EXPONENTIAL SMOOTHING
AN EXTENSION

189-66

Christopher R. Sprague

EXPONENTIAL SMOOTHING -- AN EXTENSION

Christopher R. Sprague

I. Introduction

Exponential smoothing in its various forms is a commonly-used technique in demand forecasting. This paper proposes a method for extending the usefulness of the technique where long lead times are encountered.

Section II describes a technique for estimating errors at an arbitrary lead time.

Section III describes a computer program incorporating this technique.

Section IV gives a brief summary of results, which is offered only as indication without any claim of statistical significance.

II. The Exponential Smoothing Model of Demand -- Evaluation of Errors

A wide class of demand-generating processes seem to be well modeled by

$$d(t) = (s + rt) \cdot f(t) + u(t) \qquad (1)$$

where $d(t)$ is demand at time t

s is "average" demand at time 0

r is a linear trend term

f is a seasonal factor

f normalized so that $\sum_{t=1}^{m} f(t) = m$, where m is number of periods in a seasonal cycle

and $u(t)$ is a random variable normally distributed with zero mean and variance independent of t.

While this model is conceptually neat, one sad fact of life is that, to remain valid, it must admit of changes (albeit slow changes) in the values of s, r, and the f's. Normally it is expected that the changes in model parameters are sufficiently slow to be swamped by the u's in any given period.

One useful way of estimating the model parameters so as to use them for predictive purposes is known as Exponential Smoothing with Linear Trend and Ratio Seasonals[1] in which the parameters are recursively estimated as follows:

$$Q(I) = D(I)/F(I) \tag{2a}$$

$$S(I) = A*Q(I) + (1.-A) * (S(I-1) + R(I-1)) \tag{2b}$$

$$R(I) = B * (S(I) - S(I-1)) + (1.-B) * R(I-1) \tag{2c}$$

$$F(I+M) = C * (D(I)/S(I)) + (1.-C) * F(I) \tag{2d}$$

where:

D(I) is actual demand in period I

Q(I) is deseasonalized demand as of period I

D(I) is actual demand in period I

F(I) is estimated seasonal factor for period I

S(I) is estimated average deseasonalized demand as of period I (not period 0)

R(I) is estimated trend for period I

M is the number of periods in one seasonal cycle (e.g., 12 for most monthly data)

and A, B, and C are "smoothing constants" between 0 and 1.

It is evident from the above that the process needs S(0), R(0), and F(1) ... F(M) to begin, but thereafter needs only D(I) for each period to continue.

The proof of the pudding, however, is in the prediction, and the appropriate prediction for month I + L is:

$$P(I+L) = (S(I) + L * R(I)) * F(I+L) \tag{3}$$

where P is a predicted demand, and the F's are reused, i.e.,

$$F(I+L = F(I+L-M), \quad M < L \leq 2M \tag{4}$$

Now, if the model (1) is valid, then the "best" possible estimates of S. R, and the F's will give the "best" possible prediction. These parameters are obviously dependent on the choice of A, B, and C, as well as on the measured demands D, so that the problem is to choose appropriate values of A, B, and C to produce the "best" estimates of the parameters. Typically this is done by establishing a set of smoothing constants and running the smoothing process against historical data for periods L = 1 ... I, predicting 1 period ahead, and forming the sum:

$$E = \sum_{j=1}^{I} (((D(J) - P(J)) ** 2) * G ** (I-J)) \qquad (5)$$

where G is a weighting factor ≤ 1.

This associates a squared-error term with a set of smoothing constants. If G is less than 1.0, the most recent periods are emphasized in the sum, reflecting the idea that errors long-past are not so important as errors in the recent past.

Since we can associate an error term with any set of smoothing constants, we can obviously find an "approximately optimum" set by an exhaustive search or by hill-climbing or by some combination of both. This is called "Adaptive Smoothing".

The remainder of this paper is insensitive to which method of the two above is used, so we here leave the point.

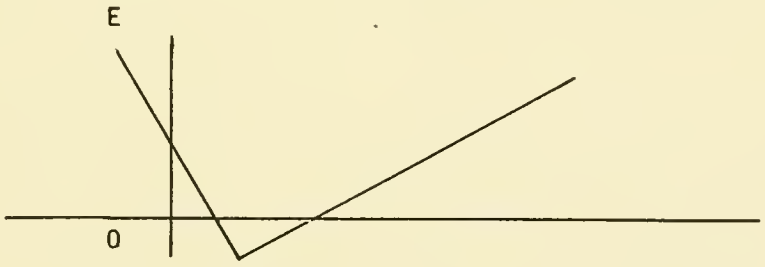
It is widely held that the procedure outlined above results in a set of parameters which will give the "best" predictions of future demand. While "nonsense" is perhaps too strong a word to apply to this belief, let us at least try to cast some doubt (or some stones). There are at least four reasons why this approach ought to be examined very carefully before being adopted:

1. The model (1) is only approximately valid. If it were exact, then we could indeed conclude that the set of smoothing constants which minimized prediction error 1 period in advance was the set which would yield the "best" estimates of model parameters.

