International Journal of Forecasting 36 (2020) 212–216

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Discussion

Correlation analysis of forecasting methods: The case of the M4 competition

Pantelis Agathangelou, Demetris Trihinas, Ioannis Katakis [*]

Department of Computer Science University of Nicosia, Cyprus

ARTICLE INFO

Keywords:
Forecasting
M4
Makridakis
Machine learning
Statistics

ABSTRACT

This commentary introduces a correlation analysis of the top-10 ranked forecasting methods that participated in the M4 forecasting competition. The “M" competitions attempt to promote and advance research in the field of forecasting by inviting both industry and academia to submit forecasting algorithms for evaluation over a large corpus of real-world datasets. After performing the initial analysis to derive the errors of each method, we proceed to investigate the pairwise correlations among them in order to understand the extent to which they produce errors in similar ways. Based on our results, we conclude that there is indeed a certain degree of correlation among the top-10 ranked methods, largely due to the fact that many of them consist of a combination of well-known, statistical and machine learning techniques. This fact has a strong impact on the results of the correlation analysis, and therefore leads to similar forecasting error patterns.

1. M4 competition data and forecasts

The M4 competition extended the “M” competition series by introducing a large dataset comprising 100,000 time series, organized in six distinct categories based on the sampling interval (hourly, daily, weekly, monthly, quarterly, and yearly) and originating from a diverse range of domains (e.g., finance, demographics, industry). Given a set of known (training) data for each of the 100,000 time series, the competitors were requested to provide a specific number of forecasts for each one. We evaluate and rank the submitted methods using the overall weighted average (OWA) of two accuracy measures: the mean absolute scaled error (MASE) and the symmetric mean absolute percentage error (sMAPE). These measures are calculated as follows:

$$\mathrm{sMAPE} = \frac{1}{h}\sum_{t=1}^{h}\frac{2|Y_t - \hat{Y}_t|}{|Y_t| + |\hat{Y}_t|} \tag{1}$$

$$\mathrm{MASE} = \frac{1}{h}\frac{\sum_{t=1}^{h}|Y_t - \hat{Y}_t|}{\frac{1}{n-m}\sum_{t=m+1}^{n}|Y_t - Y_{t-m}|}, \tag{2}$$

where Y_t is the actual value at point t, $\hat{Y}_t$ is the estimated forecast, h is the forecasting horizon and m is the data frequency (M4 Team, 2018).

Table 1 presents the results of the top-10 methods along with two evaluation metrics that denote the percentage improvement of the respective method over the M4 benchmark, which was a combination of three statistical models (simple exponential smoothing, Holt’s exponential smoothing, and damped exponential smoothing). In the rest of this commentary, we refer to each method as ⟨rank⟩⟨first_author_name⟩ for ease of communication of the name of each method along with its performance.

[*] Corresponding author.
E-mail addresses: agathangelou.p@live.unic.ac.cy (P. Agathangelou), trihinas.d@unic.ac.cy (D. Trihinas), katakis.i@unic.ac.cy (I. Katakis).

https://doi.org/10.1016/j.ijforecast.2019.03.020

Table 1
The top-10 ranked methods according to Makridakis, Spiliotis, and Assimakopoulos (2018), with % denoting the improvement relative to the M4 benchmark.

id	Rank	Code name	Author(s)	Affiliation	sMAPE	OWA
118	1	1-Smyl	Smyl, S.	Uber Technologies	9.4%	8.6%
245	2	2-Manso	Montero-Manso, P., Talagala, T., Hyndman, R. J. & Athanasopoulos, G.	University of A Corua & Monash University	6.6%	6.7%
237	3	3-Pawlikowski	Pawlikowski, M., Chorowska, A. & Yanchuk, O.	ProLogistica Soft	5.7%	6.3%
72	4	4-Jaganathan	Jaganathan, S. & Prakash, P.	Individual	6.8%	6.2%
69	5	5-Fiorucci	Fiorucci, J. A. & Louzada, F	University of Brasilia & University of São Paulo	5.7%	6.1%
36	6	6-Petropoulos	Petropoulos, F. & Svetunkov, I.	University of Bath & Lancaster University	5.3%	5.6%
78	7	7-Shaub	Shaub, D.	Harvard Extension School	4.3%	4.2%
260	8	8-Zampeta	Zempeta-Legaki, N. & Koutsouri, K.	National Technical University of Athens	4.5%	4.1%
238	9	9-Doornik	Doornik, J., Castle, J. & Hendry, D.	University of Oxford	5.0%	3.7%
39	10	10-Pedregal	Pedregal, D. J. & Trapero, J.	University of Castilla-La Mancha	3.5%	3.2%

