89-07

Useful Criteria to Select Which Forecast or Estimation Method is Better, By Benjamin Klugh, Jr., Gary Keough, and William Pratt, Research and Applications Division and Estimates Division, National Agricultural Statistics Service, U.S. Department of Agriculture, Washington, D.C. 20250, Staff Report No. _____________, September 1989.

DRAFT

ABSTRACT

This paper provides objective decision criteria to select between two methods of forecasting or estimation, (usually a current method and a new method). Three new selection criteria are compared to three standard criteria using simulated data. No universal test is uncovered that performs best across all sample sizes and correlations, however, a decision procedure is recommended. For samples of size 36 or larger a preliminary test approach using the correlation between forecast or estimate errors to select the best method is recommended. For sample sizes smaller than 36, the sample correlation statistic is unreliable, therefore one of the new statistics, NES-1, with a critical limit of 0.9 is suggested. Two agricultural statistics data series examples are included to demonstrate the decision techniques.

```
*********************************************************
*                                                       *
*  This paper was prepared for limited distribution to  *
*  the research community outside the U.S. Department   *
*  of Agriculture.  The views expressed herein are not  *
*  necessarily those of NASS or USDA.                   *
*                                                       *
*********************************************************
```

TABLE OF CONTENTS

TABLES AND FIGURES

TABLES IN THE APPENDIX

SUMMARY

This paper provides criteria to select between two methods of forecasting or estimation. Three new selection criteria are compared to three standard criteria using simulated data. No

universal test is uncovered that performs best across all sample sizes and correlations. However, several reasonable tests are suggested. For samples of size 36 or larger a preliminary test approach using the correlation between forecast or estimate errors to select the test method is recommended. Plotting the errors from the two methods is also recommended to identify influential data points or a possible nonlinear relationship. Reasonable type 1 and type 2 errors are constructed for these test methods. However, samples of size 60 or larger are desired.

For sample sizes smaller than 36, the sample correlation statistic is unreliable. One of the new statistics, a modified Thiel's U statistic with a critical limit of 0.9 is suggested. Type 2 errors under this statistic are often under 5 percent. Unfortunately, type 1 errors under this statistic can be as large as 50 percent. This penalty is paid with small sample sizes to keep the type 1 error small.

When no correlation existed between forecast or estimate errors, mean square error performed best among the three standard statistics evaluated. Under moderate correlation, mean absolute error or mean absolute percent error performed better. All three criteria performed well under high correlation.

Simulation techniques were used to develop historical data series for evaluation. The simulation procedure consisted of randomly generating a final estimate and two unbiased preliminary estimates, an old and a new, for each time period. Errors in the estimates for the new method were generated for 15 different error conditions with values from 50% smaller to 50 % larger than the old method. For each error condition 500 different replicates were created consisting of 12, 36, or 60 time periods. The above process was repeated for 3 different correlations between the errors for the new and old method: 0.0, 0.6, and 0.95. The errors were generated assuming a uniform distribution or a normal distribution. A total of 4,860,000 periods of data were generated for this analysis using SAS on a Compaq 386. If we considered these data periods to be months, then 4,050 centuries of data were produced.

Finally, the recommended procedures are applied in two examples. The first example with small sample sizes is an evaluation of remote sensing estimates versus June Enumerative Survey Estimates of planted acreage for six crops. The second example with large sample size compares the preliminary manufactured Minnesota/Wisconsin milk price estimates to forecasts from Box-Jenkins time series transfer function models.

DRAFT

Useful Criteria to Select Which Forecast or Estimation Method is Better

By Benjamin Klugh, Jr., Gary Keough and William Pratt

INTRODUCTION

The criteria investigated in this paper are aimed at answering the question: Which of two forecast or estimation methods is better? This question arises when a new survey replaces an old survey, when a new survey procedure replaces an old procedure, when an estimate or forecast from a model replaces a subjective evaluation of survey data, or when a ratio estimator based upon check data replaces a more frequent periodic survey estimate. Examples of this last practice can be found in the published midmonth livestock price and in California monthly milk production estimates.

If a judgement between methods is based upon a one time parallel test, then criteria to consider are ease of implementation, cost of procedure, variance of the estimate, and statistical significance between the estimates from each method. However, this evaluation is a one-time test with a sample size of one. The new method may be better and lose, or poorer and win. One time tests may not be extremely conclusive. If on the other hand, we could parallel test or simulate the results of a parallel test over a longer period of time, we could produce a stronger statistical statement about whether a newer or older method of estimation or forecasting is better.

Many criteria have been suggested for evaluation, such as mean error, mean absolute error, mean absolute percent error, and mean square error. Each of these statistics provides a description of how a procedure performs, but none of these statistics is truly a testing comparison. After calculating the statistic, the reviewer or analyst is still left with a decision as to which procedure is better. In this paper we attempt to suggest three statistics that incorporate strengths from each of the above. In addition, the suggested statistics are defined so that they are mean invariant; therefore, the size of the estimate being analyzed has little effect on the value of the statistic. Probability tables are provided in appendix A for the new statistics and discussed in the results section. The new statistics are compared with three of the standard comparison statistics: mean absolute error, mean absolute percent error, and mean square error. Next, the standard and new statistics are employed in joint tests to see if one joint test exists that is universally best or generally better. A strategy is proposed for using these statistics, so that when the value of a comparison statistic is calculated type 1 and type 2 errors are controlled. Finally, the recommended procedures are applied in two examples. The first example compares remote sensing estimates and June Enumerative Survey Estimates of planted acreage for six crops. This set of data suffers from small sample sizes. The second example has a large sample size. In this example, the preliminary manufactured milk price estimates for the Minnesota/Wisconsin milk price series are compared to forecasts from Box/Jenkins time series transfer function models.

METHODOLOGY

Three new evaluation statistics are proposed in this section. The first statistic is a modification of Theil's U statistic ( ). Theil's U statistic is given as follows:

$$\left(\left(\sum_{i=1}^{n-1} ((N_i - F_i)/F_{i-1})^2 \right) / \left(\sum_{i=1}^{n-1} ((F_{i-1} - F_i)/F_{i-1})^2 \right) \right)^{1/2}$$

where

n = sample size for evaluation
N_i = New forecast method
F_{i-1} = Naive forecast method
F_i = Final estimate.

Theil suggested this criteria as an evaluation technique of a forecast procedure. He recommended that the new forecast procedure be compared against a naive forecast procedure where the series forecast equals the previous value of the series. He created a statistic to suggest that one forecast was better than some other forecast procedure, even if performance was judged against only a naive forecast. Theil's statistic was a step away from such descriptive criteria as mean deviation, mean absolute percent error, or root mean square error. The evaluator at least had a belief that if he failed Theil's criteria his new method was no better than simply assuming the previous value of the series. If Theil's statistic produces a value of 1, the naive method and the new method perform equally well.If the statistic produces a value less than one then the new method performs better than a naive forecast. If the statistic produces a value larger than one then the new method performs worse than the naive forecast.

