REVUE FRANÇAISE D'AUTOMATIQUE, D'INFORMATIQUE ET DE RECHERCHE OPÉRATIONNELLE. RECHERCHE OPÉRATIONNELLE

STEVEN C. WHEELWRIGHT
SPYROS MAKRIDAKIS

Foregasting with adaptive filtering

Revue française d'automatique, d'informatique et de recherche opérationnelle. Recherche opérationnelle, tome 7, n° V1 (1973), p. 31-52.

<http://www.numdam.org/item?id=RO_1973__7_1_31_0>

R.A.I.R.O.
(7e année, V-1, 1973, p. 31-52)

FORECASTING WITH ADAPTIVE FILTERING

by Steven C. WHEELWRIGHT (1) and Spyros MAKRIDAKIS (2)

Abstract. — *During the past decade Regression Analysis has gained wide acceptance as a method for preparing medium and long range forecasts for time series. However, for a short-term forecasting situation or when the number of observations is small, regression analysis is costly and often impractical. Exponential smoothing is the forecasting method most often used in these latter situations, but it has some major shortcomings too. Rather than trying to distinguish some underlying pattern from the noise (randomness) included in observed data, exponential smoothing simply « smooths » the extreme values in preparing a forecast, which in many cases is not completely suitable. Thus there are a number of medium range forecasting situations and cases for which not much data is available where neither regression analysis nor exponential smoothing methods are appropriate.*

This paper briefly examines the general class of forecasting methods that are based on a weighting of past observations and then presents the theoretical and practical aspects of adaptive filtering, a method for determining an appropriate set of weights. Adaptive Filtering, a technique previously developed in telecommunications engineering, is attractive in many forecasting situations involving time series because it does discriminate between noise and an underlying pattern, it is conceptually appealing and easy to apply, it can be used with a relatively small amount of data, and the accuracy and reliability of its forecasts compare very favorably with other techniques.

Some Existing Techniques for Forecasting

There are numerous situations which arise in the operation of a business that require the development of a forecast for a time series. One of the most common of these involves the area of production scheduling and inventory control. In order to control out-of-stock costs and keep inventory costs within reason, firms must forecast demand for individual products and groups of products and then use those forecasts in making production decisions. Similarly, in the areas of finance, budgeting and marketing, forecasts must be prepared for working capital, cash flow, prices and other time series. While most of these situations involve short or medium term forecasts, firms also are faced with requirements for longer term projections in areas such as capacity utilization, capital requirements, and market growth.

(1) Harvard Business School, Boston, Massachusetts.
(2) INSEAD, Fontainebleau, France.

In order to meet these forecasting requirements, a number of methods have been developed for managers. These have been adopted to varying degrees, based largely on the manager's evaluation of their accuracy, their cost, and his ability to understand what they actually do (1). The majority of these methods are based on the idea that past observations contain information about some underlying pattern of the time series. The purpose of the forecasting method is then to distinguish that pattern from any noise (randomness) that also may be contained in past observations and then to use that pattern to predict future values in the series. A general class of widely used forecasting methods that attempts to deal with both causes of fluctuations in a time series is that of smoothing. Specific techniques of this type assume that the extreme values in a series represent the randomness and thus by « smoothing » these extremes, the basic pattern can be identified. The two methods of this type that are used most often are moving averages and exponential smoothing.

The technique of moving averages consists of taking the n most recent observations of a time series, finding the average of those values, and using that average as a forecast for the next value in the series. That is (2),

$$s_{t+1} = \frac{1}{n}[x_t + x_{t-1} + \ldots + x_{t-(n-1)}]$$

where

s_{t+1} = the moving average forecast for period $t + 1$ based on the previous n observations

n = the number of observations included in the average

x_i = the observed value in period i ($i = 1, 2, \ldots t$).

This approach to short term forecasting is referred to as moving averages because n is held constant and for each new forecast, t is incremented by 1 and the average is recomputed by dropping the oldest observation and picking up a new observation. The value of n determines how much of the fluctuations in observed values is carried into the smoothed value, s_{t+1} : a larger value of n giving a more smoothed forecast than a smaller value of n.

A major drawback of moving averages is that it assigns equal weight to each of the past n observations and no weight to observations before that. It can often be argued that the most recent observations in a series contain more information than the older values. Following this line of reasoning, many managers have adopted the technique of exponential smoothing which gives decreasing importance (smaller weights) to older observations.

---

(1) As has become evident during the past few years, the ease with which a manager can understand a forecasting method is a major factor in determining its use in practice.

(2) The notation used throughout this paper is that lower case letters represent scalar quantities and upper case letters represent vectors. The only exception to this is that φ is used to represent a single cross correlation, $\Phi(x, d)$ is used to represent a vector, and $[\Phi(x, x)]$ is used to represent a matrix of these coefficients. Finally, where the range of values for a summation index is not given, it is from $t - n + 1$ to t.

Exponential smoothing can be described mathematically as

$$s_{t+1} = \alpha x_t + (1 - \alpha)s_t$$

where

s_{t+1} = the exponentially smoothed value to be used as a forecast for period $t + 1$

α = the smoothing constant ($0 \leqslant \alpha \leqslant 1$)

x_i = the observed value in period i ($i = 1, 2, \ldots t$).

This general equation can be expanded by replacing s_t with its computed value. Carrying out this expansion gives

$$s_{t+1} = \alpha x_t + \alpha(1 - \alpha)x_{t-1} + \alpha(1 - \alpha)^2 x_{t-2} + \alpha(1 - \alpha)^3 x_{t-3} + \ldots$$

From this expanded form it can be seen that since α is between 0 and 1, decreasing weights are being given to older observations and the size of α determines the relative value of these weights. A larger α (close to 1) gives most of the weight to very recent observations whereas a small α (close to 0) does not give much weight to any single observation, thus giving a much more smoothed value for s_{t+1}.

