International Journal of Forecasting 38 (2022) 1519–1525

Contents lists available at ScienceDirect

International Journal of Forecasting

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

Fathoming empirical forecasting competitions' winners

Azzam Alroomi [b], Georgios Karamatzanis [a], Konstantinos Nikolopoulos [a,*], Anna Tilba [a], Shujun Xiao [a]

[a] Durham University, United Kingdom
[b] Arab Open University, Kuwait

ARTICLE INFO

Keywords:
Forecasting
Competitions
Performance
Machine learning
Benchmarks

ABSTRACT

The M5 forecasting competition has provided strong empirical evidence that machine learning methods can outperform statistical methods: in essence, complex methods can be more accurate than simple ones. Regardless, this result challenges the flagship empirical result that led the forecasting discipline for the last four decades: keep methods sophisticatedly simple. Nevertheless, this was a first, and we can argue that this will not happen again. There has been a different winner in each forecasting competition. This inevitably raises the question: can a method win more than once (and should it be expected to)? Furthermore, we argue for the need to elaborate on the perks of competing methods, and what makes them winners?

1. Introduction

The M5 Competition has provided strong empirical evidence that machine learning (ML) methods can outperform statistical methods. This result largely contradicts the major finding of previous forecasting competitions (Petropoulos, Apiletti, Assimakopoulos, Babai, et al., 2022), which was as follows:

simple methods are on par with (more) complex ones,

with the debatable exception of M4 where hybrid approaches and combinations prevailed (Barker, 2020; Gilliland, 2020).

Nevertheless, this is a first for pure ML methods, after many respective competitions,[1] and as such, one could argue this will not happen again. In fact, in each M Competition, there has been a different winner, which inevitably raises the question: can a method win more than once?

Even more frustrating, forecasters frequently do not understand why their methods are successful. Thus, in future forecasting competitions, we may be tempted to ask in advance competitors to submit a plausible explanation as to why their method is expected to work well on the specific dataset.

The M5 competition was another attempt to examine new features of time series forecasting methods performance, following recommendations from many scholars in the field, including Hong (2020) and Fry and Brundage (2020). M5 focused on forecasting the hierarchical high-frequency unit sales of 42,840 time series of Walmart.

It is very important for the discipline that a large number of participants took part, and that an undergraduate student won it rather than a seasoned forecaster. Furthermore, the good news is that ML methods had their first clear win, with a LightGBM claiming the victory (Makridakis, Spiliotis, & Assimakopoulos, 2020b). This came to

* Corresponding author.
E-mail address: kostas.nikolopoulos@durham.ac.uk (K. Nikolopoulos).

[1] ML methods were not tested in all M competitions, so there is a school of thought arguing that, nowadays, ML is only properly tested in empirical forecasting competitions, illustrating their full potential.

https://doi.org/10.1016/j.ijforecast.2022.03.010

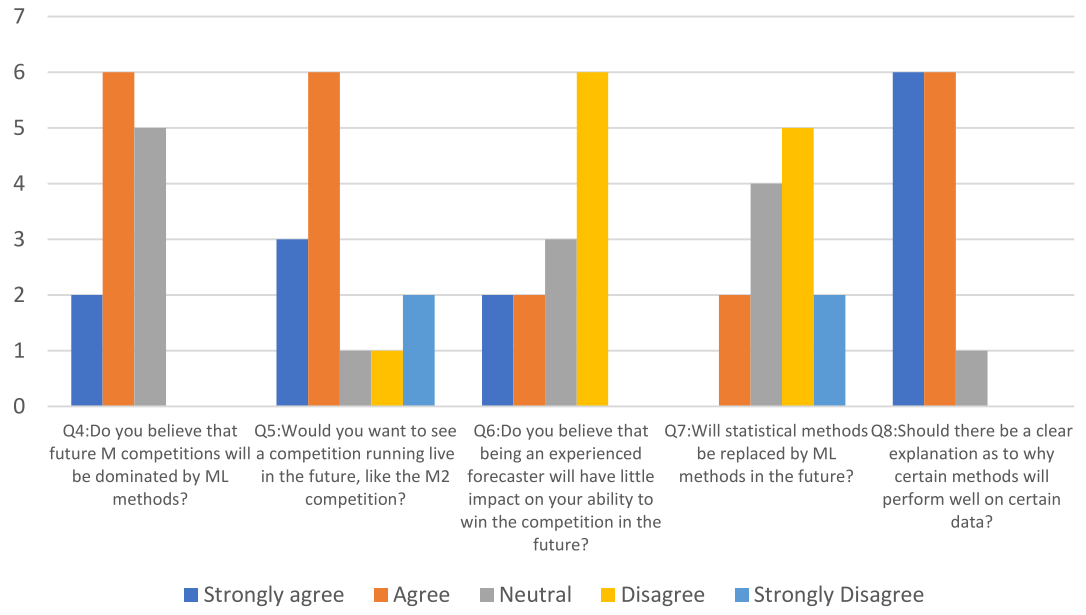

Fig. 1a. Answers for each 5-points Likert scale question (from Questions 4–8).

