straight-line interpolations themselves. Now introduce some correlation between the movements in Y and X. Until the correlation reaches $\frac{1}{2}\sigma_v/\sigma_u$, it serves only to offset part of the damage done by the uncorrelated variation in Y; the net effect is an improvement only when the correlation exceeds $\frac{1}{2}\sigma_v/\sigma_u$.

It may be worth emphasizing that the relevant correlation is that between u and v, not that between X and Y. The former may be quite low even though the latter is quite high because of high serial correlation between the successive values of X and Y (see the formulas in Appendix Note 1). This is one of the major reasons why graphic inspection of time series plotted in their original form may be extremely misleading in judging the value of a series as an interpolator. Sad experience persuades me that it is not easy to find an interpolator for which the relevant correlation is above the critical value. In view of the widespread use of method (1), and the rather casual way in which interpolators are often chosen, I would not be at all surprised to find that in practice the use of related series generally gives larger errors than straight-line interpolation.

III. CORRELATION METHODS

A. *Method (b) and the Errors Associated with It*

Both M_1 and M_0 can be regarded as special cases of a more general method, which we may call method (b), or M_b, and which consists of estimating u by the following:

$$_bu^* = bv. \tag{22}$$

If $b = 1$, this is M_1; if $b = 0$, this is M_0.

The error involved in using M_b is

$$_bd = {_b}u^* - u = bv - u, \tag{23}$$

so the mean error is

$$\mu_{bd} = \mu_{bv-u} = E(bv - u) = b\mu_v - \mu_u, \tag{24}$$

which will, of course, be zero if $\mu_v = \mu_u = 0$,[10] and the mean square error is

$$\text{M.S.E. } (M_b) = E(_bd^2) = E(bv - u)^2 = \sigma_{bv-u}^2 + \mu_{bv-u}^2$$

$$= \sigma_u^2 + b^2\sigma_v^2 - 2b\rho_{uv}\sigma_u\sigma_v + \mu_u^2 + b^2\mu_v^2 - 2b\mu_u\mu_v. \tag{25}$$

The results of the preceding section can, of course, all be derived from these formulas by setting b equal to zero and 1, respectively.

We may now ask what value of b (say, β) is optimum in the sense of minimizing the mean square error. Differentiating the right-hand side of (25) and setting the derivative equal to zero gives

$$2\beta\sigma_v^2 - 2\rho_{uv}\sigma_u\sigma_v + 2\beta\mu_v^2 - 2\mu_u\mu_v = 0, \tag{26}$$

---

[10] But not, as in M_1, if $\mu_v = \mu_u \neq 0$.

from which

$$\beta = \frac{\rho_{uv}\sigma_u\sigma_v + \mu_u\mu_v}{\sigma_v^2 + \mu_v^2}.$$ (27)

Inserting this value of b in (25) and simplifying, we can write the minimum mean square error attainable with M_b as

$$\text{M.S.E. } (M_\beta) = \sigma_u^2 + \mu_u^2 - \frac{(\rho_{uv}\sigma_u\sigma_v + \mu_u\mu_v)^2}{\sigma_v^2 + \mu_v^2}$$ (28)

or, as a ratio to the mean square error of linear interpolation,

$$\frac{\text{M.S.E. } (M_\beta)}{\text{M.S.E. } (M_0)} = 1 - \frac{(\rho_{uv}\sigma_u\sigma_v + \mu_u\mu_v)^2}{(\sigma_v^2 + \mu_v^2)(\sigma_u^2 + \mu_u^2)}.$$ (29)

For the case in which we are primarily interested, namely, that for which $\mu_u = \mu_v = 0$, (27) and (29) become

$$\beta(\mu_u = \mu_v = 0) = \rho_{uv}\frac{\sigma_u}{\sigma_v},$$ (30)

$$\left[\frac{\text{M.S.E. } (M_\beta)}{\text{M.S.E. } (M_0)}\right]_{(\mu_u = \mu_v = 0)} = 1 - \rho_{uv}^2.$$ (31)

This method can be regarded as estimating u by a weighted average of the estimates given by M_0 and M_1, with the estimate given by M_0 weighted by $(1 - \beta)$, and that given by M_1, by β. Apart from a scale factor (σ_u/σ_v), the weight given to the related series is greater the higher the correlation between the movements of the original and related series.