2. The manager who is on the line for a sales forecast does not care whether the parameters are correct. He wants to know that his predictions are good for some meaningful lead time (perhaps 6 to 30 periods). It may be useful to note here that after some time, the F's may no longer sum to M, and S and R may be under or over-stated proportionately. This great change in model parameters makes absolutely no difference in the prediction!

3. Assuming that the desired prediction lead-time is greater than 1, the process as described yields no estimate of error (or cost) at the actual lead-time to be used.

4. Where did the sum-of-squared-errors term arise, anyway? It came from statistics, least-squared fits, and the like, where its virtue is computational simplicity (later in this paper we exploit this in another context). In this case, however, there is no great computational advantage and the error (or, more properly, cost) function might just as well be absolute deviation or any arbitrary thing like:



D(J) - P(J) (6)

Before proceeding, we should probably dispose of objection (4) above. While there is nothing ideal about the sum-of-squared-error formulation, it

does have a number of advantages. First, and probably controlling, if we divide such a sum by the sum of weights $\sum_{J=1}^{I}$ (G ** (I.J)) we obtain a number like a variance, which can be viewed as an estimate of the variance of u (see Equation (1)). While this analogy cannot be pushed too far,[2] it is useful for people with some "feel" for statistical concepts. Second, the square root of the "variance" is at least of the same order of magnitude as the expected absolute error in each prediction, and this is a useful number for the user. Third, and last, if the error (or cost) function is even mildly discontinuous, there is little hope of finding an acceptable "approximately optimum" set of parameters by any method short of exhaustive search on an extremely fine grid.

So, having raised four objections and backed off from one, we pursue the other three, all of which seem to be easily answered by simply evaluating predictions at the desired lead time rather than at a lead time of 1.

As with many brilliantly simple solutions, this one has rather grave flaws, for if we have data for periods 1 ... N and desire a prediction for period N + L, our error function will contain terms for predicted vs. actual demands in periods L ... N, i.e., based on parameter values as of periods 0 ... N-L. Thus, instead of having N terms, the error function will contain N-L + 1 terms. Worse yet, the most recent of those will be L periods old. Perhaps some numbers are in order.

Suppose we have 30 periods 1 ... 30 of real data and wish to predict period 42; so L is 12. If we were predicting with L = 1, our error function would contain P(1) - D(1) through P(30) - D(30) based on S(0), R(0), F(1) ... F(M) through S(29), R(29), F(30) ... F(29+M). Instead with L = 12, we will use P(12) - D(12) through P(30) - D(30), based on S(0), R(0), F(1) ... F(M) through S(18), R(18), F(19) ... F(18 + M).

The fact that each sum of errors has 11 fewer terms than it otherwise would is serious indeed, but it pales by comparison with the fact that the most recent such term is always 11 periods too old (relatively). Clearly another way out is desirable.

We now propose a method of estimating error at an arbitrary lead time. It is based on the assumption that the expected squared error at any lead time L is a linear function of L itself, given a set of smoothing constants and the initial estimates S(0), R(0), and F(1) ... F(M). Let us adopt some new notation:

Z(I,L) = (S(I) = L * R(I)) * F(I+L); the prediction of period I + L made at the end of period I. (7)

E2(I,L) = (Z(I,L) - D(I+L)) ** 2; the square of the difference between the actual demand in period I + L and the prediction made for the same period at the end of period I. (8)

W(I) = G ** (N-I), G $\leq$ 1.; a weighting factor corresponding to period I's set of predictions for periods I + L, where L = 1 ... N - 1. (9)

We now seek coefficients U and V such that the double sum

$$\sum_{I=0}^{N-1} \left(W(I) \sum_{L=1}^{N-1} (E2(I,L) - U - V * L) ** 2 \right) \qquad (10)$$

