# Forecasting in social settings: the state of the art

**Spyros Makridakis[a], Rob J. Hyndman[b] and Fotios Petropoulos[c]**

[a]Institute for the Future (IFF), University of Nicosia, Cyprus
[b]Department of Econometrics and Business Statistics, Monash University, Australia
[c]School of Management, University of Bath, United Kingdom

*"There's no chance that the iPhone is going to get any significant market share"*
Steve Ballmer, CEO
Microsoft
April 2007

**Abstract**
This paper provides a non-systematic review of the progress of forecasting in social settings. It is aimed at someone outside the field of forecasting, wanting to appreciate the results of the M4 Competition by reading a survey paper to get informed about the state of the art of this discipline. It discusses the recorded improvements over time in forecast accuracy, the need to capture forecast uncertainty, and what can go wrong with predictions. Subsequently, the review distinguishes knowledge achieved over the last years into (i) what we know, (ii) what we are not sure about, and (iii) what we don't know. In the first two areas, we explore the difference between explanation and prediction, the existence of an optimal model, the performance of machine learning methods on time series forecasting tasks, the difficulties with predicting non-stable environments, the performance of judgment, and the value-added of exogenous variables. The article concludes with the importance of (thin and) fat tails, the challenges and advances in causal inference, and the role of luck.

## 1. The facts

### 1.1 A brief history of forecasting

In terms of human history, it is not too long ago since forecasting moved from the religious, the superstitious and even the supernatural (Scott, 2014) to the more scientific. But even today, old fortune-telling practices still hold among people who pay to receive the "prophetic" advice of "expert", professional forecasters, including those who claim they can predict the stock market and make them rich§by following their advice. In the emerging field of "scientific" forecasting, there is absolute certainty about two things. First, **no one** possesses prophetic powers, even though many pretend to do so, and second, **all**

predictions are uncertain; often the only thing varying among such predictions is the **extent** of such uncertainty.

The field of forecasting outside the physical sciences started at the end of the nineteenth century with attempts to predict economic cycles and continued later with efforts to forecast the stock market. Later, it was extended to predictions concerning business, finance, demography, medicine, psychology and other areas of social sciences. The young field achieved considerable success after the Second World War with Robert Brown's work (1959 and 1963) to predict the demand for thousands of inventory items stored in navy warehouses. Given the great variety of forecasts required as well as the computational needs for doing so, the work had to be carried out easily, using the mechanical calculators of the time. Brown's achievement was to develop various forms of exponential smoothing that were sufficiently accurate for the problems faced and computationally light. Interestingly, in the Makridakis and Hibon (1979) study, and the M1 and M2 competitions, his simple, empirically developed models were found to be more accurate than the highly sophisticated ARIMA models of Box & Jenkins (Box et al., 2008).

As computers became faster and cheaper, the field expanded with econometricians, engineers and statisticians proposing various advanced forecasting approaches, believing that greater sophistication would improve forecasting accuracy. There were two faulty assumptions with such beliefs. First, it was assumed that the model that best fitted the available data (model fit) would also be the most accurate one for forecasting beyond such data (post-sample predictions). But the effort to minimise model fit errors contributed to over parameterisation and overfitting. Simple methods that captured the dominant features of the generating process were both less likely to overfit and likely to be at least as accurate as statistically sophisticated ones (see Pant and Starbuck, 1990). The second faulty assumption was that of constancy of patterns/relationships, assuming that the future will be an exact continuation of the past. Although history repeats itself, it never does so in precisely the same way. Simple methods tend to be less affected by changes in the data generating process, resulting in smaller post-sample errors.

Starting in the late 1960s, significant efforts were made through empirical and other studies and competitions to evaluate forecasting accuracy and establish some objective findings of our ability to predict the future and assess the extent of uncertainty associated with these predictions. Today, after many such studies/competitions, we have a good idea of the accuracy of the various predictions in business, economic and social fields and, lately, involving climate changes as well as the uncertainty associated with them. Most importantly, we have witnessed considerable advances in the field of forecasting that have been adequately documented in the past in two published papers. Makridakis (1986) surveyed the theoretical and practical developments in the field of forecasting and discuss the findings of empirical studies and their implications until that time. Another pioneering paper was published twenty years later by Armstrong (2006), aimed at "summarizing what has been learned over the past quarter century about the accuracy of forecasting methods" (p.583) while covering new developments, including neural networks that were in their infancy at that time. The purpose of the present paper is to provide an updated survey for

the non-forecasting expert who wants to be informed of the state of the art of forecasting in social sciences and to understand the findings/conclusions of the M4 Competition better.

Some of the conclusions of these earlier surveys have been overturned with additional evidence. For example, Armstrong (2006) found that neural nets and Box-Jenkins methods fared poorly against alternatives. Yet both have now been shown to be competitive. For neural nets, good forecasts have been obtained when there are enormous collections of available data (Salinas et al., 2017). For Box-Jenkins methods, improved identification algorithms (Hyndman & Khandakar 2008) have led to them being competitive with (and sometimes better than) exponential smoothing methods. Other conclusions have stood the test of time: for example, that combining forecasts improves accuracy.

## 1.2 When predictions go wrong

Although forecasting in the physical sciences can attain amazing levels of accuracy (see section 1.3) this is not the case in social contexts where practically all predictions are uncertain while a good number of them can be unambiguously wrong, particularly when binary decisions are involved like the decision facing the U.S. Federal Reserve of whether to raise or lower interest rates, given the competing risks of inflation and unemployment. The big problem is that some wrong predictions can affect not only a firm or a small group of people, but also whole societies like those involving global warming, or can be detrimental to our health. Ioannidis, a medical professor at Stanford, has devoted his life to study health predictions. His findings are disheartening, articulated in an article published in *PLOS ONE* entitled "Why Most Published Research Findings Are False" (Ioannidis, 2005[1]). A popular piece on a similar theme in *The Atlantic* entitled "Lies, Damned Lies, and Medical Science" (Freedman, 2010) is less polite. It summarised such findings as "Much of what medical researchers conclude in their studies is misleading, exaggerated, or flat-out wrong". The article concluded with the question "why are doctors—to a striking extent—still drawing upon misinformation in their everyday practice?"

A recent example exemplifying Ioannidis' conclusions is the findings of two studies eight years apart, whose results were contradictory making it impossible to know what advice to follow to benefit from medical research. In 2010, a meta-analysis was published (Micha et al., 2010) that reviewed six studies evaluating the effects of meat and vegetarian diets on mortality, involving a total of more than 1.5 million people. It concluded that all-cause mortality was higher for those who eat meat, mainly red or processed meat, daily. A new study, however, published in 2018 (Mente and Yusuf, 2018), using a large sample of close to 220,000 people, found that eating red meat and cheese reduced cardiovascular disease by 22% and decreased the risk of early death by 25% (with such large sample sizes all differences are statistically significant). If conflicting medical predictions, based on substantial sample sizes and hundreds of millions spent on designing and carrying them out, are widespread, what are we to surmise about studies in other disciplines that are less well funded, utilising small sample sizes, or basing their predictions on judgment and opinion? Moreover, if the conclusions of a medical study are reversed in a period of just

[1] Ioannidis paper is one of the most viewed/downloaded published in *PLOS ONE* with more than 2.3 million views and more than 350K downloads.

eight years, how can we know that those of new studies will not produce the same contradictions? Recommendations about the treatment of disease are based on the findings of medical research, but how such findings can be trusted when most of them, according to Ioannidis, are false? Clearly, there is a predictability problem, extending beyond medicine to practically all fields of social science including economics (Dewald et al., 1986 and Camerer et al., 2016). Fortunately, empirical studies in the field of forecasting have provided us with some objective evidence to allow us to know both the accuracy of predictions and estimating the level of uncertainty.

There are some famous examples of forecasting errors, including Ballmer's forecast about the iPhone, possibly the most successful of all products ever marketed. In 1798, Malthus predicted that we were confronted by mass starvation as the population was growing geometrically while food production was increasing only arithmetically. Today's material abundance and decreases in population growth in most advanced countries have been moving in the opposite direction of his predictions. In 1943, Thomas Watson, IBM's president made his infamous prediction: "I think there is a world market for maybe five computers", missing it by a billion times if all computers, including smartphones, are counted (see also Schnaars, 1989). But even recent predictions by professional organisations specialising in forecasting, using modern computers and well-trained PhDs, can go wrong as was the case with the complete inability of these organisations to predict the great 2007/2008 recession and its grave implications. The same has been true with technological forecasting that failed to predict, even a few decades beforehand, the arrival and widespread usage of the three major inventions of our times: The computer, the Internet, and the mobile phone. Judgmental predictions have been evaluated by Tetlock (2006) who has compared the forecasts of experts in different macroeconomic fields to forecasts made by well-informed laity or those based on simple extrapolation from current trends. His conclusions: not only are most experts not more accurate, but they also find it more difficult to change their minds when new evidence becomes available.