With this method, the use of the related series yields an improvement over linear interpolation whenever there is any correlation between the movements in the two series, which is clearly what a satisfactory method should do. The improvement is, of course, small if the correlation is low and increases as the size of the correlation increases.

It will be noticed that β as given in (30) is simply the slope of the regression of u on v. This is as it should be, for we have been traversing familiar ground by a somewhat unfamiliar route suggested by the form of our particular problem.

As noted earlier, the basic problem is to estimate an (unknown) value of u from a known value of v, where u and v are two correlated variables. The estimate of u from v with minimum variance is given by the regression of u on v which is

$$(u - \mu_u) = \rho_{uv}\frac{\sigma_u}{\sigma_v}(v - \mu_v),$$ (32)

or, if $\mu_u = \mu_v = 0$,

$$u = \rho_{uv}\frac{\sigma_u}{\sigma_v}v.$$ (33)

15

The practice of using the regression $u = v$ (for this is, of course, what M_1 amounts to) instead of (33) has continued to be so widespread partly because the problem is so seldom stated in this form and partly, as already noted, because more information is required to use (33).

The earlier statements about the desirability of using seasonally adjusted data can be derived from (32). The use of method (b) with seasonally unadjusted data and with b set equal to $(\rho_{uv}\sigma_u/\sigma_v)$ is equivalent to using (33) instead of (32) even though μ_u and μ_v are not equal to zero. If $\mu_u = \mu_v$, but their common value is not zero, (32) and (33) will give the same results only if $(\rho_{uv}\sigma_u/\sigma_v)$ is set equal to or replaced by 1, which is why it makes no difference whether M_1 is applied to seasonally adjusted or unadjusted data when the seasonal in Y is used to estimate the seasonal in X, while it does make a difference whether M_b, with $b \neq 1$, is applied to seasonally adjusted or unadjusted data. Since (32) is the regression that yields minimum variance, it yields better results than (33) when μ_u and μ_v are not equal to zero. But (32), with $\mu_u = \mu_v$, is equivalent to first removing the seasonal from Y, then using the adjusted Y to interpolate a seasonally adjusted X, then putting the seasonal computed from Y back into X.

All of the results of linear regression theory apply to our special case and hence we know that if ρ_{uv}, σ_u, and σ_v are known and if $\mu_u = \mu_v = 0$, (33) is the "best" estimate of u from v in a number of senses of "best."[11]

In practice, of course, neither condition is strictly satisfied. It remains, therefore, to consider the estimation of ρ_{uv}, σ_u, and σ_v; and the effect of nonzero means of u and v.

B. *Estimating Needed Parameters*

The estimation of the parameters needed to use M_β is closely connected with the selection of a related series for interpolation in the first instance. The crucial desideratum in selecting an interpolator is that its deviations from a trend for the dates for which interpolation is to be done have a high correlation with the corresponding deviations for the series being interpolated. But of course the values of the series being interpolated are unknown for those dates, else interpolation would be unnecessary, so the size of the correlation cannot be computed directly. It is generally estimated indirectly by examining the correlation either for other dates or time intervals for which both series are known or for a similar but not identical pair of series for the same dates and then assuming that what is true for other dates or time intervals or series is also true for the interpolation dates or time intervals or interpolated series. (See examples in Section I).

---

[11] The estimate of u from (33) under the stated conditions is clearly unbiased and has minimum variance. If the distribution of u and v is a normal bivariate distribution, it is also the maximum likelihood estimate, uniformly most powerful, etc., etc. In some of these other senses of the term "best," it may not be the "best" estimate for some nonnormal distributions.

