The Combination of Forecasts

J. M. BATES and C. W. J. GRANGER

Department of Economics, University of Nottingham

Two separate sets of forecasts of airline passenger data have been combined to form a composite set of forecasts. The main conclusion is that the composite set of forecasts can yield lower mean-square error than either of the original forecasts. Past errors of each of the original forecasts are used to determine the weights to attach to these two original forecasts in forming the combined forecasts, and different methods of deriving these weights are examined.

INTRODUCTION

OUR INTEREST is in cases in which two (or more) forecasts have been made of the same event. Typically, the reaction of most statisticians and businessmen when this occurs is to attempt to discover which is the better (or best) forecast; the better forecast is then accepted and used, the other being discarded. Whilst this may have some merit where analysis is the principal objective of the exercise, this is not a wise procedure if the objective is to make as good a forecast as possible, since the discarded forecast nearly always contains some useful independent information. This independent information may be of two kinds:

(i) One forecast is based on variables or information that the other forecast has not considered.

(ii) The forecast makes a different assumption about the form of the relationship between the variables.

The second case in particular does not necessarily lead to a situation in which a combined forecast improves upon the better individual forecast, though there are occasions when it can, as is shown in Section 1 of the Appendix.

It should be noted that we impose one condition on the nature of the individual forecasts, namely that they are unbiased. A set of forecasts that consistently overestimate the true values would, if combined with a set of unbiased forecasts, lead to forecasts which were biased; in all likelihood the combined forecasts would have "errors" rather larger than the unbiased forecasts. The first step therefore is to check that the individual sets of forecasts are unbiased, and if biased to correct for the average percentage (or absolute) bias.

Before the discussion of different ways in which forecasts could be combined, an empirical justification is given by making a crude combination of two forecasts. The forecasts chosen were of the international airline passenger data, for which (amongst others) Brown, and Box and Jenkins have made monthly forecasts for one period ahead. The forecasts are published in an article by Barnard,[1] who says, "the forecasting methods ... developed by Professor Box ...

and Dr. (now Professor) Jenkins . . . have proved . . . so successful . . . that we are now searching for processes . . . (for which) it is possible to find alternative methods which forecast better". The combination illustrated is the arithmetic mean of the two individual forecasts, with Table 1 giving the details for 1953.

TABLE 1. ERRORS IN FORECASTS (ACTUAL LESS ESTIMATED) OF PASSENGER MILES FLOWN, 1953

Month	Brown's exponential smoothing forecast errors	Box–Jenkins adaptive forecasting errors	Combined forecast ($\frac{1}{2}$ Brown + $\frac{1}{2}$ Box–Jenkins) errors
Jan	1	−3	−1
Feb.	6	−10	−2
March	18	24	21
April	18	22	20
May	3	−9	−3
June	−17	−22	−19·5
July	−24	10	−7
Aug.	−16	2	−7
Sept.	−12	−11	−11·5
Oct.	−9	−10	−9·5
Nov.	−12	−12	−12
Dec.	−13	−7	−10
Variance of errors	196	188	150

An enumeration of these and other forecasts of these data is made at a later stage. For the moment it may merely be noted that for the period 1951–60 the variance of errors in the three forecasts mentioned were 177·7 (Brown), 148·6 (Box–Jenkins), and 130·2 (combination with equal weights to each of the individual forecasts). Thus, even though Brown's forecasts had a larger variance than that of Box–Jenkins's forecasts, they were clearly of some value.

Work by Stone *et al.*[7] has made use of ideas rather similar to these, though their work related solely to making improved estimates of past national income figures for the U.K. and did not tackle forecasting problems.

CHOICE OF METHOD FOR DETERMINING WEIGHTS

Though the combined forecast formed by giving equal weights to each of the individual forecasts is acceptable for illustrative purposes, as evidence accumulated one would wish to give greater weight to the set of forecasts which seemed to contain the lower (mean-square) errors. The problem was how best to do this. There are many ways of determining these weights, and the aim was to choose a method which was likely to yield low errors for the combined forecasts.

Our first thought for a method was derived in the following way. It was assumed that the performance of the individual forecasts would be consistent over time in the sense that the variance of errors for the two forecasts could be denoted by σ_1^2 and σ_2^2 for all values of time, t. It was further assumed that both forecasts would be unbiased (either naturally or after a correction had been applied). The combined forecast would be obtained by a linear combination of the two sets of forecasts, giving a weight k to the first set of forecasts and a weight $(1-k)$ to the second set, thus making the combined forecast unbiased. The variance of errors in the combined forecast, σ_c^2 can then be written:

$$\sigma_c^2 = k^2 \sigma_1^2 + (1-k)^2 \sigma_2^2 + 2\rho k \sigma_1 (1-k) \sigma_2,$$

where k is the proportionate weight given to the first set of forecasts and ρ is the correlation coefficient between the errors in the first set of forecasts and those in the second set. The choice of k should be made so that the errors of the combined forecasts are small: more specifically, we chose to minimize the overall variance, σ_c^2. Differentiating with respect to k, and equating to zero, we get the minimum of σ_c^2 occurring when:

$$k = \frac{\sigma_2^2 - \rho\sigma_1\sigma_2}{\sigma_1^2 - \sigma_2^2 - 2\rho\sigma_1\sigma_2}. \tag{1}$$

