Available online at www.sciencedirect.com

SCIENCE DIRECT®

International Journal of Forecasting 21 (2005) 397–409

www.elsevier.com/locate/ijforecast

The M3 competition: Statistical tests of the results

Alex J. Koning[a,1], Philip Hans Franses[a,*], Michèle Hibon[b,2], H.O. Stekler[c,3]

[a]Econometric Institute, Erasmus University Rotterdam, P.O. Box 1738, NL-3000 DR Rotterdam, The Netherlands
[b]INSEAD, Boulevard de Constance, 77305 Fontainebleau, France
[c]Department of Economics, George Washington University, Washington, DC 20052, United States

Abstract

The main conclusions of the M3 competition were derived from the analyses of descriptive statistics with no formal statistical testing. One of the commentaries noted that the results had not been tested for statistical significance. This paper undertakes such an analysis by examining the primary findings of that competition. We introduce a new methodology that has not previously been used to evaluate economic forecasts: multiple comparisons. We use this technique to compare each method against the best and against the mean. We conclude that the accuracy of the various methods does differ significantly, and that some methods are significantly better than others. We confirm that there is no relationship between complexity and accuracy but also show that there is a significant relationship among the various measures of accuracy. Finally, we find that the M3 conclusion that a combination of methods is better than that of the methods being combined was not proven.

Keywords: Forecasting competitions; M3 competition; Multiple comparisons; Analysis of ranks

1. Introduction

There have been many forecasting competitions to determine which forecasting procedures outperform other methods. The latest competition, M3, is based on an analysis of the way 24 methods forecast 3003 time series. The results of this competition have been discussed extensively, but Stekler (2001) suggested that a formal evaluation was required to determine whether there was a statistically significant difference[4] in the forecasting accuracy of these methods. This paper undertakes such an analysis focusing on the rankings of the various methods. In addition to examining the rankings of the various methods and determining which ones differ significantly, we also

[*] Corresponding author. Tel.: +31 10 4081273; fax: +31 10 4069162.

E-mail addresses: koning@few.eur.nl (A.J. Koning), franses@few.eur.nl (P.H. Franses), michele.hibon@insead.edu (M. Hibon), hstekler@gwu.edu (H.O. Stekler).

[1] Tel.: +31 10 4081268; fax: +31 10 4069162.

[2] Tel.: +33 1 60 72 91 18; fax: +33 1 60 74 55 00.

[3] Tel.: +1 202 994 61 50; fax: +1 202 994 61 47.

[4] It should be noted that a statistically significant difference does not imply that the difference is important. That determination can be made only in the context of the decision that will be made using the forecast.

0169-2070/$ - see front matter © 2004 International Institute of Forecasters. Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.ijforecast.2004.10.003

examine the other major conclusions of the M3 competition.

There were four main conclusions of the M3 competition. “(1) Statistically sophisticated or complex methods do not necessarily produce more accurate forecasts than simpler ones. (2) The rankings of the performance of the various methods vary according to the accuracy measure being used. (3) The combination of various methods outperforms, on average, the specific methods being combined and does well in comparison with other methods. (4) The performance of the various methods depends upon the length of the forecasting horizon.” (Makridakis & Hibon, 2000, pp. 458–459). These conclusions were derived from analyses of descriptive statistics with no formal statistical testing.

After presenting the statistical methodology, we first determine whether, in fact, there is a statistical difference in the forecasting accuracy of all of the methods. The test that we use is based on the average rankings of the various methods at each and every horizon. This test is concerned with the null hypothesis that a single ranking does not differ from a random ranking. We then compare rankings at a given horizon H with both the best method and with average performance. A comparison with the best method allows us to determine which methods are significantly worse than the best method. The comparison with average performance enables us to determine which methods were statistically better (worse) than the average forecasting method. This analysis is applied at all horizons for the monthly, quarterly and yearly series to determine whether the relative performance of the forecasting methods is consistent.[5]

Finally, we examine the other conclusions of the Makridakis–Hibon study, i.e., (1) whether there is any relationship between the complexity of the techniques and their accuracy and (2) whether the five descriptive statistics that were used in the M3 competition to measure accuracy yield similar results.

We choose to use nonparametric statistical methodology concerning the reported rankings. Rankings are easy to understand, and in addition, they are distribution-free. In addition, as we will show below, this methodology facilitates the comparison of many methods, as well as the comparison of methods with the best or the worst method.

2. Methodology: Ranking tests

We consider the following situation. There are K methods (k=1, 2, ... K) which have been applied to N time series (n=1,2, ...N) to forecast for H periods (h=1, 2, ...H). For each of these K methods and for each h, we have a ranking in terms of root-mean-squared prediction error (RMSPE) or some other measure like mean absolute percent error (MAPE) averaged over the N time series. It can be of interest to compare rankings across H, and it can also be important to see if a single ranking differs significantly from a random ranking.

Next, we present three test statistics, their asymptotic distributions and two illustrations from the M3 competition.

2.1. Overall test

Let A_{nk} denote the accuracy of method k for time series n, as measured by RMSPE or some other measure. Suppose that for each method k, the average rank $\bar{R}_k$ is the average of the ranks $R_{1k}, R_{2k}, \ldots, R_{Nk}$, where R_{nk} is the rank of A_{nk} among $A_{n1}, A_{n2}, \ldots, A_{nK}$. We shall assume that $A_{n1}, A_{n2}, \ldots, A_{nK}$ in fact have been obtained by monotone transforming (unknown) independent random variables $U_{n1}, U_{n2}, \ldots, U_{nK}$, for each series. That is, $A_{nk}=\varphi_n(U_{nk})$, where φ_n is a strictly increasing function. Observe that R_{nk} coincides with the rank of U_{nk} among $U_{n1}, U_{n2}, \ldots, U_{nK}$. One may think of U_{nk} as the accuracy of method k for time series n, as measured by some latent measure which depends on the time series n.

