Makridakis Competitions

or, the State of the Art of Forecasting in Social Setting

Rafał Kucharski

University of Economics in Katowice, Katowice, Poland

Forecasting is hard

There's no chance that the iPhone is going to get any significant market share

Steve Ballmer, CEO Microsoft, April 2007[1]

[1] after [Makridakis, Hyndman, and Petropoulos, 2020]

Genesis

Makridakis, Hibon, and Moser [1979]:

> *simple methods perform well in comparison to the more complex and statistically sophisticated ones*

Criticism motivated the subsequent M, M2 and M3 Competitions that prove beyond the slightest doubt those of the Makridakis, Hibon and Moser study.

M-Competition – M1 (1982)

- 1001 time series,
- 15 forecasting methods (+9 nine variations)

Main conclusions in Makridakis et al. [1982]:

- Statistically sophisticated or complex methods do not necessarily provide more accurate forecasts than simpler ones.
- The relative ranking of the performance of the various methods varies according to the accuracy measure being used.
- The accuracy when various methods are combined outperforms, on average, the individual methods being combined and does very well in comparison to other methods.
- The accuracy of the various methods depends on the length of the forecasting horizon involved.

The findings of the study have been verified and replicated through other competitions and new methods by other researchers.

M2 (1987-1991)

- The purpose of the M2-Competition was to simulate real-world forecasting better in the following respects:
  - Allow forecasters to combine their statistically based forecasting method with personal judgment.
  - Allow forecasters to ask additional questions requesting data from the companies involved in order to make better forecasts.
  - Allow forecasters to learn from one forecasting exercise and revise their forecasts for the next forecasting exercise based on the feedback.
- conducted on a real-time basis
- only 29 time series: 6 macroeconomic series (US data)
- 23 from the four collaborating four companies

The results of the competition were claimed to be statistically identical to those of the M1 [Makridakis et al., 1993].

Fildes and Makridakis [1995]:

Despite the evidence produced by these competitions, the implications continued to be ignored, to a great extent, by theoretical statisticians.

M3 (1999)

Intended to both replicate and extend the features of the M-Competition and M2-Competition, through the inclusion of more methods and researchers (particularly researchers in the area of neural networks) and more (3003) time series:

Time interval	Micro	Industry	Macro	Finance	Demogr.	Other	Total
Yearly	146	102	83	58	245	11	645
Quarterly	204	83	336	76	57	0	756
Monthly	474	334	312	145	111	52	1428
Other	4	0	0	29	0	141	174
Total	828	519	731	308	413	204	3003

Minimum thresholds were set for the number of observations:
14 for yearly series, 16 for quarterly series, 48 for monthly series,
and 60 for other series.

M3 (1999) cont.

Measures used to evaluate the accuracy of forecasts:

- symmetric mean absolute percentage error (sMAPE)
- Average Ranking
- median symmetric absolute percentage error (mdsAPE)
- Percentage Better
- median RAE

Comparisons are published in [Makridakis and Hibon, 2000].

M3 (199) cont.

The two best methods were not obviously "simple"

- Best method was "Theta", described by Assimakopoulos and Nikolopoulos [2000] in a highly complicated and confusing manner.
- Later Hyndman and Billah [2003] showed that the Theta method is equivalent to an average of a linear regression and simple exponential smoothing with drift.
- The 2nd best method was the commercial software package ForecastPro. The algorithm used is not public, but enough information has been revealed that we can be sure that it is not simple. The algorithm selects between an exponential smoothing model and an ARIMA model based on some state space approximations and a BIC calculation [Goodrich, 2000].
- The Box-Jenkins' ARIMA models did much better than in the previous competitions

M3 (1999) cont.