The qualification "from v" is included because it may be that a better estimate can be constructed by using more information. In particular, if we return to our original variables, one can describe the problem as that of estimating x_1 from x_0, x_2, y_0, y_1, y_2. Assume that all means, standard deviations, and correlation coefficients are known. Then the "best" estimate of x_1 would be given by the regression of x_1 on the other five variables. See Appendix Note 3 for the conditions under which (33) will give the same results as this multiple regression.

The problem can, of course be generalized still further by taking into account either more values of X and Y, or additional related series.

16

The important point is that there exists some pair of "test" series, the correlation between which is regarded as an estimate of the correlation between the interpolator and the values to be estimated by interpolation. This pair of "test" series then also provides a basis for constructing estimates of ρ_{uv}, σ_u, and σ_v. Values analogous to u and v can be computed for the test series. (E.g., if annual data are used to judge monthly interpolators, the difference between each annual observation of one series—when it has been put in the form, e.g., logarithms, to be used—and the straight-line trend connecting the preceding and succeeding year is a value analogous to u; the corresponding difference for the other series is a value analogous to v.) From these, the required estimates can be constructed.[12]

Another way of constructing estimates of ρ_{uv}, σ_u, and σ_v is to compute estimates of the relevant parameters of the original X and Y series (or the series taken as representative of them) and then use the formulas in Appendix Note 1 to convert these into estimates of ρ_{uv}, σ_u, and σ_v. Of course, if precisely the same data are used as in the preceding method, the results will be identical, aside from arithmetical errors or errors introduced by rounding. The advantage of this method is that it may be possible to use more extensive or more suitable data, since data may be available for some parameters of the original series that are not available for the parameters of the transformed u and v series. The disadvantage of this method is that it is likely to involve combining evidence from different sets of series and hence may introduce errors arising from the heterogeneity of the different sets of data. The question of the circumstances under which this method will yield good results clearly needs more study. Though my initial hunch was that this method would often be useful and generally decidedly superior to the preceding method, the limited experiments I have made point in the opposite direction.[13]

One point of a somewhat different kind may perhaps be mentioned here, though it is relevant to other issues as well as to the estimation of parameters needed for interpolation. The series for which values must be interpolated is frequently a component of a broader series. For example, data may be available monthly on employment in a sample of firms and annually on employment in all firms. Then the series to be interpolated is employment in the firms that are not in the sample, not in all firms. Deposit data may be available monthly for all weekly-reporting banks, at irregularly spaced "call dates" for all member banks, and annually for all banks. Then nonweekly-reporting

---

[12] The assumption that $\mu_u = \mu_v = 0$ is likely to have as much basis for the test series as for the basic series. If so, it may be desirable to modify the usual statistical formulas for standard deviations and correlation coefficients by using deviations from zero rather than from the observed means of u and v, and dividing sums of squares by the number of observations rather than that number less one. The advantage is, of course, gaining one degree of freedom.

[13] Note that the additional data available with this method include the basic X, Y series for the period to be interpolated, since these can be used, alone or together with other data, to estimate:

$$\rho_{x_0 x_2}, \; \rho_{y_0 y_2}, \; \rho_{x_0 y_2}, \; \rho_{x_2 y_0}, \text{ as well as } \sigma_{x_0}, \sigma_{x_2}, \sigma_{y_0}, \sigma_{y_2}.$$

Indeed, if assumptions (vi), (vii), and (viii) of Appendix Note 1 can be accepted, external data are required solely to estimate the correlation coefficients involving x_i and y_i. Even these might be obtained from the basic series themselves by interpolation between the other correlation coefficients. The conditions in Appendix Note 3 suggest how to interpolate the correlation coefficients. Some of the problems involved in combining different bodies of data are dealt with in T. W. Anderson, "Maximum Likelihood Estimates for a Multivariate Normal Distribution When Some Observations are Missing," *Journal of the American Statistical Association*, 52(1957), 200–3.