Next, we assume that the latent measure is special in the sense that there exist continuous cumulative distribution functions $F_1, F_2, \ldots, F_N$ such that

$$P(U_{nk} \le x) = F_n(x - \tau_k) \tag{1}$$

where τ_k is the unknown additive method effect contributed by the kth method; note that F_n only depends on the time series n, and τ_k only depends on

[5] The M3 competition analyzed 24 methods and 3003 series. This paper examines only 22 of these methods because the AAM1 and AAM2 methods did not provide forecasts for yearly series. We report only the results obtained from the yearly, quarterly and monthly series. This corresponds to a total of 2829 series.

the method k. For each time series n, this implies the existence of i.i.d. random variables V_{n1}, V_{n2},…, V_{nK} with common cumulative distribution function F_n such that U_{nk} is equal to $V_{nk}+\tau_k$. Observe that R_{nk} coincides with the rank of $V_{nk}+\tau_k$ among $V_{n1}+\tau_1$, $V_{n2}+\tau_2$,…, $V_{nK}+\tau_n$. If any method effect, τ_j say, is smaller than any other method effect, then U_{nj} will be smaller than any of the other U_{nk} 's on average over all time series, and hence, the accuracy of method j will in general be better than the accuracy of any other method.

The Friedman test, see Friedman (1937, 1939), Wallis (1939), Kendall and Smith (1939) and also Hollander and Wolfe (1999, p. 272), is a well-known nonparametric test of the null hypothesis

$$H_0 : \tau_1 = \tau_2 = \ldots = \tau_{K.} \tag{2}$$

If the null hypothesis holds, then the rankings R_{n1}, R_{n2},…, R_{nK} have been obtained by ranking i.i.d. random variables $V_{n1}+\tau_1$, $V_{n2}+\tau_2$,…, $V_{nK}+\tau_{n.}$ for each time series n. The null hypothesis is tested against the alternative hypothesis

$$H_1 : \tau_1, \tau_2, \ldots, \tau_K \text{ not all equal.}$$

If the alternative hypothesis holds, then the random variables $V_{n1}+\tau_1$, $V_{n2}+\tau_2$,…, $V_{nK}+\tau_n$ are still independent but may differ in location with respect to each other. The Friedman test is based on the test statistic

$$S = \frac{12N}{K(K+1)} \sum_{k=1}^{K} \left(\bar{R}_k - \frac{K+1}{2} \right)^2 \tag{3}$$

where $\bar{R}_k$ is the average rank assigned to method k after applying the methods to N different times series, where we use the notation as in Hollander and Wolfe (1999). Critical values of S are found in Table A.22 in Hollander and Wolfe (1999). Under the null hypothesis (1), S converges in distribution to a Chi-square random variable with K-1 degrees of freedom, as N tends to infinity. This test has already been used in the forecasting literature, see Stekler (1991, 2001), Kolb and Stekler (1996) and De Gooijer and Zerom (2001) among others.

2.2. Multiple comparisons with the best (MCB)

When a null hypothesis is not rejected by a statistical test, it is generally agreed that all hypotheses implied by that hypothesis (its "components") must also be considered as not rejected. In contrast, when the hypothesis does get rejected, the question is which components may also be rejected. Statistical methods that address this issue are called multiple-comparison procedures. To our knowledge, the following methods so far have never been used in the forecasting literature.

McDonald and Thompson (1967, 1972) developed a multiple-comparisons procedure for the null hypothesis (1). Note that in this case, the component hypotheses take the form

$$H_0, k_1k_2 \; : \; \tau_{k_1} = \tau_{k_2} \tag{4}$$

where k_1=1, 2,…, k_2-1 and k_2=1, 2,…, K. Each component hypothesis H_{0,k_1k_2} is rejected if and only if

$$|\bar{R}_{k_1} - \bar{R}_{k_2}| \geq r_{\alpha,K,N}$$

where the critical value $r_{\alpha,K,N}$ is chosen to make the experiment-wise error rate, as defined in Hsu (1996, Definition 2.2.2, p. 27), equal to α. Hence, $r_{\alpha,K,N}$ is the largest constant such that

$$P_{Ho}((\max \bar{R}_k) - (\min \bar{R}_k) \geq r_{\alpha,K,N}) \leq \alpha \tag{5}$$

Note that $r_{\alpha,K,N}$ also depends on N. Values of $Nr_{\alpha,K,N}$ are given in Table A.24 in Hollander and Wolfe (1999). For large N, Hollander and Wolfe (1999, Eq. (7.27)) show that

$$r_{\alpha,K,N} \approx q_{\alpha,K} \sqrt{\frac{K(K+1)}{12N}} \tag{6}$$

where $q_{\alpha,K}$ is the upper α percentile point of the range of K independent standard normal variables. For tables of $q_{\alpha,K}$, see Table 1 in Harter (1960) and Table A.17 in Hollander and Wolfe (1999).

To visualize the results of the multiple-comparison procedures, one can use plots, which are constructed as follows. For each method k, an interval is drawn with length $r_{\alpha,K,N}$ and centered at R_k. If the intervals for method k_1 and k_2 do not overlap, then H_{0,k_1k_2} is rejected. Moreover, a reference line can be drawn at the height of the upper boundary of the interval of the "best" method, which naturally has the lowest upper boundary. This reference line in fact indicates the "unconstrained multiple-comparison procedure with the best, deducted from all pairwise comparisons," as described in paragraph 4.2.1.1 in Hsu (1996). This implies that all methods with confidence intervals above the reference line perform significantly worse than the best method and hence should preferably not be used in the considered situation.