The main goal of this commentary is to investigate and analyze the results of the M4 competition in order to provide a comprehensive understanding of its outcome and the potential impact on the forecasting field. More specifically, we investigate the degree of pairwise correlation among the top-10 methods in order to help us to understand the extent to which they produce errors in similar patterns and whether these patterns are the by-product of the underlying ensemble of models that were selected to form part of each competitor's forecasting module. This study concentrates on point predictions and does not consider prediction intervals.

The data that we exploit for the analysis are as follows:

1. The dataset of time series used for the M4 competition.[1]
2. The forecasts of the participating methods (made available to us by the M4's organizers).
3. The future data for all time series from the competition dataset (also publicly available).

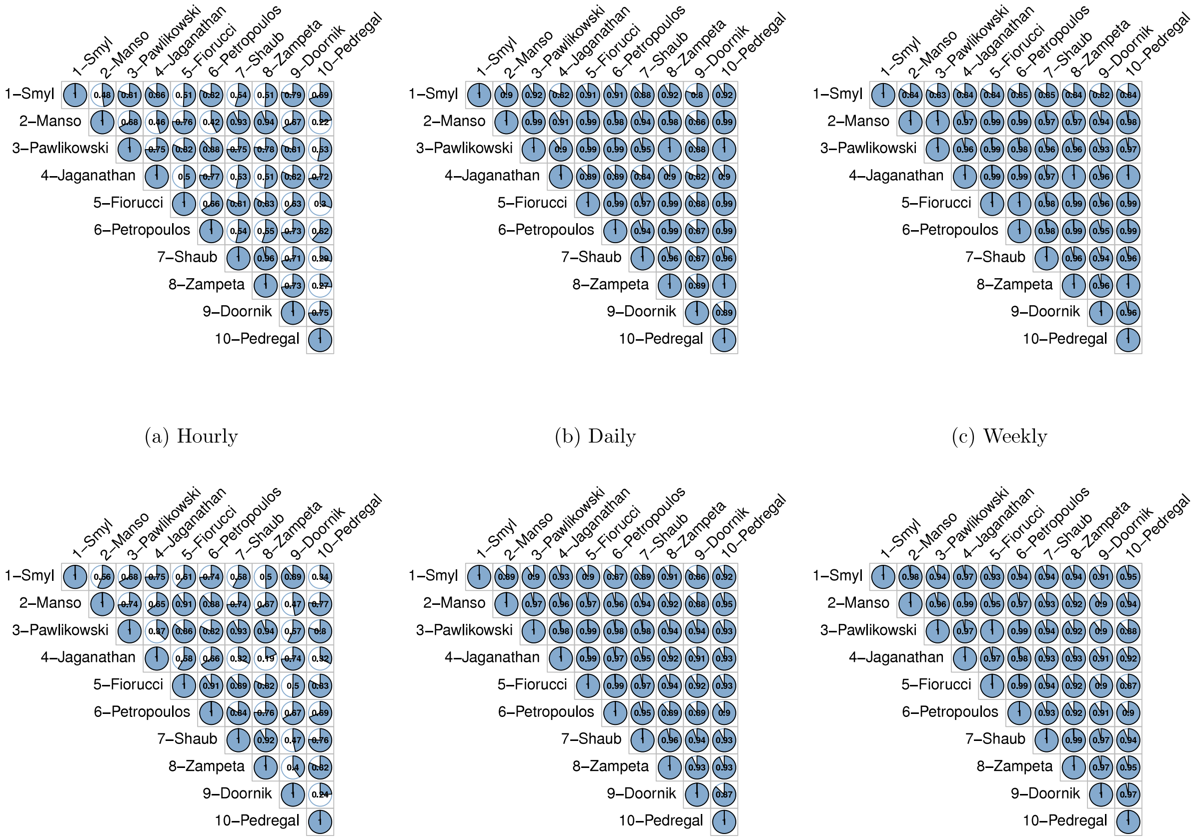
Before performing the correlation analysis, we began by introducing a pre-processing step. In particular, for the top-10 methods, denoted by M_i, the datapoint error percentage[2] for the respective forecasting horizon, denoted by H, was derived from the future data comprising the N time series per category (e.g., hourly, monthly). This outputs an $M_i^{err} = N \times H$ forecasting error percentage matrix for each method and time series category. Having compiled the forecasting error percentage matrices for the top-10 methods, the Pearson correlation coefficient is then derived in order to measure the average bi-variate correlation among the methods. Fig. 1 depicts the results of the correlation analysis visually for each time series category.