the surprise of many scholars in the field, for example, Green and Armstrong (2015) who have claimed that

> "*no matter what type of forecasting methods used, complexity harms accuracy" (p. 1684).*

2. What makes a winner?

In the M4 Competition, only 17 of the top 50 top-performing methods shared information about their forecasting approaches (Makridakis, Spiliotis, & Assimakopoulos, 2020a). As stated in Armstrong and Green (2019), the participants should explain the principles and methods used which would allow the organizers, commentators, and researchers to understand what makes them perform better (or not).

There is also the long-standing question of why each competition has a different winner? Common sense suggests that no one method is likely to be dominant in all circumstances. We can argue, and the experts we interviewed were aligned to this view, that we cannot expect the results of all empirical competitions and studies will be the same: methods are improving, new ideas are introduced, and therefore new methods prevail. For example, in the case of the M4, cross-learning and cross-validation made the difference. Equally interesting is whether a broad category of methods tends to perform significantly better than others in given contexts, and why?

One might even argue what the M competitions have to offer, apart from testing "new" (and retesting "old") methods against a set of new data; at the end of the day, the top-performing methods in the competition might not be the best method available on the market, but the best at capturing the nature of the data, metrics, and benchmarks (Bontempi, 2020),

3. Our (new) empirical qualitative data on the M5 competition

We decided to collect new primary data for our paper to capture the views of the field on the matter, rather than just providing solely our views for the results of the M5.

We conducted an online survey consisting of 20 questions (Table 1). The instrument includes one question (the first) to grant consent, fifteen 5-point Likert scale questions ("strongly agree", "agree", "neutral", "disagree", and "strongly disagree"), and four open-ended questions (100–150 words).

An invitation was sent via email on 2nd February 2021 to a purposeful – but quite diverse – sample of 73 experts who contributed to the state-of-the-art review paper, Forecasting Theory and Practice (Petropoulos et al., 2022). A total of 17 responses were received resulting in a response rate of 23.3%.

To complement our survey, we also conducted a series of seven structured interviews from a convenience sample of leading experts in the field, including past Editors of the International Journal of Forecasting (Appendix); such interviews are particularly suitable in situations where the researcher seeks to understand the interviewee's perspective on the topic (Easterby-Smith, Thrope, & Lowe, 2002). Our qualitative approach was inductive and interpretive in nature (Hudson & Ozanne, 1988; Saunders, Lewis, & Thornhill, 2016). Interviews took place between the 15th of February and the 3rd of March 2021 online via Zoom. Each interview lasted around 20 minutes. All interviews were recorded and transcribed.

4. Empirical findings from the survey

Our survey results are presented in Figs. 1a–1c. We hereafter discuss the ones we find more interesting,