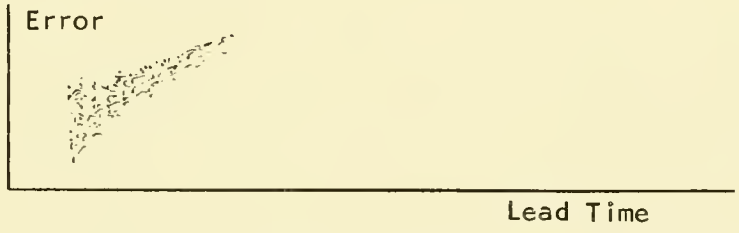
is minimized. In essence, we are going to perform a weighted curve-fitting with lead time the independent variable, squared prediction error the dependent variable, and weights falling geometrically with age of prediction. We will produce N * (N+1)/2 "observations" (or, more properly, points to be fitted) as follows:

Using data through period 0 (i.e., the initial estimates), we produce predictions for periods 1 ... N, i.e., Z(0,1) ... Z(0,N). The squared difference between each prediction and its corresponding data item is E2(0,1) ... E2(0,N) and the associated lead time is L = 1 ... N. The weight for all these points is G ** N.

Similarly, using data through period 1, we produce predictions for periods 2 ... N, i.e., Z(1,1) ... Z(1,N-1), corresponding errors E2(1,1) ... E2(1,N-1) and lead times L = 1 ... N-1. The weight for all these is G ** (N-1).

This process continues until, using data through period N-1, we predict period N, i.e., Z(N-1,1), error E2(N-1,1), at lead time L=1, with weight G.

Ignoring the weights for a moment, a scatter diagram of these points would look something like Figure 2:



Methods for obtaining such a "best linear fit" are well known, so we state only the equations used to find U and V

$$X0 = Y[1] \tag{12a}$$

$$X1 = Y[L]/X0 \tag{12b}$$

$$X2 = Y[E2(1,L)]/X0 \tag{12c}$$

$$X11 = Y[L ** 2]/X0 - X1 ** 2 \tag{12d}$$

$$X22 = Y[E2(1,L) ** 2]/X0 - X2 ** 2 \tag{12e}$$

$$X21 = Y[L * E2(1,L)]/X0 - X1 * X2 \tag{12f}$$

$$V = X21/X22 \tag{12g}$$

$$U = X2 - V * X1 \tag{12h}$$

where the special symbol Y[H] is taken to mean

$$Y[H] = \sum_{I=0}^{N-1} \left(W(I) \sum_{L=1}^{N=1} H \right) \qquad (13)$$

Now we obtain an estimate of the expected squared error at any lead time L as

$$E = U + L * V \qquad (14)$$

This answer to the first three objections stated on page 4 has some flaws, of which any potential user should be aware:

1. It is expensive. Each set of smoothing constants tested requires doing a linear regression computation involving N * (N + 1)/2 observations. This works out to approximately 4N times the calculations required for simple evaluation of error at L = 1. Computer time is rapidly becoming a cheap commodity, however, and a practical problem (N = 96, L = 12, 125 different combinations of A, B, and C) takes about 3 minutes of 7094 time. This is not excessive for, say, a monthly forecast of corporate sales.

2. It could be wrong -- especially where the actual demand-generating process at work is very close to the model given in (1), where, presumably, the simple scheme of evaluation at L = 1 would be better. In certain pathological cases (such as using the wrong value of M)[3] V might be negative -- a clear indication of trouble, since for some L, this would imply negative squared errors. This did not occur in several trials with "dirty" data, however.

3. It is very sensitive to the choice of G, S(0), R(0), and F(1) ... F(M). The simple scheme is also sensitive to these, but, because the relationships are simpler, obvious mistakes are more easily detected.

4. It rests on some shaky foundations -- notably the assumption that

squared error of prediction is a linear function of lead time. This is not much more ridiculous than the rest of the assumptions which underlie exponential smoothing, a technique which is well established and demonstrably useful.

In summary, we have described a method for choosing a set of smoothing constants on the basis of expected error at some arbitrary prediction lead time. It attempts to accomplish three objectives: overcome minor invalidities in the model underlying Exponential Smoothing with Linear Trend and Ratio Seasonals; emphasize quality of prediction rather than quality of parameters; and give a meaningful estimate of error at arbitrary lead time rather than at the normal, nearly-meaningless 1 period. These objectives are achieved at some cost: increased computation; increased risks of error due to poor initial conditions of strange data; and shaky theoretical underpinning. While the model has performed well in limited tests with real data, it will be some time before any determination of its relative advantage can be made.

III. Design of a Computer Program Using the Method for Choosing Smoothing Constants

As occasionally happens, the method of selecting smoothing constants described above was incorporated into an operating computer program somewhat before it had been conceptualized. It is useful, perhaps, to describe the program from the point of view of a use..