In the case where $\rho = 0$, this reduces to:

$$k = \sigma_2^2/(\sigma_1^2 + \sigma_2^2). \tag{2}$$

It can be shown that if k is determined by equation (1), the value of σ_c^2 is no greater than the smaller of the two *individual* variances. The algebra demonstrating this is recorded in Section 2 of the Appendix.

The optimum value for k is not known at the commencement of combining forecasts. The value given to the weight k would change as evidence was accumulated about the relative performance of the two original forecasts. Thus the combined forecast for time period T, C_T, is more correctly written as:

$$C_T = k_T f_{1,T} + (1-k_T) f_{2,T},$$

where $f_{1,T}$ is the forecast at time T from the first set and where $f_{2,T}$ is the forecast at time T from the second set.

Equations (1) and (2) are used as a basis for some of the methods that follow shortly. Thought, however, was given to the possibility that the performance of one of the forecasts might be changing over time (perhaps improving) and that a method based on an estimate of the error variance since the beginning of the forecasts might not therefore be appropriate. In consequence we have also constructed two methods which give more weight to recent errors than to those of the past (see methods (iii) and (iv) below).

DESIRABLE PROPERTIES OF METHODS

Good methods (defined by us as those which yield low mean-square forecast error) are likely to possess properties such as:

(a) The average weight of k should approach the optimum value, defined by (1), as the number of forecasts increased—provided that the performance of the forecasts is stationary.

(b) The weights should adapt quickly to new values if there is a lasting change in the success of one of the forecasts.

(c) The weights should vary only a little about the optimum value.

This last point is included since property (a) is not sufficient on its own. If the optimum value for k is 0·4, one may still obtain poor combined forecasts if k takes two values only, being 0 on 60 per cent of occasions and 1·0 on the remaining 40 per cent.

In addition to these properties, there has been an attempt to restrict methods to those which are moderately simple, in order that they can be of use to businessmen.

Five methods have so far been examined, and in each of them the combined forecast at time T, C_T, has been derived from giving a weight k_T to the forecast for time T from the first set and a weight $1-k_T$ to the second forecast for time T; the weights k_T have in all cases been determined from past (known) errors of the two series denoted as $e_{1,1}, e_{1,2}, \ldots, e_{1,T-1}$ and $e_{2,1}, e_{2,2}, \ldots, e_{2,T-1}$, except for k_1 which has been arbitrarily chosen as 0·5 for all methods.

The methods are:

(i) Let us denote $\sum_{t=T-v}^{T-1}(e_{2,t})^2$ by E_2, and a similar summation of the first set of forecast errors by E_1. Then:

$$k_T = \frac{E_2}{E_1+E_2}.$$

(ii)
$$k_T = xk_{T-1}+(1-x)\quad \frac{E_2}{E_1+E_2},$$

where x is a constant of value between zero and one.

(iii) Let us denote $\sum_{t=1}^{T-1} w^t(e_{2,t})$ by S_2^2. w is seen to be a weight, which for $w>1$ gives more weight to recent error variances than to distant ones.

$$k_T = \frac{S_2^2}{S_1^2+S_2^2}.$$

(iv) Let us denote the weighted covariance by C, where:

$$C = \sum_{t=1}^{T-1} w^t e_{1,t} e_{2,t},$$

$$k_T = \frac{S_2^2-C}{S_1^2+S_2^2-2C}.$$

(v)
$$k_T = xk_{T-1}+(1-x)\quad \frac{|e_{2,T-1}|}{|e_{1,T-1}|+|e_{2,T-1}|}.$$

The differences between methods are not the only factors, and suitable choice of the parameters, v, x and w, are also of importance. Despite this, our presentation concentrates upon the differences between the various methods, since we would expect these to be the more important.

It is pertinent to note that method (v) fails to satisfy criterion (a). It scores well on criteria (b) and (c), and for combining forecasts which are almost equally good it does very well. However, since it underestimates the weight to give to the better forecast it is to some extent an unsatisfactory method. A further disadvantage is that the weight k_t is as much affected by a given ratio arising from two small errors in the individual forecasts as it is from the ratio arising from two large errors. Thus if past errors have averaged 10, k_t will be just as much altered if $e_{1,t-1} = 0{\cdot}01$ and $e_{2,t-1} = 0{\cdot}02$ as if the errors were 10 and 20. None of the other methods need suffer from this disadvantage. Further, they all satisfy criterion (a), though some may not be so obviously satisfactory in respect of criterion (b). It is perhaps easier to comment further on the methods after referring to empirical work undertaken.