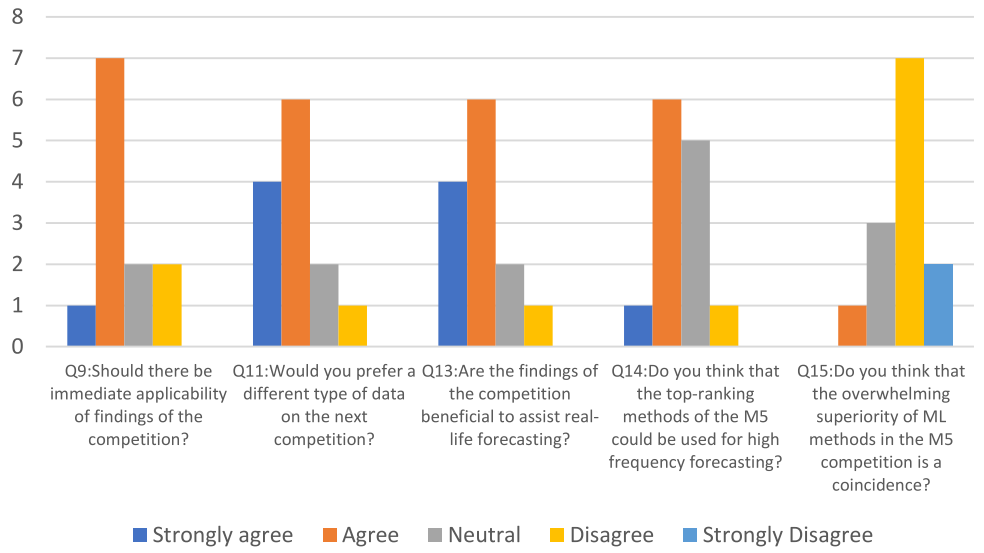

Fig. 1b. Answers for each 5-points Likert scale question (from Questions 9–15).

Table 1
Online survey questions.

(1)	I consent for the answers that I provide for this survey to be analyzed and used as part of this discussion paper
(2)	Why do you think machine learning (ML) methods performed better than statistical methods for the first time in the M5 competition? (100–150 words)?
(3)	Why do you think there is a different winner in every M competition (100–150 words)?
(4)	Do you believe that future M competitions will be dominated by ML methods?
(5)	Would you want to see a competition running live in the future, like the M2 competition?
(6)	Do you believe that being an experienced forecaster will have little impact on your ability to win the competition in the future?
(7)	Will statistical methods be replaced by ML methods in the future?
(8)	Should there be a clear explanation as to why certain methods will perform well on certain data?
(9)	Should there be immediate applicability of findings of the competition?
(10)	Should there be immediate applicability of findings of the competition and if so, how? (100–150 words)
(11)	Would you prefer a different type of data for the next competition?
(12)	If you would prefer a different type of data, what type of data would it be? (100–150 words)
(13)	Are the findings of the competition beneficial to assist real-life forecasting?
(14)	Do you think that the top-ranking methods of the M5 could be used for high-frequency forecasting?
(15)	Do you think that the overwhelming superiority of ML methods in the M5 competition is a coincidence?
(16)	Do you believe the pure ML methods could be used in other areas?
(17)	Do you think ML methods are the best way for applying "cross-learning" in forecasting applications?
(18)	Do you believe "cross-learning" is the precondition for the optimization of ML methods?
(19)	Do you agree that the success of ML methods in the M5 is due to the specific dataset?
(20)	Do you believe that with the prevalent use of computer science, ML methods will become more important in identifying and extrapolating data patterns?

nevertheless, all results are included in the three tables for the readers' further consideration.

In Fig. 1a, at Q5 (*Would you want to see a competition running live in the future, like the M2 competition?*), the experts pose different views as to whether a live competition should be running in the future. Although most experts are in favor (6 experts "agree" and 3 experts "strongly agree" which account for 46.15% and 23.08%, respectively), there are 2 experts that "strongly disagree"; "disagree" and "neutral" gained 1 vote, respectively.

In Q7 (*Will statistical methods be replaced by ML methods in the future?*), although answers are more spread, still there is a good deal of disagreement between experts in the field.

For Q6 (for expertise been not that important for forecasting) and Q8 (on the need for forecasters to explain how their methods work), there was more consensus as per the following figure.

There is a good level of consensus for all questions Q9 to Q15 as per Fig. 1b with an absence of experts strongly disagreeing (and strongly agreeing in Q15). So, a good representative sample of experts in the field, by and large, believe that the result of forecasting competitions should be applied in practice (Q9 and Q13), would like different types of data used in each competition (Q11), think that the winners of the M5 could deal with high-frequency data (Q14), and do not think that the win of ML methods is a one-off (Q15 and Q4).

In Fig. 1c, there is a good level of consensus for questions Q16 & Q20 (on the use and importance of ML in other application areas) while for Q17 and Q18 (ML and