Exponential smoothing has been widely used by managers because it is easy to understand, inexpensive to apply and intuitively appealing because the manager has some control over the weights through assignment of a value for α. However, a major drawback of this method is that there is no easy way to determine the most appropriate value of α. Some work on this problem has been done under the title of adaptive smoothing, aimed at examining alternative rules that might be used to determine when and by how much the value of α should be varied [1]. Another author, Brown, has also looked at this problem and has developed rules that can be used to trade off the cost of variance in the forecast with the cost of response time to changes in the underlying pattern [2].

To further improve on this smoothing technique, higher forms of exponential smoothing have been developed. These higher forms can handle time series models other than the constant model assumed in simple exponential smoothing. (For example, double exponential smoothing assumes a trend model.) However, even with these additions, exponential smoothing is still not completelya dequate in many forecasting situations because it does smooth the observed values rather than explicitly looking for the underlying pattern.

An approach to forecasting that is based on a weigthing of past observations but avoids some of the weaknesses of exponential smoothing is polynomial fitting. (Although this method has only been widely used in the area of satellite tracking, it will be discussed briefly here because it illustrates the relationship between smoothing techniques and adaptive filtering.) The method of polynomial fitting consists of taking the $n + 1$ most recent observations and fitting

an n^{th} degree polynomial to these values. A few examples will give a better idea of the advantages and disadvantages of this technique.

The simplest form is for $n = 1$, in which case the forecast is based on a single observation,

$$s_{t+1} = x_t.$$

In the case of $n = 2$, a straight line is fitted to the two most recent observations to give

$$\begin{aligned} s_{t+1} &= x_t + (x_t - x_{t-1}) \\ &= 2x_t - x_{t-1}. \end{aligned}$$

To fit a polynomial to three points ($n = 3$), the method of first differences can be used to obtain a parabola

$$s_{t+1} = 3x_t - 3x_{t-1} + x_{t-2}.$$

One can continue in a similar manner for larger values of n. As can be seen from these few examples, this method gives an exact fit to the n most recent observations, taking no account of older observations that may be available or of the randomness (noise) that may be present in the observed values. Thus while smoothing techniques are often unacceptable because they smooth extreme values rather than trying to identify a unique underlying pattern, polynomial fitting in its standard form may be unacceptable because it does no smoothing of randomness but treats the observed values as being exact in their representation of the underlying pattern.

Each of the three methods for forecasting time series described above is based on a weighted sum of past observations which in general can be written as,

$$s_{t+1} = \sum_{i=t-n+1}^{t} w_i x_i \qquad (1)$$

where

s_{t+1} = the forecast for period $t + 1$
w_i = the weight assigned to observation i
x_i = the observed value in period i
n = the number of observations (and weights) used in preparing the forecast.

It can readily be seen that each method consists simply of a rule or set of rules for determining the weights, w_i. During recent years a number of additional methods, many of which have been both technically complex and statisticaly rigorous, have been developed for computing the most appropriate set of weights [2, 3, 4, 5]. These various methods not only differ in their ability to predict a range of underlying patterns, the assumptions that must be made in applying each of them, and the ease with which they can be used in practice,

but also in the degree to which they can be easily understood by management. This last point is particularly important since it is the most common reason why many technically sophisticated and rigorous methods have failed to gain widespread management acceptance. Smart managers simply do not base decisions on techniques they don't comprehend.

An Adaptive Process for Weighting Past Observations

Adaptive filtering is a procedure that can be used to determine the value of a set of weights for use in time series forecasting. As will be shown, this method is not only technically sound, but in addition it can be applied in a wide range of situations and can be explained in a manner that is intuitively appealing to management. The remainder of this paper will focus on the theoretical developments of adaptive filtering and its practical application.

The original work on filter design was done by Norbert Wiener [6] in the forties. Wiener focused on the design of *linear* filters for noise elimination and for predicting and smoothing *statistically stationary* signals. Using the procedures he developed gives results that are optimal in terms of least squares when the series is in fact statistically stationary.

Following Wiener's work, various authors including Kalman and Bucy have developed procedures that give optimal *time-variable* linear filters for *non-stationary* time series [7]. When such a series exists, the Kalman-Bucy approach can give substantially better results (in terms of least squares) than the simpler Wiener approach.

The drawback of both the Wiener and Kalman-Bucy procedures is that the filters must be designed on the basis of *a priori* information or assumptions about the statistics of the time series involved. In practice, these two filtering approaches only give optimal results when the statistical characteristics of the series in fact match the *a priori* information on the basis of which the filters were designed. When the *a priori* statistical characteristics are not known perfectly, these approaches do not give optimal results.

The adaptive filtering approach to be described here bases its design of the filter on *estimated* statistical characteristics of the time series. The statistics are not measured explicitly and then used to design the filter, but rather the process of estimation and design go on in a single cycle, using an algorithm that continuously updates the estimates as the design process is carried out.

It can be argued that the form of adaptive filtering to be described here is almost as simple to apply as the Wiener filter, and should perform almost as well as the Kalman-Bucy filter given perfect *a priori* information. When the statistical characteristics are not known perfectly *a priori*, it is quite possible that the adaptive filter will outperform both the Wiener filter and the Kalman-Bucy filter. When little or no *a priori* information is available, the use of an adaptive filter may be the only reasonable possibility.

While it may be instructive to undertake an extensive comparison of these and other forecasting techniques using empirical data, that is not the intent of this paper. Rather, the purpose is to present the theoretical development of an adaptive filtering approach to forecasting and to demonstrate the application of that technique in practice. As a starting point for doing this, one can consider the approach illustrated in figure 1 and suggested by Widrow [8] and Pertuz [9].