PERFORMANCE OF DIFFERENT METHODS

The empirical work undertaken so far has been rather limited, in that we have examined forecasts only of the airline passenger data referred to above. To reach definite conclusions about most of the methods is therefore impossible. There are, however, a number of tentative conclusions that do emerge.

Five forecasts of the airline passenger data have been examined, all being monthly forecasts, made for one-period ahead. In addition to those of Brown and Box–Jenkins for the years 1951–60, there are a set of forecasts prepared by P. J. Harrison for the years 1954–9 and two sets of forecasts for 1951–60 prepared by one of us. Those prepared by Harrison are similar to the "Seatrend" ones published by him in *Applied Statistics*,[5] the sole difference being that the present forecasts are based on three parameters as opposed to two for the published ones. The two forecasts made by ourselves are referred to as the Constant Seasonal and the Changing Seasonal forecasts. The basis of both forecasts concerns two stages, the first being the estimation of a simple seasonal component which, together with the summation of data for recent months, is used to make a crude forecast for the next month. The second step is to "correct" this crude forecast by utilizing information of the last month's crude forecast error. The only difference between the two forecasts is that the Changing Seasonal forecast analyses if the seasonal components are changing over the years, and if so makes some allowance before deriving crude forecasts. The two forecasts are described more fully in another article.[2]

Since Brown's "exponential smoothing" and Box–Jenkins's "adaptive forecasting" methods are fairly familiar to readers, the results of combining these two forecasts will be described first. The Brown forecasts are obtained by analysing data from 1948 onwards, fitting a first-order harmonic as an estimate of the seasonal, and then making estimates of the level and trend factors in the usual exponential smoothing way. For details consult Harrison.[5] The Box–Jenkins forecasts are derived from forecasting 12-monthly moving sums, 11 months of which are known. The forecast for the sum of 12 months ending in month T, F_T, is dependent upon the forecast error in F_{T-1}, the change in error between F_{T-2} and F_{T-1} and the sum of all past errors. For general details of the method consult Box and Jenkins.[3]

The overall variances of the two individual forecast errors were 177·7 and 148·6; together with a correlation coefficient of 0·60, this implied that a value of $k = 0{\cdot}39$ throughout [see equation (1)] would lead to an optimal combined forecast if the ratio of the two individual forecast error variances were constant throughout the whole period. Equation (2), which takes no account of the correlation between the two sets of errors, gives $k = 0{\cdot}46$. Using $k = 0{\cdot}46$ would lead to combined forecasts not quite as good, but which might be thought to be the likely lower limit for methods (i)–(iii) and (v).

The results for combining different methods are recorded in Table 2, both for the absolute errors and the proportionate errors. In many ways the use of proportionate errors is a more appropriate measure, since both the mean and the absolute errors increase considerably over time, but it is to be noted that both these forecasts were prepared from an analysis based on absolute figures.†

It may be noted that all the methods except the fourth were capable (with appropriate choice of parameters) of yielding forecasts which were better than the ones thought to be "optimal". This is indeed surprising, since the value of the constant weights is calculated using the sum of error variances for all 10 years: in other words, the forecasts derived from the constant weights have been possible only by the use of hindsight. In contrast, each of the five methods used only past errors in making a combined forecast for any period. Why does this result occur? The absolute errors of Brown's set of forecasts were not consistently higher than Box–Jenkins's in the ratio 177·7 : 148·6 for all time periods. At times the Brown forecasts were good, and if the forecast error at time t was small it was likely that the error in month $t+1$ would also be small; if, however, there was a change in trend, Brown's exponentially smoothed forecast not only failed to discern it immediately (as is inevitable), but took some time before it fully corrected for it. In consequence Brown's forecast errors are highly serially correlated, yielding a Durbin–Watson statistic of 0·83 for the period 1951–60, a value which is significant at the 1 per cent level. In addition, the seasonal model assumed by Brown fails adequately to reflect the seasonal pattern at certain times of the year. These two reasons are partly responsible for the success of the methods in yielding lower forecast errors than

† In contrast it is interesting to note that a more recent set of forecasts by Box, Jenkins and Bacon are based on the logarithms of the data. The basis for deriving these forecasts is published in a book edited by Harris.[4] The error variance of these forecasts is 160.