member banks are to be interpolated monthly between call dates; nonmember banks, monthly between years. In such cases, the component available for shorter intervals is frequently used to interpolate the component available less frequently. Symbolically, the final series desired is $A+B$, A is available at the desired time intervals, B is available only at longer intervals, and A is used (or to be tested for use) to interpolate B. In practice, A is often used to interpolate $A+B$ rather than B alone. For some methods of interpolation (e.g., M_1 when the data are in the form of ratios to linear trends or M_β, when the value of β is correctly adapted to the series being interpolated) the results are arithmetically identical whether A is used to interpolate $A+B$ or to interpolate B. But for other methods of interpolation (e.g., M_b for $b \neq 0$, when the same value of b is used for $A+B$ as for B alone), the results are not identical. Even more important, the correlation that is relevant in judging whether A is a good interpolator is between A and B, not between A and $A+B$. The latter is almost bound to be higher than the former, since it involves correlating A partly with itself; and it may be near unity when the correlation between A and B is near zero.[14]

A similar pitfall arises when a series that is not a component of the final series desired is used as an interpolator for one of the components. Call it C. Then the use of C to interpolate $A+B$ will almost always give results that are different from (and obviously inferior to) those obtained by using C to interpolate B alone. And again the correlation that is relevant in judging C as an interpolator is between C and B, not between C and $A+B$.

We may summarize this point in the form of an important practical maxim (maxim III): *Perform interpolation only on the part of a series that is unknown for the dates for which interpolation is to be done; never on a broader total, part of which is known for those dates.*

C. Effects of Errors in Estimates of Needed Parameters If Means of u and v Are Zero

Suppose that estimates of ρ_{uv}, σ_u, σ_v are available and are used to obtain an estimate $\hat{b}$ of β, namely

$$\hat{b} = r_{uv} \frac{s_u}{s_v} \cdot \tag{34}$$

The sampling error of this estimate can, of course, be approximated by the usual statistical procedures. There remains the question how sensitive our preceding conclusions are to the error made in estimating β, i.e., to the deviation of $\hat{b}$ from β. In particular, how large an error can be tolerated without making the mean square error of M_b greater than that of M_1 or than that of M_0?

These questions can readily be answered from (25), which is valid for any value of b. We can therefore compare the mean square error for M_b with that for M_0 and for M_1 and determine the values of $\hat{b}$ for which each method is bet-

---

[14] These comments seem so obvious that I am tempted to apologize for making them. But the problem seldom appears in the naked form in which I present it. In consequence, I suspect that one of the major factors responsible in practice for the use of poor interpolators is the tendency to correlate (numerically or by graphic inspection) A with $A+B$.

ter or worse than the others. Let us restrict ourselves to cases for which $\hat{b}$ is positive (i.e., positive observed correlation between u and v) and for which we can take $\mu_u = \mu_v = 0$. Let the symbol $M_{b_1} > M_{b_2}$ indicate that method b_1 is better (i.e., yields a smaller mean square error) than method b_2. The results depend, as might be expected, on the value of β, i.e., the "optimum" value of b, and can be summarized as follows:

$$M_1 \gtreqless M_0 \text{ according as } \frac{1}{2} \lesseqgtr \beta$$

$$M_{\hat{b}} \gtreqless M_0 \text{ according as } \frac{1}{2}\, b \lesseqgtr \beta \qquad\qquad (35)$$

$$M_{\hat{b}} \gtreqless M_1 \text{ according as } \frac{1}{2}\,(\hat{b}+1) \gtreqless \beta \qquad \text{for } \hat{b} < 1$$