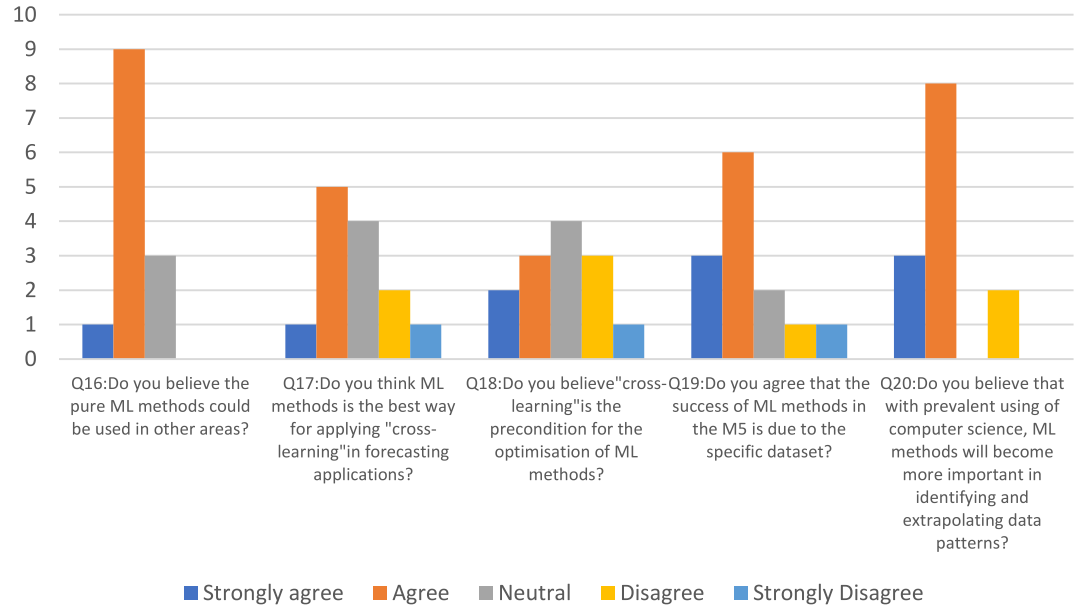

Fig. 1c. Answers for each 5-points Likert scale question (from Questions 16–20).

Table 2
Frequency of keywords in the answers of the open-ended questions of the survey.

	Q2: Why do you think machine learning (ML) methods performed better than statistical methods for the first time ever in the M5 competition?	Q3: Why do you think there is a different winner in every M competition?	Q10: Should there be immediate applicability of findings of the competition and if so, how?	Q12: If you would prefer different types of data, what type of data would it be?
Dataset	60%	44%	40%	
Statistical methods	20%	25%		
No best method for all situations	10%	06%		
Improvement of methods	10%	06%		
Random elements/other elements		19%		
Accuracy/uncertainty			20%	
Decision-making/planning			40%	
Hierarchical series				25%
Mix type data				12%
Financial aspect				25%
Sports, election forecasting				12%
Classification problems				13%
High-frequency data				13%

cross-learning), and Q19 (M5 dataset favoring ML methods) views are varying.

In Table 2, we attempt a basic frequency count of "keywords" appearing in our open-ended survey questions (Q2, Q3, Q10, and Q12). The most interesting result is that the most frequent keyword appearing in the answers of Q2 (*Why do you think machine learning (ML) methods performed better than statistical methods for the first time ever in the M5 competition?*) and Q3 (*Why do you think there is a different winner in every M competition?*) is "**dataset**". Thus, the overwhelming majority of experts believe the reason why ML methods performed better in M5, and why each competition has a different winner, is due to the specific dataset used in each competition. Furthermore, some experts mentioned that the methods over the years have been improved, and there is no unique method fit for every forecasting task.

5. Empirical findings from the interviews

This section presents an empirical analysis of our interviews, with the caveat that, since our protocol had just

five questions, one could argue that maybe a follow-up study could achieve a more in-depth assessment of our intended quest. Content analysis was used in our quest and the respective analysis of the interview transcripts.

To ensure the good quality of data analysis, techniques similar to those used by Eisenhardt (1989) and Tilba and McNulty (2013) were employed. The analysis was done in several steps. Firstly, transcript data files and experts' responses were grouped around each interview question. Secondly, these grouped responses were coded based on emerging themes for each interview question. For example, content associated with Question 1 about characteristics of winning criteria for M competitions was grouped according to "data" and "methods" themes. Thirdly, to ensure the credibility of our analysis (Shah & Corley, 2006), throughout the process, the authors discussed coding, cross-referencing, and emerging themes. We aimed to ensure that the data was linked with the research questions during the interpretation. These emerging conceptual themes were organized into the overarching themes that informed our main findings.

This section presents the findings from all our interviews, which were held to complement our survey findings and also help identify practical issues with regard to the M5 and future competitions.

Q1: Why is there a different winner in every M competition?