TABLE 2. OVERALL VARIANCES OF FORECAST ERRORS, 1951–60

Forecasts		Weight constant throughout		Weights varying according to method										
				(i)		(ii)			(iii)		(iv)		(v)	
		Wt.	Variance	v	Var.	v	x	Var.	w	Var.	w	Var.	x	Var.
				1	125·6	1	−0·3	123·3	1	130·9	1	134·1	−0·3	125·1
				2	127·1	1	0	125·7	2	125·5	2	131·0	0	126·0
Brown	178	0·39	128·6	3	129·2	1	0·6	129·2	3	125·6	3	134·2	0·6	128·7
Box–Jenkins	149	0·16	129·2	6	128·8	7	−0·6	122·2					1	130·2
				12	130·8	7	0	125·7						
$r = 0{\cdot}60$						7	0·2	127·5						
Percentage errors:														
				1	14·3	1	0	14·4	1	14·4	1	14·6	0·3	14·3
				2	14·7	1	0·25	14·2	1·5	14·1	1·5	15·0	0·3	14·18
Brown	19·1	0·47	14·23	3	14·4	1	0·5	14·2	2	14·1	2	15·0	0·6	14·2
Box–Jenkins	16·7	0·42	14·19	6	14·16	7	0	14·1	2·5	14·2	2·5	15·2	1	14·3
				12	14·3	7	0·2	14·2						

Note: Starting values are $k_1 = 0{\cdot}5$ everywhere.

For methods (i)–(iv) k_2 was usually taken as the average of k_1 and the value suggested by the strict application of the methods as described above. Reasons for this are discussed in the next section.

would have resulted from the weight k being constant throughout; they may also help to explain the success of parameters which enable the weight to be determined by the most recent error only (see, in particular, method (1) for the absolute errors). This phenomenon one would expect to be the exception rather than the rule perhaps, since an *average* of recent forecast errors would normally be regarded as a more reliable measure of "error" than the last forecast error.

The poor performance of method (iv) deserves some comment. The weighted correlation coefficient often becomes very high for high values of w and can result in k becoming 0 or 1, thus failing to meet desired property (c) (values of k should vary only slightly about the average value). This method tends therefore to score badly when positive serial correlation of residuals is present in one of the forecast errors, since the weight w would best be high in computing $S_2^2/(S_1^2+S_2^2)$, but is best rather lower for computing C. If the nature of forecast errors is such that serial correlation of residuals is observed or expected, then method (iv) should either not be used or be altered to something similar to:

$$k_T = \frac{S_2^2 - zC}{S_1^2 + S_2^2 - 2zC},$$

where $0 \leqslant z \leqslant 1$. One crude possibility might be to make $z = 1/w$.

The results for the combination of Brown and Box–Jenkins forecasts are somewhat untypical, and it is appropriate to examine the results for other combined forecasts. The principal features of the individual forecasts are given in Section 4 of the Appendix, together with the results for different combinations of forecasts (see Tables A3 and A4). There are five noteworthy features of these results. First, there is little to choose between the performance of the different methods of generating the weights k_t. Second, with one exception the combined forecasts have considerably lower error variance than the better of the individual forecasts used in the combination. The one exception is for the combination of the Constant and Changing forecasts, where the Changing forecast is based entirely on the same estimate of the seasonal as the Constant *except* for an allowance for a changing seasonal pattern partway through the period (1956). In all other respects these two forecasting methods are identical. Third, optimal values for the parameter x are not always found to be positive. This occurs in the combination of Brown and Changing forecasts, and also, rather surprisingly, in method (ii) for combining Box–Jenkins with both Constant and Changing. Our surprise exists because it cannot be explained by the presence of serial correlation of residuals, since none exists in any of the original series. Interestingly enough, the optimal values for x are positive for all combinations with the Harrison forecasts. Fourth, in contrast to the results obtained in combining Brown's and Box–Jenkins's forecasts, an average of a number of past errors is often a better measure of "error" than the last error only. Fifth, in methods (iii) and (iv) "high" values of w are often found to perform better than low values. Few people with experience of use of "discount" factors ($=1/w$) would expect that $w \geqslant 2$ would perform well.

Where these oddities occur, there might be a temptation to lay the blame on the original forecasts, saying that these forecasts possess certain undesirable features, rather than attribute the oddities to characteristics of the methods. This, however, would be unconstructive. The objective is to devise robust methods for combining different forecasts. There is no reason to believe that the forecasts examined are untypically poor; quite the contrary indeed! The purpose in commenting upon the characteristics of the original forecasts is to see if it is possible to learn why the methods for combining forecasts do well or badly; analysis of, and speculation about, the reasons for the performances observed may well add useful information to that derived from a listing of the empirical results. (We hope soon to add to the empirical results by examining forecasts made for other variables.) In general, then, the objective is to discover a robust method which will combine satisfactorily the forecasts that exist, whatever their characteristics; further, it is not to be expected that the characteristics of these forecasts will always be known—hence the need for robustness.

MINOR MODIFICATIONS TO METHODS

One modification is a delay in giving full weight to methods (i) to (iv). The results given in the above section and in the Appendix use weights for time period $t=2$ which are in most cases an arithmetic average of the weight suggested by the method and a weight of 0·5. The reason is that the methods are able to use only one error for each forecast and are therefore likely to give unreliable estimates of the errors to be expected for each series. In the third part of the Appendix is shown the distribution of:

$$\frac{\sum_t (e_{2,t})^2}{\sum_t (e_{2,t})^2 + \sum_t (e_{1,t})^2},$$

that is obtained when two sets of sample errors are taken from populations having the same variance, which itself is unchanging over time. It can be seen that the 95 per cent limits are very different from 0·5 for small sample sizes.

