Content list available at JournalPark

Turkish Journal of Forecasting

Journal Homepage: tjforecast.xyz

A Note on the Robustness of Performance of Methods and Rankings for M4 Competition

E. Egrioglu[1,*], R. Fildes[2]

[1]Giresun University, Faculty of Arts and Sciences, Department of Statistics, Gure Campus, 28200 Giresun, Turkey
[2]Lancaster University, Management Science School, Department of Management Science, Marketing Analytics and Forecasting Research Centre Lancaster, United Kingdom

ARTICLE INFO

Article history:

Received	16	December	2020
Revised	22	December	2020
Accepted	23	December	2020
Available online	27	December	2020

Keywords:

M4 forecast competition
Error metrics
Direction accuracy

DISCUSSION

ABSTRACT

M4 forecasting competition provided some useful information to forecasting literature. The provided information is based on calculated error metrics for ranking methods in the competition. The organizers of the competition calculated arithmetic mean as a descriptive statistic for evaluating the performance of all competitors. In this paper, the effect of different descriptive statistics for ranking of methods is investigated. It is found that the distribution of error metrics for competitor forecasting methods is not symmetric. Thus, the arithmetic mean descriptive statistic is not a good metric to determine the centre of non-symmetric distributions and it will not well present centre of the distribution. In this study, it is showed that the median will well present centre of distribution for error metrics of competitor forecasting methods. When the median is used as a descriptive statistic for ranking methods, the ranks of methods is different form ranks which are calculated according to the arithmetic mean descriptive statistics. Moreover, the direction accuracy metric is calculated for the best ten methods in the competition. So, the forecasting methods are ranked according to direction accuracy and it is showed that the ranks are different from the competition results.

1. Introduction

The recent M4 competition (its results are presented in [9]) has become the biggest forecast competitions among forecasting competitions [8]. The objectives of this latest competition were declared as considering more number of time series, more data frequencies, prediction intervals and using statistically-robust error measures. A key element was to establish robust rankings of the many methods considered. The focus on this note it explore this. The M-Competition series considered a wide range of error measures. [3] discussed about M-competitions and suggested using of Mean absolute percentage error (MAPE), Median absolute percentage error (MdAPE), % Better, average rankings, Geometric relative absolute average, Median relative absolute error in the next competitions. [10] considered five accuracy measures symmetric MAPE (SMAPE), Average Ranking, Median symmetric absolute percentage error, Percentage Better, and Median relative absolute error to analyze the performance of the various methods in M3 Competition. [12] emphasised that no error measure is perfect. [1] analysed correlations between methods. Many discussions have been made on error measures. [5] found that SMAPE has been shown to be asymmetric in its treatment of positive and negative forecast errors. [2] recommended the average relative geometric mean absolute percentage, which is the geometric mean of ratios of mean absolute errors, as an alternative to the

[*] Corresponding author.
E-mail addresses: erol.egrioglu@giresun.edu.tr (Erol Egrioglu), r.fildes@lancaster.ac.uk (Robert Fildes)

mean absolute scaled error (MASE). In the M4 competition organizers considered using of the SMAPE, the MASE and overall weighted average (OWA). [11] emphasis that the correlation between SMAPE and MASE is about 0.90 and this shows a strong relation between two error metrics. Despite strong relationships, SMAPE and MASE can produce different rankings. Because of this, OWA was computed to find final rankings by using arithmetic means of SMAPE and MASE values in M4 competition. [6] stated that it would clearly be interesting in the future to see how robust the findings are to alternative error measures and to the alternative loss functions that they assume implicitly. [7] suggested to use full predictive densities instead of point forecasts and well-known error measures. [11] declared that instead of full predictive densities, prediction intervals with different confidence level will be able to ask in the future competition. The some well-known and the most common error metrics are given in Table 1.