Broadly, the interviewees suggested that both data, as well as methods used in each competition, were the key factors in producing a different winner of the competition every time. The experts also referred to the importance of the domain, hierarchical structure of the data, as well as using different algorithms was also important in producing a different winner each time in the competition. The following selected interview quotes[2] demonstrate this difference in opinion:

> *"One factor is the number of time series. In particular from 3,000 time series to 100,000 time series, from M3 to M4, … allow cross-learning to perform more… Another factor is that the frequency of the data in M5 is completely different. The third factor, of course, is that the first four (M-) competitions were not domain-specific, while the last one was. I would also say a fourth factor is the structure of the data, the hierarchical structure of the data…"* (Expert 1)

> *"Because the methods have changed… And combined with other methods as well, new combinations, so I think it is purely the 'methods'. The second reason might be that the latest competition was purely based on Walmart data, so purely retail data."* (Expert 2)

Q2: Why are the machine learning methods dominating M5?

Opinions differed about why the machine learning methods dominated the M5 competition. Expert 1 suggested that it has to do with the LightGBM method and the cross-learning effect, while experts 2 and 3 highlighted that methods have been improving and getting more sophisticated over the years. One expert even claimed that overall ML methods did not dominate. The following interview extracts highlight some of these points:

> *"The LightGBM method was implemented by many participants in M5, and did perform well … But again, you have to look out why it works well, and I think the reason behind that is the cross-learning effect…"* (Expert 1)

> *"I think methods are getting more sophisticated… I assume the methods are improving."* (Expert 2)

Q3: How did a student manage to win the accuracy competition?

The experts had mixed feelings that a student managed to win the competition as opposed to a "seasoned" professional:

> *"21 years ago, when Theta method won the competition, one of the competitors, Kostas Nikolopoulos back then was a Ph.D. student. Now, the top performer of M5 is an MSc student. I think this is not the only time we will observe something like that… I think it has to do a little bit of luck here too"* (Expert 1)

> *"That is a big surprise for us, that is a student that was just taking a course. His methods (being) more accurate than 5000 experienced data scientists and forecasting managers…"* (Expert 3)

Some experts believed that it has to do with luck as well:

> *"Also, another thing is that the number 4 approach of the M4 competition, managed to achieve a position of 47 in the M5 competition. 47 is not as good as 2 or 3, but 47 out of 5000, is still on the top 1%.. For me, when it comes to the M5, the first position is as good as the fifth position or the tenth position… "* (Expert 1)

> *"I think there are several reasons. One of course is luck. I think the top 3 or 4 methods were very close, and if you toss the dice differently … or randomise (their initial values) in (a slightly) different way, someone else might win… "* (Expert 2)

Q4: Why do complex machine learning methods achieve greater accuracy than simple statistical methods in the last two M competitions?

Overall, there seems to be a consensus amongst our experts about the ability of ML methods to account for the greater complexity of information. For example, expert 2 highlighted the use of a wider range of information and combined it as follows:

> *"…it is quite clear that ML method use a wider range of information; can also deal with non-linearity; this used to be the great advantage of judgmental forecasting over algorithmic methods… the ML methods are in any way imitating the human brain…"* (Expert 2)

[2] *Quotes and excerpts have been slightly modified in order to convert oral everyday speech to a written form.*

Q5: Where is the future of the M competitions heading?

While elaborating on the future of the M competitions, the experts wanted to see live and real-time competitions, as well as using different sources of data, and even (as in M2) re-examining and exploring human judgment:

> *"I'd like to say explore judgement again; they did that in the M2 competition... Because you need people who are experts in their companies. Judgmental forecasting is so widely used, almost every company, just forecasts (based) on pure judgement. … And the second thing is, we need not (to) always focus on point forecasting; the literature is dominated by point forecasting. We need (to) do much more on understanding how to produce forecasts capturing uncertainty"* (Expert 2)

Furthermore, some experts also believed that more forecasters need to participate in forecasting competitions. Expert 1 notably argues that as follows:

> "*you shouldn't call yourself a forecaster if you don't participate. You have to have skin in the game…you cannot just talk and do nothing*"[3] (Taleb, 2018).

6. Conclusion, limitations, and the future

We believe we provided a good range of possible explanations for why ML performed well in the M5 competition, as well as why every forecasting competition has a different winner, among many other smaller insights (as per the 20+5 questions researched in our survey and interviews).

