the monthly amounts of evapotranspiration may vary because of several factors, the dominant controls are those associated with the time of year. One would therefore expect a uniform progression from month to month through the course of the year. The erratic variations represent, it is believed, errors because the figures are the residuals between relatively large quantities.

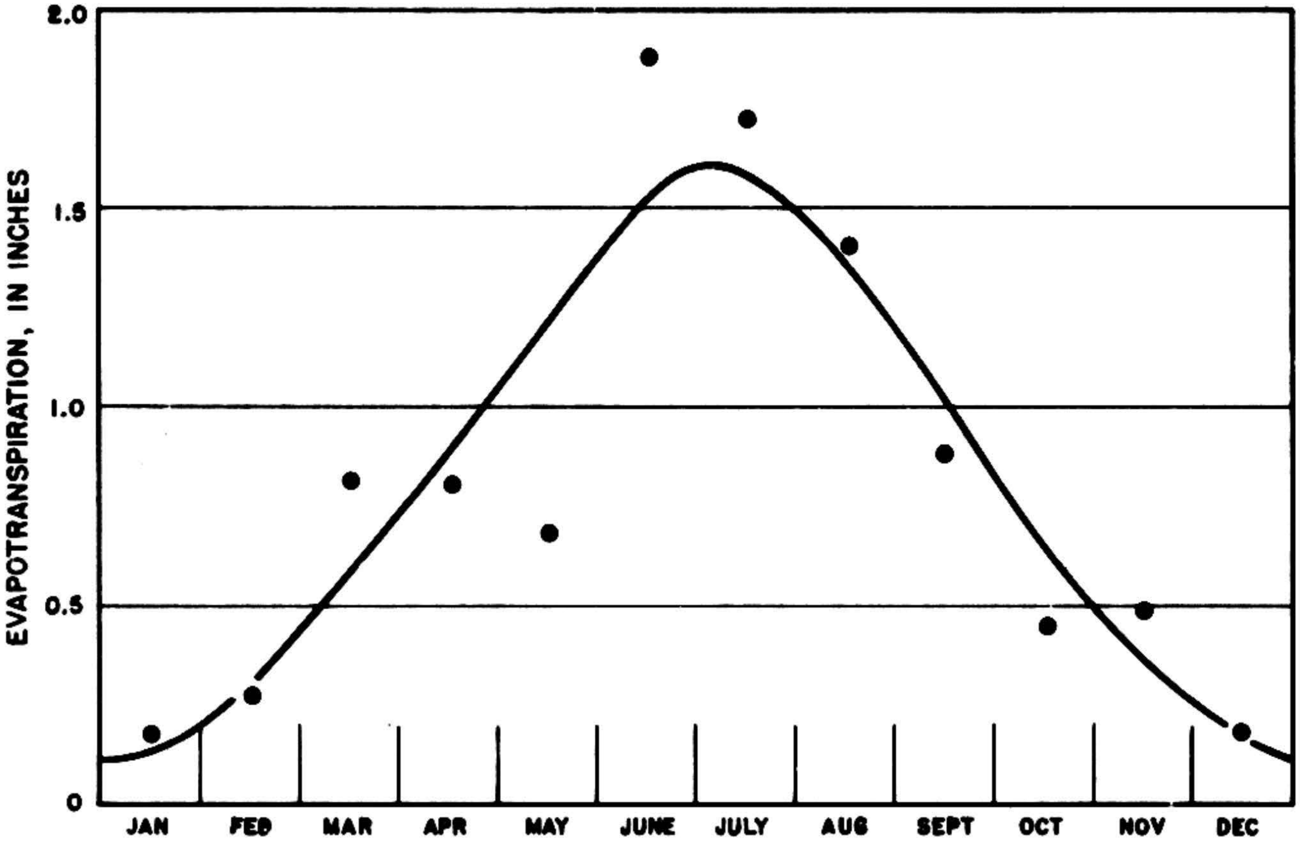

Figure 3.--Result of fitting a cycle graph by double integration.

A 5-point moving arc could be used, but where the cycle is of a sinusoid nature, double integration will do the job more simply.

The first data column in table 4 lists the given values of computed evaportranspiration. The second column lists the deviations from the mean of the data. The third column shows the cumulative summation of the deviations, and completes the first integration, sometimes known as the mass residual. The process is repeated in the fourth and fifth columns and a second integration made as shown in the fifth data column. Deviations for the mean of the second integration are shown in the sixth column. These figures are divided by the constant $-(12/2\pi)^2$ in which the numerator 12 corresponds to p the number of items in the cycle, and the quotients are listed in column 7. These quotients in the seventh column represent the smoother or average deviations from the mean of the data, in this example 0.81. Adding these to the mean gives the final smoothed results, which are shown on figure 3, which represents a reasonable interpretation of the original data.

Table 4. - Example of double integration applied to monthly data.