Though the results quoted here are for weights modified in time period 2 only, it is possible to modify the weights for a number of time periods up to time period A, say, which may be chosen arbitrarily. An appropriate formula for this might be:

$$k_T = \left(\frac{T-1}{A}\right) 0{\cdot}5 + \frac{A-(T-1)}{A} \hat{k}_T, \quad \text{for } T-1 \leqslant A,$$

where $\hat{k}_T$ is the value of k_T derived from the straightforward application of the method.

A second minor modification was to restrict k_t to lie between 0 and 1. "Likely" (i.e. positive) values of the parameters would not generate values of k_t outside this range, but negative values of x have been examined in methods (ii) and (v). The restriction on the k_t's is then deemed advisable. The unfortunate effect of a negative weight for k_t is easily imagined; if two forecasts are 80 and 100 and the respective weights are -1 and $+2$ the combined forecast is 120.

COMBINING FORECASTS FROM THE OUTSET

Section 4 has given results of combining forecasts using different methods and using different values of the parameters x, w and v. In a forecasting situation, one has the problem of choosing from the outset the method by which the forecasts will be combined and the parameters of the method.

One simple way of proceeding would be the following:

(a) k_1 (the weight given to the first forecast in period 1) $= 0{\cdot}5$ unless there is good reason to believe that one forecast is superior to the other.

(b) Use method (i) to determine weights for $k_2, k_3, \ldots, k_6$, summing $\sum_{t=t-v}^{t-1}(e_{2,t})^2$ and $\sum(e_{1,t})^2$ over as many terms as exist. For k_1 and beyond, sum $\sum(e_{2,t})^2$ and $\sum(e_{1,t})^2$ over the last six terms only.

The choice of the method to use is somewhat arbitrary. We have made our choice on grounds of simplicity. In so far as we have any evidence of its performance, method (i) is not obviously inferior to the others. Table A5 of the Appendix records the rankings of the performance of the different methods on combining the airline forecasts. Method (ii) also scores well, but involves the choice of an additional parameter, a choice not easy to make. The recommendation to sum over the last six terms is based partly on the performance of $v = 6$ (see Tables A3 and A4).

There are other possibilities which are likely to yield satisfactory results, one such being the use of method (iii) with $w = 1{\cdot}5$ or 2. It is, however, doubtful whether there is much improved performance in any of these other possibilities, none of which are quite as easy to administer.

The suggestion made above must be regarded as tentative until further work has been done.

COMMENTS

It seems appropriate to emphasize what is our methodology in this work. The methods suggested utilize observed errors of the forecasts and are to this extent purely automatic; at no stage is an attempt made to analyse the reasons for errors in either set of individual forecasts, nor has any assumption been made about the underlying model of the time series. Such work is properly the responsibility of the initial forecaster(s). A moment's reflection will reveal why this is so, and why it would be fruitless to attempt to build a model of the way an economic time series is generated in order to derive the theory of how to combine forecasts. If a model of a time series can be accurately specified then it is possible to derive individual forecasts which could not be bettered. The combining of forecasts would have no part to play in such a situation. This is further commented upon at the end of Section A2 of the Appendix. Whether such a state is commonly achieved is left to the reader to judge.

There may, however, be some point in analysing certain characteristics of the initial forecast errors, since there is some suggestion that these characteristics have some effect on the relative performance of the different methods of combining, and quite considerable effect on the relative performance of different parameters within any given method. It would be unwise to base general conclusions on results only for airline passenger tickets sold, but the presence of serial correlation of residuals in one of the initial forecasts may well have implications for the choice of parameters and also may require modifications to at least one of the methods.

Finally, there may be some use in comparing individual forecasts with the combined forecast. A combined forecast with a significantly lower error variance than either individual forecast implies that the models used for the individual forecasts are capable of some improvement. One may thus obtain hints of a truer model which would be of analytic value. It should be noted, however, that this "truer model" may be of two kinds; it may be a basically different model incorporating a new variable or a different treatment of a variable, or it may be simply an acknowledgement of non-stationarity of the parameters of the model.

CONCLUSIONS

A number of methods of combining two sets of forecasts have been presented, and it is to be noted that, providing the sets of forecasts each contain some independent "information", the combined forecasts can yield improvements.

One unexpected conclusion is that, though the methods suggested for combining forecasts allow the weights to change, this can often lead to better forecasts than those that would have resulted from the application of a constant weight determined *after* noting all the individual forecast errors.