What we can conclude by surveying the successes and failures in forecasting is that there is a significant amount of uncertainty in all our predictions and that such uncertainty is greatly underestimated for two reasons. First, by our attitude to extrapolate in a linear fashion from the present to the future and second by our fear of the unknown and our psychological need to reduce the anxiety associated with such fear by believing we can control the future by accurately predicting it (known as the illusion of control, Langer, 1975). It becomes, therefore, imperative to be aware of the difficulties of accurate predictions and the underestimation of uncertainty associated with them, to be able to minimise this bias. The field of quantitative forecasting has a possible advantage that it is possible to assess the accuracy of forecasts and the level of uncertainty surrounding them by utilising information from empirical and open forecasting competitions.

## 1.3 Improving forecasting accuracy over time

A scientific approach to forecasting in the physical sciences began with Halley's comet predictions in the early 1700s (Halleio 1704), which turned out to be remarkably accurate. Other forecasts followed including the somewhat less successful meteorological forecasts

of Beaufort and FitzRoy in the late 1850s (Burton 1986). These were highly controversial at the time, and FitzRoy, in particular, was heavily criticised and subsequently committed suicide. Nevertheless, he left a lasting legacy including the word "forecast" which he had coined for his daily weather predictions. Over the ensuing 150 years, there has been extraordinary progress in forecast accuracy in meteorology (Kistler et al. 2001; Saha et al. 2014) and other physical sciences as the underlying physical processes have become better understood, the volume of observations has exploded, computing power has increased, and the ability to share information across connected networks has become available.

The social sciences are different. First, there is usually limited theoretical or quantitative basis to represent a causal or underlying mechanism. So we rely on statistical approximations that roughly describe what we observe, but may not represent a causal or underlying mechanism. Second, despite the deluge of available data today, much of this information does not directly concern what we want to forecast. For example, we may wish to predict GDP next quarter, and we have an enormous amount of daily stock market data available, but no daily data on expenditure on goods and services. Third, what we are trying to forecast is often affected by the forecasts themselves. For example, central banks might forecast next year's housing price index, and then raise interest rates leading the index to be lower than the forecast. Such feedback does not occur in astronomical or weather forecasts.

For these reasons, social science forecasts are unlikely ever to be as accurate as forecasts in the physical sciences, and the potential improvements in accuracy are somewhat limited. Nevertheless, increased computing power, and a better understanding of how to separate signal from noise should lead to some improvements in forecast accuracy. However, at least in the case of macroeconomic forecasting, this does not appear to have been the case (Fildes and Stekler 2002; Stekler 2007; Heilemann and Stekler 2013).

On the other hand, time series forecasting has demonstrably improved over the last 30 years. We can measure the change through the published accuracy of forecasting competitions over the past 40 years, beginning with the first Makridakis competition (Makridakis et al. 1982), then the M3 competition (Makridakis and Hibon 2000), and finally the recent M4 competition (Makridakis, Spiliotis, and Assimakopoulos 2018a). In measuring forecast accuracy improvement, we have applied the best-performing methods from each competition to the data from previous competitions. In this way, we can see how the methods have improved over time.

These comparisons are not straightforward because the forecast accuracy measure used was not consistent between competitions. In fact, there is still no agreement on the best measure of forecast accuracy. We will compare results using the MAPE (used in the first competition), the sMAPE (used in the M3 competition) and the MASE. The M4 competition used a weighted average of the sMAPE and MASE values. All measures are defined and discussed by Hyndman & Koehler (2006) and Hyndman & Athanasopoulos (2018).

In the first Makridakis competition (Makridakis et al. 1982), the best performing method overall (as measured by MAPE) was simple exponential smoothing applied to deseasonalized data, where the deseasonalization used a classical multiplicative

decomposition (Hyndman and Athanasopoulos 2018); this is denoted by DSES. For non-seasonal data, DSES is equivalent to simple exponential smoothing.

In the M3 competition, the best method (as measured by sMAPE) which is in the public domain was the Theta method (Assimakopoulos and Nikolopoulos 2000). When applying the Theta method, we use the `thetaf()` implementation from the forecast package for R (Hyndman et al. 2018), to ensure consistent application to all data sets.

In the M4 competition, the best performing method (as measured by a weighted average of sMAPE and MASE) for which we had R code available was the FFORMA method (Montero-Manso et al. 2019) which came second in the competition.

In addition to these methods, for comparison, we also include the popular `auto.arima()` and `ets()` methods (Hyndman et al. 2002; Hyndman and Khandakar 2008) as implemented in Hyndman et al. (2018), along with a simple average of the forecasts from these two methods (denoted "ETSARIMA"). We also include two simple benchmarks, naive and naive on the seasonally adjusted data (naive 2).

When we apply all of these methods to the data from all three competitions, we can see how forecast accuracy has changed over time, as shown in Table 1. Note that the mean values of MAPE, sMAPE and MASE have been calculated by simultaneously applying arithmetic mean across series and horizons. Other ways of averaging the results can lead to different conclusions due to greater weight being placed on some series or some horizons. It is not always obvious in the published competition results how these calculations have been done in the past, although in the case of the M4 competition, the code has been made public to help eradicate such confusion.

**Table 1**: *Comparing the best method from each forecasting competition against each other and benchmark methods.*

| | M1 competition | | | M3 competition | | | M4 competition | | |
|---|---|---|---|---|---|---|---|---|---|
| **Method** | **MAPE** | **sMAPE** | **MASE** | **MAPE** | **sMAPE** | **MASE** | **MAPE** | **sMAPE** | **MASE** |
| FFORMA | **15.9** | **14.4** | **1.28** | 18.4 | **12.6** | **1.11** | **14.3** | **11.8** | **1.17** |
| ETSARIMA | 17.4 | 15.3 | 1.32 | 18.7 | 13.1 | 1.13 | 14.9 | 12.3 | 1.22 |
| ETS | 17.7 | 15.6 | 1.35 | 18.7 | 13.3 | 1.16 | 15.6 | 12.8 | 1.27 |
| ARIMA | 18.9 | 16.3 | 1.38 | 19.8 | 14.0 | 1.18 | 15.2 | 12.7 | 1.24 |
| Theta | 20.3 | 16.8 | 1.41 | **17.9** | 13.1 | 1.16 | 14.7 | 12.4 | 1.30 |
| DSES | 17.0 | 15.4 | 1.46 | 19.2 | 13.9 | 1.31 | 15.2 | 12.8 | 1.41 |
| Naive 2 | 17.7 | 16.6 | 1.52 | 22.3 | 15.8 | 1.40 | 16.0 | 13.5 | 1.44 |
| Naive | 21.9 | 19.4 | 1.79 | 24.3 | 16.6 | 1.50 | 17.5 | 14.7 | 1.70 |

There are several interesting aspects to this comparison.

- Based on the MAPE and sMAPE, DSES did well on the M1 data, and is competitive with other non-combining methods on the M3 and M4 data, but it does poorly based on the MASE.
- While Theta did well on the M3 data (winner of that competition), it less competitive on the M1 and the M4 data.
- The most recent method (FFORMA) outperforms the other methods on every measure for the M1 and M4 competitions, and on all but the MAPE measure for the M3 competition.
- The ETSARIMA method (averaging the ETS and ARIMA forecasts) is almost as good as the FFORMA method in terms of MASE, and is easier and faster to compute.
- Using the MASE criterion, the results are relatively clear-cut across all competitions (in the order displayed). With other accuracy criteria, the results are less clear.

While there is some variation between periods, the good performance of FFORMA and ETSARIMA is relatively consistent across data sets and frequencies. Clearly, progress in

forecasting methods has been uneven, but the recent M4 competition has helped advance the field considerably in several ways including: (1) encouraging the development of several new methods; and (2) providing a large set of data to allow detailed comparisons of various forecasting methods over different time granularities.

## 1.4 The importance of being uncertain

No forecasts are exact, and so it is important to provide some measures of forecast uncertainty. Without expressing such uncertainty clearly and unambiguously, forecasting is not far removed from fortune-telling.

The most general approach for expressing uncertainty is to estimate the "forecast distribution" — the probability distribution of future observations conditional on the information available at the time of forecasting. A point forecast is usually the mean (or sometimes the median) of this distribution, and a prediction interval is usually based on the quantiles of this distribution (Hyndman and Athanasopoulos 2018). Consequently, forecasting has two primary tasks:

1. To provide point forecasts which are as accurate as possible;
2. To specify or summarise the forecast distribution.

Little attention was paid to forecast distributions, or measures of forecast distribution accuracy, until relatively recently. There was no measure of distributional forecast uncertainty used in the M1 and M3 competitions, for example, and it is rare to see such measures used in Kaggle competitions.