$$\frac{1}{2}\,(\hat{b}+1) \lesseqgtr \beta \qquad \text{for } \hat{b} > 1$$

The various combinations obtained by putting these results together are summarized on Figure 1. The area indicated by cross-hatching is that for which M_0 is the best of the three methods; that by shading, for which M_1 is; and the white area, that for which $M_{\hat{b}}$ is. The $\hat{b} = \beta$ line shows the "optimum" value of b. In general, substantial errors are tolerable in $\hat{b}$ without rendering $M_{\hat{b}}$ worse than the other two methods. If $\beta \leq \frac{1}{2}$, the error in estimating β must exceed 100 per cent of β to render M_b worse than either of the other two methods; if $\beta \geq \frac{1}{2}$, the error in estimating $(1-\beta)$ must exceed 100 per cent of $(1-\beta)$ to do so. These are narrow ranges only when β is in the neighborhood of 0 and 1, i.e., when the interpolator has zero correlation or (if $\sigma_u = \sigma_v$) a very high correlation with the series being interpolated.

D. *Effect of Nonzero Means of Deviations from Trend*

We have so far supposed that μ_u and μ_v can be taken equal to zero. It will be recalled that this case, while it may appear highly special, is in fact extremely general since it includes the case in which the values of μ_u and μ_v are known and interpolation is used to estimate the deviations of u and v from these known values.[15] It will be recalled also that if it is considered impossible to specify anything about either the values of μ_u and μ_v or the relation between them, then v (or the series Y from which it is derived) is hardly to be considered eligible as an interpolator of u (or the series X).

There remains the case when μ_u and μ_v are considered unknown, but something is considered known about the relation between them. In particular, we may suppose that the two means are equal, i.e., $\mu_u = \mu_v$. It is hard to conceive of any practical example of this situation. In general, evidence that v (or Y) is a "good" interpolator of u (or X) must be based on data for a period of time which may or may not be the same period as that to be interpolated. The acceptance of such evidence implies that values of v and u are in some sense to be

---

[15] In turn, this includes the case in which μ_u is taken equal to μ_v and the latter is estimated from the known values of v for the various dates, as well as such additional methods of estimating μ_u and μ_v as using the "test" series.