Finally, though the comments in this paper have related solely to combining two forecasts, there is every reason to combine more than two forecasts (where they exist). Work at Nottingham is now being directed towards operational ways of doing this.

REFERENCES

[1] G. A. Barnard (1963) New methods of quality control. *Jl R. statist. Soc.* **A 126**, 255.
[2] J. M. Bates (1970) A short-term forecasting method for seasonal data. (To be published.)
[3] G. E. P. Box and G. M. Jenkins (1962) Some statistical aspects of adaptive optimization and control. *Jl R. statist. Soc.* **B 24**, 297.

[4] G. E. P. Box, G. M. Jenkins and D. W. Bacon (1967) Models for forecasting seasonal and non-seasonal time series. In *Spectral Analysis of Time Series*, p. 271 (Ed. B. Harris). John Wiley, New York.

[5] P. J. Harrison (1965) Short-term sales forecasting. *Appl. Statist.* **XIV**, 102.

[6] M. G. Kendall and A. Stuart (1961) *The Advanced Theory of Statistics*, Vol. II, pp. 18, 19. Griffin, London.

[7] R. Stone, D. G. Champernowne, and J. E. Meade (1942) The precision of national income estimates. *Rev. Econ. Stud.* **9**, 111.

APPENDIX

A1. *Combining an arithmetic and a logarithmic forecast*

As an example of the usefulness of combining forecasts, we give below arithmetic and logarithmic forecasts of the output indices for the gas, electricity and water sector. Column 2 gives the indices, with 1958 = 100, as published in *National Income and Expenditure*, 1966. The third column records linear forecasts made in the following way. The forecast for year t is made by extrapolating the regression line formed by a least-squares fit of the actual figures for 1948 until the year $t-1$. The fourth column records the logarithmic, or exponential, forecasts obtained from fitting equations of the form log (output) $= a+b$ (time), once again utilizing actual figures for all previous years back to 1948.

Table A1. Individual forecasts of output indices for the gas, electricity and water sector

Year	Actual index (1958 = 100)	Linear forecast	Exponential forecast
1948	58		
1949	62		
1950	67	66	66·3
1951	72	71·3	71·9
1952	74	76·5	77·4
1953	77	79·2	80·3
1954	84	81·9	83·2
1955	88	89·0	88·6
1956	92	91·6	93·7
1957	96	96·0	98·5
1958	100	100·2	103·2
1959	103	104·3	107·8
1960	110	108·1	112·1
1961	116	112·9	117·4
1962	125	118·0	123·3
1963	133	124·2	130·2
1964	137	130·9	137·8
1965	145	137·0	145·0
Sum of squared errors		263·2	84·2

The results for combining forecasts in different ways are:

Weight constant throughout		Weights varying according to method											
		(i)		(ii)			(iii)		(iv)		(v)	
Wt.	Sum of squares	v	S. of sq.	v	x	S. of sq.	w	S. of sq.	w	S. of sq.	x	S. of sq.
0·16	77·3	1	44·7	1	0	44·6	1	101·1	1	101·5	0	54·7
0·5	106·6	2	55·7	1	0·2	50·4	1·5	74·8	1·5	69·6	0·2	60·3
		3	76·0	1	0·4	60·5	2	64·1	2	58·7	0·5	77·5
		6	101·3									
		10	97·8									

Certain features are worthy of note. First, a constant weight of 0·16 (or indeed anything between just above zero and $\frac{1}{3}$) can yield a lower sum of squared errors than the exponential forecast on its own; despite this there is the possibility of obtaining a higher sum of squared errors as is shown by the application of the constant weight of 0·5. Second, the methods which allow the weights to vary according to recent forecast errors can yield considerable improvements upon the optimum constant weight; given the nature of the individual forecasting methods, such a result is perhaps not altogether surprising.

The example has shown that combining forecasts of linear and exponential forecast can be profitable. A word of warning is, however, necessary here. One should not assume that combining two forecasts is bound to yield an improvement upon the better of the two individual forecasts. Many of the output indices for other sectors of the economy are so clearly following an exponential trend that the best that one can achieve by a combined forecast is to give all the weight to the exponential forecast.

A2. *The relationship between the combined forecast variance and the variances of the original forecast errors*

The combined forecast at time t, C_t, is derived from giving a weight k to the "first" forecast and a weight $(1-k)$ to the "second" forecast. If the variance of errors for these two forecasts can be correctly denoted as σ_1^2 and σ_2^2 for all time periods, and if both forecasts are unbiased, then the optimum value for k is given by:

$$k = \frac{\sigma_2^2 - \rho\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}. \tag{1}$$

Under these assumptions (stationarity and unbiasedness), the variance of the combined forecast errors, σ_c^2, becomes:

$$\sigma_c^2 = \frac{\sigma_1^2\sigma_2^2(1-\rho^2)}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}.$$

Then:

$$\sigma_c^2 - \sigma_1^2 = \frac{\sigma_1^2(\sigma_1 - \rho\sigma_2)^2}{(\sigma_1 - \rho\sigma_2)^2 + \sigma_2^2(1-\rho^2)},$$

which is clearly $\geqslant 0$.