**Prediction interval evaluation**

The simplest approach to summarising the uncertainty of a forecast distribution is to provide one or more prediction intervals with specified probability coverage. It is well-known that these intervals are often narrower than they should be (Hyndman et al, 2002) — that is, that the actual observations fall inside the intervals less often than the nominal coverage implies. For example, 95% prediction intervals for the ETS and ARIMA models applied for the M1 and M3 competition data, obtained using the automatic procedures in the forecast package for R, yield coverage percentages as low as 76.8%, and never higher than 95%. Progress has been made in this area too, with the recent FFORMA method (Montero-Manso et al. 2019) providing average coverage of 94.5% for these data sets. Figure 1 shows the coverage for nominal 95% prediction intervals, for each method and forecast horizon when applied to the M1 and M3 data. ARIMA models do particularly poorly here.

It is also evident from Figure 1 that there are possible differences in the two data sets, with percentage coverage lower for the M1 competition than for the M3 competition.

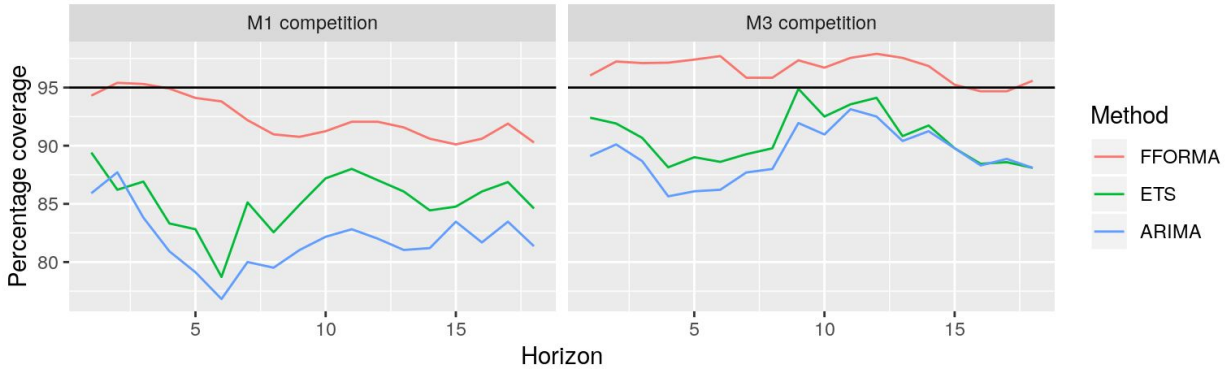


**Figure 1**. *Actual coverage achieved by nominal 95% prediction intervals.*

There are at least three reasons for the underestimation of uncertainty of standard statistical models.

1. Probably the biggest factor is that model uncertainty is not taken into account. The prediction intervals are produced assuming that the model is "correct", which is clearly never the case.
2. Even if the model is correctly specified, the parameters must be estimated, and parameter uncertainty is also rarely accounted for in time series forecasting models.
3. Most prediction intervals are produced assuming Gaussian errors. When this assumption is not true, especially when the errors have a fat-tailed distribution, the prediction interval coverage will usually be underestimated.

In contrast, some modern forecasting methods do not use an assumed data generating process to compute prediction intervals. The prediction intervals from FFORMA are produced using a weighted combination of the intervals from its component methods, where the weights are designed to give appropriate coverage while also taking account of the length of the interval.

Coverage is important, but it is not the only requirement for good prediction intervals. A good prediction interval is as small as possible while maintaining the specified coverage. Winkler proposed a scoring method to enable comparisons between prediction intervals that takes account of both coverage and width of the intervals. If the $100(1-\alpha)\%$ prediction interval for time $t$ is given by $[l_t, u_t]$, and $y_t$ is the observation at time $t$, then the Winkler (1972) score is defined as the average of

$$W(l_t, u_t, y_t) = \begin{cases} (u_t - l_t) & l_t < y_t < u_t \\ (u_t - l_t) + \frac{2}{\alpha}(l_t - y_t) & y_t < l_t \\ (u_t - l_t) + \frac{2}{\alpha}(y_t - u_t) & y_t > u_t. \end{cases}$$

This both penalises for wide intervals (since $u_t - l_t$ will be large), and penalises for non-coverage with observations well outside the interval being penalised more heavily. Although this was proposed in 1972, it has received very little use until recently, when a

scaled version of it was used in the M4 competition. The lower the score, the better the forecasts are. For a discussion of some of the problems with interval scoring, see Askanazi et al. (2018).

**Forecast distribution evaluation**

To our knowledge, the only forecasting competitions that have evaluated whole forecast distributions have been the GEFCom2014 and GEFCom2017 energy forecasting competitions (Hong et al. 2016). Both used percentile scoring as an evaluation measure.

For each time period $t$ throughout the forecast horizon, the participants provided the percentiles $q_{i,t}$ where $i = 1, 2, \ldots, 99$. Then the percentile score is given by the pinball loss function:

$$L(q_{i,t}, y_t) = \begin{cases} (1 - i/100)(q_{i,t} - y_t) & y_t < q_{i,t} \\ (i/100)(y_t - q_{i,t}) & y_t \geq q_{i,t}. \end{cases}$$

To evaluate the full predictive densities, this score is then averaged over all percentiles and all time periods. If the observations follow the forecast distribution, then the average score will be the smallest value possible. If the observations are more spread out, or in some other way deviate from the forecast distribution, then the average score will be higher. Other distribution scoring methods are also available (Gneiting and Raftery 2007).

Without a history of forecast distribution evaluation, it is not possible to explore how this area of forecasting has improved over time. However, we recommend future forecast evaluation studies to include forecast distributions, especially in areas where the tails of the distribution are of particular interest, such as in energy and finance.

# 2. What we know

## 2.1 On explaining the past versus predicting the future

Forecasting is about predicting the future, but this can only be done based on information from the past, raising the issue of how to select the most appropriate information and corresponding model for predicting the future. For a long period, and for lack of a better alternative, it was believed that such model should be chosen by how well it could explain, that is, fit, the available past data (somewhat like asking a historian to predict the future). Makridakis and Hibon (1979), for instance, had difficulties explaining in the presentation of their paper to the Royal Statistical Society in London their findings that single exponential smoothing was more accurate than the Box-Jenkins approach and that the combination of methods was more accurate than the individual methods being combined. Theoretically, with the correct model and assuming that the future is the same as the past, these findings would not be possible. However, this theoretical posture does not necessarily hold for the future that could be quite different from the past. The superiority of combining as well as the higher accuracy of exponential smoothing methods in comparison to ARIMA models was proven again with the M1 and M2 Competitions. Some statisticians, however, were still

unwilling to accept the empirical evidence, arguing that theory was more important than empirical competitions, powerfully expressed by Priestley stating "we must resist the temptation to read too much into the results of the analysis" (Makridakis and Hibon, 1979, p. 127).

The debate ended with the M3 Competition (Makridakis and Hibon, 2000) with its 3,003 time series. Results showed, one more time, the value of combining and the superior performance of some simpler methods (such as the Theta method) in comparison to other more complicated methods (notably one particular neural networks application). Slowly but steadily this evidence is being accepted by a new breed of academic forecasters and well-informed practitioners interested in improving the accuracy of their predictions. Moreover, in the M3 and M4 competitions ARIMA models have improved their accuracy considerably surpassing that of exponential smoothing methods where model selection used Akaike's information criterion (Akaike, 1977).

Emphasis has shifted, therefore, from arguing about the value of competitions to learning as much as possible from the empirical evidence to improve the theoretical and practical aspects of forecasting. The M4 Competition, covered in detail in this Special Issue, is the most recent evidence of such a fundamental shift in the attitudes toward forecasting and the considerable learning that has been taking place in the field. A number of academic researchers have guided this shift in universities. Determined practitioners from companies like Uber, Amazon, Google, Microsoft and SAS, among others, present every year in the International Symposium on Forecasting (ISF) their advances. They are focused on improving the forecasting accuracy and harness its benefits while also being concerned about measuring the uncertainty in their predictions.

## 2.2 On the (non) existence of a best model

Many forecasting researchers have been on the quest of identifying the best forecasting model for each particular case. This quest is often viewed as the "holy grail" in forecasting. While earlier studies have studied the concept of aggregate selection (Fildes, 1989), meaning selecting one model for all the series within a data set, later studies have suggested that such an approach can only work for highly homogeneous data sets. In fact, as Fildes and Petropoulos (2015) have shown, if we had the means to identify which models would perform best for each series correctly, we could observe savings of up to 30% compared to using the best (but same) model on all series.

