Available online at www.sciencedirect.com
SCIENCE DIRECT
International Journal of Forecasting 21 (2005) 397–409
international journal of forecasting
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}, . . ., R_{Nk}, where R_{nk} is the rank of A_{nk} among A_{n1}, A_{n2},. . .,A_{nK}. We shall assume that A_{n1}, A_{n2},. . ., A_{nK} in fact have been obtained by monotone transforming (unknown) independent random variables U_{n1}, U_{n2},. . ., 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},. . ., 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,. . ., F_N such that

$$P(U_{nk} \leq x) = F_n(x - \tau_k) \quad (1)$$

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.