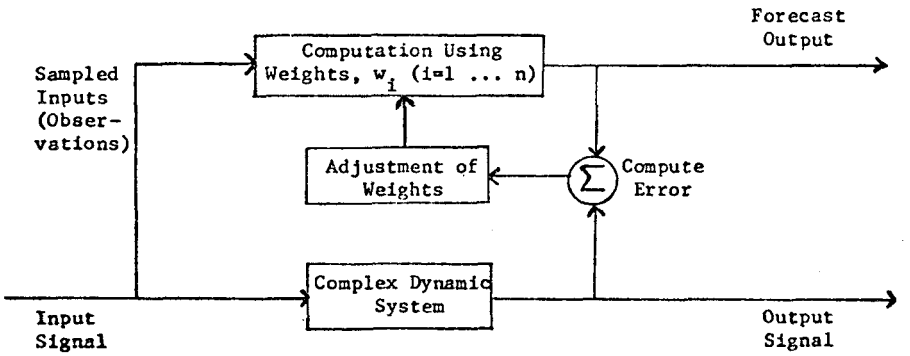

Figure 1

A Model for Determining the Weights in a Time-Series Forecast

In terms of figure 1, we would like to develop a method for adjusting the weights that will distinguish between an underlying pattern and noise by eliminating as much noise as possible from the observed series of values. The criterion we will use for comparing alternative sets of weights is the expected value of the error squared (the mean square error).

In applying the technique of adaptive filtering in determining the most appropriate set of weights for forecasting, a basic assumption is that there exists some underlying pattern (signal) that can be represented as a weighted sum of past observations. Thus a wide range of functional forms, such as a constant, a linear trend, a seasonal pattern, or any polynomial form, can be identified and predicted using this technique.

The process of determining the weights is an iterative one with a cycle consisting of taking a set of n observations, computing a forecast for the next observation based on a set of n weights, then comparing that forecast with the observed value (using the mean square error) and finally revising the weights in such a way that the mean square error will be reduced. Obviously the key to the effectiveness of adaptive filtering is in the rule used to adapt the weights at the end of each cycle. This rule can be developed by first examining the criterion of mean square error.

By definition, the error is the difference between the actual value and the forecast value.

$$e_{t+1} \overset{\Delta}{=} x_{t+1} - s_{t+1}$$
$$= x_{t+1} - \sum_{i=t-n+1}^{t} w_i x_i \qquad (2)$$

where

e_{t+1} = the error in the forecast for period $t+1$
x_{t+1} = the actual value (observed) for period $t+1$
s_{t+1} = the forecast for period $t+1$
w_i = the weight for the i^{th} observed value ($i = t-n+1, t-n+2, \ldots t$)
x_i = the observed value for period i ($i = t-n+1, t-n+2, \ldots, t$)
n = the number of weights (and observations) used in the forecast.

To obtain the expected value of the error squared, we first square (2) giving

$$e_{t+1}^2 = \left(x_{t+1} - \sum_{i=t-n+1}^{t} w_i x_i \right)^2$$
$$= x_{t+1}^2 - 2 \sum_{i=t-n+1}^{t} w_i x_i x_{t+1} + \sum_{i=t-n+1}^{t} \sum_{j=t-n+1}^{t} w_i w_j x_i x_j. \qquad (3\,a)$$

Since we always will be dealing with the error at period $t+1$ and since all summations will be taken from $t-n+1$ to t, we can simplify the above notation by writing,

$$e^2 = d^2 - 2 \sum_i w_i x_i d + \sum_i \sum_j w_i w_j x_i x_j \qquad (3\,b)$$

where $d \overset{\Delta}{=} x_{t+1}$, the value being forecast.

We can now take the expected value of (3 *b*) to obtain the mean square error,

$$\overline{e^2} = \overline{d^2} - 2 \sum_i w_i \varphi(x_i, d) + \sum_i \sum_j w_i w_j \varphi(x_i, x_j) \qquad (4)$$

where

$\overline{e^2}$ = mean square error for period $t+1$
$\overline{d^2}$ = expected value of the squared observation for period $t+1$
w_i = the weight assigned to the i^{th} observation
$\varphi(x_i, d) \overset{\Delta}{=}$ the correlation, $\overline{x_i d}$
$\varphi(x_i, x_j) \overset{\Delta}{=}$ the correlation, $\overline{x_i x_j}$ (1).

---

(1) In order for these expectations to be identically equal to the correlations, it is necessary that the data be normalized. This simply involves transforming the original data to X^* using $x_i^* = \frac{x_i - \bar{x}}{\sigma}$. See William Feller, *Introduction to Probability and its Applications*, Volume I, New York : John Wiley & Sons, 1963, pp. 215-222.

An examination of (4) shows that if we have a stationary series (one where the underlying pattern is stable and thus the correlations do not change) then the mean square error is a second order function of the weights, w_i. Thus the mean square error performance function can be visualized as a bowl shaped surface, a parabolic function of the weight variables. The aim of adaptive filtering is to seek the minimum mean square error (the bottom of the performance surface) by revising the weights through the iterative process mentioned earlier.

We can now examine the process for determining the weights. However, before doing so, it is important to realize that an underlying assumption in this development is that we have a stable (static) pattern in our data. After we have dealt with this stable case, we can then consider how the process might be adjusted for the non-stable situation. (A continuous shifting in the basic pattern in the data — a specific non stable case — can be visualized as a bowl-shaped error surface, where the bottom of the bowl is continuously moving.)

The search procedure that we will use to find the best set of weights is the method of steepest descent. The details of this approach have been described by Wilde [10]. Essentially, it consists of selecting a starting point on the performance surface and then moving towards the bottom of the surface by following an iterative procedure. In order to do this we must be able to compute (or estimate) for any point on the performance surface the direction in which the minimum point on the surface lies. We can then adjust our weights in such a manner that our new weights represent a point on the error surface that is closer to the optimum set of weights (the bottom of the bowl) than were our old weights. The method of steepest descent does this by using the following rule to adjust the weights :

$$W' = W - k\overline{\nabla e^2} \tag{5}$$

where

W' = the revised vector of weights

W = the old weight vector

k = a constant factor (> 0)

$\overline{\nabla e^2}$ = the gradient vector of $\overline{e^2}$.