Approaches for individual selection of the best model per series (or even per combination of series/horizon) include information criteria (Hyndman et al., 2002), validation and cross-validation approaches (Tashman, 2000), approaches that use knowledge obtained from the data to find temporary solutions to the problems faced (Fildes and Petropoulos, 2015), approaches based on time series features and expected errors (Wang et al., 2009; Petropoulos et al., 2014) and approaches based on expert rules (Adya et al., 2001). However, all the approaches above are limited with regards to their input: They are attempting to identify the best model for the future conditioned on the information from the past. However, as the previous section has highlighted, explaining the past is not the same as predicting the future. When dealing with real data, well-specified "data generation

processes" do not exist. The future might be completely different than the past, and the "best" models may no longer be appropriate. Even if we could identify the best model, we would be limited to appropriately estimating its parameters.

In fact, there exist three types of uncertainties when dealing with real forecasting situations: model uncertainty, parameter uncertainty and data uncertainty (Petropoulos et al., 2018a). Such types of uncertainties are dealt with in practice through the combining of models/methods. As George Box has put it, "all models are wrong, but some are useful." Combinations have again and again been proved to bring benefits in forecasting accuracy (Makridakis, 1989; Clemen, 1989; Timmermann, 2006) while also decreasing the variance of the forecasts (Hibon and Evgeniou, 2005) thus rendering operational settings more efficient. Current approaches to forecast combinations include, among others, combinations based on information criteria (Kolassa, 2011), the use of multiple temporal aggregation levels (Andrawis et al., 2011; Kourentzes et al., 2014; Athanasopoulos et al., 2017), bootstrapping for time series forecasting (Bergmeir et al., 2016) and forecast pooling (Kourentzes et al., 2018).

Approaches based on combinations have dominated the rankings in the latest instalment of the M Competitions. It is important to highlight that one element of the success of forecast combinations is carefully selecting the appropriate pool of models and their weights. One way to explain the good performance of combinations is that the design of the M competitions requires that the nature and history of the series is concealed. This reduces the amount of background information that can be applied to the forecasting problem and may advantage combinations relative to models that are individually selected per series. In fact, as Fildes and Petropoulos (2015) have shown, model selection can outperform forecast combinations in certain situations such as when a dominant method exists or under stable environments. Finally, evidence in the M4 results suggests that hybrid approaches, which are based on combining simple time-series techniques with modern machine-learning methods on a conceptual level (rather than a forecast level), perform very well.

## 2.3 On the performance of machine learning

The hype publicizing the considerable achievements in artificial intelligence (AI) also extends to machine learning (ML) forecasting methods. There were high expectations that hedge funds utilising ML techniques would outperform the market (Satariano and Kumar, 2017). New evidence has shown, however, that even though the potential is enormous, the track record is mixed (Asmundsson, 2018).

Although some publications have claimed to show excellent accuracy of ML forecasting methods, very often they have not been compared against sensible benchmarks. For stock-market data, for example, it is essential to include a naive benchmark, yet this is often not done (see, for example, Wang and Wang, 2017). In addition, some studies claim high accuracy by hand-selecting examples where the proposed method happens to do well. Even when there is a reasonably large set of data used for the empirical evaluation, and the time series have not been chosen to favour the proposed approach, it is essential to consider the

statistical significance of any comparisons made. Otherwise, conclusions can be drawn from random noise (Pant and Starbuck, 1990).

One advantage of large forecasting competitions is that they provide a collection of data against which new methods can be tested, and for which published accuracy results are available. The data sets are also large enough that the statistical significance of results should be achievable for any meaningful improvements in forecast accuracy.

One disadvantage is that they are a heterogeneous mix of frequencies, lengths and categories so that the circumstances where individual methods shine or fall down is extracted with some difficulty from the raw results.

In time series forecasting, the hype has been moderated as studies have shown that application of ML methods leads to poor performance in comparison to statistical methods (some ML supporters still argue about the validity of the empirical evidence). We are not supporters or critics of one approach or another, and we believe that there is considerable overlap between the statistical and the ML approaches to forecasting. Moreover, they are complementary in the sense that ML methods are more vulnerable to excessive variance while statistical ones to higher bias. At the same time, the empirical evidence to date shows a clear accuracy superiority of the statistical methods in comparison to ML ones when applied to individual time series, or large collections of heterogeneous time series. In a study using 1045 of the monthly M3 series (the same as those utilised by Ahmed et al. 2010) whose number of observations was larger than 81, Makridakis, Spiliotis and Assimakopoulos (2018b) found, using accepted practices to run the methods, that the most accurate of the ML methods were less accurate than the least accurate statistical one. Moreover, fourteen ML methods were less accurate than naive 2.

ML methods did not do well in the M4 Competition either with most of them doing worse than the naive 2 benchmark (for more details see the M4 paper this issue). We believe that it is essential to figure out the reasons for such poor performance of the ML methods. One possibility is the relatively large number of parameters associated with ML methods compared to statistical methods. Another is the number of important choices related to the design of ML, usually made using validation data, as there is no standardised ML approach. The time series used in these competitions are generally not particularly long, with a few hundred observations at most. This is simply insufficient to build a complicated nonlinear, nonparametric forecasting model. Even if the time series is very long (at least a few thousand observations), there are difficulties with data relevance as the dynamics of the series may have changed, and the early part of the series may bias the forecasting results.

Machine learning methods have done well in time-series forecasting when forecasting an extensive collection of homogeneous data. For example, Amazon uses deep learning neural networks to predict product sales (Salinas et al., 2017; Wen et al., 2017) by exploiting their vast database of sales history on related products, rather than building a separate model for the sales of each product.

We expect that future research effort will contribute to making these methods more accurate. Both the first and the second-best methods of the M4 Competition used ML ideas to improve accuracy, and we would expect that additional, innovative notions would be found in the future to advance their utilisation.

# 3. What we are not sure about

## 3.1 On predicting recessions/booms/non-stable environments

One area in forecasting that has attracted much attention is that of extreme events, which include but are not limited to economic recessions/booms and natural disasters. Such events have a significant impact from a socioeconomic perspective, but also are notoriously tricky to predict, some of them being "Black Swans" (events with no known historical precedence).

Take as an example the great recession of 2008. At the end of December 2007, Business Week reported that only 2 out of 34 forecasters predicted a recession for 2008. Even when the symptoms from the recession became more evident, Larry Kudlow (an American financial analyst and the Director of the National Economic Council under the Trump administration) insisted that there was no recession. Similarly, the Federal Open Market Committee failed to predict the 2008 recession (Stekler and Symington, 2016). Interestingly, after the recession, most economic analysts, victims of their hindsight, were able to provide detailed explanations and reasons behind the recession. The few "prophets" that were indeed able to predict the great recession did not offer equally good predictions for other extreme events, as if their prophetic powers were lost overnight.

Two recent studies have made some first steps towards predicting market crashes and bubble bursts. Gresnigt et al. (2015) model financial market crashes as seismic activity and create medium-term probability predictions which consequently feed an early warning system. Franses (2016) proposes a test to identify bubbles in time series data as well as to indicate whether a bubble is close to bursting.

## 3.2 On the performance of humans versus models

Judgment has always been an integral input to the forecasting process. Earlier studies focused on the comparative performance of judgmental versus statistical forecasts, when judgment was used to produced forecasts directly. The results of such studies have been inconclusive. For instance, while Lawrence et al. (1985) and Makridakis et al. (1993) found that unaided human judgment can be as good as the best statistical methods of the M1 forecasting competition, Carbone and Gorr (1985) and Sanders (1992) found that judgmental point forecasts are less accurate than statistical methods. The reason behind these results is the fact that well-known biases govern judgmental forecasts, such as the tendency of forecasters to dampen the trends (Lawrence and Makridakis, 1989; Lawrence et al., 2006), anchoring and adjustment (O'Connor et al., 1993) and confusion of the signal with the noise (Harvey, 1995; Reimers and Harvey, 2011). On the other hand, statistical methods are consistent and can handle a vast number of time series seamlessly. Still,

judgment is the only option for producing estimates for the future when data are not available.

Judgmental biases apply even to forecasters with domain or technical expertise. As such, the Expert Knowledge Elicitation (EKE, Bolger and Wright, 2017) literature has examined many ways of designing methods to reduce the danger of biased judgments from experts. Strategies to mitigate humans' biases include decomposing the task (Edmundson, 1990; Webby et al., 2005), offering alternative representations (tabular versus graphical formats, Harvey and Bolger, 1996) and the provision of feedback (Petropoulos et al., 2017).