It is evident immediately from this figure that the top-10 methods present different pairwise correlations for the hourly, monthly, and (partly) daily time series, which featured the longest forecasting horizons ($h_{hourly} = 48$, $h_{monthly} = 18$, $h_{daily} = 14$), resulting in the methods returning a wide range of different forecasts for these particular time series. This urged us to investigate further how correlated methods structured their forecasting modules. Specifically, 2-Manso is correlated strongly with 3-Pawlikowski, 7-Shaub, 5-Fiorucci and 10-Pedregal. This is due to 2-Manso utilizing a forecasting module of a combination of nine statistical and neural network models (Hyndman & Athanasopoulos, 2013), including an ARIMA and a Theta-based model, which are also adopted by 7-Shaub. Similarly, 2-Manso is correlated strongly with 3-Pawlikowski and 5-Fiorucci, both of which also adopt combinations of several purely statistical models. In turn, 5-Fiorucci and 8-Zampeta are also correlated strongly, with 8-Zampeta embracing a purely statistical forecasting approach by utilizing a Box–Cox transformation and a Theta-based model. In the same way, 3-Pawlikowski presents strong correlations with 7-Shaub and 8-Zampeta due to it embracing both an ARIMA and a Theta-based model, which, as has been mentioned, are used by 7-Shaub and 8-Zampeta respectively.

Moreover, 1-Smyl is correlated with 4-Jaganathan, 6-Petropoulos and 9-Doornik, due to the fact that these methods all exploit forecasting approaches that model the seasonal trends that are highly evident in the hourly and monthly time series. Nonetheless, 1-Smyl appears to be correlated weakly with all of the other top-10 methods for the weekly, quarterly and monthly time series. This is due to the unique approach adopted in 1-Smyl, where not all of the data points comprising these time series were exploited. As the author says, "Stepping through 300 years of data to forecast six years seems excessive". Therefore, 1-Smyl only considered a 60-year history for the yearly time series, a 20-year horizon for the monthly time series and 40 years for the quarterly time series. On the other hand, 10-Pedregal appears strongly correlated with different methods depending on the time series (e.g., hourly, yearly). This is due to the forecasting approach adopted in 10-Pedregal, where different models are embraced depending on the time series type. For the yearly and quarterly time series, a Theta-based model is adopted, while an ARMA model is used subsequently on the residuals. As such, 10-Pedregal is correlated strongly with 2-Manso and 8-Zampeta for these time series, whereas for the monthly, weekly and daily time series, it takes the normalized mean from the

[1] Publicly available at https://www.m4.unic.ac.cy/the-dataset/

[2] Denoted by $(y - \hat{y})/y$.



(a) Hourly (b) Daily (c) Weekly

(d) Monthly (e) Quarterly (f) Yearly

Fig. 1. The average correlation matrix for all methods.

best benchmarks introduced by the competition organization, which are pure statistical models, and therefore is correlated strongly with 2-Manso, 5-Fiorucci and 7-Shaub. In turn, 10-Pedregal adopts a seasonal model for the hourly time series that is also embraced in the forecasting modules of 1-Smyl, 4-Jaganathan and 9-Doornik.