- International Journal of Forecasting Special Issue (2000)
- R package `Mcomp`: Data from the M-Competitions:
  - The 1001 time series from the M-competition
  - and the 3003 time series from the IJF-M3 competition

The M3 data have continued to be used since 2000 for testing new time series forecasting methods. In fact, unless a proposed forecasting method is competitive against the original M3 participating methods, it is difficult to get published in the IJF.[2]

[2] Hyndman [2017]

2017 – Im Westen nichts Neues

(. . .) two colleagues and myself submitted a paper[3] *for publication in Neural Networks. The paper was rejected without sent to referees, and we received the following report by the Action Editor: "Based on the contents of the paper, I think it does not contain enough contribution to be sent to possible reviewers. The paper basically presents a comparison of standard models, from the so-called machine learning group, with statistical models in forecasting time series benchmarks.* **There are many new machine learning models that have proved to overcome the results provided by statistical models, in many competitions, using the same benchmark datasets**. *Therefore, I recommend that the paper should be rejected." (. . .) I would like to thank "Neural Networks" for motivating me to start the M4 Competition*

Spyros Makridakis

[3] Makridakis, Spiliotis, and Assimakopoulos [2018]

M4 (2018)

- Announced in November 2017
- The competition started in Jan 1 2018 and ended in May 31 2018
- Initial results were published in the IJF on June 21, 2018.
- 100,000 real-life series selected randomly* from a database of 900,000 ones on December 28, 2017
- The minimum number of observations: 13 for yearly, 16 for quarterly, 42 for monthly, 80 for weekly, 93 for daily and 700 for hourly series,
- mainly from the Economic, Finance, Demographics and Industry areas, while also including data from Tourism, Trade, Labor and Wage, Real Estate, Transportation, Natural Resources and the Environment
- Forecasting Horizons: 6 for yearly data, 8 for quarterly, 18 for monthly, 13 for weekly, 14 for daily and 48 for hourly.

M4 (2018) – benchmark methods

- **well known, readily available, straightforward to apply**
- **whose computational requirements are minimal**

1. Naïve 1 ($\hat{y}_{T+h} = y_T$)
2. Seasonal Naïve
3. Naïve 2 (Naïve 1 seasonally adjusted)
4. Simple Exponential Smoothing (S)
5. Holt's Exponential Smoothing (H)
6. Dampen Exponential Smoothing (D)
7. Combining S-H-D The arithmetic average of methods 4, 5 and 6.
8. Theta
9. MLP – a perceptron of a very basic architecture and parameterization[4]
10. RNN – a recurrent network of a very basic architecture and parameterization[5]

[4] developed in Python + Scikit v0.19.1

[5] developed in Python + Keras v2.0.9 + TensorFlow v1.4.0

M4 (2018) cont.

Accuracy measures:

- OWA – Overall Weighted Average $= (sMAPE + MASE)/2$
- MASE – Mean Absolute Scaled Error[6]

$$MASE = \frac{1}{h} \frac{\sum_{i=1}^{h} |y_t - \hat{y}_t|}{\frac{1}{n-m} \sum_{t=m+1}^{n} |y_t - y_{t-m}|},$$

  - y_t – value of the time process at time t
  - $\hat{y}_t$ – estimated forecast of y_t
  - h – forecasting horizon
  - m – frequency of the data (eg. 12 for monthly series)
- The accuracy measures are computed for each horizon separately and then combined to cover, in a weighted fashion, all horizons together for each accuracy measure

[6] Hyndman and Koehler [2006]

M4 (2018) cont.

Two additions to the previous competitions:

- Participants are required to submit a detailed description of their approach, and **a source or execution file for reproducing the forecasts** (benchmark R code)
- Participants are encouraged (but not required) to produce prediction intervals evaluated using Mean Scaled Interval Score[7] (MSIS):

$$\frac{1}{h}\frac{\sum_{t=1}^{h}\left[(U_t - L_t) + \frac{2}{\alpha}(L_t - y_t)\mathbb{1}(y_t < L_t) + \frac{2}{\alpha}(y_t - U_t)\mathbb{1}(y_t > U_t)\right]}{\frac{1}{n-m}\sum_{t=m+1}^{n}|y_t - y_{t-m}|}$$

where

- $[L_t, U_t]$ is the $100(1-\alpha)\%$ prediction interval for time t,
- y_t is the observation at time t, $t = 1, \ldots, h$.

The competition used 95% prediction intervals, so $\alpha = 0.05$.