Hence σ_c^2 is not less than σ_1^2, nor (by symmetry) σ_2^2.

The combined forecast yields a big improvement in all cases except where the individual forecasts are almost equally good and are highly correlated (ρ close to unity). It is only fair to note that this may well be a common occurrence.

If the larger variance is denoted by σ_2^2, then σ_2^2 can be re-written as $a \, . \, \sigma_1^2$, where $a \geqslant 1$. It should also be noted that if $\rho = 1/\sqrt{a}$, $\sigma_c^2 = \sigma_1^2$. In particular, if σ_1^2 is the minimum variance unbiased forecast then ρ can be shown to be equal to $1/\sqrt{a}$ and hence no improvement is possible in this case by combining the individual forecasts. For details see Kendall and Stuart.[6] There, denoting the efficiency of each forecast (i.e. the variance of each forecast relative to the minimum variance forecast) by E_1 and E_2, ρ is shown to be bounded by:

$$\sqrt{E_1 E_2} \pm \sqrt{(1-E_1)(1-E_2)}.$$

It is to be noted that $\rho > 0$ if $E_1 + E_2 > 1$.

A3. *The distribution of* $\sum(e_{2,t}^2)/[\sum(e_{2,t}^2) + \sum(e_{1,t}^2)]$

TABLE A2. CHARACTERISTICS OF THE STATISTIC $s_2^2/(s_2^2 + s_1^2)$

Sample size n	S.E.	Confidence limits 95 % Lower	95 % Upper	99 % Upper
1	0·354	0·006	0·994	0·9998
2	0·289	0·050	0·950	0·990
3	0·250	0·097	0·903	0·967
4	0·224	0·135	0·865	0·941
5	0·204	0·165	0·835	0·916
6	0·189	0·189	0·811	0·894
7	0·177	0·209	0·791	0·875
8	0·167	0·225	0·775	0·858
12	0·139	0·271	0·729	0·806
24	0·100	0·336	0·664	0·727

These results only relate to the case in which two independent samples are taken from normal distributions: for many types of non-normal distribution the confidence limits would be even wider. The chief point of note is the magnitude of the likelihood of getting values of the statistic close to zero or one for small samples, even when taken from populations with equal variance: for small samples a value close to zero or one is not indicative of a difference between the population error variances.

If two samples (of errors) are taken from normally distributed populations both having zero mean and the same variance then:

$$\sum_{t=-n}^{-1} (e_{2,t}^2) \Big/ \left[\sum_{-n}^{-1} (e_{2,t}^2) + \sum_{-n}^{-1} (e_{1,t}^2) \right] = \frac{F}{1+F},$$

where F is the F statistic.

If we put $X = F/(1+F)$ then the probability that the value of X will fall in the internal $(x, x+\mathrm{d}x)$ is given by:

$$\mathrm{d}P = \frac{x^{\frac{1}{2}(n-2)}(1-x)^{\frac{1}{2}(n-2)}}{B(\frac{1}{2}n, \frac{1}{2}n)}\,\mathrm{d}x \tag{3}$$

Equation (3) is a beta distribution, with an expected value, or mean, of 0·5 and a standard error of $1/[2\sqrt{(n+1)}]$. Recorded below are the standard errors and the 95 and 99 per cent confidence limits for the distribution for a number of different values of n.

A4. *Results of combining forecasts of the airline passenger data*

The empirical work in this paper relates to combining forecasts made of the sales of airline passenger tickets. Five individual forecasts have been used: the forecasts and their principal features are:

Brown. An exponentially smoothed forecast. The residuals are serially correlated. The estimates of some of the seasonal components are subject to some error.

Box–Jenkins. No short-term serial correlation of errors. Avoids a direct estimate of the seasonal components by forecasting 12-monthly totals, thus introducing a slight negative correlation of error in month t with month $t-12$.

Harrison. Some short-term serial correlation of errors. Good seasonal estimates; changing seasonal factors.

Constant. No short-term serial correlation. "Constant" crude seasonals fitted. Since the data exhibit definite changes in the seasonal proportions, this results in positive serial correlation of month t with month $t-12$.

Changing. No serial correlation. Changing crude seasonal factors estimated.

Detailed results for different pairings of forecasts are given in Tables A3 (1951–60) and A4 (1954–9 only).

The performance of the different methods is summarized in Table A5 which records their rankings as judged by the "best" choice of parameters, a rank of 1 indicating the best performance. The ranks for 1951–60 are recorded without brackets, for 1954–9 only with brackets.