Fig. 2 provides supplemental information from the above correlation study. It depicts the visualization of the average forecasting error percentage matrices, introduced earlier in the pre-processing step. Overall, the similarities and differences generally align with the comments made above for the top-10 ranked methods and the respective categories.

2. Conclusion and future work

As has been mentioned, most of the top-10 ranked methods in the M4 competition employed either a combination of purely statistical models or a combination of statistical models with machine learning. However, the method which made the most significant improvement over the M4 benchmark, as indicated by the (OWA) evaluation (see Table 1), was 1-Smyl.

1-Smyl's model architecture and implementation were based on an innovative formula that consisted of two main parts: an appropriate neural network and a combination of statistical models. In general, such model implementations encompass two machine learning characteristics. The neural part ensures global feature extraction and the statistical part achieves local feature extraction. The collaboration of these two approaches in a single unit resulted in an overall improvement on the forecasting task relative to other, mainly pure statistical or pure neural, models. We attribute this improvement to the fact that global features took into account patterns that happened in the past, outside the scope of local features. An analysis of the error results over certain time series showed that there do exist such pattern combinations, past with local, that can reduce the accumulated error if exploited appropriately. The truncated input series technique of 1-Smyl was in line with this spirit. The author exploited a valid "past" along with local patterns in a way that reduced the time step error and contributed to an overall improvement on the forecasting task.

Overall, since prediction is all about reducing the uncertainty, such hybrid modules, which exploit the

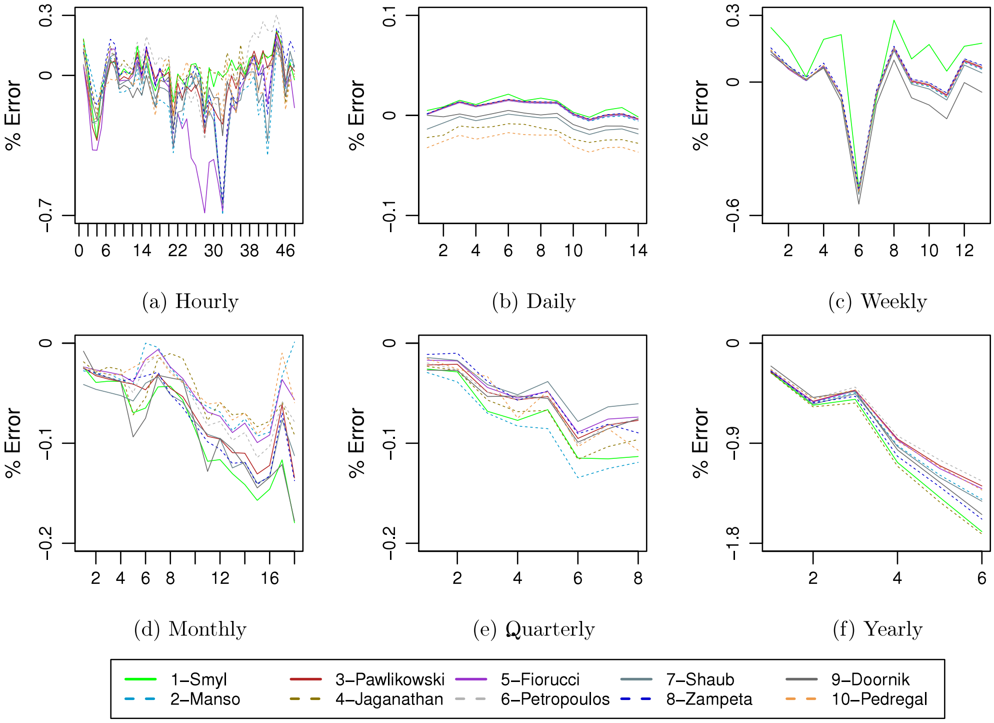


Fig. 2. Average forecasting time step error percentage for all categories and the top-10 methods. The *x*-axis represents the forecasting horizon, and a negative error percentage indicates that the method forecasts a value lower than the actual.

advantages of both deep learning and well-known statistical methods, may potentially help to obtain better forecasting solutions.