Month	1 Computed evapo-trans-piration	2 Dev. from mean	3 1st Integra-tion	4 Dev. from mean	5 2nd Integra-tion	6 Dev. from mean	7 Dev. $-(12/2\pi)^2$	8 Smoothed values
Jan	0.18	-0.63						
			-0.63	-0.57				
Feb	.27	-.54			-0.57	+1.91	-0.52	0.29
			-1.17	-1.11				
Mar	.82	+.01			-1.68	+.80	+.22	.59
			-1.16	-1.10				
April	.80	-.01			-2.78	-.30	+.08	.89
			-1.17	-1.11				
May	.68	-.13			-3.89	-1.41	+.39	1.20
			-1.30	-1.24				
June	1.88	+1.07			-5.13	-2.65	+.73	1.54
			-.23	-.17				
July	1.72	+.91			-5.30	+2.82	+.77	1.58
			+.68	+.74				
Aug	1.40	+.59			-4.56	-2.08	+.57	1.38
			+1.27	+1.33				
Sept	.88	+.07			-3.22	-.74	+.20	1.01
			+1.34	+1.40				
Oct	.44	-.37			-1.82	+.66	-.18	.63
			+.97	+1.03				
Nov	.48	-.33			-.79	+1.69	-.46	.35
			+.64	+.71				
Dec	.18	-.63			-.08	+2.40	-.66	.15
			+.01	+.08				
Jan					0	+2.48	-.68	.13
Means	.81	0	-.06	0	-2.48	0	0	.81

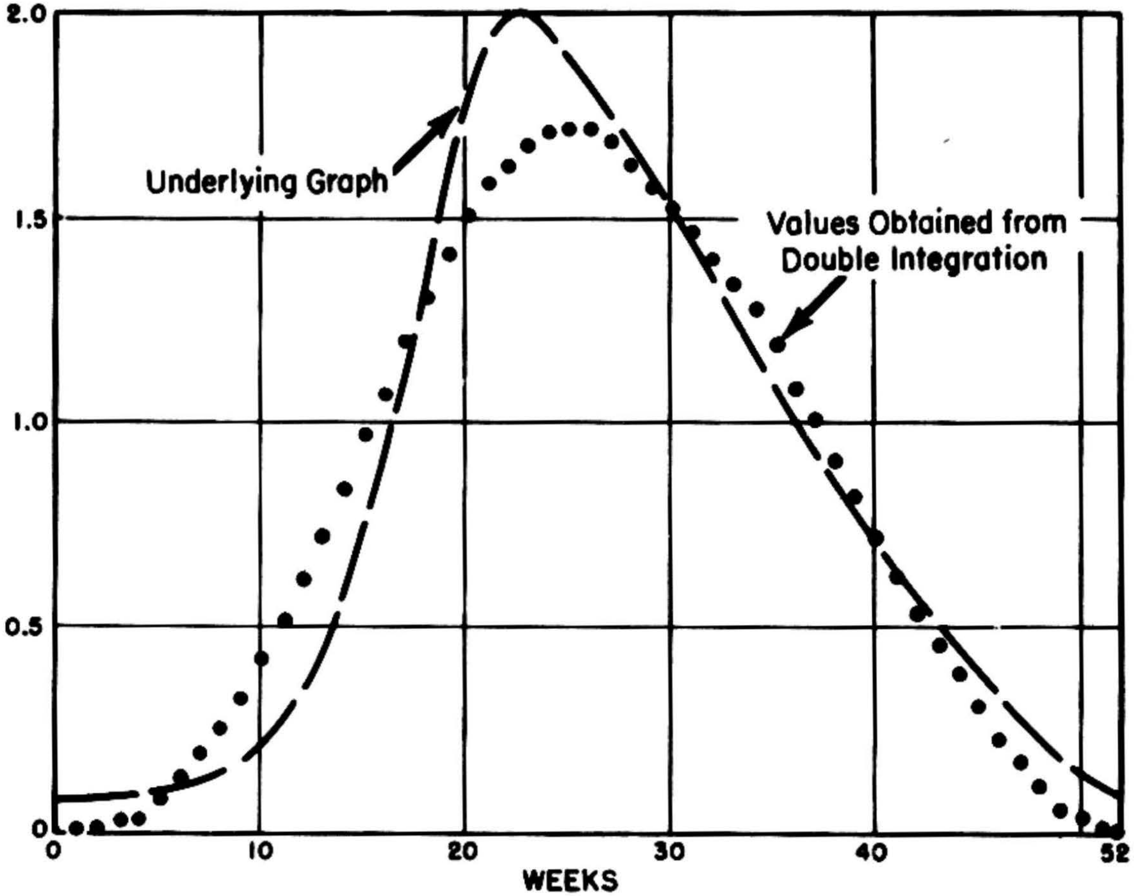

Figure 4.--Double integration method applied to the example of Figure 2.

Figure 4 shows the same method applied to a previous example. In that example the moving arc method gave a closer approximation to the original curve, since it provides a local fit. The double integration method fits a sinusoidal curve to the data as a group.

Summary. The moving arc methods are quite general. The parabolic weights should not extend beyond a reversal in curvature, otherwise quartic weights should be used. The double integration method provides a simpler solution to those problems where the cycle is of a sinusoidal form.

References

Powell, R. W., Successive integration as a method for finding long-period cycles, Annals. of Math. Statistics, pp. 123-136, May 1930.

Sasuly, Max, Trend analysis of statistics, Brookings Institute, 1934.

Whittaker, E.T., and Robinson, G., The calculus of observations, Blackie & Son, London, 1932.