$\hat{b}$

$\hat{b} = 2\beta$ $\hat{b} = 2\beta - 1$

$\hat{b} = \beta$

$M_1 > M_0 > M_{\hat{b}}$

$M_0 > M_1 > M_{\hat{b}}$

$M_1 > M_{\hat{b}} > M_0$

$M_{\hat{b}} > M_1 > M_0$

1

$M_0 > M_{\hat{b}} > M_1$

$M_{\hat{b}} > M_1 > M_0$

$M_1 > M_{\hat{b}} > M_0$

$\frac{1}{2}$

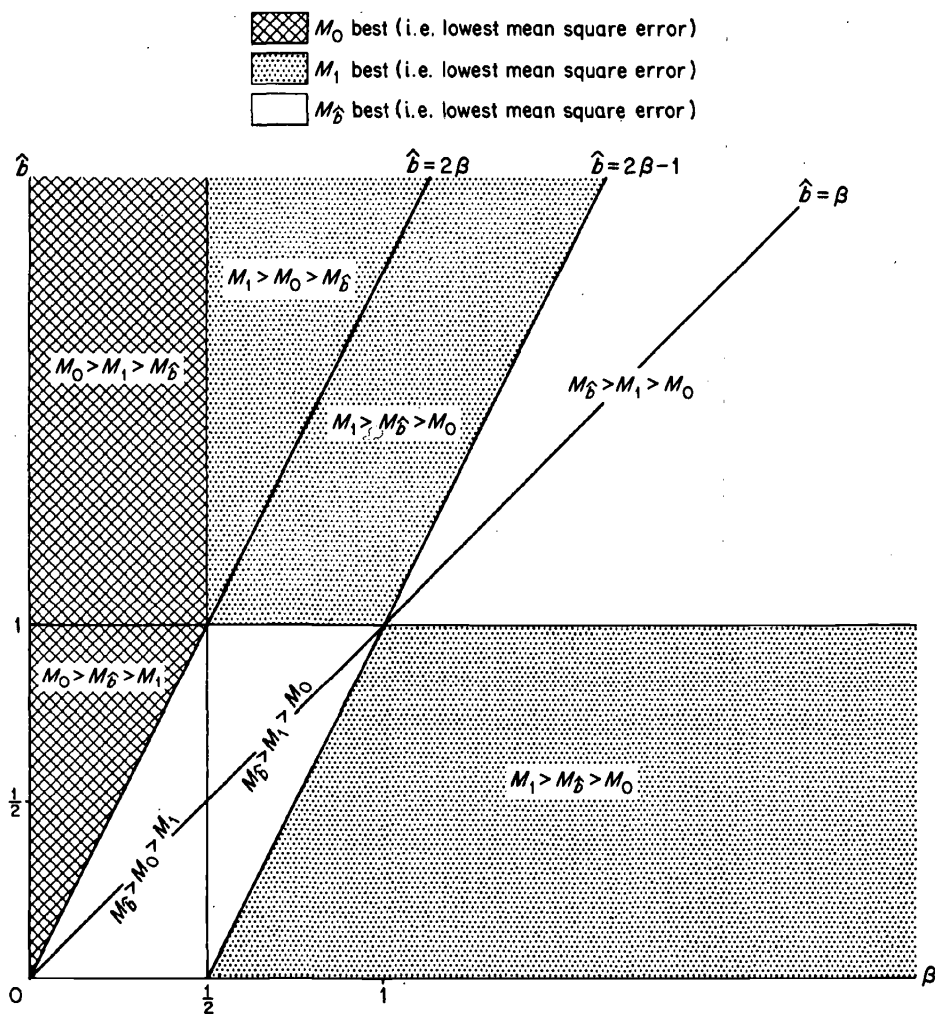$M_{\hat{b}} > M_0 > M_1$

O $\frac{1}{2}$ 1

β

FIG. 1. Ranking of Three Interpolation Methods by Mean Square Error.

regarded as homogeneous over time. It is hard to conceive of circumstances under which the homogeneity would apply to the correlation between u and v but not to their means. But values of v are available over the period to be interpolated and can be averaged to give an estimate of μ_v and, hence, indirectly of μ_u. Perhaps, however, μ_u and μ_v might be regarded as varying from observation to observation in the basic series but not the test series, while ρ_{uv} is regarded as the same throughout and $\mu_u = \mu_v$ for each observation separately.

Under these assumptions, strictly held, it is impossible to say anything about the relative size of the mean square errors of the different variants of M_b, since, as can be seen from (25), these now depend on the size of the common mean of u and v. M_1 has, however, one unique virtue: it is unbiased, as can be seen from (24); that is, if method (1) is used, the average of the errors of interpolation will be zero, whereas it will not be if b is set equal to any other value.

20

To put it differently, the estimation equation (32) cannot be used without an estimate of μ_u and μ_v. Under the assumptions so far made, the best estimates of μ_u and μ_v for any particular date are given by the value of v for that date. But if these estimates are inserted in (32), the estimate of u for that date is given by v. Method (1) is therefore, under these assumptions, the "best" method of estimation.

Even in this case, however, if we relax the assumptions slightly, we may be able to do better than M_1 because the bias introduced by using a value of $b \neq 1$ may be more than balanced by the associated reduction in the variance of the estimate. Suppose M_b is used with b set equal to

$$\beta' = \rho_{uv} \frac{\sigma_u}{\sigma_v}, \tag{36}$$

where an estimate of β' is computed, say, from the test series and this estimate can be accepted as correct. In other words, suppose we interpolate by assuming μ_u and μ_v to be equal to zero even though this assumption is considered in some sense a poor approximation. Write μ for μ_u and μ_v. From (25)

$$\text{M.S.E. } (M_1) = \sigma_u^2 + \sigma_v^2 - 2\rho_{uv}\sigma_u\sigma_v, \tag{37}$$

and

$$\text{M.S.E. } (M_{\beta'}) = \sigma_u^2(1 - \rho_{uv}^2) + \mu^2\left(1 - \rho_{uv}\frac{\sigma_u}{\sigma_v}\right)^2. \tag{38}$$