Table 1. Well-known error metric formulas in the forecast competitions

$MdAPE = median\left\{\frac{x_t - \hat{x}_t}{x_t} \; ; t = 1,2,\ldots,ntest\right\}$
$MAPE = \frac{100}{ntest}\sum_{i=1}^{ntest}\left\lvert\frac{x_t - \hat{x}_t}{x_t}\right\rvert$
$SMAPE = 200 \times \frac{1}{ntest}\sum_{t=1}^{ntest}\frac{\lvert x_t - \hat{x}_t\rvert}{\lvert x_t\rvert + \lvert\hat{x}_t\rvert}$
$MASE = \frac{1}{ntest}\sum_{t=1}^{ntest}\left\lvert\frac{x_t - \hat{x}_t}{\frac{1}{n-1}\sum_{i=2}^{n}\lvert x_t - x_{t-1}\rvert}\right\rvert$

We here explore the robustness of the results. Previous research into the various competition results has shown that rankings were not robust to the choice of error measures. In this paper, we consider alternative error measures that take into account the distribution of the error statistics. When the empirical distributions of SMAPE and MASE are examined for all methods, it is found that arithmetic mean is not an appropriate descriptive statistics the distributions are heavily skewed so that reliance of the arithmetic mean is potentially misleading. Instead of using the arithmetic mean, median statistics give a different picture, demonstrating the results are not robust as claimed. In the next section of the paper, we examine the distribution of various error statistics including SMAPE and MASE and show where the results differ substantially and where the claims made for M4 seem well-founded. The final section revisits some of the earlier methodological conclusions in an attempt to resolve one of the controversies that have long surrounded the methodology of forecasting competitions evaluation.

2. Data Analysis

We focus our analysis on the best 10 methods and 2 benchmarks: combination and ETS. Moreover, different kind of rankings and evaluations are proposed. In the forecast competitions, the direction of forecasts is ignored to report. In this paper, the performance of the best ten methods is investigated in terms of direction accuracy. Moreover, the first ranked methods are determined for each series and computed the first ranked proportions for the best ten methods. It is expected to see any method can be always preferred to solve any series in the M4 competition. The best ten methods of M4 competition are listed in Table 2.

In the M4 competition, competitors' methods are ranked by using OWA operator. [9] summarized the calculation of ranks with the following two sentences. "We compute the OWA of SMAPE and MASE by first dividing their total value by the corresponding value of Naïve 2 to obtain the relative SMAPE and the relative MASE, respectively, and then computing their simple arithmetic mean. Note that SMAPE and MASE are first estimated for each series by averaging the error computed for each forecasting horizon, then averaged again across all time series to compute the average value for the entire dataset. On the other hand, OWA is computed only once at the end of the evaluation process for the whole sample of the series. Thus, although OWA is relative in nature, it is more indicative and robust than typical relative measures and measures based on relative errors.".

It is clear that the OWA values were calculated arithmetic mean of SMAPE and MASE criteria values for 100.000 time series. It can be seen box-plot graphs of SMAPE and MASE. In figure 1, box-plot graphs of SMAPE and MASE for Smyl, S. Method and ETS are given. The distributions are clearly non-symmetric and arithmetic mean is not an appropriate descriptive statistic for them. Same graphs can be seen for other methods.

Table 2. The best ten method according to organizers ranking

Rank	User ID	Team Members	Affiliation	Type of Method
1	118	Smyl, S.	Uber Technologies	Hybrid
2	245	Montero-Manso, P., Talagala, T., Hyndman, R. J. & Athanasopoulos, G.	University of A Coruña & Monash University	Combination (S & ML)
3	237	Pawlikowski, M., Chorowska, A. & Yanchuk, O.	ProLogistica Soft	Combination (S)
4	72	Jaganathan, S. & Prakash, P.	Individual	Combination (S & ML)
5	69	Fiorucci, J. A. & Louzada, F.	University of Brasilia & University of São Paulo	Combination (S)
6	36	Petropoulos, F. & Svetunkov, I.	University of Bath & Lancaster University	Combination (S)
7	78	Shaub, D.	Harvard Extension School	Combination (S)
8	260	Legaki N. Z. & Koutsouri K.	National Technical University of Athens	Statistical
9	238	Doornik, J., Hendry, D. & Castle, J.	University of Oxford	Combination (S)
10	39	Pedregal, D.J., Trapero, J. R., Villegas, M. A. & Madrigal, J. J.	University of Castilla-La Mancha	Combination (S)