2.3. Illustration of MCB

Consider the rankings in Table 1, which are part of the M3 competition results (Makridakis & Hibon, 2000). Applying the MCB procedure, we note that N=1428 is much too large to be covered by Table A.24 in Hollander and Wolfe (1999). We resort to the asymptotic approximation (5) to determine $r_{\alpha,K,N}$ for α=0.05, K=22 and N=1428. With $K(K+1)/(12N)$=0.0295 and $q_{\alpha,K}$=5.081, we obtain $r_{\alpha,K,N}$=0.873. The average rank belonging to the best forecasting method Theta (at least for this case) is 10.4, and hence, the interval for Theta has endpoints 10.4–0.5*0.87=9.96 and 10.4+0.5*0.87=10.84. Because Theta is the best performing method in the given situation, the upper boundary 10.84 becomes the reference value.

In Fig. 1, the intervals for all 22 methods, as well as this reference value, are displayed. The intervals of the forecasting methods Naive2, Single, Dampen, B-J automatic, Autobox2, Autobox3, RobustTrend, AutomatANN, Flores–Pearce1, Flores–Pearce2, PP-autocast, SmartFcs, Theta-sm and ForcX all are above the reference value without any overlap, and hence, these methods perform significantly worse than the best forecasting method. The explicit trend methods Holt, Winter and Comb S-H-D, the ARIMA/ARARMA methods Autobox1 and ARARMA, the expert systems ForecastPro and RBF do not perform significantly worse than the best forecasting method Theta.

2.4. Multiple comparisons with the mean (ANOM)

Halperin, Greenhouse, Cornfield, and Zalokar (1955) were the first to propose a statistical procedure

Table 1
Comparison of forecast methods using rank intervals and comparison with the best method

Method	Category	$\bar{R}$	Rank interval	Significantly worse than the best
Naive2	Naïve	12.9	(12.463,13.337)	Yes
Single	Naïve	12.6	(12.163,13.037)	Yes
Holt	Explicit trend	11.0	(10.563,11.437)	No
Dampen	Explicit trend	11.7	(11.263,12.137)	Yes
Winter	Explicit trend	11.0	(10.563,11.437)	No
Comb S-H-D	Explicit trend	10.7	(10.263,11.137)	No
B-J automatic	ARIMA/ARARMA	11.7	(11.263,12.137)	Yes
Autobox1	ARIMA/ARARMA	11.2	(10.763,11.637)	No
Autobox2	ARIMA/ARARMA	11.4	(10.963,11.837)	Yes
Autobox3	ARIMA/ARARMA	11.6	(11.163,12.037)	Yes
RobustTrend	Explicit trend	11.6	(11.163,12.037)	Yes
ARARMA	ARIMA/ARARMA	11.1	(10.663,11.537)	No
AutomatANN	Neural network	12.0	(11.563,12.437)	Yes
Flores–Pearce1	Expert system	11.9	(11.463,12.337)	Yes
Flores–Pearce2	Expert system	11.9	(11.463,12.337)	Yes
PP-Autocast	Explicit trend	11.6	(11.163,12.037)	Yes
ForecastPro	Expert system	10.6	(10.163,11.037)	No
SmartFcs	Expert system	11.9	(11.463,12.337)	Yes
Theta-sm	Explicit trend	12.0	(11.563,12.437)	Yes
Theta	Decomposition	10.4	(9.963,10.837)	No
RBF	Expert system	10.7	(10.263,11.137)	No
ForcX	Expert system	11.5	(11.063,11.937)	Yes

Average ranks $\bar{R}$, derived rank intervals and multiple comparisons with the best (MCB) results for 22 forecasting methods in six categories over all 1428 monthly series in the M3 competition, horizon 12.

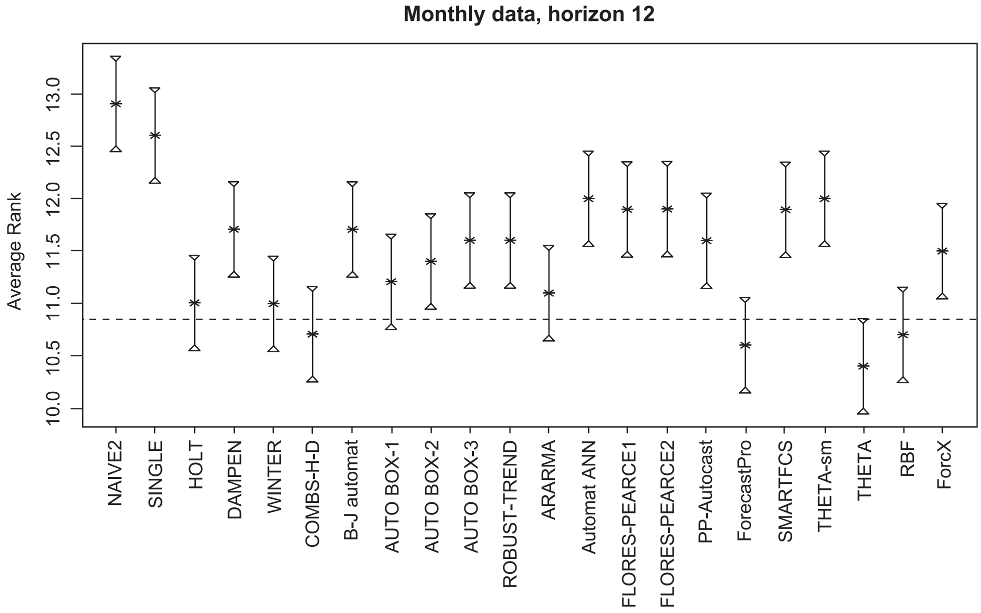

Fig. 1. Average ranks of 22 methods over all 1428 monthly series, horizon 12, MCB intervals.

for multiple comparisons with the mean. During the 1960s, the procedure was reinvented in the area of industrial statistics under the name analysis of means [ANOM], see Ott (1958, 1967) and Ryan (2000, Chapter16).