[7] Gneiting and Raftery [2007]

M4 (2018) – datasets & code

Frequency	Demogr.	Finance	Industry	Macro	Micro	Other	Total
Yearly	1 088	6 519	3 716	3 903	6 538	1 236	23 000
Quarterly	1 858	5 305	4 637	5 315	6 020	865	24 000
Monthly	5 728	10 987	10 017	10 016	10 975	277	48 000
Weekly	24	164	6	41	112	12	359
Daily	10	1 559	422	127	1 476	633	4 227
Hourly	0	0	0	0	0	414	414
Total	8 708	24 534	18 798	19 402	25 121	3 437	100 000

- Links on the webpage:
  https://www.mcompetitions.unic.ac.cy/the-dataset/
- R package: https://github.com/carlanetto/M4comp2018
- Methods code supplied by participants:
  https://github.com/M4Competition/M4-methods

M4 (2018) – critical remarks

- Rather than prediction intervals, participants could have been asked to provide full forecast distributions (e.g., by submitting the percentiles from 1% to 99%), and a probability scoring method such as CRPS could be used for evaluation, as was done in the GEFCom 2014 (Global Energy Forecasting Competition), for example.
- Even if we just stick to intervals, at least we could have a range of probability coverages (e.g., 50%, 80%, 95%, 99%) to give some more detailed idea of the forecast distribution in each case.
- It does not appear that there will be multiple submissions allowed over time, with a leaderboard tracking progress (as there is, for example, in a Kaggle competition ⓘ). This is unfortunate, as this element of a competition seems to lead to much better results. See Athanasopoulos and Hyndman [2011].

[Hyndman, 2017]

M4 (2018) – prizes & winners

- Best performing method according to OWA: 9000€
  Slawek Smyl (Uber Technologies) [Smyl, 2020]
- 2nd-best performing method according to OWA: 4000€
  Pablo Montero-Manso & team (University of Coruna & Monash)
  [Montero-Manso et al., 2020]
- 3rd-best performing method according to OWA, 2000€
  Maciej Pawlikowski (ProLogistica)
  [Pawlikowski and Chorowska, 2020]
- Best performing method according to MSIS, 5000€
  Prediction Intervals Prize: Slawek Smyl
- The Uber Student Prize, 5000€
  Pablo Montero-Manso
- The Amazon Prize 2000€
  The best reproducible forecasting method: Slawek Smyl

M4 (2018) – conclusions & legacy

- use sophisticated method to combine simple methods
- read International Journal of Forecasting Special Issue (2020)

"The "M" competitions organized by Spyros Makridakis have had an enormous influence on the field of forecasting.
They focused attention on what models produced good forecasts, rather than on the mathematical properties of those models.
For that, Spyros deserves congratulations for changing the landscape of forecasting research through this series of competitions."

[Hyndman, 2017]

source: https://mofc.unic.ac.cy/

M5 (2020)

- The competition will start on February 1, 2020
- The participants are asked to submit their forecasts no later than June 31, 2020 before midnight.
- Hierarchical sales data provided by Walmart

Area	California	Texas	Wisconsin	Total
Stores	4	3	3	10
Departments	28	21	21	70
SKUs[8]	39 965	29 900	29 988	99 853
Total	39 998	29 925	30 013	99 937

- Information on explanatory variables
- Forecast: point forecast, 4 prediction intervals and median
- Competition platform: Kaggle

source: https://mofc.unic.ac.cy/m5-competition/

[8] Stock Keeping Unit

Thank you!

References

Jon Scott Armstrong. *Long-range forecasting : from crystal ball to computer*. Wiley, 1978. ISBN 0471822604.

J.Scott Armstrong and Fred Collopy. Error measures for generalizing about forecasting methods: Empirical comparisons. *International Journal of Forecasting*, 8(1):69–80, jun 1992. ISSN 0169-2070. doi: 10.1016/0169-2070(92)90008-W. URL https://www.sciencedirect.com/science/article/abs/pii/016920709290008W.

V. Assimakopoulos and K. Nikolopoulos. The theta model: a decomposition approach to forecasting. *International Journal of Forecasting*, 16(4):521 – 530, 2000. ISSN 0169-2070. doi: https://doi.org/10.1016/S0169-2070(00)00066-2. URL http://www.sciencedirect.com/science/article/pii/S0169207000000662.