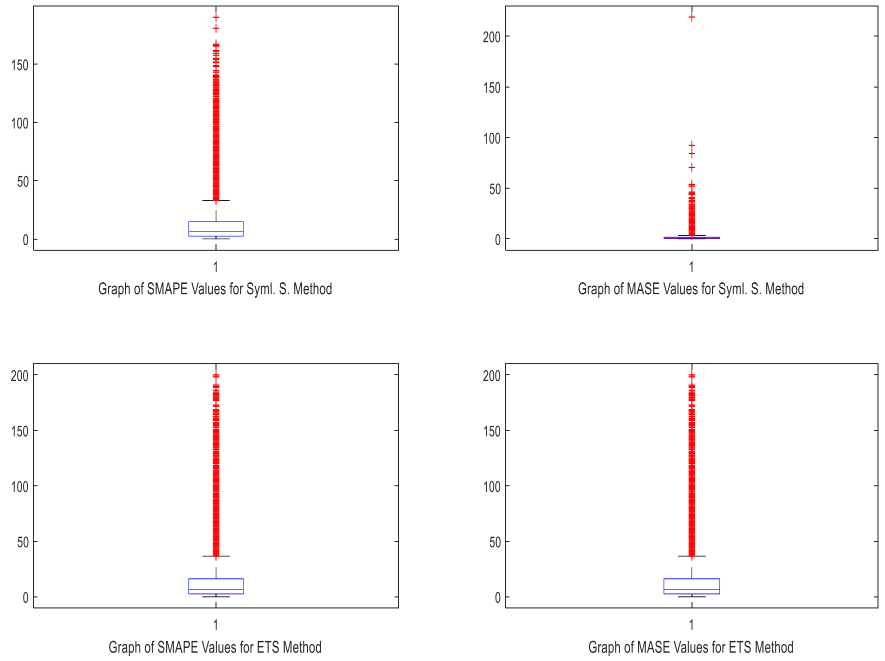

Figure 1. Box-Plot of SMAPE and MASE values from Smyl S. and ETS methods for 100.000 time series

For this kind of non-symmetric distributions, the descriptive statistics should be robust like median. Moreover, the mean and median will be dramatically different from each other for a non-symmetric empirical distribution. In Table 3, median statistics for SMAPE values are given for the best ten methods and two benchmarks. Moreover, the ranking is calculated according to mean and median statistics in the last two columns. It is clear that all rankings are changed. Moreover, the winner method is changed. The last row of the table, Spearman's rho correlation coefficients (their p-value comparisons in the bracket) between mean and median of SMAPE values for best ten methods are given. The insignificant correlations presented with "*" mark. The formula of the Spearman's rho correlation coefficients is given below and *d* presents the difference between rank numbers in the formula.

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \tag{1}$$

It is really interesting that the correlations are insignificant for Monthly, Weekly and Hourly data sets. These means that the rankings are completely changed for Monthly, Weekly and Hourly data. It should be noted that the rankings can be still changed in significant correlation situations

Table 3. Median statistics of SMAPE values from the best ten competitors for M4 data: forecast horizon