This commentary has analysed and studied the top-10 ranked forecasting methods that participated in the M4 competition from a correlation analysis perspective. We observed that the greatest similarities were noted at the shorter forecasting horizons, whereas significant differences were seen at the longest. We provided a detailed analysis of the structures of the methods and their respective similarities and differences. Finally, we provided some high-level insights that we believe could possibly inspire the forecasting community in future research activities.

References

Hyndman, R. J., & Athanasopoulos, G. (2013). *Forecasting: principles and practice*. OTexts.

M4 Team (2018). *M4 competitor's guide: prizes and rules*. See https://www.m4unic.ac.cy/wp-content/uploads/2018/03/M4-CompetitorsGuide.pdf.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). The M4 competition: Results, findings, conclusion and way forward. *International Journal of Forecasting*, *34*, 802–808.

Pantelis Agathangelou holds an Msc in Information Systems from the Faculty of Pure and Applied Sciences, Open University of Cyprus. His research interests include Data Mining, Pattern Classification, Sentiment Analysis, Mining Social Networks and Artificial intelligence. He has co-developed DidaxTo (http://deixto.com/didaxto/) a tool that implements an unsupervised approach for discovering patterns that will extract a domain-specific dictionary from product reviews. He is currently a Ph.D. Candidate at the University of Nicosia Dept of Computer Science.

Demetris Trihinas is currently a Faculty Member at the Department of Computer Science, University of Nicosia. Previously, I was a Postdoctoral Fellow at the Computer Science Department, University of Cyprus. My research interests include Distributed and Internet Computing with particular focus in Big Data Analytics, Data Visualization, and Cloud and Edge Computing. My Ph.D. dissertation targeted developing low-cost probabilistic and adaptive learning models for approximate monitoring in order to improve energy-efficiency and reduce both the volume and velocity of data generated and consumed by IoT services offered through the cloud. For the research conducted during my doctoral studies, I was selected by the Heidelberg Laureate Forum as one of the 100 young promising researchers in Computer Science for 2015. I additionally hold a Computer Science MSc from the University of Cyprus and a Dipl-Ing in Electrical and Computer Engineering from the National Technical University of Athens (NTUA). I have extensive experience in European research and innovation projects where I participated as a Work Package Leader and Senior Researcher in multiple projects (e.g., Unicorn H2020, PaaSport FP7, CELAR FP7) funded under the European Commission FP7 and H2020 schemes. Additionally, I'm the developer of the JCatascopia cloud monitoring system, the AdaM framework for IoT devices and also an active member of the Cloud Application Management Framework (CAMF) which is an official Eclipse Foundation project. My work is published in IEEE/ACM journals and conferences such as Trans. Services Computing (TSC), Trans. Cloud Computing (TCC), ICDCS, INFOCOM, BigData, CCGrid, and ICSOC.

Dr. Ioannis Katakis is an Associate Professor and co-Director of the Artificial Intelligence Lab () at Sthe Computer Science Department of the University of Nicosia. Born and raised in Thessaloniki, Greece, studied Computer Science and holds a Ph.D. in Machine Learning for Text Classification (Aristotle University, 2009). After his post-graduate studies, he served multiple universities as a lecturer and a senior researcher. His research interests include Mining Social, Web and

Urban Data, Smart Cities, Sentiment Analysis and Opinion Mining, Deep Learning, Data Streams and Multi-label Learning. He published papers in International Conferences and Scientific Journals related to his areas of expertise (ICDE, CIKM, ECML/PKDD, IEEE TKDE, ECAI), organized multiple workshops (at KDD, ICML, ECML/PKDD, EDBT/ICDT), and special issues (DAMI, InfSys) and is an Editor at the journal Information Systems. His research has been cited more than 5000 times in the literature. Recently, Ioannis has been included in the list of top young Greek scientists based on the impact of his work. He has extensive experience in European research projects where he participated as a Quality Assurance Coordinator and a Senior Researcher. He regularly serves the program committee of international conferences (ECML/PKDD, WSDM, AAAI, IJCAI) and evaluates articles in top-tier journals (TPAMI, DMKD, TKDE, TKDD, JMLR, TWEB, ML).