This equation states that we determine the adjusted weights by starting with our old weight vector and correcting it by a constant factor (k) multiplied by the negative of the gradient vector. Simply speaking, the negative of the gradient vector tells us in which direction the minimum of the performance surface lies and the constant factor, k, determines how far we will move in that direction. In order to use (5) in finding the best set of weights, we need to know the value of the gradient for a given weight vector, W. In theory, this value

can be found by differentiating the mean square error function in (4) with respect to the weights. This gives for each weight, w_i,

$$\overline{\frac{\partial e^2}{\partial w_i}} = -2\varphi(x_i, d) + 2\sum_j w_j\varphi(x_i, x_j). \qquad (6\ a)$$

The entire gradient vector can be written as

$$\overline{\nabla e^2} = -2\Phi(x, d) + 2W[\Phi(x, x)] \qquad (6\ b)$$

where

$\overline{\nabla e^2}$ = the gradient vector

$\Phi(x, d)$ = the vector of cross correlations between the observed values, x_i, and the desired value, d

W = the vector of weights, w_i

$[\Phi(x, x)]$ = the matrix of cross correlations between each pair of observed values, (x_i, x_j).

To find the optimal set of weights that minimizes the mean square error, we want $\overline{\nabla e^2} = 0$. Thus using (6 *b*) gives

$$\Phi(x, d) = W_{\text{LMS}}[\Phi(x, x)] \qquad (7\ a)$$

where

$\Phi(x, d)$ and $[\Phi(x, x)]$ are as before

W_{LMS} = the vector of weights that gives the least mean square (LMS) error.

This can be written as

$$W_{\text{LMS}} = \Phi(x, d)[\Phi(x, x)]^{-1} \qquad (7\ b)$$

To implement this approach for finding W_{LMS} requires a knowledge of the cross correlations represented by $\Phi(x, d)$ and $[\Phi(x, x)]$. Unfortunately these are often difficult if not impossible to determine. Thus to be of real use to the practitioner in forecasting, what is needed is an alternative means for finding, or at least approximating, W_{LMS}.

The method developed by Widrow for doing this utilizes measured gradient estimates based on an approximation for $\overline{\nabla e^2}$ ([1]). We can find such an estimate by first using e^2 as an approximation for $\overline{e^2}$. Admittedly this is a very crude estimate of $\overline{e^2}$ and one may wonder why an average of several values of e^2 is not used instead. The reason is that, as pointed out earlier, the real power of adaptive filtering is when one has little or no *a priori* information on the statis-

(1) WIDROW, *op. cit.*

tical characteristics of a time series. If one were to use an estimate of $\overline{e^2}$ based on several values of e^2, it would limit the usefulness of this approach, and as will be shown later, the use of e^2 to approximate $\overline{e^2}$ is sufficiently accurate in many cases to give very reliable results. Thus we can approximate the components of the gradient vector by

$$\overline{\frac{\partial e^2}{\partial w_i}} \cong \frac{\partial e^2}{\partial w_i} = 2e \frac{\partial e}{\partial w_i}. \tag{8}$$

Using the definition of e given by (2), we have

$$\frac{\partial e}{\partial w_i} = -x_i$$

and (8) can be rewritten as

$$\overline{\frac{\partial e^2}{\partial w_i}} \cong -2ex_i. \tag{9}$$

Thus the approximation of the entire gradient vector is

$$\overline{\nabla e^2} \cong -2eX \tag{10}$$

where $X =$ the vector of observed values, x_i.

Substituting (10) into equation (5) gives us a means of adjusting our weights in an iterative fashion as we search for those which will minimize the mean square error. That is,

$$W' = W + 2keX \tag{11}$$

In order to use this approach for adjusting the weights, we need to specify both the number of weights, n, and the adjustment constant, k. We can then « train » a set of weights by taking a series of observed values, computing the error resulting from the use of the initial set of weights, and then updating our weights using (11). As this process is repeated it will move towards the minimum mean square error on the performance surface (the bottom of the bowl). The rate at which one moves towards the best set of weights, W_{LMS}, is determined by the value of the adjustment constant, k. The larger the value of k, the greater the adjustment in the weights at each iteration. This rate of adjustment can be thought of as the « learning speed » of the system. Thus k is often called a learning constant.

One way to better understand the importance and effect of the learning speed is to define and compute μ, the fraction of the error that is corrected on each iteration. Using the following definition,

$$\Delta e \triangleq -\mu e \tag{12}$$

where

Δe = the change in error resulting from adjustment of the weights,

μ = a positive error reduction factor (the minus sign is necessary in order for μ to be positive when the error is reduced)

we can solve for μ, obtaining

$$\mu = \frac{-\Delta e}{e}. \tag{13}$$

Now using the definition of Δe and (2) we can write

$$\Delta e = -(W' - W)X \tag{14}$$

and since we know $(W' - W)$ from (11), we then have

$$\Delta e = -(2ekX^T)X = -2keX^TX$$

which is a scalar since X^TX is the dot production of two vectors. Substituting into (13) gives

$$\mu = \frac{-\Delta e}{e} = 2kX^TX. \tag{15}$$

The importance of being able to compute the error reduction for each iteration is that it can be used to determine when the adaption process has leveled off. That is, one would expect that after several iterations the error reduction would become very small and thus going through additional iterations would not have much effect on the weights. (This is shown in a practical application in the next section.)

An important aspect of the use of adaptive filtering in forecasting is specifying a value of k that will ensure that the adaption process will converge to the set of weights that will minimize the mean square error, W_{LMS}. Widrow has shown that a necessary and sufficient condition for stability of the steepest-descent adaptation process is

$$\frac{1}{\lambda_{\max}} > k > 0$$

where $\lambda_{\max}$ = the maximum eigenvalue of $[\Phi(x, x)]$ (1).