The component hypotheses take the form

$$H_{0,k_1} : \tau_{k_1} = \bar{\tau} \qquad (7)$$

where k_1=1, 2,…, K, and $\bar{\tau}$ is the average method effect. Let $\bar{\bar{R}}$ denote (K+1)/2, the mean of the average ranks. Each component hypothesis H_{0,k_1} is rejected if and only if

$$|\bar{R}_{k_1} - \bar{\bar{R}}| \geq r'_{\alpha,K,N} \qquad (8)$$

where the critical value $r'_{\alpha,K,N}$ is chosen to make the experiment-wise error rate equal to α. Hence, $r'_{\alpha,KN}$ is the largest constant such that

$$PH_o\left(\max|\bar{R}_{k_1} - \bar{\bar{R}}| \geq r'_{\alpha,K,N}\right) \leq \alpha. \qquad (9)$$

Under the null hypothesis, the statistic $\sqrt{12N/K(K+1)}\max_k|\bar{R}_k - \bar{\bar{R}}|$ converges in distribution to the maximum absolute deviate as $N \to \infty$. The maximum absolute deviate is defined as $\max_k|Z_k - \bar{Z}_k|$ where Z_1, …, Z_K are independent standard normal random variables, and $\bar{Z}_K = K^{-1}\sum_{k=1}^{K} Z_k$. Thus,

$$r'_{\alpha,K,N} \approx H_{\alpha,K}\sqrt{\frac{K(K+1)}{12N}} \qquad (10)$$

where $H_{\alpha,K}$ is the corresponding critical value of the maximum absolute deviate. Tables of $H_{\alpha,K}$ are found in Halperin et al. (1955), see also Nelson (1983), Soong (2001) and Farnum (2004). Alternatively, $H_{\alpha,K}$ may be determined by algorithms for evaluating rectangular multivariate normal probabilities as in Genz (1992), see also Bretz, Genz and Hothorn (2001).

ANOM results are usually visualized by a plot of $\bar{R}_k$ versus k on which three horizontal lines are drawn. The center line is drawn at height R, and the upper and lower control lines run at a distance $r_{\alpha,K,N}$ from the center line. Typically, the plotted points are connected by a vertical line to the centerline. This implies that all methods outside the control lines perform significantly differently than the average.

2.5. Illustration of ANOM

Applying the ANOM procedure, we again use the asymptotic approximation (10) to determine $r'_{\alpha,K,N}$ for

Table 2
Comparison of forecast methods using multiple comparisons with the average ranked method

Method	Category	$\bar{R}$	Significantly worse than average	Significantly better than average
Naive2	Naïve	12.9	Yes	No
Single	Naïve	12.6	Yes	No
Holt	Explicit trend	11.0	No	No
Dampen	Explicit trend	11.7	No	No
Winter	Explicit trend	11.0	No	No
Comb S-H-D	Explicit trend	10.7	No	Yes
B-J automatic	ARIMA/ARARMA	11.7	No	No
Autobox1	ARIMA/ARARMA	11.2	No	No
Autobox2	ARIMA/ARARMA	11.4	No	No
Autobox3	ARIMA/ARARMA	11.6	No	No
RobustTrend	Explicit trend	11.6	No	No
ARARMA	ARIMA/ARARMA	11.1	No	No
AutomatANN	Neural network	12.0	No	No
Flores–Pearce1	Expert system	11.9	No	No
Flores–Pearce2	Expert system	11.9	No	No
PP-Autocast	Explicit trend	11.6	No	No
ForecastPro	Expert system	10.6	No	Yes
SmartFcs	Expert system	11.9	No	No
Theta-sm	Explicit trend	12.0	No	No
Theta	Decomposition	10.4	No	Yes
RBF	Expert system	10.7	No	Yes
ForcX	Expert system	11.5	No	No

Average ranks $\bar{R}$ and multiple comparison with the mean results for 22 forecasting methods in six categories over all 1428 monthly series in the M3 competition, horizon 12.

α=0.05, K=22 and N=1428. It follows that $H_{\alpha,K} \approx 2.973$ and, hence, that $r'_{\alpha,K,N} \approx 0.511$. It also follows that each component hypothesis H_{0,k_1} is rejected if and only if $\bar{R}_{k1}$falls outside the interval with endpoints $\bar{\bar{R}} \pm r'_{\alpha,K,N} \approx 11.5 \pm 0.511$. That is, methods for which $\bar{R}_{k_1}$>11.5+0.511=12.11 are performing significantly

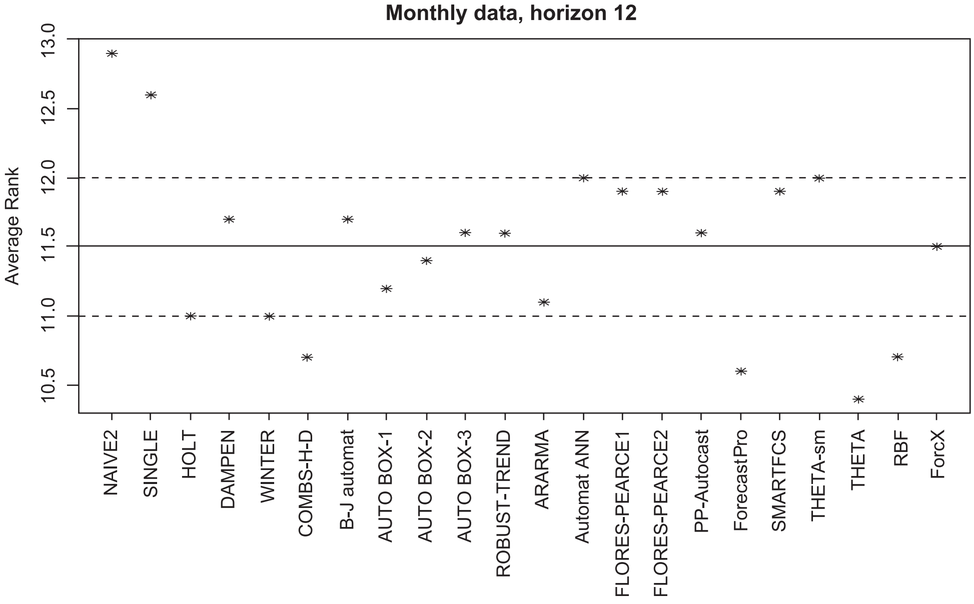

Fig. 2. Average ranks of 22 methods over all 1428 monthly series, horizon 12, analysis of means.

less than average, and methods for which $R_{k_1}<11.5-0.511=10.89$ are performing significantly better than average.