User ID	**Yearly**	**Quarterly**	**Monthly**	**Weekly**	**Daily**	**Hourly**	**Total**	**General Rank (Median)**	**General Rank (Mean)**
118	7,851	5,384	6,814	4,648	2,099	3,205	6,265	2	1
245	8,112	5,374	6,679	4,348	2,042	2,915	6,259	1	3
237	8,328	5,480	6,780	4,477	1,429	2,903	6,361	6	5
72	8,219	5,513	6,779	5,179	2,005	3,370	6,356	5	2
69	8,088	5,477	6,815	5,129	1,979	3,243	6,303	3	4
36	8,143	5,443	6,859	4,577	1,998	3,519	6,355	4	6
78	7,973	5,867	7,080	6,194	2,177	3,295	6,490	7	9
260	7,969	5,850	7,188	5,030	1,989	3,455	6,547	9	8
238	8,772	5,686	7,097	4,584	1,997	3,022	6,670	10	7
39	8,488	5,641	7,034	5,148	1,967	3,449	6,536	8	10
Comb.	**8,799**	**5,748**	**7,040**	**5,098**	**1,972**	**8,800**	**6,627**		
ETS	**8,966**	**5,608**	**6,995**	**5,060**	**1,991**	**5,734**	**6,614**		
Correlations	**0,855** **(p<0.05)**	**0,915** **(p<0.05)**	**0,745*** **(p>0.05)**	**0,442*** **(p>0.05)**	**0,842** **(p>0.05)**	**0,527*** **(p>0.05)**	**0,770** **(p<0.05)**		

* presents insignificant correlations at the 0.05 level (two-sided)

In Table 4, median statistics for MASE values are given for the best ten methods. Moreover, the ranking is calculated according to mean and median statistics in the last two columns. It is clear that many rankings are changed. The winner method too changes. Spearman's rho correlation coefficients are presented in the last row of Table in the same manner in Table 3. It is clear that all correlations are significant.

Table 4. Median statistics of MASE values from the best ten competitors for M4 data

User ID	**Yearly**	**Quarterly**	**Monthly**	**Weekly**	**Daily**	**Hourly**	**Total**	**General Rank (median)**	**General Rank (mean)**
118	2,142	0,864	0,693	1,573	2,507	0,736	0,933	3	1
245	2,144	0,862	0,699	1,548	2,390	0,727	0,932	2	3
237	2,157	0,877	0,708	1,440	1,653	0,730	0,927	1	5
72	2,192	0,874	0,696	1,561	2,353	0,744	0,940	4	2
69	2,119	0,871	0,711	1,790	2,319	0,898	0,948	5	4
36	2,158	0,859	0,715	1,430	2,327	1,094	0,949	6	6
78	2,104	0,933	0,725	1,950	2,502	1,203	0,979	8	9
260	2,080	0,943	0,761	1,905	2,318	1,624	1,005	10	8
238	2,325	0,902	0,729	1,591	2,360	0,656	0,975	7	7
39	2,250	0,907	0,743	1,906	2,305	0,955	0,983	9	10
Comb.	**2,259**	**0,918**	**0,757**	**1,839**	**2,322**	**1,915**	**1,004**		
ETS	**2,329**	**0,886**	**0,736**	**1,666**	**2,336**	**1,065**	**0,980**		
Correlations	**0,891** **(p<0.05)**	**0,963** **(p<0.05)**	**0,988** **(p<0.05)**	**0,915** **(p<0.05)**	**0,891** **(p<0.05)**	**0,976** **(p<0.05)**	**0,842** **(p<0.05)**		

OWA values are calculated according to median statistics of SMAPE and MASE, and they are given in Table 5. In the calculation of OWA, median values of SMAPE and MASE for Naive2 were used similar to original calculation. When table 4 is examined, the winner method is changed and many of ranking is changed in the new rankings. As a result of the calculations, Montero-Manso, P.et al. method is the new winner of the competition. Moreover, all correlations are significant for OWA like MASE results.