The previous discussion has focused on judgmental forecasts produced directly. However, the judgment in forecasting can also be applied in the form of interventions (adjustments) on the statistical forecasts that are produced from a forecasting support system. Model-Based forecasts are frequently adjusted by experts on operations/supply chain settings (Franses and Legerstee, 2009b; Fildes et al., 2009). Such revised forecasts differ often significantly from the statistical ones (Franses and Legerstee, 2009a); however small adjustments are also observed and are linked with the sense of ownership of the forecasters (Fildes et al., 2009). Experts tend to adjust upwards more often than downwards (Franses and Legerstee, 2009b) which is attributed to the optimism bias (Trapero et al., 2013); such upwards adjustments are far less effective (Fildes et al., 2009). Empirical evidence also suggests that experts can reduce the forecasting error when adjustment size is not too large (Trapero et al., 2013).

Another part of the forecasting process where judgment can be applied is that of model selection. Assuming that many alternative models are offered by a modern forecasting software systems, managers often rely on their judgment to select the most suitable one rather than pushing the magic button labelled "automatic selection" (which selects between models based on algorithmic/statistical approaches, for example, using an information criterion). The study by Petropoulos et al. (2018b) is the first to offer some empirical evidence on the performance of judgmental versus algorithmic selection. When the task follows a decomposition approach (selection of the applicable time series patterns which is then translated to the selection of the respective forecasting model), the judgmental selection is on average as good as selecting via statistics, while humans have the advantage more often of avoiding the worst of the candidate models.

Two strategies are particularly useful in enhancing the judgmental forecasting performance. The first strategy is a combination of statistics and judgment (Blattberg and Hoch, 1990). This can be intuitively applied to the cases where the statistical and judgmental forecasts have been produced independently, but it works even in cases where the managerial input could be affected by the model output, as in judgmental adjustments. Several studies have shown that adjusting the adjustments can lead to increased forecasting performance (Fildes et al., 2009; Franses and Legerstee, 2011) but also better inventory performance (Wang and Petropoulos, 2016). The second strategy is the mathematical aggregation of the individual judgments that have been produced independently, also known as "wisdom of crowds" (Surowiecki, 2005). In the study by Petropoulos et al. (2018b) concerning model selection, the aggregation of the selections of

five individuals led to forecasting performance that is significantly superior to that of algorithmic selection.

In summary, we observe that the research focus has shifted over time from directly producing judgmental forecasts to adjusting statistical forecasts and judgmentally selecting between forecasts. The value-added of judgment in the forecasting process increases as we shift further from merely producing a forecast judgmentally.

However, given the exponential increase in the number of series needing to be forecast by a modern organisation (for instance, the number of stock keeping units in a large retailer may very well exceed 100,000), it is not always possible nor practical to allocate resources for manually managing each series.

## 3.3 On the value of explanatory variables

Using exogenous explanatory variables would seem to be an obvious approach to improving forecast accuracy. Rather than only relying on the history of the series we wish to forecast, we can utilise other relevant and available information as well.

In some circumstances, the data from explanatory variables can significantly improve forecast accuracy. One such situation is electricity demand forecasting where current and past temperatures can be used as explanatory variables (Ben Taieb and Hyndman 2014). Electricity demand is highly sensitive to the ambient temperature, with hot days leading to the use of air-conditioning, and cold days leading to the use of heating. Mild days (with temperatures around 20C) tend to have the lowest electricity demand.

However, often the use of explanatory variables is not as helpful as one might imagine. First, the explanatory variables themselves may need to be forecast. In the case of temperatures, good forecasts are available from meteorological services up to a few days ahead, and these can be used to help forecast electricity demand. But in many other cases, forecasting the explanatory variables may be just as difficult as forecasting the variable of interest. For example, Ashley (1988) argues that forecasts of many macroeconomic variables are so inaccurate that they should not be used as explanatory variables. Ma et al. (2016) demonstrate that the inclusion of competitive promotional variables as explanatory variables for retail sales is of limited value, but that the addition of focal variables leads to substantial improvements over time series modelling with promotional adjustments

A second problem arises due to the assumption that the relationships between the forecast variable and the explanatory variables will continue. When this assumption breaks down, there is model misspecification.

A third issue is that the relationship between the forecast variable and the explanatory variables needs to be strong and precisely estimated (Brodie et al., 2001). If the relationship is weak, there is little value in including the explanatory variables in the model.

It is possible to assess the value of explanatory variables and to test if either of these problems is prevalent by comparing the forecasts from three separate approaches. First,

we can use a purely time series approach, ignoring any information available in explanatory variables. Second, an ex-post forecast, building a model using explanatory variables but then using the future values of those variables when producing an estimate. Third, an ex-ante forecast, using the same model but substituting the explanatory variables with their forecasts.

Athanasopoulos et al. (2011) carried out this comparison in the context of tourism data, as part of the 2010 tourism forecasting competition. The explanatory variables, in this case, included relative CPI and prices between the source and destination countries. Not only were the purely time series forecasts better than the models including explanatory variables, but the ex-ante forecasts were also better than the ex-post forecasts. This suggests that the relationships between tourism numbers and the explanatory variables changed throughout the study. Further supporting this conclusion is the fact that time-varying parameter models did better than fixed parameter models. However, the time-varying parameter models did not do as well as the purely time series models, showing that the forecasts of the explanatory variables were also problematic.

To summarise, explanatory variables can be useful, but only under two specific conditions: (1) when there are accurate forecasts of the explanatory variables; and (2) when the relationships between the forecasts and the explanatory variables are likely to continue into the forecast period. Both conditions are satisfied for electricity demand. Neither condition is satisfied for tourism demand. Unless both conditions are satisfied, time series forecasting methods are better than using explanatory variables.

# 4. What we don't know

## 4.1 On thin/fat tails and Black Swans

Another misconception that prevailed in statistical education for a long time was that normal distributions could approximate practically all outcomes/events, including the errors of statistical models. Furthermore, there was little or no discussion of what could be done when normality could not be assured. Now it is accepted that Gaussian distributions, although extremely useful, are of limited value to approximate some areas of applications (Cooke, Nieboer, Misiewicz, 2014; Makridakis and Taleb, 2009), in particular in those referring to forecast error distributions, describing the uncertainty in forecasting. In this paper, we have emphasised the critical role of uncertainty, expressing our conviction that providing forecasts without specifying the level of uncertainty associated with them, amounts to nothing more than fortune telling. It is one thing, however, to identify uncertainty and another to get prepared to face it realistically and effectively. Furthermore, it must be clear that dealing with uncertainty cannot be done without incurring a cost or accepting lower opportunity benefits.

Table 2 distinguishes four types of events following Rumsfeld's classification. In Quadrant I, the known/knowns category, the accuracy of forecasting depends on the variance (randomness) of the data and can be assessed from past information. Moreover, uncertainty is well defined and can be measured, usually following a normal distribution

with thin tails. In Quadrant II (known/unknowns), including events like recessions, the accuracy of forecasting cannot be assessed as the timing of a recession, crisis or boom cannot be known while their consequences can vary widely from one recession to another. Uncertainty in this quadrant is considerably greater while its implications are much harder to assess than in Quadrant I. It is characterised by fat-tails, extending well beyond the three sigmas of the normal curve. A considerable problem amplifying the level of uncertainty is that a forecast, like the sales of some product, during a recession moves from Quadrant I to II increasing uncertainty considerably, making it much harder to prepare to face it.

**Table 2.** *Accuracy of Forecasting, Type of Uncertainty and Extent of Risk*

| | | | |
|---|---|---|---|
| **UNCERTAINTY** | KNOWN | **I. Known/Known**<br>(Law of large numbers, independent events, e.g. sales of tooth brush, shoes or beer)<br><br>Forecasting: Accurate (depending on variance)<br>Uncertainty: Thin tailed and measurable<br>Risks: Manageable, Can be minimized | **III. Unknown/Known**<br>(Cognitive biases, Strategic moves, e.g. Uber re-introducing AVs in a super-competitive industry)<br><br>Forecasting: Accuracy depends on several factors<br>Uncertainty: Extensive and hard to measure<br>Risk: Depends on extent of biases, strategy success |
| | UNKNOWN | **II. Known/Unknown**<br>(Unusual/special conditions, e.g., the effects of the 2007/2008 recession on the economy)<br><br>Forecasting: Inaccuracy can vary considerably<br>Uncertainty: Fat-tailed, hard to measure<br>Risks: Can be substantial, tough to manage | **IV. Unknown/Unknown (Black Swans)**<br>(Black Swans: Low probability high impact events, e.g. implications of a total collapse in global trade)<br><br>Forecasting: Impossible<br>Uncertainty: Unmeasurable<br>Risks: Unmanageable except through costly antifragile strategies |
| | | KNOWN | UNKNOWN |
| | | **FORECAST EVENTS** | |

Things can get further uncertain in Quadrant III for two reasons. First, Judgmental biases influence events for instance people fail to address obvious high impact dangers before they spiral out of control (Wucker, 2016). Second, by the inability to predict the implications of self-fulfilling and self-defeating prophecies caused by the actions and

reactions of market players. This category includes strategy and other important decisions where the forecast or the anticipation of action or plan can modify the future course of events, mainly when there is a zero-sum game where the pie is fixed. Finally, in Quadrant IV, any form of forecasting is by definition difficult, requiring analysis and evaluation of past data to determine the extent of uncertainty and risk involved. Taleb, the author of Black Swan (Taleb, 2007), is more pessimistic stating that the only way to be prepared to face Black Swans is by having established antifragile strategies that would allow one to dampen the negative consequences of Black Swans. Although there have been other writers suggesting insurance and robust strategies to cope with uncertainty and risk, Taleb has brought renewed attention to the issue of highly improbable, high stakes events and has contributed to making people aware of the need to be prepared to face them and for instance, having enough cash reserves to suffer a significant financial crisis as that of the 2007-08 or have invested in adequate capacity to handle a boom.