George Athanasopoulos and Rob J. Hyndman. The value of feedback in forecasting competitions. *International Journal of Forecasting*, 27 (3):845 – 849, 2011. ISSN 0169-2070. doi: https://doi.org/10.1016/j.ijforecast.2011.03.002. URL http://www.sciencedirect.com/science/article/pii/S0169207011000495.

Robert Fildes and Spyros Makridakis. The Impact of Empirical Accuracy Studies on Time Series Analysis and Forecasting. *International Statistical Review*, 63(3):289, dec 1995. ISSN 03067734. doi: 10.2307/1403481. URL https://www.jstor.org/stable/1403481?origin=crossref.

Tilmann Gneiting and Adrian E Raftery. Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, 102(477):359–378, mar 2007. ISSN 0162-1459. doi: 10.1198/016214506000001437. URL http://www.tandfonline.com/doi/abs/10.1198/016214506000001437.

Robert L. Goodrich. The Forecast Pro methodology. *International Journal of Forecasting*, 16(4):533–535, oct 2000. ISSN 0169-2070. doi: 10.1016/S0169-2070(00)00086-8. URL https://www.sciencedirect.com/science/article/abs/pii/S0169207000000868.

Rob J. Hyndman. M4 forecasting competition, 2017. URL https://robjhyndman.com/hyndsight/m4comp/.

Rob J. Hyndman and Baki Billah. Unmasking the theta method. *International Journal of Forecasting*, 19(2):287 – 290, 2003. ISSN 0169-2070. doi: https://doi.org/10.1016/S0169-2070(01)00143-1. URL http://www.sciencedirect.com/science/article/pii/S0169207001001431.

Rob J. Hyndman and Anne B. Koehler. Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4): 679–688, oct 2006. ISSN 0169-2070. doi: 10.1016/J.IJFORECAST.2006.03.001. URL https://www.sciencedirect.com/science/article/abs/pii/S0169207006000239.

S. Makridakis, A. Andersen, R. Carbone, R. Fildes, M. Hibon, R. Lewandowski, J. Newton, E. Parzen, and R. Winkler. The accuracy of extrapolation (time series) methods: Results of a forecasting competition. *Journal of Forecasting*, 1(2):111–153, apr 1982. ISSN 02776693. doi: 10.1002/for.3980010202. URL http://doi.wiley.com/10.1002/for.3980010202.

Spyros Makridakis and Michele Hibon. The m3-competition: results, conclusions and implications. *International Journal of Forecasting*, 16 (4):451 – 476, 2000. ISSN 0169-2070. doi: https://doi.org/10.1016/S0169-2070(00)00057-1. URL http://www.sciencedirect.com/science/article/pii/S0169207000000571.

Spyros Makridakis, Michele Hibon, and Claus Moser. Accuracy of forecasting: An empirical investigation. *Journal of the Royal Statistical Society. Series A (General)*, 142(2):97–145, 1979. ISSN 0035-9238. doi: https://doi.org/10.2307/2345077. URL http://www.jstor.org/stable/2345077.

Spyros Makridakis, Chris Chatfield, Michèle Hibon, Michael Lawrence, Terence Mills, Keith Ord, and LeRoy F. Simmons. The M2-competition: A real-time judgmentally based forecasting study. *International Journal of Forecasting*, 9(1):5–22, apr 1993. ISSN 0169-2070. doi: 10.1016/0169-2070(93)90044-N. URL https://www.sciencedirect.com/science/article/abs/pii/016920709390044N.

Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. Statistical and Machine Learning forecasting methods: Concerns and ways forward. *PLOS ONE*, 13(3):e0194889, mar 2018. ISSN 1932-6203. doi: 10.1371/JOURNAL.PONE.0194889. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0194889.

Spyros Makridakis, Rob J. Hyndman, and Fotios Petropoulos. Forecasting in social settings: The state of the art. *International Journal of Forecasting*, 36(1):15–28, jan 2020. ISSN 01692070. doi: 10.1016/j.ijforecast.2019.05.011. URL https://linkinghub.elsevier.com/retrieve/pii/S0169207019301876.