The nature of the dataset and the evolution of methods are the main reasons the experts believe that make every completion a unique one, with a new and gradually more advanced (and possibly more complex) winner. In the M5, this winner came from the family of the ML methods; probably next time (M6 onwards) a deep-learning one?[4]

Arguably you cannot expect that the results of all competitions/empirical studies will be the same. Things are changing, methods are improving, and new ideas are introduced (otherwise we would not need the competitions in the first place…). In the case of the M4, cross-learning and cross-validation made the difference, and (as a result) these were widely used in the M5 too, making the distance between ML methods to the simpler ones, much bigger. It also came out forcefully that it would be helpful for forecasters to explain why they think their methods will be relatively accurate when submitting their forecasts to competitions like the M5. If we do not have testable explanations for performance, then we have "dustbowl empiricism" and we are unable to assess the extent to which the findings can be generalized.

We tried to provide some empirical insights to answer some difficult long-standing questions, and to that end, we decided to rely upon the expert views of seasoned forecasters, and not just our view, experience, and expertise – such sampling of expert views comes with the usual caveat and respective limitations of small samples, but this might just be a good starting point for future research…

[3] This is, however, a provocative to many point of view that is not widely shared in the forecasting community.

[4] This is a sophisticated guess as consensus of the authoring team.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix

Interview Protocol

Interview Protocol: Introduction

- Brief introduction of the research project and interviewee's role in the project
- Ensure confidentiality and anonymity
- Ask permission to record the interview
- Ensure that the recording is for the academic use only. Let the respondent know that it is possible to stop the recording at any time.
- Discuss briefly the issues that will be covered during the interview
- Any questions?

START RECORDING

Q1: Why is there a different winner in every M competition?

Q2: Why are the machine learning methods dominating M5?

Q3: How did a student manage to win the M5?

Q4: Why do machine learning methods seem to achieve greater accuracy than statistical methods in the last two M competitions?

Q5: Where is the future of the M competitions heading?

Conclusion

Additional: Is there anything you would like to mention or highlight regarding on the forecasting competitions, the past, the future one, or the M5?

- **Thank you for your time.**

SWITCH OFF RECORDING

References

Armstrong, J. S., & Green, K. C. (2019). Why didn't experts pick M4-competition winner? Available at: https://repository.upenn.edu/marketing_papers/431.

Barker, J. (2020). Machine learning in M4: What makes a good unstructured model? *International Journal of Forecasting*, *36*, 150–155.

Bontempi, G. (2020). Comments on M4 competition. *International Journal of Forecasting*, *36*, 201–202.

Easterby-Smith, M., Thrope, R., & Lowe, A. (2002). *Management research an introduction. London, Thousand oaks*. New Delhi: SAGE Publications.

Eisenhardt, K. M. (1989). Building theories from case study research. *Academy of Management Review*, *14*, 532–550.

Fry, C., & Brundage, M. (2020). The M4 forecasting competition – A practitioner's view. *International Journal of Forecasting*, *36*, 156–160.

Gilliland, M. (2020). The value added by machine learning approaches in forecasting. *International Journal of Forecasting*, *36*, 161–166.

Green, K. C., & Armstrong, J. S. (2015). Simple versus complex forecasting: The evidence. *Journal of Business Research*, *68*(8), 1678–1685.

Hong, T. (2020). Forecasting with high frequency data: M4 competition and beyond. *International Journal of Forecasting*, *36*, 191–194.

Hudson, L. A., & Ozanne, J. L. (1988). Alternative ways of seeking knowledge in consumer research. *Journal of Consumer Research*, *14*(4), 508–521.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2020a). The M4 competition: 100, 000 time series and 61 forecasting methods. *International Journal of Forecasting*, *36*, 54–74.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2020b). The M5 accuracy competition: results, findings and conclusions. *International Journal of Forecasting*, *2020*, 10, Available at: https://www.researchgate.net/publication/344487258_The_M5_Accuracy_competition_Results_findings_and_conclusions.

Petropoulos, F., Apiletti, D., Assimakopoulos, V., Babai, M. Z., et al. (2022). Forecasting: theory and practice. *International Journal of Forecasting*, in press, https://www.sciencedirect.com/science/article/pii/S0169207021001758.

Saunders, M., Lewis, P., & Thornhill, A. (2016). *Research methods for business students* (7th ed.). Essex: Pearson Education Limited.

Shah, S. K., & Corley, K. G. (2006). Building better theory by bridging the quantitative-qualitative divide. *Journal of Management Studies*, *48*(80), 1821–1835.

Taleb, N. N. (2018). *Skin in the game: Hidden asymmetries in daily life*. Incerto.

Tilba, A., & McNulty, T. (2013). Engaged versus disengaged ownership: The case of pension funds in the UK. *Corporate Governance: An International Review*, *21*, 165–182.