An alternative method for ensuring convergence which is easier to use than the above involves μ. It can be shown that if

$$2 > \mu > 0 \tag{16}$$

(1) WIDROW, *op. cit.*, pp. 29-34.

then the use of (11) to adapt the weight vector will always converge to W_{LMS} (1). Substituting the value of μ from (15) into (16), one finds that this condition will always be met if

$$\frac{1}{(X^TX)_{\max}} > k > 0 \tag{17}$$

where the relevant X vector is the one with maximum size,

$$(X^TX)_{\max} \text{ (2)}.$$

The vector of maximum size can generally be approximated after a visual inspection of the observations and a value of k can then be specified to be used in adjusting the weights. As a practical matter one can always select a small value of k to insure convergence, realizing that by doing so it will take additional iterations to reach a set of weights that are arbitrarily close to W_{LMS}, since as k is decreased, the positive error reduction, μ, on each iteration is decreased also. The authors have found in forecasting a wide range of situations that if the vector of observations is first normalized by dividing each value by the largest value in the series, a good rule of thumb is to then let k equal $1/n$ where n is the number of weights used. (This gives a k value which satisfies (17), and as will be shown in the next section, generally k need only fall within a range of values to give near optimal results.)

Using Adaptive Filtering in Practice

The previous section has outlined the theoretical development of a general scheme for forecasting based on the concept of using a weighted sum of past observations. There are several features of adaptive filtering, the method for setting the weights, that make it attractive to the manager. First is the fact that it utilizes the « information » contained in past observations to find the best set of weights. Perhaps equally important is its simplicity. The adaptation of the weights involves only a single equation (11). This equation is not only easy to use, but it allows the manager to adjust the procedure to fit his own situation and data by allowing him to alter the number of observations to be used in setting the weights and to specify the rate at which the weights are adapted.

An illustration of how adaptive filtering can be used as a forecasting technique in a specific situation should serve to highlight its usefulness. Consider the case of a French wine company who as part of their planning process desire to forecast champagne sales in France on a monthly basis. They have available from industry sources actual monthly sales values from January 1962 through September 1970 (105 months). These values are shown in table 1.

(1) See Widrow, pp. 28-29, 34.
(2) It should be noted that a k value satisfying (17) is a sufficient condition for convergence, but *not* a necessary condition.

TABLE 1. — *Monthly Champagne Sales (in 1000's of bottles)*

Year	Month	Sales	Year	Month	Sales	Year	Month	Sales
1970			1967	Dec	13.916	1964	Dec	9.254
				Nov	10.803		Nov	7.614
				Oct	6.873		Oct	5.211
	Sept	5.877		Sept	5.222		Sept	3.528
	Aug	1.431		Aug	1.821		Aug	1.573
	July	4.298		July	3.523		July	3.260
	June	5.312		June	4.677		June	3.986
	May	4.618		May	4.968		May	3.937
	April	4.788		April	4.276		April	3.523
	March	4.577		March	4.510		March	4.047
	Feb	3.564		Feb	3.957		Feb	3.006
	Jan	4.348		Jan	4.016		Jan	3.113
1969	Dec	12.670	1966	Dec	11.331	1963	Dec	8.357
	Nov	9.851		Nov	9.858		Nov	6.838
	Oct	6.981		Oct	6.922		Oct	4.474
	Sept	5.951		Sept	5.048		Sept	3.595
	Aug	1.659		Aug	1.723		Aug	1.759
	July	4.633		July	3.965		July	3.028
	June	4.874		June	4.753		June	3.230
	May	5.010		May	4.647		May	3.776
	April	4.676		April	4.121		April	3.266
	March	4.286		March	4.154		March	3.031
	Feb	3.162		Feb	4.292		Feb	2.475
	Jan	3.934		Jan	3.633		Jan	2.541
1968	Dec	13.076	1965	Dec	10.651	1962	Dec	7.132
	Nov	9.842		Nov	8.314		Nov	5.764
	Oct	6.424		Oct	5.428		Oct	4.301
	Sept	5.221		Sept	4.739		Sept	2.922
	Aug	1.738		Aug	1.643		Aug	2.212
	July	4.217		July	3.663		July	2.282
	June	3.986		June	4.539		June	3.036
	May	2.927		May	4.520		May	2.946
	April	3.740		April	4.514		April	2.721
	March	3.370		March	3.718		March	2.755
	Feb	2.899		Feb	3.088		Feb	2.672
	Jan	2.639		Jan	5.375		Jan	2.851

As pointed out in the previous section, the use of adaptive filtering in preparing a forecast involves two distinct phases. The first is the training (or adapting) of a set of weights using historical data and the second is the use of these weights to prepare a forecast. For purposes of this example, all 195 historical observations of monthly champagne sales will be used in training the set of weights.

In order to start the training phase, it is necessary to first specify the number of weights, n, and the learning constant, k. Since a brief visual inspection of the historical data in table 1 indicates that champagne sales follow a cyclical

pattern of length 12 months, the use of 12 weights would seem appropriate. Essentially this says that while the weights will be trained using several years of data, a forecast for a single month will only be based on the sales for the 12 preceding months. As a starting value for k, we might select a value of $k = 08$ (1).

With these parameters specified, the set of 12 weights can be trained using equation (11) and an initial value for each of the 12 weights. (We will arbitrarily let each of the weights have an initial value of 0.085). The first training cycle consists of taking the first 12 observations of the 105 available), computing a forecast for month 13 using

$$S_{13} = \sum_{i=1}^{12} w_i x_i,$$

computing the error of this forecast, $e = (x_{13} - s_{13})$, and then revising the weight vector using :

$$W' = W + 2keX$$

where

W' = the new vector of 12 weights
W = the old (initial) vector of 12 weights
$k = .08$
$e = x_{13} - s_{13}$
X = the vector of the first 12 observations.

The forecast for month 14 can then be computed by using the observed values for months 2 to 13 (12 values), after which the process of updating the weights can be repeated. When this process has been followed up through the forecasting of month 105, one can then start over again with the first 12 observations. Each of these series of revisions of the weights which is made by going through the entire string of observed data can be referred to as a training iteration. The number of iterations that need to be run depends on the nature of the series being studied, the adaption rate, k, and the number of observations available for training. Figure 2 shows the results of running 80 such iterations on the 105 months of champagne sales data. Even after this number of iterations, it can be seen that the adjusted weights give a forecast value that is quite close to the actual values as illustrated by the mean square error for the 80th iteration.