Pablo Montero-Manso, George Athanasopoulos, Rob J. Hyndman, and Thiyanga S. Talagala. FFORMA: Feature-based forecast model averaging. *International Journal of Forecasting*, 36(1):86–92, jan 2020. ISSN 01692070. doi: 10.1016/j.ijforecast.2019.02.011. URL https://linkinghub.elsevier.com/retrieve/pii/S0169207019300895.

Maciej Pawlikowski and Agata Chorowska. Weighted ensemble of statistical models. *International Journal of Forecasting*, 36(1):93–97, jan 2020. ISSN 01692070. doi: 10.1016/j.ijforecast.2019.03.019. URL https://linkinghub.elsevier.com/retrieve/pii/S0169207019301190.

Slawek Smyl. A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting. *International Journal of Forecasting*, 36(1):75–85, jan 2020. ISSN 01692070. doi: 10.1016/j.ijforecast.2019.03.017. URL https://linkinghub.elsevier.com/retrieve/pii/S0169207019301153.

sMAPE

Symmetric mean absolute percentage error:

$$\text{sMAPE} = \text{mean}\left(\frac{2|y_t - \hat{y}_t|}{|y_t| + |\hat{y}_t|}\right).$$

- Armstrong [1978], p. 348
- Hyndman and Koehler [2006] recommend to not use sMAPE ←

Average Ranking

For each series, the "Average Rankings" are computed by sorting, for each forecasting horizon, the symmetric absolute percentage error of each method from the smallest (taking the value of 1) to the largest. Consequently, once the ranks for all series have been determined, the mean rank is calculated for each forecasting horizon, over all series. An overall average ranking is also calculated by averaging the ranks over six, eight or 18 forecasts, for each method. ←

Percentage Better

The "Percentage Better" measure counts and reports the percentage of time that a given method has a smaller forecasting error than another method. Each forecast made is given equal weight. ←

Median symmetric absolute percentage error

The median symmetric absolute percentage error is found and reported for each method/forecasting horizon. Such a measure is not influenced by extreme values and is more robust than the average absolute percentage error. In the case of the M3-Competition the differences between symmetric MAPEs and Median symmetric APEs were much smaller than the corresponding values in the M-Competition as care has been taken so that the level of the series not be close to zero while, at the same time, using symmetric percentage errors which reduce their fluctuations. ←

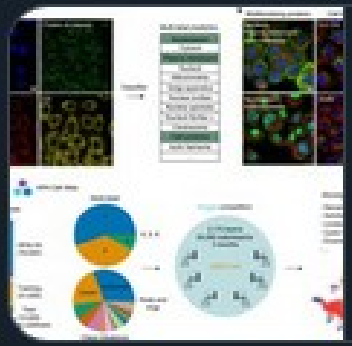
Median relative absolute error

The RAE is the absolute error for the proposed model relative to the absolute error for the Naive2 (no-change model). It ranges from 0 (a perfect forecast) to 1.0 (equal to the random walk), to greater than 1 (worse than the random walk). The RAE is similar to Theil's U2, except that it is a linear rather than a quadratic measure. It is designed to be easy to interpret and it lends itself easily to summarizing across horizons and across series as it controls for scale and for the difficulty of forecasting. The Median RAE (MdRAE) is recommended for comparing the accuracy of alternative models as it also controls for outliers (for information on the performance of this measure, see [Armstrong and Collopy, 1992]. ←

Kaggle

Bojan Tunguz
@tunguz

I am proud to announce that @naturemethods has published the paper based on the @Kaggle Human Protein Atlas competition. I want to again congratulate and thank my wonderful teammates, competition organizers, and Kaggle on this great achievement:



Analysis of the Human Protein Atlas Image Classification com...
The 2018 Human Protein Atlas Image Classification competition sought to improve automated classification of protein
nature.com

5:23 pm · 28 Nov 2019 · Twitter Web App

←

Advertising

- Rob J. Hyndman and George Athanasopoulos, *Forecasting: Principles and Practice*
  - "version 2": base + forecast
  - in developement: fpp3 (fable, feasts, tsibble)
- DataCamp: Forecasting Using R