Table 5. OWA values from the best ten competitors by using median statistics for M4 yearly data

User ID	Yearly	Quarterly	Monthly	Weekly	Daily	Hourly	Total	General Rank (median)	General Rank (mean)
118	0,710	0,787	0,851	0,854	1,060	0,727	0,8048	2	1
245	0,722	0,785	0,846	0,819	1,021	0,681	0,8037	1	2
237	0,734	0,800	0,857	0,804	0,710	0,680	0,8082	3	3
72	0,735	0,801	0,850	0,903	1,003	0,754	0,8136	5	4
69	0,717	0,797	0,861	0,957	0,990	0,788	0,8135	4	5
36	0,726	0,789	0,867	0,811	0,996	0,896	0,8174	6	6
78	0,709	0,854	0,887	1,101	1,078	0,901	0,8392	7	7
260	0,705	0,857	0,915	0,977	0,992	1,070	0,8540	10	8
238	0,782	0,827	0,890	0,853	1,003	0,672	0,8487	9	9
39	0,757	0,825	0,895	0,989	0,984	0,838	0,8438	8	10
Comb.	**0,772**	**0,838**	**0,903**	**0,967**	**0,989**	**1,958**	**0,859**		
ETS	**0,791**	**0,814**	**0,888**	**0,918**	**0,996**	**1,213**	**0,847**		
Correlations	**0,903** **(p<0.05)**	**0,970** **(p<0.05)**	**0,939** **(p<0.05)**	**0,891** **(p<0.05)**	**0,924** **(p<0.05)**	**0,855** **(p<0.05)**	**0,915** **(p<0.05)**		

In the M4 forecast competition, organizers did not investigate the accuracy of forecast directions. In this study, the direction accuracy metric values are calculated and the new rankings are constituted according to direction accuracy. The formula of direction accuracy metric is given below:

$$DA = \frac{1}{n}\sum_{i=1}^{n} a_i \quad , a_i = \begin{cases} 1 \quad if\ (y_{i+1} - y_i)(\hat{y}_{i+1} - y_i) > 0 \\ 0 \qquad\quad , \qquad otherwise \end{cases} \tag{1}$$

In the formula, x_i and $\hat{x}_i$ represents actual and forecasted values, respectively. These criteria measure the ratio of correct forecast directions. The summation of a_i elements present the total number of forecasts which have the same direction with the actual value. In table 6, the mean of direction accuracy metric values is given. The empirical distributions for DA metric are symmetric and the mean statistic is an appropriate descriptive statistics for DA metric.

Table 6. Mean of DA values for the best ten methods in the M4 competition

User ID	Yearly	Quarterly	Monthly	Weekly	Daily	Hourly	Total	General Rank
118	0,574	0,557	0,556	0,610	0,470	0,688	0,558	1
245	0,549	0,563	0,552	0,580	0,471	0,684	0,552	3
237	0,515	0,544	0,545	0,590	0,594	0,700	0,541	7
72	0,532	0,562	0,551	0,590	0,471	0,681	0,547	4
69	0,505	0,554	0,546	0,548	0,472	0,655	0,536	8
36	0,532	0,566	0,545	0,596	0,471	0,658	0,545	5
78	0,508	0,525	0,542	0,572	0,469	0,656	0,528	9
260	0,467	0,524	0,533	0,542	0,470	0,636	0,514	10
238	0,566	0,563	0,545	0,585	0,471	0,698	0,552	2
39	0,554	0,556	0,539	0,541	0,470	0,668	0,545	6

Table 6 shows, the best forecast direction accuracy is again achieved by Syml method. The surprising result is achieved by Doornik et al. method. Although Doornik et al. method was ranking 9th row according to OWA, the method has 2nd rank according to DA metric. The rank correlation with the earlier error measures is around.

Actually, calculating some descriptive statistics can be discussed because this strategy does not show us which method is the best for which series. To see this reality, the methods are ranked according to SMAPE per series and percentage to be the winner for all methods are given in Table 7.