It is evident that the parameter k is of critical importance in adaptive filtering. This constant determines how rapidly the weights are adjusted and

(1) This value of k was chosen based largely on equation (16). Since the champagne data was normalized in this example before using adaptive filtering, the largest single value in the series was 1.0. Thus as an upper bound on the maximum vector ($X_T X$), one can use a vector whose length is 12 (this corresponds to the number of weights used) and whose values are all 1.0. Using (16), this indicates that a value of k between 1/12 and 0 will guarantee convergence of the algorithm. Since a larger value of x gives more rapid convergence than a smaller value, the authors chose $k = .08$ for this example.

Fig. 2 TRAINING THE WEIGHTS FOR FORECASTING

CHAMPAGNE SALES

SERIES 1 WEIGHT STRING LENGTH 12 FORECASTING HORIZON 1 PERIOD(S)
ADAPTATION CONSTANT 0.0800 80 TRAINING ITERATIONS

LEARNING PERFORMANCE

ITERATION	MEAN-SQUARE ERROR	% ERROR - MEAN	% ERROR - VARIANCE	ERROR REDUCTION
1	86.83972	-151.97128	91723.313	0.0
11	0.84335	-4.85539	616.168	0.049119
21	0.64193	-3.54261	478.208	0.015300
31	0.59168	-3.98917	440.006	0.004372
41	0.57779	-2.88744	427.696	0.001320
51	0.57340	-2.78774	422.836	0.000461
61	0.57173	-2.75837	420.473	0.000195
71	0.57094	-2.74221	419.098	0.000104
72	0.57088	-2.74131	418.992	0.000096
73	0.57083	-2.74048	418.889	0.000092
74	0.57078	-2.73979	418.790	0.000088
75	0.57073	-2.73911	418.694	0.000087
76	0.57068	-2.73851	418.602	0.000082
77	0.57064	-2.73796	418.512	0.000078
78	0.57060	-2.73757	418.428	0.000070
79	0.57056	-2.73720	418.345	0.000071
80	0.57052	-2.73686	418.264	0.000065

OPTIMAL WEIGHTS

WEIGHT NO.	SERIES 1
1	1.018527
2	0.071629
3	-0.074096
4	0.072836
5	-0.095556
6	0.086222
7	-0.093924
8	0.034412
9	-0.090613
10	0.053638
11	-0.101789
12	0.070522

TABLE 2. — *Adaptive Filtering Forecasts for Actual Champagne Sales in France*

Computer Run	Number of Weights	Final Weight Values	Value of k	Number of Training Iterations	Mean Square Error on Final Iteration
a	12	.9754 .0991 -.0683 .0787 -.1089 .0885 -.0709 .0433 -.0982 .0630 -.0910 .1053	.04	80	.5971
b	12	1.0185 .0716 -.0741 .0728 -.0956 .0862 -.0939 .0344 -.0906 .0536 -.1018 .0705	.08	80	.5705
c	12	1.0230 .0680 -.0730 .0703 -.0926 .0841 -.0946 .0311 -.0896 .0506 -.1017 .0649	.09	80	.5696
d	12	1.0343 .0584 -.0688 .0630 -.0841 .0779 -.0944 .0211 -.0868 .0421 -.1006 .0496	.12	80	.5733

thus the amount of error reduction achieved on each iteration. Figure 2 indicates that with $k = .08$, the mean square error is within 5 % of its final value after 30 iterations and within 1 % of that value after 50 iterations. The fact that this value of k guarantees stability and convergence also means that the error reduction will never increase with additional iterations and thus once the mean square error improvement levels off, there is little reason to run additional iterations in a practical application.

In order to determine the effects of k on the number of iterations required and the error reduction on each iteration, the results using values of k from 04 to .12 for the champagne series data are shown in table 2. From this it can be seen that for 80 iterations, the optimal value of k is around .09. However, even for k values as small as .04 and as large as .12, the mean square error is within 6 % of its value at .09. The relationship between k and the mean square error for this series is shown in figure 3.

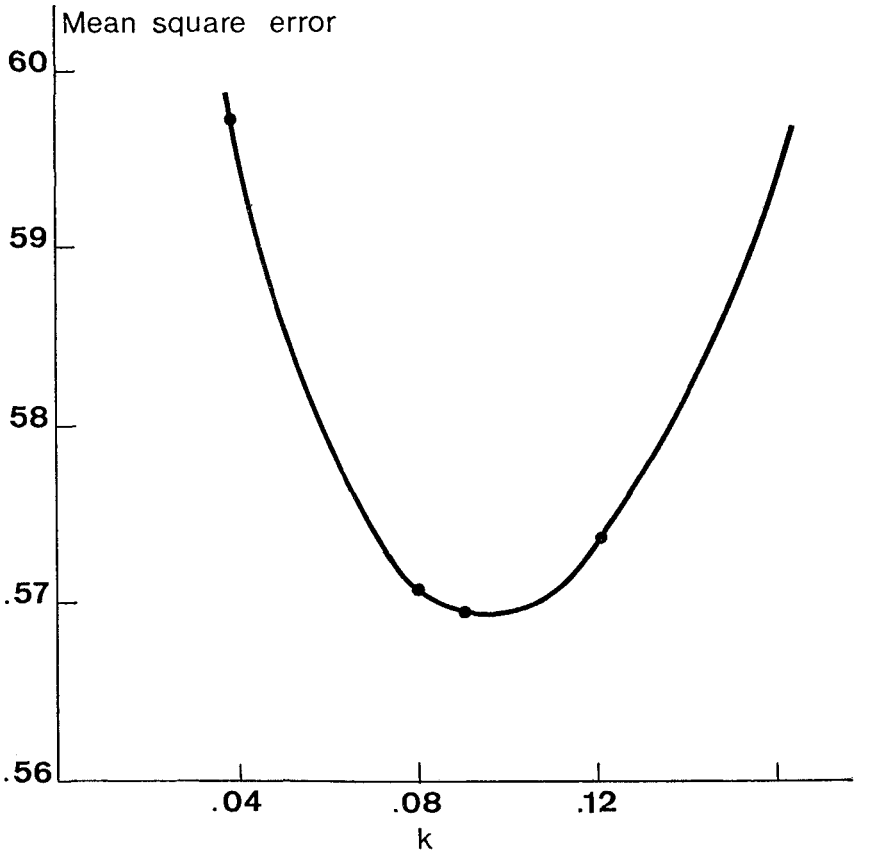

Figure 3

Behavior of Mean Square Error-Champagne Series (80 iterations)

In all of the series the authors have examined so far, the effect of changes in k have been similar, indicating that in general it is not necessary to find the optimal value of k to get good results, but one need simply be in the vicinity of this optimum. In the case of the champagne series it was found that as k was given values greater than .12, the adaptation process began to oscillate, indicating that it was reacting to random fluctuations in the series. At a k value of .25, the adaptation process failed to converge for this series.