Again, we consider the same 22 methods for monthly data and forecast horizon of $H=12$. The results of the ANOM method appear in Table 2 and Fig. 2. Taking the evidence into account, we can conclude that Naïve2 and Single perform significantly worse than average. The explicit trend method Comb S-H-D, the expert systems ForecastPro and RBF and the decomposition method Theta have average ranks, which exceed the lower boundary 11.5-0.511=10.89 and thus are performing significantly better than average.

3. Results: testing the rankings

In the M3 competition, each of the 22 methods predicted the value of each postsample observation of the 2829 time series. For each method at each horizon, the error (RMSPE) of each individual time series was calculated, and the 22 methods were ranked according to their forecast accuracy. We first determine whether the differences in accuracy among the methods are significantly different. We then compare the methods with the best and with the average performer.

Table 3
Values of Chi-square statistic from testing the rankings of 22 methods

Forecasting horizons	Yearly data	Quarterly data	Monthly data
1	331.93	326.31	521.87
2	291.70	188.43	299.03
3	228.07	134.11	377.94
4	206.96	126.22	377.60
5	209.41	133.75	230.96
6	221.49	159.57	230.29
7		163.51	290.91
8		202.95	310.21
9			323.08
10			278.38
11			351.52
12			281.76
13			246.88
14			366.43
15			360.67
16			398.60
17			326.13
18			302.76

See Eq. (2) in the text for the test statistic.

Table 4
Multiple comparisons with the best method

Method	Category	Monthly	Quarterly	Yearly
Naive2	Naïve	18	8	6
Single	Naïve	18	8	6
Holt	Explicit trend	2	6	4
Dampen	Explicit trend	15	2	0
Winter	Explicit trend	6	6	3
Comb S-H-D	Explicit trend	2	0	0
B-J automatic	ARIMA/ARARMA	17	4	6
Autobox1	ARIMA/ARARMA	12	8	6
Autobox2	ARIMA/ARARMA	15	5	0
Autobox3	ARIMA/ARARMA	17	8	6
Robust–Trend	Explicit trend	18	3	0
ARARMA	ARIMA/ARARMA	14	6	3
AutomatANN	Neural network	16	6	5
Flores–Pearce1	Expert system	18	4	0
Flores–Pearce2	Expert system	16	8	5
PP-Autocast	Explicit trend	18	2	0
ForecastPro	Expert system	0	5	0
SmartFcs	Expert system	18	8	1
Theta-sm	Explicit trend	18	8	6
Theta	Decomposition	0	0	0
RBF	Expert system	7	3	0
ForcX	Expert system	8	3	0

Number of horizons [out of 18 for monthly data, out of 8 for quarterly data and out of 6 for yearly data] for which the estimation method performs significantly worse than the best method for monthly, quarterly and yearly data.

3.1. Are the differences in accuracy statistically significant?

Table 3 presents a selection of the calculated values of the Chi-square statistic obtained from the test statistic (2). In each and every case, the Chi-square statistics are significant beyond the 0.1% level. The null hypothesis of random rankings can be rejected accordingly, and we can conclude that there are significant differences among the methods.

3.2. Are some methods significantly worse than the best?

Since all forecasting methods were not equally accurate, we determine next whether there were any methods that were consistently worse than the best. To undertake this analysis, we use the multiple-comparison procedure (MCB). Table 4 shows that Theta and ForecastPro are the best methods for monthly data, that Theta and Comb S-H-D are the best methods for quarterly data, and that Theta, RobustTrend and Comb

Table 5
Multiple comparisons with the mean

Method	Category	Monthly		Quarterly		Yearly	
		Better	Worse	Better	Worse	Better	Worse
Naive2	Naïve	0	18	0	4	0	6
Single	Naïve	0	17	0	4	0	6
Holt	Explicit trend	11	0	0	0	0	0
Dampen	Explicit trend	1	0	3	0	0	0
Winter	Explicit trend	5	0	0	0	0	0
Comb S-H-D	Explicit trend	13	0	8	0	0	0
B-J automatic	ARIMA/ARARMA	0	2	0	0	0	0
Autobox1	ARIMA/ARARMA	1	0	0	8	0	5
Autobox2	ARIMA/ARARMA	0	0	0	0	3	0
Autobox3	ARIMA/ARARMA	0	0	0	3	0	2
RobustTrend	Explicit trend	0	4	3	0	3	0
ARARMA	ARIMA/ARARMA	0	0	0	0	0	0
AutomatANN	Neural network	1	0	0	0	0	1
Flores–Pearce1	Expert system	0	5	2	0	1	0
Flores–Pearce2	Expert system	0	0	0	3	0	2
PP-Autocast	Explicit trend	0	2	2	0	1	0
ForecastPro	Expert system	18	0	1	0	2	0
SmartFcs	Expert system	0	0	0	4	0	1
Theta-sm	Explicit trend	0	5	0	0	0	4
Theta	Decomposition	18	0	8	0	4	0
RBF	Expert system	7	3	2	2	6	0
ForcX	Expert system	3	0	2	0	4	0

Number of horizons (out of 18 for monthly data, out of 8 for quarterly data, out of 6 for yearly data) for which the estimation method performs significantly better or worse than average for monthly, quarterly and yearly data.