What needs to be emphasised is that dealing with any uncertainty involves a cost. The uncertainty that the sales forecast may be below actual demand can be dealt with by keeping enough inventories, thus avoiding the risk of losing customers. However, such inventories cost money to keep and require warehouses to be stored. In other cases, the uncertainty that a share price may decrease can be dealt with through diversification, buying baskets of stocks, thereby reducing the chance of large losses by foregoing profits, however, when individual shares increase more than the average. Similarly, antifragile actions such as keeping extra cash for unexpected crises also involves an opportunity cost as such cash could have been invested in productive areas to increase income and/or reduce costs and increase profits.

The big challenge, eloquently expressed by Bertrand Russell is that we need to learn to live without the support of comforting fairy tales, adding that is perhaps the chief achievement of philosophy “to teach us how to live without certainty, and yet without being paralysed by hesitation”. An investor should not stop investing just because of the risks involved.

## 4.2 On causality

Since the early years, humans have always been trying to answer the “why” question: what are the causal forces behind an observed result. Estimating the statistical correlation between two variables tells us little about their cause-effect relationship. Their association may be due to a lurking (extraneous) variable, unknown forces or even by chance. In the real world, there are just too many intervening, confounding, mediating variables and it is hard to access their impact using traditional statistical methods. Randomised controlled trials (RCTs) have been long considered the “gold standard” in designing scientific experiments for clinical trials. However, as with every laboratory experiment, RCTs are limited in the sense that in most cases the subjects are not observed in their natural environment (medical trials may be an exception). Furthermore, RCTs may not be at all possible in cases such as the comparison of two national economic policies.

An important step towards defining causality was made by Granger (1969) who proposed a statistical test to determine whether the (lagged) values of one time series can be used to predict the values of another series. Even if it is argued that Granger causality does not

identify true causality but only predictive causality (ability to predict a series based on another series), it can still be used to identify useful predictors, such as promotions as explanatory variables for future sales.

Structural equation models (SEMs) have also been long used for modelling the causal relationships between variables and to assess unobservable constructs. However, the linear-in-nature SEMs make assumptions with regards to the model form (which variables are included in the equations) and the distribution of the error. Pearl (2000) extends SEMs from linear to nonparametric that allow the total effect to be estimated without any explicit modelling assumptions. Pearl and Mackenzie (2018) describe how we can now answer questions about the 'how' or 'what if I do' (intervention) and 'why' or 'what if I had done' (counterfactuals). Two tools have been instrumental for these developments. One is a qualitative depiction of the model that includes the assumptions and relations between the variables of interest; such graphical depictions are called causal diagrams (Pearl, 1995). The second is the development of the causal calculus that allows for interventions by modifying a set of functions in the model (Pearl, 1993; Shpitser and Pearl, 2006; Huang and Valtorta, 2012). These tools provide the means to deal with situations where confounders and/or mediators would render impossible the methods of traditional statistics and probabilities. The theoretical developments of Pearl and his colleagues are yet to be empirically evaluated.

## 4.3 On luck (and other factors) versus skills

A few lucky investing decisions are usually enough for stock-pickers to be regarded as gurus of the stock markets. Similarly, a notorious bad weather, economic or political forecast is sometimes sufficient to destroy the careers of established professionals. Unfortunately, the human mind focuses on the salient and vibrant pieces of information, that make a story more interesting and more compelling. In such cases, we should always keep in mind that the luck of "expert" stock-pickers will eventually run out. Similarly, a single inaccurate forecast does not render us bad forecasters. Regression to the mean has taught us that an excellent landing for pilot trainees is usually followed by a worse one and vice versa. The same applies to the accuracy of forecasts.

Regardless of the convinving evidence of regression to the mean, we humans tend to attribute success to our abilities and skills and failure to bad luck. Moreover, in the event of failures, we are very skilful at inventing stories, theories and explanations of what went wrong and why we, indeed, knew what would have happened (hindsight bias). The negative relationship between actual skill/expertise and beliefs in our abilities has also been very well examined and is termed the Dunning-Kruger effect: the least skilled people over-rate their abilities.

Tetlock et al. (2015), in their Superforecasting book, enlist the qualities of "superforecasters" (individuals that consistently have a higher skill/luck ratio compared to regular forecasters). Such qualities include, among others, a 360° "dragon-fly" view, balancing under- and over-reacting to information, balancing under- and over-confidence, searching for causal forces, decomposing the problem into smaller more manageable ones,

and looking back to objectively evaluate what has happened. However, even superstars are allowed to have a bad day from time to time.

If we are in a position to provide our forecasters with the right tools and if we allow them to learn from their mistakes, then their skills will improve over time. We need to convince companies not to operate under a one-big-mistake-and-you're-out policy (Goodwin, 2017). The performance of forecasters should be tracked and monitored over time and should be compared to the accuracy of other forecasts, either statistical or judgmental. Also, directly linking motivation with improved accuracy could further aid forecast accuracy (Fildes et al., 2009); regardless how intuitive this argument might be, there are plenty of companies that still operate with motivational schemes that directly contradict the need for accuracy, as is the case of providing bonuses to salesmen that have exceeded their forecasts.

Goodwin (2017), in his book Forewarned, suggests that instead of solely evaluating the outcome of forecasts based on their resulting accuracy, we should turn our attention in evaluating the forecasting process that was used to produce the forecasts in the first place. This is particularly useful when forecast evaluation over time is either not feasible or practical, as it is in the case of one-off forecasts such as the introduction of a significant new product. In any case, even if the forecasting process is appropriately designed and implemented, we should still expect the forecasts to be 'off' in about 1 out of 20 instances assuming 95% prediction intervals, a scenario not that remote.

## 5. Conclusions

Although forecasting in the hard sciences can attain remarkable levels of accuracy this is not the case in social domains where large errors are possible and all predictions are uncertain. Forecasts are indispensable for decisions whose success depends a good deal on the accuracy of these forecasts. This paper provided a survey of the state of the art of forecasting in social science aimed at non-forecasting experts wanting to be informed on the latest developments in the field and possibly to figure out how to improve the accuracy of their predictions.

Over time forecasting has moved from the religious and the superstitious to the scientific, accumulating concrete knowledge used to improve its theoretical foundation and increase its practical value. The outcome has been enhancements in forecast accuracy and improvements in estimating the uncertainty of its predictions. A major contributor in advancing the field has been the empirical studies that provide objective evidence to compare the accuracy of the various methods and validate different hypotheses. Despite all its challenges, forecasting for social settings has improved a lot over the years.

Our discussions above suggest that more progress needs to be made in forecasting under uncertain conditions, such as unstable economic environments or when fat tails are present. Also, despite the significant advances in research around judgment, there are still many open questions, such as the conditions under which judgment is more likely to outperform statistical models and how to minimise the negative effects of judgmental heuristics and biases. More empirical studies are needed to better understand the

added-value of collecting data for exogenous variables and in which domains their inclusion into the forecasting models is likely to practically improve forecasting performance. Another research area that requires rigorous empirical investigation is that of causality and the corresonding theoretical developments.

These are areas that future forecasting competitions can focus on. We would like to see future competitions to include live forecasting tasks of high-profile economic series. We would also like to see more competitions that explore the value of exogenous variables. Competitions focusing on specific domains are also very important. We have seen in the past competitions on neural networks (Crone et al., 2011), tourism demand (Athanasopoulos et al., 2011) and energy (Hong et al., 2014; 2016; 2019); we would also like to see competitions focusing on intermittent demand and retailing, among others. Furthermore, It would be great to see more work on forecasting one-off events, in line with the Good Judgment[2] project. Last but not least, we need more understanding on how improvements in forecast accuracy translate in terms of 'profit' and how to measure the cost of forecast error.

## Acknowledgements

Thanks to Pablo Montero-Manso for providing the FFORMA forecasts for the M1 and M3 data.

## References

Adya, M., Collopy, F., Armstrong, J. S., & Kennedy, M. (2001). Automatic identification of time series features for rule-based forecasting. *International Journal of Forecasting*, *17*(2), 143–157.