In keeping with this spirit, Theil's statistic was modified. We observed that F_{i-1} actually represented the old forecasting method. If we substituted O_i, the value of a current or old forecasting method, for each value of F_{i-1} we could produce a new evaluation statistic comparing a new forecasting method to an old method given by:

$$((\sum_{i=1}^{n-k} ((N_i - F_i)/O_i)^2) \;/\; (\sum_{i=1}^{n-k} ((O_i - F_i)/O_i)^2))^{1/2}$$

where

$n-k$ = sample size for evaluation
N_i = New estimation method
O_i = Old estimation method
F_i = Final estimate.

The summation has changed in the new statistic to be over n-k values. In the new statistic, the evaluation sample size is equal to n-k; where n is the total number of observations and k equals the larger value between the two sample sizes required to make the first forecast or estimate for the old or new procedure. In Theil's statistic, k equals 1 because one sample value is required to produce the first forecast of the naive statistic.

This new evaluation statistic appears to have some excellent properties. The numerators in both ratios of the statistic are the squared errors for the new and old methods. The denominators are the value of the old statistic squared. Thus, each ratio produces a relative squared error. The value of a relative error is such that if our estimate is 1, 1,000, or 1,000,000 a relative error of 10% is the same throughout. As before a value for the modified Theil's statistic of 1 would indicate both methods are the same, a value less than 1 would suggest that the new method is better, and a value greater than 1 would suggest that the new method is worse.

Phil Kott suggested a further change. If we define relative error as:

$$RE_i = (O_i - F_i)/F_i$$

then we can rewrite the denominator of the statistic as

$$\frac{O_i - F_i}{O_i} = \frac{O_i - F_i}{F_i \{ 1 + [(O_i - F_i)/F_i] \}}$$

$$= RE_i /(1 + RE_i).$$

From this result, we see that positive and negative relative errors are treated differently. Thus the value of the old method in the denominator of each ratio was replaced by the value of the final estimate. Lastly, the statistic is designed to compare estimates or forecasts calculated for the same periods of data. We can generalize this assumption by allowing the numerator and denominator to be applied to different periods of data. This final modification is accomplished by dividing both the numerator and denominator by their respective sample sizes. We will not explore this final modification in this paper but we did want to note how the statistic could be adjusted. The first evaluation statistic (NES-1) becomes:

$$(((\sum_{i=1}^{n} ((N_i - F_i)/F_i)^2)) /((\sum_{i=1}^{n} ((O_i - F_i)/F_i)^2)))^{1/2} \qquad (1)$$

where

n = sample size for evaluation
N_i = New estimation method
O_i = Old estimation method
F_i = Final estimate.

The above statistic was derived in the spirit of Theil's statistic and squared error loss. These criteria are great for long run performance, but decision theory would suggest another avenue of consideration of equal importance, minimax. In a minimax decision we are not worried about what will happen over the long run, but what is the worst thing that could happen. Therefore, our goal is to minimize our maximum loss. Or in this case we would like to find a new method whose maximum error is never worse than the maximum error of the old method. We represent the problem graphically by examining some percentage estimates with the old method (--), the new method (_._.), and the true value (___) drawn respectively.

Figure 1: Time series graph of possible values from an old estimation method, a new estimation method, and the final true value

We can see that the new method is almost always better than the old. The new method would have a smaller root mean square error and possibly a smaller value for statistic one. However, the one extremely large error in our estimate or forecast of the final value from the new method could be disastrous enough to injure our credibility. To prevent such an error, a second evaluation statistic (NES-2) is suggested:

$$(\text{Max} \left|(N_i - F_i)/ F_i\right|) / (\text{Max} \left|(O_j - F_j)/ F_j\right|) \qquad (2)$$

where

n = sample size for evaluation
N_i= New estimation method
O_j= Old estimation method
F_i= Final estimate.

Absolute values are calculated on the relative errors before the maximum errors are determined. We assumed that either an over estimate or and under estimate were equally bad.

The last evaluation statistic is in the spirit of the second statistic and is even more critical of an extreme forecast error. In this statistic, the maximum absolute relative error of the new method is divided by the average absolute relative error of the old. This statistic is the most conservative of the three evaluation statistics. Average values of this statistic are not expected to be equal to 1, but larger than 1. Should the value of the statistic be 1 or less you can be extremely certain that the new method is superior to the old method.
The third evaluation statistic (NES-3) is:

$$n (\text{Max} \left|(N_i - F_i)/F_i\right|) / (\sum_{j=1}^{n} \left|(O_j - F_j)/F_j\right|) \qquad (3)$$

where

n = sample size for evaluation
N_i= New estimation method
O_i= Old estimation method
F_i= Final estimate.

ANALYSIS PROCEDURES

Simulation techniques were used to develop historical data series for evaluation. The simulation procedure consisted of randomly generating a final estimate and two unbiased preliminary estimates, an old and a new, for each time period. Errors in the estimates for the new method were generated for

15 different error conditions with values from 50% smaller to 50 % larger than the old method. For each error condition 500 different replicates were created consisting of 12, 36, or 60 time periods. The above process was repeated for 3 different correlations between the errors for the new and old method: 0.0, 0.6, and 0.95.

Correlation between errors and not the estimates were considered because if the errors between the two methods were perfectly correlated then one error could be expressed as a linear function of the other and an exact conclusion could be drawn as to which method is better. We can rotate and translate error pairs to the positive quadrant of the x-y axis by taking absolute values and adding constants to produce a zero intercept. If the new method is expressed as a linear relationship of the old method, then the value of the slope of the linear equation indicates which method has the smaller translated errors. The translated errors from the new method are smaller for slopes less than one, equal for slopes equal to one, and larger for slopes greater than one. Since constants were used to translate the data to a zero intercept, those same constants can be added to the data to produce the final result as to which method is better.

Figure 2: Graph of the relationship between errors from two method under perfect correlation

Finally, the errors were generated assuming a uniform distribution or a normal distribution. A total of 4,860,000 periods of data were generated for this analysis using SAS on a Compaq 386. If we considered these data periods to be months, then 4,050 centuries of data were produced.

A description of the random number generators used follows. The final estimate for each period was generated from a normal distribution with a mean of 10 and a standard deviation of 1. The random numbers used to generate the errors in the estimates for the old method and new method were produced from a continuous uniform (0,1) distribution or from a normal (0,1) distribution. Errors were first generated for the old method from the appropriate marginal distribution. Then errors for the new method were produced from the appropriate conditional distribution.

For each replicate of 12, 36, or 60 time periods, the value of each of the three new evaluation statistics presented in the methods section was calculated. In addition, values for three standard evaluation statistics: mean absolute error(4), mean absolute percent error(5), and mean square error(6), were calculated. The formula for these standard statistics follow:

$$\text{mean absolute error} = \left(\sum_{i=1}^{n} |(E_i - F_i)|/n \right) \qquad (4)$$

$$\text{mean absolute percent error} = \left(\left(\sum_{i=1}^{n} |(E_i - F_i)/F_i|\right)/n\right)*100 \qquad (5)$$

$$\text{mean square error} = \sum_{i=1}^{n} (E_i - F_i)^2)/n \qquad (6)$$

where

n = sample size for evaluation
E_i= Old or New estimation method
F_i= Final estimate.