S-H-D are the best methods for annual data. Table 4 also shows that there are some methods (Naive2, Single, Theta-sm) that are consistently worse than the best at all horizons for all three categories of series. Other methods are significantly worse than the best at only some horizons and some categories of the series.

3.3. Are some methods better (worse) than average?

In comparing the methods with respect to average performance, we use the ANOM procedure. Table 5 shows which methods perform significantly better than average for the monthly, quarterly and yearly data. Theta is better than average in all three categories, while some of the other methods are better than average for one or two of the categories but not for the other(s).[6] Moreover, there are some methods that are never better than average at any horizon in any category. These are Naïve2, Single, B-J automatic, Autobox-3, ARARMA, Flores–Pearce2, SmartFcs and Theta-sm. Finally, Table 5 indicates that Naïve2 and Single are consistently worse than average.

3.4. Summary of results: Rankings

Because Theta was better than average for only four of the six yearly horizons, there was no method that was consistently better than average at all horizons for all frequency categories. Nevertheless, we were able to find some regularity in the results. There were a number of methods that were never better than average, and Naive2 and Single were consistently worse than average. Given the finding that there were some statistically significant and consistent results, the fourth conclusion of the M3 competition is somewhat attenuated. For some methods, the forecasting performance does depend on the horizon and frequency category, but for other methods, there are definitive results. Some are better than average; others are worse.

[6] It has been argued that Robust-Trend does extremely well on annual data. While it was never worse than the best or the average in this category, it was only significantly above average for three of the six annual horizons.

4. Relative performance of each method

It is also of interest to compare the relative performance of each method with every other method and to determine which methods performed significantly better than other procedures. Although the M3 competition did not specifically make this comparison, Makridakis and Hibon (2000) did compare several methods with both Naive2 and Dampen but did not test for statistical significance. Flores and Pearce (2000) used the difference-of-two-means test to compare each method with the best performing method at each specific horizon for the three categories of time series. However, they did not compare every method with every other technique.

We compare the performance of each method with every other method. Our measure of relative accuracy is the percentage of times that Method i beats Method j. The test statistic is the binomial with p=0.5. We tabulate the number of other methods that were significantly worse than Method i. While this procedure was implemented for all horizons for each category of time series, monthly, quarterly and yearly, we present only a selected set of results.

Table 6
Relative performance of forecast methods

	Yearly data		Quarterly data			Monthly data			
	Horizon		Horizon			Horizon			
	1	6	1	4	8	1	6	12	18
Methods									
Naive2	1	0	3	0	0	1	0	0	0
Single	1	0	4	0	0	3	1	1	1
Holt	8	2	7	2	3	9	18	11	12
Dampen	9	10	16	10	6	15	4	2	3
Winter	8	2	6	5	5	6	13	11	8
Comb S-H-D	8	11	14	17	14	7	11	13	14
B-J Automatic	5	4	10	8	5	7	4	2	2
Autobox1	1	0	3	0	0	4	2	7	6
Autobox2	9	11	5	2	5	4	3	6	5
Autobox3	3	2	3	1	0	7	2	3	3
RobustTrend	17	7	3	18	18	1	1	1	1
ARARMA	6	1	5	6	3	4	2	8	4
AutomatANN	3	1	3	0	2	11	1	2	5
Flores–Pearce1	9	9	7	3	2	5	2	2	2
Flores–Pearce2	0	4	0	1	4	5	2	2	6
PP-Autocast	10	10	19	7	8	7	1	4	1
ForecastPro	8	9	15	1	2	19	14	15	14
SmartFcs	3	9	1	2	0	5	2	1	7
Theta-sm	6	2	3	2	2	2	2	2	2
Theta	9	15	13	16	19	19	15	16	15
RBF	17	11	0	4	14	0	11	15	21
ForecastX	9	19	13	2	3	15	14	4	7

Performance of each method relative to all other methods: number of methods that have significantly poorer performance than the specific method.

Table 6 shows the number of methods that had significantly poorer performance than Method i. The performance of each method relative to every other method differs both within each category of time series and across the three categories. Of particular note are the improvements in the accuracy of RBF relative to other methods as the horizon increases while the relative performance of Theta is robust regardless of the horizon and time series category. This finding suggests that a practitioner's choice of forecasting method should depend on the time series category and the horizon that are of most concern to the user.

5. Relationship between complexity and accuracy

One of the conclusions of the M3 competition was that the increased complexity of a method did not necessarily result in improved accuracy relative to simpler methods. To test this hypothesis, we ranked the methods by their complexity and also ranked the methods by their accuracy. The Spearman rank correlation coefficient was used to test this hypothesis. Three time series experts were asked to rank the complexity of the 22 methods. These three different rankings are presented in Table 7.1. The 22 methods were also ranked according to their accuracy measured by their average ranking.

The Spearman rank correlation coefficient Rho is defined by:[7]

$$T = \sum_{i=1}^{K} [R(X_i) - R(Y_i)]^2$$

$$\mathrm{Rho} = 1 - 6T/K\left(K^2 - 1\right)$$

where $R(X_i)$ is the rank of complexity of Method i, $R(Y_i)$ is the rank of accuracy of Method i, and K is the number of methods.

[7] This formula is appropriate even if there are ties in the rankings, provided that the number of ties is small relative to the number of observations.