TABLE A3. OVERALL VARIANCES OF FORECAST ERRORS, 1951–60

Forecasts	Weight constant throughout		Weights varying according to method											
			(i)		(ii)			(iii)		(iv)		(v)		
	Wt.	Variance	v	Var.	v	x	Var.	w	Var.	w	Var.	x	Var.	
Brown	178		1	156	1	0·4	153	1	152	1	156	0	153	
Constant	174	0·48 / 151	3	152	1	0·6	151	1·5	153	1·5	165	0·3	151	
			6	155	1	0·8	152	2	154	2	170	0·6	150	
			12	154	3	0·4	152					1	151	
$r = 0·71$					3	0·6	153							
Brown	178		1	125	1	−1·5	114	1	127	1	128	−1	115	
Changing	133	0·30 / 123	3	127	1	0	125	1·25	123	1·25	126	0	125	
			6	122	1	0·3	127	1·5	122	1·5	128	0·3	127	
			12	125	7	−1	117	2	123			0·6	127	
$r = 0·65$					7	0	122					1	128	
Constant	174		1	141	1	0	141	1	144	1	146	0	142	
Box–Jenkins	149	0·32 / 142	2	138	1	0·3	141	2	141	2	143	0·2	142	
			3	142	2	0	138	3	140	3	143	0·4	143	
$r = 0·79$			12	143	2	0·4	141	4	140	4	145	1	144	
Changing	133		1	123	1	0	123	1	129	1	131	0	124	
Box–Jenkins	149	0·66 / 128	3	125	1	0·3	123	2	125	2	124	0·2	124	
			6	129	1	0·6	126	3	124	3	125	0·4	125	
			12	130	7	−1	128	4	124			0·6	127	
$r = 0·83$					7	0	129							
Changing	133		1	153	1	0	153	1	143	1	135	0	149	
Constant	174	1·00 / 133	3	144	1	0·3	149	1·2	142	1·2	137	0·4	147	
			6	141	1	0·9	144	1·5	142	1·3	138	0·8	145	
			12	141	7	0	141	2	143	1·5	140	1	144	
$r = 0·88$					7	0·7	141							

TABLE A4. OVERALL VARIANCES, 1954–9 ONLY

Forecasts		Weight constant throughout		Weights varying according to method											
				(i)		(ii)			(iii)		(iv)		(v)		
		Wt.	Variance	v	Var.	v	x	Var.	w	Var.	w	Var.	x	Var.
Brown	170	0·32	122	1	121	1	0	121	1	125	1	128	0	119
Harrison	130			2	113	1	0·5	113	1·2	117	1·2	113	0·4	115
				3	116	1	0·6	114	1·4	117	1·4	113	0·5	116
				6	121	2	0	113						
				12	119	2	0·6	115						
Harrison	130	0·43	106	1	118	1	0·6	110	1	107	1	110	0	113
Box–Jenkins	119			2	111	1	1	107	1·4	107	1·3	108	0·6	109
				3	106	7	0	106	1·8	107	1·6	114	1	107
				6	106	7	0·3	106						
$r = 0{\cdot}68$				12	108	7	0·6	107						
Constant	152	0·42	107	1	126	1	0·3	117	1	110	1	118	0	116
Harrison	130			2	118	1	0·6	109	1·15	110	1·1	122	0·6	107
				3	112	1	0·9	106	1·3	111	1·2	125	0·9	106
				6	114	2	0·8	105						
$r = 0{\cdot}52$				12	109	2	0·9	105						
Changing	116	0·59	102	1	104	1	0·3	103	1	105	1	112	0	102
Harrison	130			2	102	1	0·6	102	1·15	105	1·15	114	0·4	101
				3	103	1	1	103	1·3	106	1·3	115	0·8	102
				12	106	2	0	103						
$r = 0{\cdot}67$						2	0·6	103						

It is to be noted that though the ranks for 1954–9 are often the same as for 1951–60, there are some considerable changes. The results should therefore be interpreted with some caution.

TABLE A5. PERFORMANCE RANKINGS OF METHODS

Combination of forecasts	Method				
	(i)	(ii)	(iii)	(iv)	(v)
Brown and Box–Jenkins	3 (2)	2 (3)	1 (4)	5 (1)	4 (5)
Brown and Changing	1 (2 =)	2 (4)	3 (2 =)	5 (1)	4 (5)
Brown and Constant	4 (1)	2 (3)	3 (4)	5 (5)	1 (2)
Changing and Constant	3 (3)	2 (2)	4 (4)	1 (1)	5 (5)
Changing and Box–Jenkins	2 (2)	1 (1)	4 (4)	3 (3)	5 (5)
Constant and Box–Jenkins	1 = (2)	1 = (1)	3 (3)	5 (5)	4 (4)
Harrison and Brown	(3)	(2)	(5)	(1)	(4)
Harrison and Changing	(3)	(2)	(4)	(5)	(1)
Harrison and Constant	(3)	(1)	(4)	(5)	(2)
Harrison and Box–Jenkins	(2)	(1)	(3)	(5)	(4)