The standard statistics and the new evaluation statistics are compared in the results section.

RESULTS

Let us begin by considering the probability tables in the appendix. These tables contain the probability that the new method is better than the old for specified ranges of values of the new evaluation statistic and a given correlation condition. From tables A1, A2, A4, and A5, we see that when the correlation between the errors of the two estimation or forecast methods is high (0.95) then NES-1 and NES-2 produce the same result. We would accept, with a probability equal to one, that the new method is better than the old for any value of either new evaluation statistic less than or equal to one. NES-3, tables A3 and A6, does not demonstrate such a clear cut result when the correlation is high(0.95). In fact NES-3 produces a gradual decrease in probability as the value of the statistic increases. For a given value of new evaluation statistics (1), (2), or (3), the probability that the new method is better when the errors are not correlated (r=0.0) is generally higher than when there is some correlation (r=0.6).

Further probabilities that the new method is better than the old method for NES-1 and NES-2 seem to decrease more rapidly than for NES-3 as the value of the statistic increases. Finally, the assumed error distributions do seem to influence the probabilities in tables B1 to B6. In general, the probabilities corresponding to uniformly distributed errors decrease more rapidly than the normally distributed errors as the value of the new evaluation statistics increase.

Since we now have probability tables, we could calculate the values for these new evaluation statistics to select between the two methods. However, we could also use much simpler standard criteria. Tables 1, 2, and 3 compare the new statistics with different critical values to their counterpart standard criteria in terms of type 1 and type 2 errors.

The decision table below illustrates what we mean by type 1 and type 2 errors. We can classify any decision of whether a new method is better than an old method in four ways.

Figure 3: Standard probability decision table when applying a single test criteria.

		True condition	
		Better	Worse
Decision From Test	Better	Correct Decision	Type 2 Error
	Worse	Type 1 Error	Correct Decision

As can be seen in the table, when we apply a test criteria two good decisions and two bad decisions can be made. The good decisions are to say a method is better when it is better or to say a method is worse when it is worse. We make errors when we use the test and say a method is worse when in fact the method is better (Type 1 error). We also make an error when we say a method is better when in fact the method is worse (Type 2 error). The probability of type 1 and type 2 errors are very important in how we control selection of a new method.

Most of the time when we are attempting to replace an existing method with a new method, the existing method has been adequate. We are attempting to implement a new method that may be more efficient, cheaper, less time consuming, reduce respondent burden, or interfaces better with the survey program. We generally have not had problems with the forecast or estimate that has created public demand or congressional pressure for action. In view of this situation, how should we operate in selecting a new method? We believe that we would want to guard against committing a type 2 error. We would not want to replace an existing procedure with a new procedure that is worse. We should be more willing to reject a new procedure that is better so as to prevent adopting a new procedure that is worse.

In tables 1, 2, and 3 three different critical values were used for each new evaluation statistic. The statistical test employed on standard criteria used the comparison, is the value of the standard statistic for the new method less than or equal to the value of the standard statistic for the old method. If this question was answered yes, then the new method was assumed better. If the question was answered no, then the new method was assumed worse. The comparison statistics in the three tables are as follows: in table 1 , NES-1 is compared to mean square error; in table 2, NES-2 is compared to mean absolute percent error; and in table (3), NES-3 is compared to mean absolute error. Each pair of probabilities for type 1 and type 2 error in the table represents the outcomes of 7,500 samples.

Let us first consider the results for NES-1 in table 1. For all sample sizes for both uniform and normally distributed errors with no, moderate, and strong correlation, the probability of type 1 and type 2 errors using a critical value of 1 almost matches exactly with the probability of type 1 and

Table 1: Probability of type 1 and type 2 errors for three critical values of statistic one, (NES-1), compared with mean square error for three sample sizes: 12, 36, and 60; three correlations between estimation methods: r=0.0, r=0.6, and r=0.95; and two types of error distributions.

Sample Size	Value of Statistic	Errors Uniformly Distributed						Errors Normally Distributed					
		r = 0.0		r = 0.6		r = 0.95		r = 0.0		r = 0.6		r = 0.95	
		Type 1 error	Type 2 error	Type 1 error	Type 2 error	Type 1 error	Type 2 error	Type 1 error	Type 2 error	Type 1 error	Type 2 error	Type 1 error	Type 2 error
n=12	<=1.00	.195	.182	.056	.234	.000	.000	.239	.259	.076	.345	.001	.000
	<=0.95	.257	.131	.103	.143	.248	.000	.288	.207	.113	.255	.287	.000
	<=0.90	.329	.091	.172	.080	.424	.000	.338	.173	.166	.179	.514	.000
	MSEn<=MSEo	.192	.179	.056	.234	.000	.000	.238	.258	.075	.340	.001	.000

Table 1: Probability of type 1 and type 2 errors for three critical values of statistic one, (NES-1), cont:d: compared with mean square error for three sample sizes: 12, 36, and 60; three correlations

between estimation methods: r=0.0, r=0.6, and r=0.95; and two types of error distributions.

Sample Size	Value of Statistic	Errors Uniformly Distributed						Errors Normally Distributed					
		r = 0.0		r = 0.6		r = 0.95		r = 0.0		r = 0.6		r = 0.95	
		Type 1 error	Type 2 error	Type 1 error	Type 2 error	Type 1 error	Type 2 error	Type 1 error	Type 2 error	Type 1 error	Type 2 error	Type 1 error	Type 2 error
n=36	<=1.00	.137	.089	.014	.212	.000	.000	.178	.167	.020	.331	.000	.000
	<=0.95	.215	.039	.049	.091	.248	.000	.239	.108	.047	.222	.311	.000
	<=0.90	.308	.015	.128	.023	.415	.000	.316	.067	.095	.124	.563	.000
	MSEn<=MSEo	.134	.086	.014	.213	.000	.000	.176	.162	.020	.331	.000	.000
n=60	<=1.00	.116	.067	.005	.208	.000	.000	.149	.123	.008	.290	.000	.000
	<=0.95	.202	.025	.034	.084	.144	.000	.226	.071	.026	.181	.301	.000
	<=0.90	.312	.006	.112	.017	.402	.000	.314	.036	.070	.082	.568	.000
	MSEn<=MSEo	.110	.064	.005	.209	.000	.000	.149	.121	.008	.292	.000	.000

type 2 errors from the standard criteria of mean square error. As the critical value for statistic (1) becomes smaller the probability of a type 1 error increases and the probability of a type 2 error decreases. For NES-1 and a given critical value we see a gradual reduction in the probability of a type 1 error with an increase in sample size; however we see a much greater reduction in type 2 error with an increase in sample size. This behavior is what we would expect from a reliable procedure. If we only want to calculate one of these two statistics to make a decision and if our goal is to guard against a type 2 error at the sacrifice of a larger type 1 error; then we should adopt a preliminary test strategy. By preliminary test, we mean that we will first calculate the value of the correlation (r) between the errors for the two methods and use the result of that calculation to select the test procedure to apply to our data.