Table 7.1
Ranking of the complexity of the 22 methods according to three experts

Expert 1: Keith Ord		Expert 2: Robert Fildes		Expert 3: Spyros Makridakis	
Method	Rank 1	Method	Rank 2	Method	Rank 3
Naive2	1	Naive2	1	Naive2	1
Single	2	Single	2	Single	2
Holt	3	Holt	3	Holt	3
Robust-Trend	4	Robust-Trend	4	Robust-Trend	4
Dampen	5	Winter	5	Winter	4
Winter	6	Dampen	6	Comb S-H-D	4
Theta-sm	7	PP-Autocast	6	Dampen	7
PP-Autocast	8	Theta-sm	8	PP-Autocast	8
Theta	9	Comb S-H-D	9	Theta-sm	9
Comb S-H-D	10	Theta	10	B-J automatic	10
B-J automatic	11	B-J automatic	11	Autobox1	10
Autobox1	11	Autobox1	12	Autobox2	10
ARARMA	13	Autobox2	12	Autobox3	10
SmartFcs	14	Autobox3	12	Theta	14
Flores–Pearce1	14	ARARMA	15	ARARMA	15
Flores–Pearce2	14	ForecastPro	16	ForecastPro	15
ForecastPro	17	SmartFcs	16	SmartFcs	15
Autobox2	18	Flores–Pearce1	16	Flores–Pearce1	15
Autobox3	18	Flores–Pearce2	16	Flores–Pearce2	15
RBF	20	ForecastX	16	ForecastX	15
ForecastX	21	AutomatANN	21	RBF	21
AutomatANN	22	RBF	22	AutomatANN	21

The values of the Spearman coefficients for a selected number of horizons of each category are presented in Table 7.2. The critical value of the coefficient for K=22 at the 5% level is 0.4241. None of the Spearman coefficients exceeds this level, and we conclude that the degree of complexity and the accuracy of the methods are not associated. This is in accord with the M3 conclusion that an increase in complexity does not produce greater accuracy.

6. Relationship of accuracy measures

The M3 competition presented results based on five different accuracy measures: symmetric mean absolute percentage error, average ranking, percentage better, median symmetric absolute percentage error and median relative absolute error[8]. The study concluded that the results differed depending upon which accuracy measure was used. It is possible to test this conclusion. Kendall's coefficient of concordance, W, permits one to test whether several different rankings, in this case, the four accuracy measures, are in agreement. If the rankings are in perfect agreement, the value of W would be 1. If there were no agreement at all, its value would be 0.

The W test statistic is

$$W = \frac{12\sum_{j=1}^{K} R_j^2 - 3b^2K(K+1)^2}{b^2K(K^2-1)} \tag{11}$$

where b represents the number of sets of rankings (b=4), and the R_j are the sum of the ranks of the methods according to each accuracy measure. The null hypothesis is that there is no agreement among the four rankings, and the alternative is that there is agreement.

The W statistic is related to Chi-square, i.e.,

$$\chi^2 = b(K-1)W, \tag{12}$$

and is tested for significance for tabulated values of the statistic for $K-1$ degrees of freedom (Daniel, 1990, p.390).

[8] Only four of the five methods were used in this analysis. The accuracy method that was not utilized was percentage better. It was eliminated because it is based on the same measure as average ranking but displays the result method by method.

Table 7.2
Spearman rank correlation coefficients: relationship between complexity and accuracy

	Horizons	Coefficient for rank 1	Coefficient for rank 2	Coefficient for rank 3
Yearly data	1	0.221	0.171	0.199
	6	0.318	0.291	0.303
Quarterly data	1	−0.172	−0.253	−0.230
	4	−0.071	0.078	0.002
	8	0.083	0.100	0.136
Monthly data	1	0.177	0.157	0.210
	6	0.155	0.195	0.169
	12	0.224	0.223	0.213
	18	0.340	0.376	0.331

Table 8 presents the values of W and Chi-square for a selected number of horizons of each of the time series categories. In all cases, the value of W is equal to or exceeds 0.600. The χ^2 statistic exceeds 50 in all cases. With 21 degrees of freedom, the 5% critical value of χ^2 is 32.67. The null hypothesis that there is no agreement in the rankings of the four accuracy measures is clearly rejected. This finding invalidates the M3 conclusion that the results differed depending upon which accuracy measure was used.

7. Combination of methods

One of the conclusions of the M3 competition was that a combination of methods was more accurate, on average, than the methods being combined. The only combination of methods that was examined was composed of (1) Single, (2) Holt and (3) Dampen. In Section 4, we compared the accuracy of every method with that of every other method. A summary of the results showing how the combination of the three methods did relative to each of the three individual methods is presented in Table 9. The results indicate that the combination of these three methods beat (1) Single from 59.4% to 63.5%, (2) Holt from only 47.9% to 54.4% and (3) Dampen from 49.7% to 55.0% of the time.

Table 8
Kendall's coefficients of concordance W and corresponding Chi-square values: Relationship between Various Accuracy Methods

	Horizons	W	χ^2	p-value
Yearly data	1	0.796	66.9	0.00
	6	0.755	63.5	0.00
	Average 1–6	0.745	62.6	0.00
Quarterly data	1	0.803	67.5	0.00
	4	0.677	56.9	0.00
	8	0.675	56.7	0.00
	Average 1–8	0.728	61.2	0.00
Monthly data	1	0.847	71.1	0.00
	6	0.604	50.7	0.00
	12	0.600	50.4	0.00
	18	0.664	55.8	0.00
	Average 1–18	0.680	57.2	0.00

Table 9

Percentage of time comb S-H-D is better than	Yearly data	Quarterly data	Monthly data
Single	62.6*	59.4*	63.5*
Holt	51.5	54.4*	47.9
Dampen	49.7	52.6	55.0*

* Significantly better.

In some cases, the accuracy of the combination of methods was significantly better than the individual techniques[9]; in other cases, this was not true. Because only one combination of methods was examined and because these results are conflicting, we believe that the M3 conclusion about the combination of methods being significantly superior was not proved.

8. Conclusions and implications

We have reexamined the findings of the M3 competition and attempted to determine whether the conclusions are statistically significant. We used a multiple-comparisons procedure that had not previously been used in forecast evaluations and concluded that the accuracy of the various methods does differ significantly, and that some methods are significantly better than others. On the other hand, not all of the original study's conclusions were accurate or statistically significant.