Ahmed, N. K., Atiya, A. F., Gayar, N. E., & El-Shishiny, H. (2010). An Empirical Comparison of Machine Learning Models for Time Series Forecasting. *Econometric Reviews*, *29*(5-6), 594–621.

Akaike, H. (1977). On entropy maximization principle. *Application of Statistics*, 27–41.

Andrawis, R. R., Atiya, A. F., & El-Shishiny, H. (2011). Combination of long term and short term forecasts, with application to tourism demand forecasting. *International Journal of Forecasting*, *27*(3), 870–886.

Armstrong, J. S. (2006). Findings from evidence-based forecasting: Methods for reducing forecast error. *International Journal of Forecasting*, *22*(3), 583–598.

Ashley, R. (1988). On the relative worth of recent macroeconomic forecasts. *International Journal of Forecasting*, *4*(3), 363–376.

[2] https://www.gjopen.com/

Askanazi, R., Diebold, F. X., Schorfheide, F., & Shin, M. (2018). On the Comparison of Interval Forecasts. Retrieved from https://www.sas.upenn.edu/~fdiebold/papers2/Eval.pdf

Asmundsson, J. (2018). The Big Problem With Machine Learning Algorithms. *Bloomberg News*. Retrieved from https://www.bloomberg.com/news/articles/2018-10-09/the-big-problem-with-machine-learning-algorithms

Assimakopoulos, V., & Nikolopoulos, K. (2000). The Theta model: a decomposition approach to forecasting. *International Journal of Forecasting*, *16*(4), 521–530.

Athanasopoulos, G., Hyndman, R. J., Kourentzes, N., & Petropoulos, F. (2017). Forecasting with temporal hierarchies. *European Journal of Operational Research*, *262*(1), 60–74.

Athanasopoulos, G., Hyndman, R. J., Song, H., & Wu, D. C. (2011). The tourism forecasting competition. *International Journal of Forecasting*, *27*(3), 822–844.

Ben Taieb, S., & Hyndman, R. J. (2014). A gradient boosting approach to the Kaggle load forecasting competition. *International Journal of Forecasting*, *30*(2), 382–394.

Bergmeir, C., Hyndman, R. J., & Benítez, J. M. (2016). Bagging exponential smoothing methods using STL decomposition and Box–Cox transformation. *International Journal of Forecasting*, *32*(2), 303–312.

Blattberg, R. C., & Hoch, S. J. (1990). Database Models and Managerial Intuition: 50% Model + 50% Manager. *Management Science*, *36*(8), 887–899.

Bolger, F., & Wright, G. (2017). Use of expert knowledge to anticipate the future: Issues, analysis and directions. *International Journal of Forecasting*, *33*(1), 230–243.

Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung G. M. (2015). *Time Series Analysis: Forecasting and Control* (tth ed.). New Jersey: Wiley.

Brodie, R. J., Danaher, P. J., Kumar, V., & Leeflang, P. S. H. (2001). Econometric Models for Forecasting Market Share. In J. S. Armstrong (Ed.), *Principles of Forecasting: A Handbook for Researchers and Practitioners* (pp. 597–611). Boston, MA: Springer US.

Brown, R. G. (1959). *Statistical forecasting for inventory control*. McGraw-Hill.

Brown, R. G. (1963). *Smoothing, Forecasting and Prediction of Discrete Time Series*. Courier Corporation.

Burton, J. (1986). Robert FitzRoy and the Early History of the Meteorological Office. *British Journal for the History of Science*, *19*(2), 147–176.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., … Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, *351*(6280), 1433–1436.

Carbone, R., & Gorr, W. L. (1985). Accuracy of Judgmental Forecasting of Time Series. *Decision Sciences*, *16*(2), 153–160.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, *5*(4), 559–583.

Cooke, R. M., Nieboer, D., & Misiewicz, J. (2014). *Fat-Tailed Distributions: Data, Diagnostics and Dependence*. John Wiley & Sons.

Crone, S. F., Hibon, M., & Nikolopoulos, K. (2011). Advances in forecasting with neural networks? Empirical evidence from the NN3 competition on time series prediction. *International Journal of Forecasting*, *27*(3), 635–660.

Dewald, W. G., Thursby, J. G., & Anderson, R. G. (1986). Replication in Empirical Economics: The Journal of Money, Credit and Banking Project. *The American Economic Review*, *76*(4), 587–603.

Edmundson, R. H. (1990). Decomposition; a strategy for judgemental forecasting. *Journal of Forecasting*, *9*(4), 305–314.

Fildes, R. (1989). Evaluation of Aggregate and Individual Forecast Method Selection Rules. *Management Science*, *35*(9), 1056–1065.

Fildes, R., Goodwin, P., Lawrence, M., & Nikolopoulos, K. (2009). Effective forecasting and judgmental adjustments: an empirical evaluation and strategies for improvement in supply-chain planning. *International Journal of Forecasting*, *25*(1), 3–23.

Fildes, R., & Petropoulos, F. (2015). Simple versus complex selection rules for forecasting many time series. *Journal of Business Research*, *68*(8), 1692–1701.

Fildes, R., & Stekler, H. (2002). The state of macroeconomic forecasting. *Journal of Macroeconomics*, *24*(4), 435–468.

Franses, P. H. (2016). A simple test for a bubble based on growth and acceleration. *Computational Statistics & Data Analysis*, *100*, 160–169.

Franses, P. H., & Legerstee, R. (2011). Combining SKU-level sales forecasts from models and experts. *Expert Systems with Applications*, *38*(3), 2365–2370.

Franses, P. H., & Legerstee, R. (2009a). Properties of expert adjustments on model-based SKU-level forecasts. *International Journal of Forecasting*, *25*(1), 35–47.

Franses, P. H., & Legerstee, R. (2009b). Do experts' adjustments on model-based SKU-level forecasts improve forecast quality? *Journal of Forecasting*, *36*. https://doi.org/10.1002/for.1129

Freedman, D. H. (2010). Lies, Damned Lies, and Medical Science. *The Atlantic*. Retrieved from https://www.theatlantic.com/magazine/archive/2010/11/lies-damned-lies-and-medical-science/308269/

Gneiting, T., & Raftery, A. E. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, *102*(477), 359–378.

Goodwin, P. (2017). *Forewarned: A Sceptics Guide to Prediction*. Biteback Publishing.

Granger, C. W. J. (1969). Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica: Journal of the Econometric Society*, *37*(3), 424–438.

Gresnigt, F., Kole, E., & Franses, P. H. (2015). Interpreting financial market crashes as earthquakes: A new Early Warning System for medium term crashes. *Journal of Banking & Finance*, *56*, 123–139.

Halleio, E. (1704). Astronomiae Cometicae Synopsis, Autore Edmundo Halleio apud Oxonienses. Geometriae Professore Saviliano, & Reg. Soc. S. *Philosophical Transactions of the Royal Society of London Series I*, *24*, 1882–1899.

Harvey, N. (1995). Why Are Judgments Less Consistent in Less Predictable Task Situations? *Organizational Behavior and Human Decision Processes*, *63*(3), 247–263.

Harvey, N., & Bolger, F. (1996). Graphs versus tables: Effects of data presentation format on judgemental forecasting. *International Journal of Forecasting*, *12*(1), 119–137.

Heilemann, U., & Stekler, H. O. (2013). Has The Accuracy of Macroeconomic Forecasts for Germany Improved? *German Economic Review*, *14*(2), 235–253.

Hibon, M., & Evgeniou, T. (2005). To combine or not to combine: selecting among forecasts and their combinations. *International Journal of Forecasting*, *21*, 15–24.

Hong, T., Pinson, P., & Fan, S. (2014). Global Energy Forecasting Competition 2012. *International Journal of Forecasting*, *30*(2), 357–363.

Hong, T., Pinson, P., Fan, S., Zareipour, H., Troccoli, A., & Hyndman, R. J. (2016). Probabilistic energy forecasting: Global Energy Forecasting Competition 2014 and beyond. *International Journal of Forecasting*, *32*(3), 896–913.

Hong, T., Xie, J., & Black, J. (2019). Global energy forecasting competition 2017: Hierarchical probabilistic load forecasting. *International Journal of Forecasting*.

Huang, Y., & Valtorta, M. (2012). *Pearl's Calculus of Intervention Is Complete*. *arXiv [cs.AI]*. Retrieved from http://arxiv.org/abs/1206.6831

Hyndman, R., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., … Yasmeen, F. (2018). forecast: Forecasting functions for time series and linear models. Retrieved from http://pkg.robjhyndman.com/forecast

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: principles and practice* (2nd Edition). OTexts.

Hyndman, R. J., & Khandakar, Y. (2008). Automatic Time Series Forecasting: The forecast Package for R. *Journal of Statistical Software*, *27*(3), 1–22.