Step 1: Calculate the correlation (r) between the errors for the two methods.

Step 2: Select evaluation criteria based upon r.

a) if r > or = 0.9 use MSE inequality or NES-1 with a critical value of 1,

b) if r < 0.9 and type 2 error is of moderate concern use NES-1 with a critical value of 0.95,

c) if r < 0.9 and type 2 error is of serious concern use NES-1 with a critical value of 0.90.

Step 3: Apply evaluation criteria to data and make a decision.

The preliminary test procedure provides one possible solution to our decision problem. However, we have two other alternative statistics to consider. Let us now consider our gains for the criteria presented in table 2.

In table 2, we compare NES-2 to mean absolute percent error. When the correlation between errors is high then both statistics perform about the same. When correlation is equal to zero and the errors are uniformly distributed then NES-2 with a critical value as high as 1 performs slightly better than the mean absolute percent error. Under either error distribution with no correlation between errors a critical value of 0.9 for statistic (2) can be specified to

Table 2: Probability of type 1 and type 2 errors for three values of statistic two, (NES-2), compared with mean absolute percent error for three sample sizes: 12, 36, and 60; three correlations between estimation methods: r=0.0, r=0.6, and r=0.95; and two types of error distributions.

Sample Size	Value of Statistic Two	Errors Uniformly Distributed						Errors Normally Distributed					
		r = 0.0		r = 0.6		r = 0.95		r = 0.0		r = 0.6		r = 0.95	
		Type 1 error	Type 2 error	Type 1 error	Type 2 error	Type 1 error	Type 2 error	Type 1 error	Type 2 error	Type 1 error	Type 2 error	Type 1 error	Type 2 error
n=12	<=1.00	.178	.144	.027	.464	.000	.000	.271	.366	.044	.637	.001	.000
	<=0.95	.236	.099	.037	.361	.426	.000	.312	.262	.065	.561	.520	.000
	<=0.90	.313	.066	.064	.261	.627	.000	.354	.219	.088	.487	.718	.000
	<=0.80	.492	.021	.197	.096	.856	.000	.464	.140	.164	.332	.902	.000
	MAPEn<=MAPEo	.226	.217	.172	.119	.000	.000	.248	.271	.195	.155	.005	.000
n=36	<=1.00	.138	.101	.007	.453	.000	.000	.219	.241	.011	.727	.000	.000
	<=0.95	.211	.049	.021	.339	.453	.000	.271	.187	.022	.656	.568	.000
	<=0.90	.304	.020	.051	.223	.689	.000	.326	.145	.037	.580	.770	.000
	<=0.80	.506	.002	.201	.056	.909	.000	.482	.038	.148	.210	.923	.000
	MAPEn<=MAPEo	.156	.119	.114	.049	.000	.000	.187	.179	.132	.086	.000	.000
n=60	<=1.00	.125	.085	.004	.422	.000	.000	.219	.221	.005	.778	.000	.000
	<=0.95	.215	.040	.016	.306	.452	.000	.274	.163	.010	.706	.580	.000
	<=0.90	.313	.017	.052	.197	.694	.000	.332	.119	.017	.617	.805	.000
	<=0.80	.503	.002	.207	.038	.935	.000	.473	.052	.061	.203	.961	.000
	MAPEn<=MAPEo	.136	.096	.097	.041	.000	.000	.156	.128	.112	.056	.000	.000

produce a type 2 error smaller than that for mean absolute percent error. Whenever the errors are moderately correlated, statistic (2) does not perform well at all.

Table 3 compares NES-3 to mean absolute error. NES-3 produced some of the smallest type 2 error probabilities for any evaluation criteria. However, these small type 2 error probabilities were only accomplished at the extreme expense of very large type 1 error probabilities sometimes greater than 0.5. Further the critical value of the statistic changes with the two different error distributions. The critical values under uniformly distributed errors were 1.5, 1.75, and 2.0. The critical values under the normal distribution were 2.0, 2.5, and 3.0. Therefore critical test values under a uniform

Table 3: Probability of type 1 and type 2 errors for three values of statistic three, (NES-3), compared with mean absolute error for three sample sizes: 12, 36, and 60; three correlations between estimation methods: r=0.0, r=0.6, and r=0.95; and two types of error distributions.

Sample Size	n = 12	Errors Uniformly Distributed						Errors Normally Distributed					
	Value of Statistic Three	r = 0.0		r = 0.6		r = 0.95		r = 0.0		r = 0.6		r = 0.95	
		Type 1 error	Type 2 error	Type 1 error	Type 2 error	Type 1 error	Type 2 error	Type 1 error	Type 2 error	Type 1 error	Type 2 error	Type 1 error	Type 2 error
n=12	<= C1	.170	.235	.031	.543	.223	.294	.124	.501	.006	.854	.117	.672
	<= C2	.320	.095	.101	.307	.483	.081	.175	.403	.013	.770	.201	.514
	<= C3	.518	.025	.258	.102	.789	.008	.247	.315	.034	.664	.322	.344
	MAEn<=MAEo	.227	.219	.168	.118	.000	.000	.244	.273	.194	.154	.000	.000
n=36	<= C1	.226	.051	.037	.335	.411	.074	.213	.202	.008	.737	.378	.294
	<= C2	.496	.004	.156	.100	.785	.004	.309	.124	.024	.601	.543	.114
	<= C3	.662	.000	.406	.008	.971	.000	.421	.063	.055	.434	.757	.046
	MAEn<=MAEo	.158	.116	.112	.051	.000	.000	.188	.176	.131	.083	.000	.000
n=60	<= C1	.277	.020	.047	.235	.517	.027	.300	.097	.007	.652	.554	.125
	<= C2	.522	.001	.212	.042	.895	.000	.406	.043	.027	.485	.738	.037
	<= C3	.699	.000	.485	.001	.995	.000	.535	.012	.066	.295	.889	.001
	MAEn<=MAEo	.138	.095	.097	.021	.000	.000	.157	.127	.112	.056	.000	.000

Note: C1, C2, and C3 are different critical values depending on the error distribution.
Critical values for uniform errors were 2.0, 1.75, and 1.5.
Critical values for normal errors were 3.0, 2.5, and 2.0.

error distribution will not work well for normally distributed errors. The reason for this is that the statistic behaves like a normalized or scaled value. The average of the absolute values in the denominator functions similarly to the standard deviation in a normalized value. The maximum deviation in the numerator functions like a large positive deviation from the mean in a normalized value. Finally NES-3 behaves very poorly for moderate and large correlations. Therefore, NES-3 will not be evaluated further.

In tables 1, 2, and 3, we compared the new statistics to standard criteria using Type 1 and Type 2 errors. We attempted to find a best single criteria. We discovered that NES-1 and mean square error possessed the best type 1 and type 2 errors under no correlation or under very strong correlation. Further, we found that NES-2 and mean absolute error or mean absolute percent error performed best under moderate correlation. However, none of the criteria provided small type 1 and type 2 errors across all conditions. Therefore, we must now consider joint tests as illustrated for a two way test in Figure 3.