We agree with the original M3 conclusion that there is no significant relationship between accuracy and complexity. Our findings suggest that some of

[9] The test for significance was the number of times that the combination was more accurate than the individual method. The binomial statistic with p=0.5 was used to test the hypothesis that the number of superior occurrences could have been due to chance.

the other conclusions of the original study should be modified. First, we believe that some methods are significantly better than others, and that this finding does not vary with the category of series that is examined or with the forecasting horizon. Second, while we examined only one combination of methods, the combination of Single, Holt and Dampen, it was not significantly more accurate than the individual methods. This led us to state that the M3 conclusion about the combination of methods being significantly superior was not proven. Finally, we disagreed with the M3 finding that the results differed depending upon the accuracy measure that was used. We discovered that there is a significant agreement among the results obtained from the various accuracy measures that have been used in the evaluations of the various competitions.

An important implication of this paper is that our analysis demonstrates the necessity for undertaking rigorous statistical testing in any evaluation of forecast accuracy, whether or not it is a forecasting competition. If this is not done, it is possible that one may draw erroneous conclusions from the descriptive statistics. Additionally, we believe that a statistical analysis of ranks using nonparametric methodology is very useful as it is easy to perform and its outcomes are easy to interpret.

Acknowledgements

The authors would like to thank an Associate Editor, two anonymous referees and Robert Fildes, Keith Ord, Spyros Makridakis, Benito Flores, Chris Chatfield, Hans Levenbach, Robert Clemen and Guy Mélard for their helpful comments and suggestions.

References

Bretz, F., Genz, A., & Hothorn, L. A. (2001). On the numerical availability of multiple comparison procedures. *Biometrical Journal*, *43*(5), 645–656.

Daniel, W. W. (1990). *Applied nonparametric statistics*. Boston: PWS-Kent Publishing.

De Gooijer, J. G., & Zerom, D. (2001). Kernel-based multi-step ahead predictions of the U.S. short term interest rate. *Journal of Forecasting*, *19*, 335–353.

Farnum, N. R. (2004). Analysis of means using mathematical processors. *Quality Engineering*, *16*, 399–405.

Flores, B. E., & Pearce, S. L. (2000). The use of an expert system in the M3-competition. *International Journal of Forecasting*, *16*(4), 485–496 (Special issue).

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, *32*, 675–701.

Friedman, M. (1939). A correction: The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, *34*, 109.

Genz, A. (1992). Numerical computation of multivariate normal probabilities. *Journal of Computational and Graphical Statistics*, *1*, 141–149.

Halperin, M., Greenhouse, S. W., Cornfield, J., & Zalokar, J. (1955). Tables of percentage points for the studentized maximum absolute deviate in normal samples. *Journal of the American Statistical Association*, *50*, 185–195.

Harter, H. L. (1960). Tables of range and studentized range. *The Annals of Mathematical Statistics*, *31*, 1122–1147.

Hollander, M., & Wolfe D. A. (1999). *Nonparametric statistical methods*, Wiley series in probability and statistics: Texts and references section. John Wiley and Sons, New York, second edition, A Wiley Inter-science Publication.

Hsu, J. C. (1996). *Multiple comparisons. Theory and methods*. London: Chapman and Hall.

Kendall, M. G., & Smith, B. B. (1939). The problem of *m* rankings. *The Annals of Mathematical Statistics*, *10*, 275–287.

Kolb, R. A., & Stekler, H. O. (1996). How well do analysts forecast interest rates? *Journal of Forecasting*, *15*, 385–394.

Makridakis, S., & Hibon, M. (2000). The M3-Competition: Results, conclusions and implications. *International Journal of Forecasting*, *16*(4), 451–476.

McDonald, B. J., & Thompson, W. A., Jr. (1967). Rank sum multiple comparisons in one and two way classifications. *Biometrika*, *54*, 487–497.

McDonald, B. J., & Thompson, W. A., Jr. (1972). Corrections and amendments: Rank sum multiple comparisons in one and two way classifications. *Biometrika*, *59*, 699.

Nelson, L. S. (1983). Exact critical values for use with the analysis of means. *Journal of Quality Technology*, *15*, 40–44.

Ott, E. R. (1958). Analysis of means. *Technical report*, Rutgers University.

Ott, E. R. (1967). Analysis of means: A graphical procedure. *Industrial Quality Control*, *24*, 101–109.

Ryan, Th. P. (2000). *Statistical methods for quality improvement*. (2nd ed.). New York: Wiley.

Soong, W. C. (2001). Exact simultaneous confidence intervals for multiple comparisons with the mean. *Computational Statistics & Data Analysis*, *37*(1), 33–47.

Stekler, H. O. (1991). Macroeconomic forecast evaluation techniques. *International Journal of Forecasting*, *7*(3), 375–384.

Stekler, H. O. (2001). The M3-competition: The need for formal statistical tests. *International Journal of Forecasting*, *17*(4), 576–577.

Wallis, W. A. (1939). The correlation ratio for ranked data. *Journal of the American Statistical Association*, *34*, 533–538.

Biographies: Alex J. KONING is Assistant Professor of Statistics at the Econometric Institute, Erasmus University Rotterdam. His research interests are in the area of multivariate statistics and nonparametric methods.

Philip Hans FRANSES is Professor of Applied Econometrics and Professor of Marketing Research, both at Erasmus University Rotterdam. Currently, he is the director of the Econometric Institute. His research interests are marketing research, empirical finance and applied econometrics.

Michèle HIBON is a Senior Research Fellow and Lecturer at INSEAD. She graduated from the University of Paris, holds a degree in Science and a diploma in advanced studies in Physics. Her current research interests are in the areas of forecasting accuracy of time series methods including M, M2, M3 competitions.

H.O. STEKLER is a Research Professor at George Washington University. He has written extensively on forecasting issues with particular emphasis on forecast evaluations.