It follows that

$$\left.\begin{aligned} &\text{M.S.E. } (M_1) \gtrless \text{M.S.E. } (M_{\beta'}) \\ &\text{according as} \\ &\sigma_v^2 - 2\rho_{uv}\sigma_u\sigma_v \gtrless -\rho_{uv}^2\sigma_u^2 + \mu^2\left(1 - \rho_{uv}\frac{\sigma_u}{\sigma_v}\right)^2. \end{aligned}\right\} \tag{39}$$

But (39) can be written

$$\sigma_v^2\left(1 - \rho_{uv}\frac{\sigma_u}{\sigma_v}\right)^2 \gtrless \mu^2\left(1 - \rho_{uv}\frac{\sigma_u}{\sigma_v}\right)^2, \tag{40}$$

which, for

$$\rho_{uv}\frac{\sigma_u}{\sigma_v} \neq 1,$$

is

$$\sigma_v^2 \gtrless \mu^2, \tag{41}$$

which means that $M_{\beta'} \gtrless M_1$ according as

$$1 \gtrless \frac{\mu}{\sigma_v}. \tag{42}$$

Of course, if we take the assumptions strictly, and so suppose that nothing at all can be said about the value of μ, this result cannot help us. But it may well be that enough is known to indicate whether the coefficient of variation of v is smaller or larger than unity. If it is smaller than unity, M_β', will be better than M_1, and conversely.

IV. THE FORM IN WHICH TO USE THE DATA

Each method of interpolation we have been discussing is itself a set of methods, depending on the form in which the original data are expressed— whether as arithmetic observations, logarithms of the original observations, ratios of the observations to arithmetic straight-line trends connecting values for dates at which the series to be interpolated is known, etc.

I shall not attempt to explore systematically the choice of the form in which to express the data. Rather, I shall simply list the considerations suggested by the preceding analysis that are relevant to the choice. The form should, if possible, be chosen to satisfy three conditions: (1) to assure that $\mu_u = \mu_v = 0$; (2) to render the series of values of u and v for different dates homogeneous; and (3) to facilitate the accurate estimation of the required parameters.

The primary means of satisfying condition (1) is through the choice of the trend values to be associated with each unknown value of the series to be interpolated and with the corresponding value of the interpolator. The selection of the proper trend value is precisely the problem of mathematical interpolation without the aid of related series. One requirement likely to be imposed on mathematical interpolation is that it yield an unbiased estimate, which is identical with the satisfaction of condition (1). Mathematical interpolation may therefore be regarded as a first step, yielding as a first approximation what we have called the trend value, to be improved by the use of a related series. In practice, for interpolation of monthly intervals (or other time intervals shorter than a year), what is here called the trend value includes the seasonal component.

The deviation from the trend can then be computed so as to satisfy condition (2). In general, the chief problem here will be to make the standard deviation the same for different dates. For economic data, it is generally supposed that the coefficient of variation is more likely to be the same over time than the standard deviation, which suggests a logarithmic or relative transformation.

The same transformation of the data may be used to compute the trend value and the deviation from trend as, for example, when logarithms of the original data are used throughout. But this need not be done. For example, relatives to arithmetical straight-line trends involve two different transformations. Let x_i' be the observations in the form in which they come. Then the use of relatives to trend is equivalent to the transformation

$$x_i = \frac{x_i'}{(1 - w_i)x_0' + w_i x_2'}, \tag{43}$$

so the data are combined arithmetically in computing trend values, after which