Figure 3: Standard probability decision table when applying a joint test criteria for two tests.

		True condition			
		Better		Worse	
		Decision Test 1		Decision Test 1	
		Better	Worse	Better	Worse
Decision From Test 2	Better	Correct Decision All Test	Type 1 Error Test 1 No Error Test 2 Type 1 Er. Joint Test		
	Worse	No Error Test 1 Type 1 Error Test 2 Type 1 Er. Joint Test	Type 1 Error Test 1 Type 1 Error Test 2 Type 1 Er. Joint Test		
	Better			Type 2 Error Test 1 Type 2 Error Test 2 Type 1 Er. Joint Test	No Error Test 1 Type 2 Error Test 2 No Error Joint Test
	Worse			Type 2 Error Test 1 No Error Test 2 No Error Joint Test	Correct Decision All Tests

Therefore, we will attempt to discover if a joint test, employing several criteria simultaneously, exists where the type 1 and type 2 errors are reasonable across all error correlation conditions.

In a joint test, if any of the test criteria would lead us to reject, we will reject the new method as better. We will only accept the new method as better if all tests say accept. The decisions made from the two different test statistics are identical for those decisions represented by boxes on the diagonal in the above table. Therefore the only differences between the decisions in the two tests occur for the off diagonal elements. In the joint test we see that the Type 1 error will likely increase unless both tests have no values in the off diagonal boxes. However, we can also see that the Type 2 error will likely decrease unless both tests have no values in the off diagonal boxes when the true condition is that the new test is worse. Similar decision tables can be generated for joint tests with three or four criteria.

Table 4 on the next page presents the results of all two way, three way, and four way tests that possess type 1 and type 2 errors within specified limits for seven criteria:

1. mean absolute error,
2. mean absolute percent error,
3. mean square error,
4-1. NES-2 with a critical value of 1,
4-2. NES-2 with a critical value of 0.9,
5-1. NES-1 with a critical value of 1, and
5-2. NES-1 with a critical value of 0.9.

For example, for a sample size equal to 12 and a correlation between errors equal to 0.0, we find the entry 4 in the row labeled 0.25 and the column headed 0.33. We would interpret this entry in the following way. There are four joint tests with the probability of a type 1 error in the interval, $0.2 < t_1 < 0.25$, and with the probability of a type 2 error in the interval , $0.25 < \tau_2 < 0.33$.

No joint or individual test was identified as a uniformly best test for all sample sizes and error correlations. A best test is defined as having the smallest type 1 and type 2 errors under all situations. NES-1 with a critical value of 0.9 performed well in most cases; however, the type 1 error was often large even with increased sample sizes.

Table 4 again emphasizes that few if any decisions can be made from a small sample size. In fact, about the only time a small sample size tells us much is if the correlation between errors is quite high. Unfortunately, estimating a correlation coefficient from a sample size of less than 36 observations can be very unreliable. Therefore, the best we can do for a sample of less than 36 observations is to apply NES-1 with a critical value no larger than 0.9.

Table 4: Count of the number of joint tests by upper bounds for the probability of type 1 and type 2 errors for two error correlation conditions

Sample size	Type 1 Error is less than	Correlation between errors from the two methods												Type 1 Error is less than
		(r = 0.0)						(r = 0.6)						
		Type 2 Error is less than						Type 2 Error is less than						
		0.05	0.10	0.15	0.20	0.25	0.33	0.05	0.10	0.15	0.20	0.25	0.33	
n=12	0.15												4	0.10
	0.20												1	0.15
	0.25						4				13			0.20
	0.33				2	12	2			27				0.25
n=36	0.15												6	0.05
	0.20			1	5								1	0.10
	0.25			10		1			32	2				0.15
	0.33		16	4					4	2				0.20
n=60	0.15			2									5	0.05
	0.20		5	9									1	0.10
	0.25					1		9	31					0.15
	0.33	7	12	1										0.20

Note: Probabilities are read from the left of the table for r=0.0 and from the right of the table for r=0.6.

With sample sizes restricted to 36 observations or more, the results of the simulation would suggest several possible tests that produce reasonable probabilities for type 1 and type 2 errors. However, we will still need to employ a preliminary test on the correlation between the forecast or estimation errors from the two methods.

Table 5 on the next page presents six tests, three individual and three joint tests, that perform well under at least one correlation condition. The tests are presented in pairs based upon their performance under little or no correlation. Reductions in type 2 error are often sacrificed by losses in type 1 error. If the correlation between errors is large, then mean absolute error or mean square error can be used to make the decision between methods. If correlations are moderate (r=0.6), then a joint test using MAE and NES-1 with a critical value of 0.9 is best. For small correlations (r=0.0), the user would select a test from table 5 with type 1 and type 2 error that are tolerable.

Table 5: Probability of type 1 and type 2 errors for the six best joint test from an analysis of all possible single, two-way, three-way, and four-way test from seven criteria.

Test Statistic	Sample Size = 36				Sample Size = 60			
	r = 0.0		r = 0.6		r = 0.0		r = 0.6	
	Type 1 error	Type 2 error	Type 1 error	Type 2 error	Type 1 error	Type 2 error	Type 1 error	Type 2 error
Mean Absolute Er.	.188	.176	.131	.083	.157	.127	.112	.056
Mean Square Error	.176	.162	.020	.331	.149	.121	.008	.292
MAE & NES-1 at 1.0	.214	.135	.131	.082	.180	.102	.112	.056
MAE, MAPE, MSE,& NES-1 at 1.0	.220	.127	.139	.079	.189	.095	.117	.052
NES-1 at 0.9	.316	.067	.095	.124	.313	.037	.069	.093
MAE & NES-1 at 0.9	.319	.066	.141	.074	.314	.036	.117	.048

EXAMPLE 1

From 1980 to 1987 the National Agricultural Statistics Service (NASS) implemented a pilot remote sensing project. One of the purposes of the project was to estimate planted crop acreages for six crops in selected states. These estimates were often compared to estimates from the June Enumerative Survey, a major probability acreage survey conducted in June. One of the most recent comparisons is a summary report by Allen and Hanuschak (1988) that discusses the methodology as well as the results of the project. Readers interested in more detail about the project or the remote sensing acreage estimator should refer to this report. In this example we will make only a few very short statements about the remote sensing estimator for crop acreage.

The remote sensing estimator currently employed by NASS uses information from the June Enumerative Survey (JES). The exact location of fields of different kinds of crops identified in the JES land area segments are determine on the remote sensing imagery. This ground truth data is used to teach the classification procedure. Results from this classification procedure of the total land area are used in regression type estimators to produce individual crop acreage estimates.

With improved higher resolution sensors, there exists the possibility of improved crop acreage estimates from a much smaller and cheaper ground truth data collection effort. Further research would also suggest that production estimates may be feasible some day. With these considerations in mind, we could ask ourselves how successful was the pilot project in estimating plant crop acreages. Each early season estimate of planted crop acres was compared to the final revised end of season estimate. Table 6 contains the sample size, correlation coefficient, the value of NES-1, and the test conclusion.