In addition to wanting to know how adaptive filtering can be applied in a specific situation, most managers also are concerned with how its performance compares to that of other forecasting methods. To make such a comparison, the authors applied both regression analysis (1) and seasonal time series analysis (2) to the champagne data series. These two methods were chosen because they are capable of handling a cyclical pattern and they are widely used in practice. However, it should be mentioned that from a strictly technical point of view, these two methods are not the best available for this kind of a times series.

The results of preparing monthly forecasts of champagne sales using each of these three forecasting methods are shown graphically in figure 4. The performance of these three methods can further be compared in terms of the mean square error of the forecasts developed using each one.

FORECASTING METHOD	MEAN SQUARE ERROR
Adaptive Filtering	0.5696
Regression Analysis	0.7323
Seasonal Time Series	2.0110

These results indicate that both adaptive filtering and regression give substantially better forecasts than seasonal time series. Also, one can see from figure 3 that a fairly wide range of k values give a smaller mean square error than does regression. Although one might conclude from this example that adaptive filtering and regression are comparable methods in terms of mean square error, it should be remembered that in other situations and even for champagne sales in the future, the results of such a comparison could be quite different.

---

(1) The model used for regression analysis consisted of 12 independent variables — the first being the period (1 through 105) and the other 11 being dummy variables to represent the adjustment for each month of the year. Other regression models were also examined, but this one gave the best results. A standard computerized routine was used to carry out the computations. This routine was based on the development of regression analysis presented in A. M. MOOD and F. A. GRAYBILL, *Introduction to the Theory of Statistics*, New York : McGraw-Hill, 1963, pp. 328-355.

(2) Seasonal time series analysis as used in this comparison consisted of identifying the time trend in the series using simple regression, computing a monthly adjustment factor and then basing the forecast on the product of the appropriate monthly adjustment factor and the trend value. This forecasting method is presented in detail in W. A. SPURR and C. P. BONINI, *Statistical Analysis for Business Decisions*, Homewood, Ill. : Richard D. Irwin, 1967, pp. 463-348.

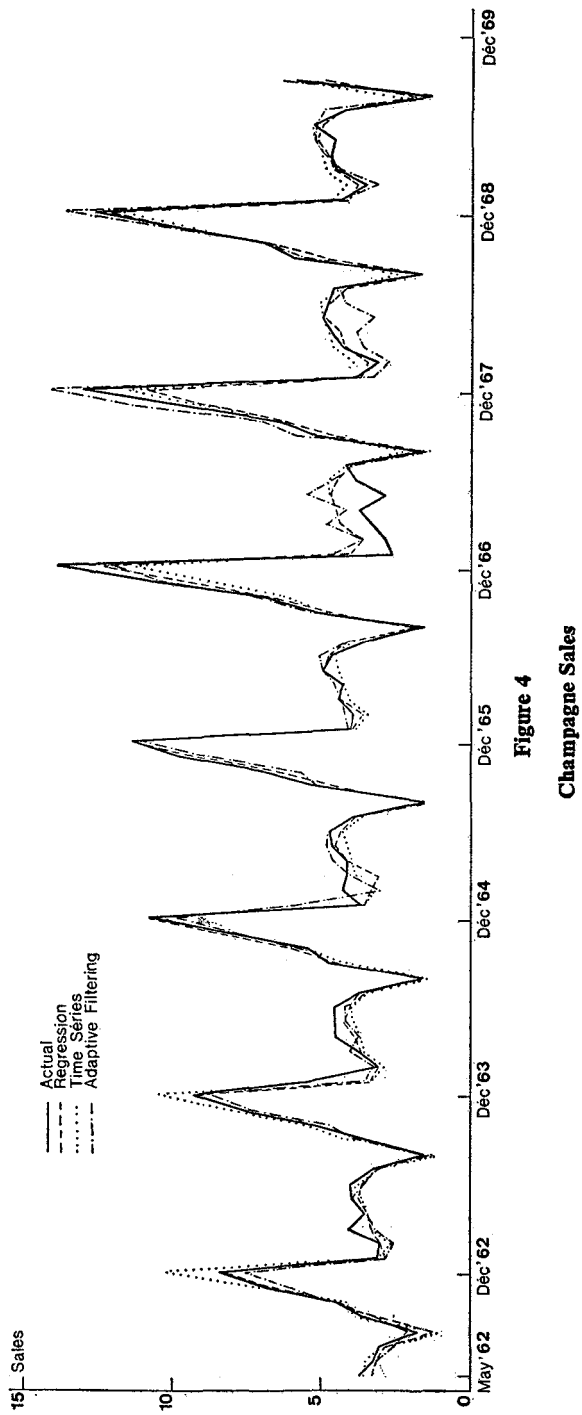

Figure 4

Champagne Sales

n° mars 1973, V-1.

Clearly a complete analysis of adaptive filtering should compare its performance with alternative methods on a number of different time series. These more elaborate numerical studies should also consider forecasting methods such as Box-Jenkins and other more sophisticated approaches that can give better results than regression and seasonal time series analysis.

