THE INTERPOLATION OF TIME SERIES BY RELATED SERIES*

MILTON FRIEDMAN
University of Chicago
and
National Bureau of Economic Research

The construction of most comprehensive economic time series involves the estimation of some components for some dates by interpolation between values ("benchmarks") for earlier and later dates. This is often done by using a related series known for all relevant dates.

In practice, the bulk of such interpolation uses only a single related series, its values for only three dates—the interpolation date, one prior date, and one later date—and the values of the given series for only the prior and later dates.

For this special case, some current methods are defective on formal grounds and all may frequently yield less accurate estimates than straight-line or other mathematical interpolation. Converting the problem into a simple bivariate regression problem suggests a generalization of current methods that takes account of the correlation between the movements of the given series and the related series.

MOST economic time series are highly manufactured products, constructed out of many bits and pieces that must be shaped and rearranged to yield the final series. One of the commonest operations performed in this process is interpolation: estimation of some component for dates for which it is not directly available from known values of that component for other dates. For example, many of the basic data used in constructing annual estimates of national income for the United States come from biennial censuses of manufactures and hence are available in the required form only for alternate years. Less comprehensive data which are available annually are used to interpolate between the biennial data. To cite a more complex example drawn from the study which gave rise to this paper, data on vault cash are available monthly for one group of banks that are members of the Federal Reserve System (weekly-reporting member banks); at irregularly spaced call dates, generally three to five times a year, for the remaining member banks; and once or twice a year, for non-

* This paper is a study by the National Bureau of Economic Research and has been approved for publication by its board of directors.

This paper is a by-product of a study (now drawing to a close) by Anna J. Schwartz and myself under the auspices of the National Bureau of Economic Research. For that study we have constructed estimates of the stock of money in the United States since 1867. Numerous problems of interpolation arose in constructing the estimates, hence my interest in the problem. An earlier draft of this paper written in 1951 has been circulated among a number of Washington agencies and has led to modifications in their procedures.

Publication has been delayed so long because I wanted to expand the paper to include some extensive numerical examples and empirical evidence on the relative merits of different interpolation procedures. I have not done so. However, since the money estimates and their derivation will soon be published in full, and this paper is essential for an understanding of our interpolation procedures, it seemed best to publish this paper even in its present incomplete form.

I am indebted to a number of persons: to William H. Kruskal for a most illuminating analysis of a statistical problem (Appendix Note 3) and for helpful and detailed comments on an earlier draft; to Millard Hastay, John W. Kendrick, Jacob Mincer, and John Tukey for helpful comments; to Phillip Cagan for some computations on test series; and most of all, to Anna Schwartz, who has not only been a helpful critic but who has also tried out the various techniques developed in this paper in the construction of our money series. My thanks also go to the National Bureau directors' reading committee: R. A. Gordon, W. Allen Wallis, and T. O. Yntema. Marie-Christine Culbert prepared the paper for press, and H. Irving Forman drew Figure 1.

member banks. A monthly series on aggregate vault cash can be compiled only by interpolation of the data for nonweekly-reporting banks.

This kind of interpolation is used at some stage in the construction of the great majority of comprehensive economic time series currently published, although it is frequently described not as interpolation but as the reverse, namely, "adjustment" of series for monthly or other shorter time units to "key" or "benchmark" data.

A closely related operation, also widely used, is the distribution of a known total for one time unit among the shorter time units of which the longer unit is composed. For example, both comprehensive annual income estimates and independently constructed but less comprehensive or less accurate quarterly estimates may be available. The problem is to distribute the difference between the annual estimate and the matching total of the initial quarterly estimates among the quarters and so to construct quarterly estimates that will add up to the annual totals. Though interpolation and distribution are closely related operations, they differ enough so that I have been unable to encompass them in full in the same analysis. This paper is concerned with interpolation alone. I have considered distribution only enough to persuade myself, on the one hand, that it cannot be directly assimilated with interpolation, and, on the other, that a similar analysis of it would yield valuable results (see Appendix Note 4).

Another closely related operation is the extrapolation of some components to more recent dates in order to get current figures prior to the available benchmark data that will later be used in interpolation. Again the results of this paper cannot be applied directly to extrapolation but the general approach can be.

Interpolation may be based solely on the series being interpolated. For example, intermediate values may be obtained by linear interpolation between known values. In this form, the problem is an ancient mathematical one common to many fields on which there is an extensive technical literature. However, this method is seldom used by itself in the construction of economic time series—indeed, too seldom, to judge by the analysis that follows. Generally, it is combined with interpolation on the basis of series that are known or assumed to be related to the series being interpolated. In this form, the problem is a special case of the more general statistical problem of prediction or estimation.[1] However, despite an immense body of experience with this kind of interpolation, it has, so far as I know, never been explored systematically in terms of the more general body of statistical techniques. Each case is typically examined anew and each investigator constructs his own *ad hoc* procedure or borrows one used in a similar problem.

Interpolation of one series by related series involves two steps: (1) selection of the related series to be used; (2) interpolation of the given series by means of the selected series. This paper deals primarily with the second step—the technique of interpolation—though its results have immediate and obvious impli-

[1] It is also a special case of a somewhat more specific problem: the replacement or estimation of missing observations. See Myron B. Fiering, "On the Use of Correlation to Augment Data," *Journal of the American Statistical Association*, March, 1962, pp. 20–32, and the references he cites.

cations for the first as well. It is, moreover, restricted to an especially simple case: interpolation when only a single related series is used, and when the only values of the given series used in the interpolation are those for one date preceding and one date following the date for which a value is to be interpolated and the only values of the related series used are for these same dates plus the interpolation date itself. Though this may seem a very special case, it has widespread importance since the bulk of all actual interpolation by related series satisfies these conditions.

This paper considers first the methods commonly employed in interpolation by related series, describing them (Section I) and then analyzing the errors of estimation associated with them (Section II). The characteristic feature of these methods is their use of a priori parameters, which is to say that none of them explicitly takes into account the degree of correlation between the interpolated and related series. For simplicity in reference, we may term them noncorrelation methods.

Section III presents a generalization of these methods that is suggested by elementary considerations in the statistical theory of correlated data and compares the errors associated with correlation and noncorrelation methods. Section IV discusses the form in which to express the data and Section V summarizes the conclusions.

I. NON-CORRELATION METHODS OF USING RELATED SERIES

Let X be the series to be interpolated; Y, the related series to be used in interpolation. Although the class of methods considered in this section does not explicitly use any information on the degree of correlation between Y and X, a particular series Y is of course chosen for use in interpolation because its intrayearly movements are believed to be highly correlated with the intrayearly movements of X. This belief may be based on nonquantitative considerations (e.g., that employment in different firms producing the same product or vault cash in different classes of banks in the same geographic area will be affected by common forces and hence are likely to move together). Alternatively, it may be based on an observed high correlation between movements in Y and in X for the time units to be interpolated (e.g., months) but for a different period than that to be interpolated; or between movements in Y and in X for different time units for the same period (e.g., a high correlation between annual observations may be taken as evidence that there is also a high correlation between the unknown intrayearly movement); or between movements in two series other than Y and X but analogous to them for the same time units and for the same period (e.g., a high correlation between vault cash in national and state weekly reporting member banks may be taken as evidence that there is also a high correlation between vault cash in weekly-reporting and nonweekly-reporting member banks).

To make the problem specific, suppose it is to convert X, which is known for one date a year, into a monthly series. The common procedure is to superimpose the intrayear movement of Y on the year-to-year movement of X, using one or another device for eliminating any difference between the year-to-year movements of the two series. Different variants arise from different ways of super-