Table 6: Comparison of planted acreage estimates from Remote Sensing and the June Enumerative Survey for six crops in selected States (1980-1987).

Crop	Sample Size	Correlation Coefficient	Value of Statistic 1	Conclusion
Corn	22	0.21	1.27	Worse
Soybeans	28	0.22	0.51	Better
Rice	11	0.67	0.96	Worse
Cotton	10	0.67	0.83	Better
Sorghum	9	0.43	1.32	Worse
Wntr Wheat	24	0.59	1.86	Worse

Sample sizes are less than 36 in all cases. We would ignore the correlation preliminary test and apply NES-1 with a critical value of 0.9. we would conclude that cotton and soybean planted crop acreage estimates are improved by using remote sensing acreage estimates. We would also note that the only other crop with a value of NES-1 less than 1 is rice. For these three crops some form of administrative data exists to eventually establish an approximation to final production. Indirectly then, if you know the final production, yield, and abandonment, you then know planted acres in a crop.

For the other three crops the remote sensing estimate does not seem to improve the planted acreage estimate. For these three crops no administrative data exists that will approximately establish production. Other researchers have explored possible reasons for poor performance in these crops. Possible problems have been traced to the classifier as well as the estimator.

Prior to the analysis in this example, NASS proposed a new pilot study in remote sensing acreage estimation. This study would use information from sensors with improved resolution. Since cotton and rice are two important specialty crops, the pilot will be conducted in several delta states. Soybeans would be included in the crop acreage estimates since it too is a common delta crop. It is nice to see that a statistical result supports expert judgement.

EXAMPLE 2

The following is quoted and paraphrased from the survey methodology bulletin (Misc. Pub. No. 1308) of the National Agricultural Statistics Service.

NASS provides monthly estimates of the average price received by farmers and the average milkfat test for milk of manufacturing grade in the two-state (Minnesota-Wisconsin) area. This series (referred to as M-W prices) is used to price fluid and surplus Grade A milk each month in Federal milk marketing orders.

Minnesota and Wisconsin produce over half of the manufacturing-grade milk marketed in the United States.....

Data for the monthly M-W price series are collected from a sample of 70 plants from a base sample of 195 plants, using a questionnaire mailed near the close of each month. This inquiry obtains information for the base month and for the first half of the succeeding month. Space is also provided on the inquiry for the plant manager's best estimate of the average fat test and milk price for the last half of the month to which the M-W estimate relates. .... The basic and preliminary data needed for the preparation of the two-State average price and fat test are collected, summarized, and analyzed by the State Statistical Offices in Minnesota and Wisconsin. The data are forwarded to Washington, D. C., for final review and consolidated into a M-W price. The report is issued the 5th of each month, or the last prior working day, from the Wisconsin SSO, and is also published and released by the AMS dairy market news service. A final revised monthly price is obtained from a complete enumeration of all unsurveyed plants at the end of the year in each State.

In this example the M-W price was compared to a forecast from a Box-Jenkins transfer function model that used the final revised two-State price as truth and used the M-W price as a possible input series. The data consisted of monthly price forecasts and estimates from January 1984 to December 1988. The first forecast was made for January 1984, and then one month ahead forecasts were made until December 1988. A total of 60 monthly forecasts were made. The transfer function model was updated or re-estimated quarterly. The final revised two-State price was considered truth. The results are contained in table 7.

Table 7: Comparison of the Monthly M-W price estimates to forecasts from a Box-Jenkins time series transfer function model where the final revised two-State price is considered as truth (January 1984 to December 1988).

Statistic	Monthly M-W Price Est.	Box-Jenkins Transfer Funct
Mean Absolute Error	0.05	0.03
Mean Abs. Percent Er.	0.4	0.3
Mean Square Error	0.05	0.04
Sample Size	60	
Correlation	0.81	
NES-1	0.678	

Since the correlation is in between moderate and large, we might consider applying the joint test in step 6 of the recommendations or applying the individual tests in step 5. Plots of the errors against each other were examined, and no influential observations or nonlinear relationships were detected. Under the joint test in step 6, we would accept the Box-Jenkins transfer function forecasts as better for both the mean absolute error test and for NES-1 with a critical value of 0.9. Our conclusion would be that the Box-Jenkins transfer function forecast improves the procedure. Under the individual tests in step 5 we would again conclude that the Box-Jenkins transfer function model forecasts are better. From probability table A1 and A4 we would also note that the probability that the new method is better, with a value of NES-1 equal to 0.678, is 1.0.

CONCLUSIONS

- The correlation between sample errors does influence the performance of evaluation criteria.

- The best test procedures for sample sizes less than 36 is to apply NES-1 with a critical value of 0.9.

- For samples of size 36 of larger a preliminary test criteria employing the correlation between estimate or forecast errors should be used.

- The estimate or forecast errors for samples size 36 or larger should be analyzed for influential observations or nonlinear relationships.

RECOMMENDATIONS

1. The agency should adopt the following evaluation procedure to determine which of two forecast or estimation procedures is better.

   Step 1 : If the evaluation sample size is larger than 35 go to Step 4.

   Step 2 : Calculate NES-1 from the data.

   Step 3 : Reject the new method as better if the value of NES-1 is larger than 0.9. Go to Step 8.

   Step 4: Calculate the correlation coefficient between the forecast or estimation errors from the two methods. Plot the errors to see if a nonlinear relationship exists. Regress the errors from the new method onto the old method to see if any influential observations exist. If the correlation is valid continue with Step 5, 6, or 7 depending on the value of the correlation.

   Step 5: For large correlations near 1, apply either the mean absolute error or the mean square error criteria. The better method will have a smaller value for either of these statistics. Go to Step 8.

   Step 6: For moderate correlations near 0.5 apply the joint test using mean absolute error and NES-1 with a critical value of 0.9. For the new method to be better, it's mean absolute error must be smaller and NES-1 will need to be less than or equal to 0.9. Go to Step 8.

   Step 7: For low correlations near 0.0 select the test from table 5 with the type 1 and type 2 error you desire.

   Step 8: Advise the appropriate branch of test result and await an implementation decision if results are favorable.

2. If the above procedure is adopted,the simulation should be expanded to define appropriate correlation intervals for test selection.

REFERENCES

1. Makridakis, S., S. Wheelwright, and V. McGee. 1983. Forecasting: Methods and Applications, 2nd Ed. New York: John Wiley & Sons, pp. 31 - 52.

2. Theil, H. 1966. Applied Economic Forecasting. Amsterdam: North-Holland Publishing Co., pp.26-32.

APPENDIX

Table A1: Probability that the new estimation method is as good or better than the old estimation method given that the value of NES-1 is in the interval shown for three sample sizes: 12, 36, and 60; and three correlations between uniformly distributed errors: r=0.0, r=0.6, and r=0.95.