1. General

The program is designed to run on an IBM 709-7090-7094 operating under the control of the 9/90/94 FORTRAN Monitor System (FMS). The operating deck consists of:

```
* (installation sign-on card)

* XEQ
  (binary deck)

* DATA
  (data cards)
```

The program is input-driven by a set of control cards. The basic scheme is to read input data until a control card is recognized, then log a message to the effect that such-and-such control card was encountered, and then proceed to take action as designated by the control card. Thus a wide variety of functions can be called for by changes in the input deck rather than the program.

2. Entering Basic Data

When the program encounters a card where the first six columns are *DEPEN, it is conditioned to accept basic demand data. It reads one more card, interpreted as follows:

Columns	1 - 60	Alphabetic description of data
	61 - 62 Period	of first piece of data entered
	63 - 64 Year	
	65 - 66 Period	of last piece of data entered
	67 - 68 Year	
	69 - 70 M, the number of periods/year (e.g., 01, 04, 12, 52)	

The first and last periods given in this card define the number of periods (200 maximum) of data to be entered. The program expects these data to follow immediately, 9 eight-column fields per card.

3. Establishing a Set of Smoothing Constants

When the program encounters a card whose first six columns are *ALPHA it reads one more card to define a set of smoothing constants as follows:

Columns	1 - 2	Number of values of A to be tried
	3 - 4	First value of A in 1/100th
	5 - 6	Second through 11th values of A
	...	
	23 - 24	Same as above
	25 - 26	Number of values of B to be tried
	27 - 48	As above
	49 - 50	Number of values of C to be tried
	51 - 72	As above

In default, the program uses 5 values each of A, B, C, specifically .1, .3, .5, .7, .9.

4. Establishing a Range of Predictions

When the program encounters a card whose first six columns are *RANGE, it reads one more card to define a set of ranges of data as follows:

Prediction Range:	Columns	1 - 2	Period	First data point
		3 - 4	Year	
		5 - 6	Period	Last data point
		7 - 8	Year	
Evaluation Range:		9 - 10	Period	First data point
		11 - 12	Year	
		13 - 14	Period	Last data point
		15 - 16	Year	
Computation Range:		17 - 18	Period	First data point
		19 - 20	Year	
		21 - 22	Period	Last data point
		23 - 24	Year	

The prediction range is the range of dates over which prediction based

on the already stored data are desired.

The evaluation range is the range of dates over which prediction error is to be estimated. It normally would be identical with the prediction range.

The computation range is the range of dates of the basic data which may be used in the predictions. It normally coincides with the range of stored data.

If we denote the range of stored data as the "data range", then the default conditions are as follows:

If one leaves unspecified	The program uses
Low date prediction range	High date data range +1
High date prediction range	Low date prediction range
Low date evaluation range	Low date prediction range
High date evaluation range	Low date evaluation range
Low date computation range	Low date data range
High date computation range	Minimum of: High date of data range and low date prediction range less 1

The *RANGE card also initiates the actual prediction procedure.

A. Force all ranges to consistency with prediction range.

B. Form initial estimates of S(0), R(0), F(1) ... F(M). The program "cheats" by using the first two years of the computation range to produce these estimates.

C. For all possible combinations of A, B, C, form an estimate of the error at the mid-point of the evaluation range. This corresponds exactly to the method discussed in section 1 except that it truncates the curve-fitting process not at N but at lead time corresponding to the difference between the high dates of the evaluation and computation ranges. (This saves some time.)

D. Using the "best" A, B, C, form terminal estimates of S(N), R(N), F(N+1) ... F(N+M).

E. Form and print estimates for the prediction range.

For purposes of testing, the program also contains facilities for analyzing differences between actual and predicted demands, and for comparing its predictions with those made by a simple linear regression against an optimally-lagged leading indicator series. These, however, are not directly relevant.

IV. Results in Brief

Tests were made on two series, both monthly ten-year sales histories of inexpensive durable goods. Predictions were made for the period covered by the last year of the series given average lead times of six (6) to thirty (30) months. Over this limited range, at least, the actual average squared error appeared to be no worse than linear with lead time. In none of the trials did the estimated average error at a specified lead time differ from the actual average error observed by more than 10%.

Footnotes

[1] See Holt, Modigliani, Muth, and Simon, Planning Production, Inventories, and Work Force (Englewood Cliffs: Prentice Hall), 1960, and R. G. Brown, Smoothing, Forecasting, and Prediction of Discrete Time Series (Englewood Cliffs: Prentice Hall), 1963

[2] Note that the model which minimizes this "variance" does not necessarily produce unbiased estimates of demand.

[3] Some closed systems tend to generate their own internal cycles. As Forrester points out, such systems are not good subjects for statistical or quasi-statistical analysis at the aggregate level.

187-66

188-66

189-66

191-66

192-66

194-66

195-66