Table 7. The percentage of series when a method is ranked first according to SMAPE

User ID	Yearly	Quarterly	Monthly	Weekly	Daily	Hourly	Total	General Rank
118	0,188	0,185	0,192	0,181	0,102	0,125	0,185	1
245	0,113	0,101	0,113	0,095	0,117	0,114	0,110	3
237	0,091	0,082	0,089	0,075	0,071	0,103	0,087	5
72	0,079	0,076	0,073	0,050	0,039	0,155	0,074	8
69	0,065	0,078	0,067	0,028	0,034	0,119	0,068	10
36	0,067	0,077	0,068	0,042	0,034	0,088	0,069	9
78	0,077	0,079	0,073	0,075	0,043	0,067	0,074	7
260	0,090	0,088	0,084	0,061	0,086	0,065	0,086	6
238	0,090	0,100	0,104	0,106	0,220	0,093	0,104	4
39	0,140	0,135	0,137	0,287	0,256	0,072	0,142	2

When Table 7 is examined, the rankings are completely changed and rankings are very close to each other except Syml method. The Syml method is the best method for %18.5 of all series.

When the same calculations are made for MASE values, the results are very similar to the results in Table 6. There is no need to calculate an OWA metric for SMAPE and MASE if you examine the percentage of series when a method is ranked first. The ranking is the same for both MAPE and MASE. The interesting result is that any method cannot be recommended as a dominant method. These results remember us to [4] comments and strengthen his idea. [4] commented on M4 competition results and it is commented that “the results certainly should guide the short-list, but we have no series statistics that would tell a practitioner whether, for their particular problem, they should (say) use combining, use ‘Smyl’, or select among Damped, Theta and ARIMA, or even an ML method.”.

3. Conclusions

The findings of this discussion can be summarized follow. Usage of the arithmetic mean as descriptive statistics for SMAPE and MASE values is not suitable. The median or robust descriptive statistics can be preferred. When the median is referred, the rankings dramatically changed. The rankings according to mean and median are more changed for SMAPE than MASE and OWA. The correlations are insignificant for Monthly, Weekly and Hourly data sets by using SMAPE. Direction accuracy can be a discriminative error metric, different methods can give successful results for direction accuracy. Any method cannot be proposed as the best method for all series. The best method is changed series by series. Each series needs its special interest.

Acknowledgements

This study is partly supported by Turkish Science and Technological Researches Foundation with Award Number:1059B191800872, Recipient: Erol Egrioglu.

References

[1] P. Agathangelou, D. Trihinas, I. Katakis, Correlation analysis of forecasting methods: The case of the M4 competition, International Journal of Forecasting, 36 (2020), 212–216.

[2] A. Davydenko, R. Fildes, Measuring forecasting accuracy: The case of judgmental adjustments to SKU-level demand forecasts, International Journal of Forecasting, 29-3 (2013), 510–522.

[3] R. Fildes, M. Hibon, S. Makridakis, N. Meade, Generalising about univariate forecasting methods: further empirical evidence. International Journal of Forecasting, 14-3, (1998), 339-358.

[4] R. Fildes, Learning from forecasting competitions, International Journal of Forecasting, 36 (2020), 186–188.

[5] P. Goodwin, Lawton, R., On the asymmetry of the symmetric MAPE, International Journal of Forecasting, 15 (1999), 405–408.

[6] P. Goodwin, Performance measurement in the M4 Competition: Possible future research, International Journal of Forecasting, 36 (2020), 189–190.

[7] S. Kolassa, Why the ‘‘best’’ point forecast depends on the error or accuracy measure, International Journal of Forecasting, 36 (2020), 208–211.

[8] R. J. Hyndman, A brief history of forecasting competitions. International Journal of Forecasting, 36-1 (2020), 7-14.

[9] S. Makridakis, E. Spiliotis, V. Assimakopoulos, The M4 Competition: 100,000 time series and 61 forecasting methods, International Journal of Forecasting, 36 (2020), 54–74.

[10] S. Makridakis, M. Hibon, The M3-Competition: results, conclusions and implications, International Journal of Forecasting, 16 (2000), 451–476.

[11] S. Makridakis, E. Spiliotis, V. Assimakopoulos, Responses to discussions and commentaries, International Journal of Forecasting, 36 (2020), 217–223.

[12] P. Petropoulos, S. Makridakis, The M4 competition: Bigger. Stronger. Better, International Journal of Forecasting, 36 (2020) 3–6.