Value of	sample size = 12			sample size = 36			sample size = 60		
Statistic	(r=0.0)	(r=0.6)	(r=0.95)	(r=0.0)	(r=0.6)	(r=0.95)	(r=0.0)	(r=0.6)	(r=0.95)
0.0 < <=0.5	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
0.5 < <=0.7	0.9693	1.0000	1.0000	0.9990	1.0000	1.0000	1.0000	1.0000	1.0000
0.7 < <=0.8	0.9028	0.9694	1.0000	0.9956	0.9978	1.0000	0.9968	1.0000	1.0000
0.8 < <=0.85	0.8005	0.8394	1.0000	0.9748	0.9810	1.0000	0.9951	0.9929	1.0000
0.85< <=0.9	0.6792	0.6888	1.0000	0.9159	0.8650	1.0000	0.9654	0.8993	1.0000
0.9 < <=0.95	0.6698	0.5544	1.0000	0.8158	0.5725	1.0000	0.8681	0.5712	1.0000
0.9 < <=1.0	0.5825	0.3755	1.0000	0.6360	0.2504	1.0000	0.6986	0.2099	1.0000
1.0 < <=1.05	0.5305	0.2270	0.0000	0.4938	0.0693	0.0000	0.4728	0.0373	0.0000
1.05< <=1.1	0.4029	0.1306	0.0000	0.3265	0.0286	0.0000	0.2797	0.0022	0.0000
1.1 < <=1.15	0.2845	0.0675	0.0000	0.1779	0.0165	0.0000	0.1218	0.0000	0.0000
1.15< <=1.2	0.2719	0.0518	0.0000	0.1133	0.0000	0.0000	0.0587	0.0000	0.0000
1.2 < <=1.25	0.2194	0.0345	0.0000	0.0408	0.0036	0.0000	0.0147	0.0000	0.0000
1.25< <=1.3	0.1632	0.0075	0.0000	0.0336	0.0000	0.0000	0.0067	0.0000	0.0000
1.3 < <=1.4	0.0967	0.0103	0.0000	0.0105	0.0000	0.0000	0.0000	0.0000	0.0000
1.4 < <=1.5	0.0686	0.0051	0.0000	0.0110	0.0000	0.0000	0.0000	0.0000	0.0000
1.5 < <=2.0	0.0258	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
> 2.0	0.0145	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

1) The above table was produced from simulation runs where the new method assumed 15 relative error values from 50% better to 50 % worse than the old method.
2) Errors in the estimation method were assumed to be uniform.
3) The table was generated from 22,500 samples of size 12, 36, and 60.

4) Statistic Used

$$\left(\left(\sum_{i=1}^{n} ((N_i - F_i)/F_i)^2 \right) / \left(\sum_{i=1}^{n} ((O_i - F_i)/F_i)^2 \right) \right)^{1/2} \qquad n = \text{sample size for evaluation}$$

where N_i = New estimation method O_i = Old estimation method F_i = Final estimate

Table A2: Probability that the new estimation method is as good or better than the old estimation method given that the value of NES-2 is in the interval shown for three sample sizes: 12, 36, and 60; and three correlations between uniformly distributed errors: r=0.0, r=0.6, and r=0.95.

Value of	sample size = 12			sample size = 36			sample size = 60		
Stat	(r=0.0)	(r=0.6)	(r=0.95)	(r=0.0	(r=0.6)	(r=0.95)	(r=0.0)	(r=0.6)	(r=0.95)
0.0 < <= 0.4	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
0.4 < <= 0.6	0.9947	0.9961	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
0.6 < <= 0.8	0.9470	0.8397	1.0000	0.9942	0.9005	1.0000	0.9947	0.9326	1.0000
0.8 < <= 0.9	0.8185	0.4689	1.0000	0.9287	0.5042	1.0000	0.9419	0.5272	1.0000
0.9 < <= 1.0	0.6731	0.2135	1.0000	0.7007	0.1789	1.0000	0.7593	0.1955	1.0000
1.0 < <= 1.1	0.4409	0.0857	0.0000	0.4185	0.0329	0.0000	0.3845	0.0217	0.0000
1.1 < <= 1.2	0.2331	0.0272	0.0000	0.1429	0.0076	0.0000	0.1278	0.0000	0.0000
1.2 < <= 1.3	0.1037	0.0063	0.0000	0.0538	0.0000	0.0000	0.0384	0.0000	0.0000
1.3 < <= 1.4	0.0821	0.0052	0.0000	0.0145	0.0000	0.0000	0.0062	0.0000	0.0000
1.4 < <= 1.5	0.0500	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1.5 < <= 1.6	0.0177	0.0000	0.0000	0.0081	0.0000	0.0000	0.0000	0.0000	0.0000
1.6 < <=1.75	0.0245	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1.75< <= 2.0	0.0078	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
> 2.0	0.0303	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

1) The above table was produced from simulation runs where the new method assumed 15 relative error values from 50% better to 50 % worse than the old method
2) Errors in the estimation method were assumed to be uniform
3) The table was generated from 22,500 samples of size 12, 36, and 60.

4) Statistic Used

$$(\mathrm{Max}\ |N_i - F_i / F_i|) / (\mathrm{Max}\ |O_j - F_j / F_j|)$$ n = sample size for evaluation

where N_i = New estimation method O_i = Old estimation method F_i = Final estimate

Table A3: Probability that the new estimation method is as good or better than the old estimation method given that the value of NES-3 is in the interval shown for three sample sizes: 12, 36, and 60; and three correlations between uniformly distributed errors : r=0.0, r=0.6, and r=0.95.

Value of	sample size = 12			sample size = 36			0.Sample size = 60		
Statistic	(r=0.0)	(r=0.6)	(r=0.95)	(r=0.0)	(r=0.6)	(r=0.95)	(r=0.0)	(r=0.6)	(r=0.95)
0.0 < < 0.8	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
0.8 < < 1.0	0.9940	0.9983	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
1.0 < < 1.2	0.9859	0.9714	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
1.2 < < 1.4	0.9510	0.8423	0.9821	1.0000	0.9947	1.0000	1.0000	1.0000	1.0000
1.4 < < 1.5	0.8880	0.6913	0.9390	1.0000	0.9473	1.0000	1.0000	0.9952	1.0000
1.5 < < 1.6	0.8128	0.5574	0.8994	0.9968	0.8800	1.0000	0.9959	0.9659	1.0000
1.6 < < 1.7	0.7360	0.4179	0.7920	0.9791	0.7281	0.9778	1.0000	0.8736	1.0000
1.7 < < 1.8	0.6843	0.3445	0.7486	0.9453	0.5566	0.9710	1.0000	0.7005	0.9945
1.8 < < 1.9	0.5940	0.2873	0.5881	0.8657	0.3796	0.9058	0.9591	0.5690	0.9706
1.9 < < 2.0	0.4486	0.1889	0.5164	0.7826	0.2722	0.7809	0.8753	0.3340	0.9061
2.0 < < 2.1	0.4102	0.1496	0.4682	0.6027	0.1588	0.6858	0.7556	0.2179	0.7867
2.1 < < 2.2	0.3481	0.0743	0.3423	0.4882	0.1014	0.5499	0.6102	0.1117	0.6696
2.2 < < 2.3	0.2995	0.0898	0.2965	0.3401	0.0670	0.4423	0.4956	0.0569	0.5021
2.3 < < 2.5	0.2178	0.0456	0.2232	0.2392	0.0259	0.2736	0.2336	0.0282	0.3200
2.5 < < 3.0	0.1156	0.0327	0.1388	0.0595	0.0090	0.0926	0.0724	0.0047	0.0899
> 3.0	0.0334	0.0000	0.0360	0.0074	0.0078	0.0177	0.0010	0.0000	0.0067

1) The above table was produced from simulation runs where the new method assumed 15 relative error values from 50% better to 50 % worse than the old method
2) Errors in the estimation method were assumed to be uniform
3) The table was generated from 22,500 samples of size 12, 36, and 60.