There is one other aspect of adaptive filtering that in many situations makes it clearly preferable to other methods. This is the small amount of data that is required to initially use adaptive filtering. The reason that much less data is required with this method is that the sequence of weights is independent of the specific time period (eg., the month) being forecast. The effect of this characteristic can be illustrated with the champagne series by supposing that one had only the first 20 observations in the series and wished to forecast the twenty-first value. If the parameters (weights) were identified with specific months, the technique could not really be used in this situation since for some months there would only be one observation and for the others there would be only two observations. However, in applying adaptive filtering to these 20 observations and training 12 weights there were 7 different cycles that could be made in adapting the weights. After several iterations through these cycles, the weights were quite similar in value to those determined using 105 observations (1).

Summary

The purpose of this paper has been to present the theoretical basis of adaptive filtering, to show how it can be used in time series forecasting, and to briefly compare its performance with other well known forecasting techniques. The real power of adaptive filtering over other forecasting techniques comes from the fact that it requires no *a priori* information (or assumptions) concerning the statistical characteristics of the time series involved and it is intuitively appealing to the practicing manager and straightforward to apply. It also has the additional advantage that it can be used when only a limited amount of historical data is available.

The type of situation in which it can profitably be applied is one where the manager is confronted with a time series which is relatively new to him (and therefore largely unknown) and where the potential value of a forecast is substantial. The use of adaptive filtering allows him to prepare forecasts that are generally as good as, if not better than, those resulting from the use of other techniques. To do this he need only specify three factors : the number of weights, the learning constant and the number of iterations to be used in

(1) It should be noted that while adaptive filtering can be used with a relatively small set of observations, as the sample gets smaller the weights will be more likely to represent some of the randomness in the sample as well as the underlying pattern than would be the case with a larger set of observations. Also, if the underlying pattern is changing over time, it is important to revise the weights as new observations become available.

training the weights (1). Although from a theoretical standpoint each weight is a parameter in the adaptive filtering model that must be determined, the manager applying this method is only required to specify three factors.

However, the fact that for a time series with a 12-month seasonal pattern, 12 weights must be trained is a drawback of adaptive filtering. Where the value of the forecast is high, maintaining 12 weights in computer storage is insignificant. But when several thousand items must be forecast, this storage requirement may become an important criteria in selecting an alternative method.

One final advantage of adaptive filtering is that since the basic underlying pattern of most time series is evolving over time, a forecasting technique must take such changes into account if it is to continue to give accurate forecasts and to maintain the confidence of the manager who uses these forecasts. By its very nature, adaptive filtering is such a technique.

Clearly there is still much that should be done to investigate the application of this technique. First there is a need for further research on situations where the basic underlying pattern in the data is changing over time (dynamic). One way of handling this problem is to use a relatively small value of k and to update the weights (i.e., go through the adaptation process) periodically as additional data become available. However, it should be possible to develop more precise and more effective decision rules for these situations.

Another area for further study would be the comparison of this method of forecasting to other approaches such as exponential smoothing, time series analysis and regression analysis for a range of practical situations. This paper has done it for a single situation, but obviously there are many other types of situations that deserve similar study.

Equally important as the comparison of alternative forecasting methods would be research on what determines the best number of weights, size of k, number of iterations needed, and frequency of revisions in the weights. These would be of great help to the practitioner, making adaptive filtering easier to use for forecasting.

One final area that deserves further investigation is the use of adaptive filtering with multiple series of data. For example, rather than basing a sales forecast only on information contained in past sales data, one could also consider the information contained in related series of data such as in an industrial index, GNP figures or sales in a complementary industry. (This is often done with multiple regression.) It is possible to use adaptive filtering on several series of data by determining and using weights for those series as well as for the basic series being forecast. Although the authors have been successful in one such application of adaptive filtering, the limitations and possibilities for doing it in general have not been examined.

---

(1) In place of specifying the number of iterations to be performed in training one can specify the level of error reduction (on a single iteration) that is to be achieved.

REFERENCES

[1] MONTGMOMERY Douglas C., « Adaptive Control of Exponential Smoothing Parameters by Evolutionary Operation », *AIIE Transactions*, September 1970.

ROBERTS S. D. and REED R., « The Development of a Self-Adaptive Forecasting Technique », *AIIE Transactions*, December 1969.

TRIGG D. W. and LEACH A. D., « Exponential Smoothing with an Adaptive Response Rate », *Operations Research Quaterly*, March 1967.

WHYBARK D. Clay, « A Comparison of Adaptive Forecasting Techniques », Paper No. 302, Graduate School of Industrial Administration, Purdue University, Lafayette, Indiana, March, 1971.

[2] BROWN Robert G., *Smoothing Forecasting and Prediction of Discrete Time Series*, Englewood Cliffs, New Jersey : Prentice-Hall, 1963.

[3] BRENNER J. L. et al., « Difference Equations in Forecasting Formulas », *Journal of the Institute of Management Sciences*, vol. 15, No. 3, November 1968, pp. 141-159.

[4] MORRISON Norman, *Introduction to Sequential Smoothing and Prediction*, New York, McGraw-Hill Book Co., 1969.

[5] BOX George E. P. and JENKINS Gwilym M., *Time Series Analysis*, San Francisco, California : Holden-Day, 1970.

[6] WIENER Norbert, *Extrapolation, Interpolation and Smoothing of Stationary Time Series with Engineering Applications*, New York : Wiley, 1949.

[7] KALMAN R. E. and BUCY R. S., « New Results in Linear Filtering and Prediction Theory », *Journal of Basic Engineering (Trans. ASME)*, vol. 83 D, 1961.

[8] WIDROW Bernard, « Adaptive Filters 1 : Fundamentals, « SU-SEL-66-126. Systems Theory Laboratory, Stanford University, Stanford, California, December 1966.

[9] PERTUZ Alexis, « Adaptive Time Series Forecasting », Unpublished Term Paper, Stanford Business School, Stanford, California, June 1968.

[10] WILDE Douglas J. and BEIGHTER Charles S., *Foundations of Optimization*, Englewood Cliffs, N.J. : Prentice-Hall, 1964 (pp. 271-339).