4) Statistic Used

$n \; (\mathrm{Max} \; |N_i - F_i)/F_i \; | \;) \; / \; (\; \Sigma_{j=1} \; ((O_j - F_j)/F_j \;) \;)$ n = sample size for evaluation

where N_i = New estimation method O_i = Old estimation method F_i = Final estimate

Table A5: Probability that the new estimation method is as good or better than the old estimation method given that the value of NES-2 is in the interval shown for three sample sizes: 12, 36, and 60; and three correlations between normally distributed errors : r=0.0, r=0.6, and r=0.95.

Value of	sample size = 12			sample size = 36			sample size = 60		
Statistic	(r=0.0)	(r=0.6)	(r=0.95)	(r=0.0)	(r=0.6)	(r=0.95)	(r=0.0)	(r=0.6)	(r=0.95)
0.0 < <= 0.4	0.9774	0.9984	1.0000	0.9940	1.0000	1.0000	1.0000	1.0000	1.0000
0.4 < <= 0.6	0.8885	0.8747	1.0000	0.9660	0.8991	1.0000	0.9849	0.9088	1.0000
0.6 < <= 0.8	0.7258	0.5504	1.0000	0.8327	0.5061	1.0000	0.8751	0.4942	1.0000
0.8 < <= 0.9	0.6156	0.3605	1.0000	0.6944	0.2404	1.0000	0.7075	0.1878	1.0000
0.9 < <= 1.0	0.5196	0.2640	1.0000	0.5583	0.1664	1.0000	0.5575	0.0768	1.0000
1.0 < <= 1.1	0.4702	0.1660	0.0005	0.4334	0.0671	0.0000	0.4495	0.0391	0.0000
1.1 < <= 1.2	0.4077	0.1264	0.0000	0.3426	0.0445	0.0000	0.3125	0.0166	0.0000
1.2 < <= 1.3	0.3747	0.1059	0.0029	0.2584	0.0150	0.0000	0.2713	0.0278	0.0000
1.3 < <= 1.4	0.3018	0.0833	0.0057	0.2343	0.0122	0.0000	0.1886	0.0000	0.0000
1.4 < <= 1.5	0.2705	0.0421	0.0000	0.1346	0.0000	0.0000	0.1541	0.0000	0.0000
1.5 < <= 1.6	0.2857	0.1311	0.0000	0.1202	0.0000	0.0000	0.0934	0.0000	0.0000
1.6 < <=1.75	0.2518	0.0000	0.0000	0.1181	0.0000	0.0000	0.0485	0.0000	0.0000
1.75< <= 2.0	0.1621	0.0645	0.0000	0.0672	0.0000	0.0000	0.0455	0.0000	0.0000
> 2.0	0.0869	0.0000	0.0000	0.0428	0.0000	0.0000	0.0364	0.0000	0.0000

1) The above table was produced from simulation runs where the new method assumed 15 relative error values from 50% better to 50 % worse than the old method
2) Errors in the estimation method were assumed to be normal
3) The table was generated from 22,500 samples of size 12, 36, and 60.

4) Statistic Used

$(\text{Max} \; |N_i - F_i / F_i|) / (\text{Max} \; |O_j - F_j / F_j|)$ n = sample size for evaluation

where N_i = New estimation method O_i = Old estimation method F_i = Final estimate

Table A6: Probability that the new estimation method is as good or better than the old estimation method given that the value of NES-3 is in the interval shown for three sample sizes: 12, 36, and 60; and three correlations between normally distributed errors : r=0.0, r=0.6, and r=0.95.

Value of	sample size = 12			sample size = 36			sample size = 60		
Statistic	(r=0.0)	(r=0.6)	(r=0.95)	(r=0.0)	(r=0.6)	(r=0.95)	(r=0.0)	(r=0.6)	(r=0.95)
0.0 < < 0.8	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
0.8 < < 1.0	0.9700	0.9976	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
1.0 < < 1.2	0.9545	0.9824	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
1.2 < < 1.4	0.9053	0.9033	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
1.4 < < 1.5	0.8383	0.7735	1.0000	1.0000	0.9860	1.0000	1.0000	1.0000	1.0000
1.5 < < 1.6	0.8151	0.6905	0.9358	0.9861	0.9773	1.0000	1.0000	0.9959	1.0000
1.6 < < 1.7	0.8000	0.6013	0.9283	1.0000	0.9172	1.0000	1.0000	0.9932	1.0000
1.7 < < 1.8	0.7526	0.5461	0.8629	0.9940	0.8614	1.0000	1.0000	0.9465	1.0000
1.8 < < 1.9	0.7629	0.4450	0.8012	0.9730	0.7892	1.0000	1.0000	0.9256	1.0000
1.9 < < 2.0	0.6643	0.3791	0.7420	0.9502	0.6674	0.9853	1.0000	0.8372	1.0000
2.0 < < 2.1	0.6364	0.3568	0.6970	0.9204	0.5678	0.9556	0.9878	0.7277	1.0000
2.1 < < 2.2	0.5377	0.2544	0.6444	0.8933	0.4282	0.9427	0.9718	0.6275	1.0000
2.2 < < 2.3	0.5321	0.2776	0.6128	0.8745	0.3816	0.9209	0.9662	0.5061	0.9873
2.3 < < 2.5	0.4973	0.1904	0.5242	0.7891	0.2805	0.7943	0.9195	0.3912	0.9500
2.5 < < 3.0	0.4311	0.1462	0.4173	0.6265	0.1510	0.6241	0.7648	0.1571	0.7615
> 3.0	0.2217	0.0449	0.2900	0.2335	0.0335	0.3782	0.2749	0.0233	0.4196

1) The above table was produced from simulation runs where the new method assumed 15 relative error values from 50% better to 50 % worse than the old method

2) Errors in the estimation method were assumed to be normal

3) The table was generated from 22,500 samples of size 12, 36, and 60.

4) Statistic Used

$$n \, (\text{Max} \, |N_i - F_i)/F_i \, | \,) \; / \; (\Sigma_{j=1} ((O_j - F_j)/F_j \,) \,)$$ n = sample size for evaluation

where N_i = New estimation method O_i = Old estimation method F_i = Final estimate