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, *22*(4), 679–688.

Hyndman, R. J., Koehler, A. B., Snyder, R. D., & Grose, S. (2002). A state space framework for automatic forecasting using exponential smoothing methods. *International Journal of Forecasting*, *18*(3), 439–454.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124.

Kistler, R., Kalnay, E., Collins, W., Saha, S., White, G., Woollen, J., … Fiorino, M. (2001). The NCEP–NCAR 50-Year Reanalysis: Monthly Means CD-ROM and Documentation. *Bulletin of the American Meteorological Society*, *82*(2), 247–268.

Kolassa, S. (2011). Combining exponential smoothing forecasts using Akaike weights. *International Journal of Forecasting*, *27*(2), 238–251.

Kourentzes, N., Barrow, D., & Petropoulos, F. (2018). Another look at forecast selection and combination: Evidence from forecast pooling. *International Journal of Production Economics*.

Kourentzes, N., Petropoulos, F., & Trapero, J. R. (2014). Improving forecasting by estimating time series structural components across multiple frequencies. *International Journal of Forecasting*, *30*(2), 291–302.

Langer, E. J. (1975). The illusion of control. *Journal of Personality and Social Psychology*, *32*(2), 311–328.

Lawrence, M., Goodwin, P., O'Connor, M., & Önkal, D. (2006). Judgmental forecasting: A review of progress over the last 25 years. *International Journal of Forecasting*, *22*(3), 493–518.

Lawrence, M. J., Edmundson, R. H., & O'Connor, M. J. (1985). An examination of the accuracy of judgmental extrapolation of time series. *International Journal of Forecasting*, *1*(1), 25–35.

Lawrence, M., & Makridakis, S. (1989). Factors affecting judgmental forecasts and confidence intervals. *Organizational Behavior and Human Decision Processes*, *43*(2), 172–187.

Ma, S., Fildes, R., & Huang, T. (2016). Demand forecasting with high dimensional data: The case of SKU retail sales forecasting with intra- and inter-category promotional information. *European Journal of Operational Research*, *249*(1), 245–257.

Makridakis, S. (1986). The art and science of forecasting An assessment and future directions. *International Journal of Forecasting*, *2*(1), 15–39.

Makridakis, S. (1989). Why combining works? *International Journal of Forecasting*, *5*(4), 601–603.

Makridakis, S., Chatfield, C., Hibon, M., Lawrence, M., Mills, T., Ord, K., & Simmons, L. F. (1993). The M2-competition: A real-time judgmentally based forecasting study. *International Journal of Forecasting*, *9*(1), 5–22.

Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., … Winkler, R. (1982). The accuracy of extrapolation (time series) methods: Results of a forecasting competition. *Journal of Forecasting*, *1*(2), 111–153.

Makridakis, S., & Hibon, M. (1979). Accuracy of Forecasting: An Empirical Investigation. *Journal of the Royal Statistical Society. Series A*, *142*(2), 97–145.

Makridakis, S., & Hibon, M. (2000). The M3-Competition: results, conclusions and implications. *International Journal of Forecasting*, *16*(4), 451–476.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018a). The M4 Competition: Results, findings, conclusion and way forward. *International Journal of Forecasting*.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018b). Statistical and Machine Learning forecasting methods: Concerns and ways forward. *PloS One*, *13*(3), e0194889.

Makridakis, S., & Taleb, N. (2009). Decision making and planning under low levels of predictability. *International Journal of Forecasting*, *25*(4), 716–733.

Mente, A., & Yusuf, S. (2018). Evolving evidence about diet and health. *The Lancet. Public Health*, *3*(9), e408–e409.

Micha, R., Wallace, S. K., & Mozaffarian, D. (2010). Red and processed meat consumption and risk of incident coronary heart disease, stroke, and diabetes mellitus: a systematic review and meta-analysis. *Circulation*, *121*(21), 2271–2283.

Montero-Manso, P., Athanasopoulos, G., Hyndman, R. J., & Talagala, T. S. (2019). FFORMA: Feature-Based Forecast Model Averaging. *International Journal of Forecasting*.

O'Connor, M., Remus, W., & Griggs, K. (1993). Judgemental forecasting in times of change. *International Journal of Forecasting*, *9*(2), 163–172.

Pant, P. N., & Starbuck, W. H. (1990). Innocents in the Forest: Forecasting and Research Methods. *Journal of Management*, 16(2), 433–460.

Pearl, J. (1993). Comment: Graphical Models, Causality and Intervention. *Statistical Science: A Review Journal of the Institute of Mathematical Statistics*, *8*(3), 266–269.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, *82*(4), 669–688.

Pearl, J. (2000). Causality: Models, reasoning, and inference, *384*. Retrieved from https://psycnet.apa.org/fulltext/2000-07461-000.pdf

Pearl, J., & Mackenzie, D. (2018). *The Book of Why: The New Science of Cause and Effect*. Allen Lane.

Petropoulos, F., Goodwin, P., & Fildes, R. (2017). Using a rolling training approach to improve judgmental extrapolations elicited from forecasters with technical knowledge. *International Journal of Forecasting*, *33*(1), 314–324.

Petropoulos, F., Hyndman, R. J., & Bergmeir, C. (2018a). Exploring the sources of uncertainty: Why does bagging for time series forecasting work? *European Journal of Operational Research*.

Petropoulos, F., Kourentzes, N., Nikolopoulos, K., & Siemsen, E. (2018b). Judgmental selection of forecasting models. *Journal of Operations Management*, *60*, 34–46.

Petropoulos, F., Makridakis, S., Assimakopoulos, V., & Nikolopoulos, K. (2014). `Horses for Courses' in demand forecasting. *European Journal of Operational Research*, *237*, 152–163.

Reimers, S., & Harvey, N. (2011). Sensitivity to autocorrelation in judgmental time series forecasting. *International Journal of Forecasting*, *27*(4), 1196–1214.

Saha, S., Moorthi, S., Wu, X., Wang, J., Nadiga, S., Tripp, P., … Becker, E. (2014). The NCEP Climate Forecast System Version 2. *Journal of Climate*, *27*(6), 2185–2208.

Salinas, D., Flunkert, V., & Gasthaus, J. (2017). *DeepAR: Probabilistic Forecasting with Autoregressive Recurrent Networks*. *arXiv [cs.AI]*. Retrieved from http://arxiv.org/abs/1704.04110

Sanders, N. R. (1992). Accuracy of judgmental forecasts: A comparison. *Omega*, *20*(3), 353–364.

Satariano, A., & Kumar, N. (2017). The Massive Hedge Fund Betting on AI. *Bloomberg News*. Retrieved from https://www.bloomberg.com/news/features/2017-09-27/the-massive-hedge-fund-betting-on-ai

Schnaars, S. P. (1989). *Megamistakes: Forecasting and the Myth of Rapid Technological Change* (29th ed. edition). The Free Press.

Scott, M. (2015). *Delphi: A History of the Center of the Ancient World*. Princeton University Press.

Shpitser, I., & Pearl, J. (2006). Identification of conditional interventional distributions. In *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence, UAI 2006* (pp. 437–444).

Stekler, H. O. (2007). The future of macroeconomic forecasting: Understanding the forecasting process. *International Journal of Forecasting*, *23*(2), 237–248.

Stekler, H., & Symington, H. (2016). Evaluating qualitative forecasts: The FOMC minutes, 2006–2010. *International Journal of Forecasting*, *32*(2), 559–570.

Surowiecki, J. (2005). *The Wisdom of Crowds: Why the Many Are Smarter Than the Few* (New Ed edition). Abacus.

Taleb, N. N. (2007). *The Black Swan: The Impact of the Highly Improbable*. Penguin.

Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: an analysis and review. *International Journal of Forecasting*, *16*(4), 437–450.

Tetlock, P. E. (2006). *Expert Political Judgment: How Good Is It? How Can We Know?* (New Ed edition). Princeton University Press.

Tetlock, P., Gardner, D., & Richards, J. (2015). *Superforecasting: The Art and Science of Prediction* (Unabridged edition). Audible Studios on Brilliance.

Timmermann, A. (2006). Forecast combinations. In C. W. J. G. G. Elliott & A. Timmermann (Eds.), *Handbook of Economic Forecasting* (Vol. 1, pp. 135–196). Elsevier.

Trapero, J. R., Pedregal, D. J., Fildes, R., & Kourentzes, N. (2013). Analysis of judgmental adjustments in the presence of promotions. *International Journal of Forecasting*, *29*(2), 234–243.

Wang, J., & Wang, J. (2017). Forecasting stochastic neural network based on financial empirical mode decomposition. *Neural Networks: The Official Journal of the International Neural Network Society*, *90*, 8–20.

Wang, X., & Petropoulos, F. (2016). To select or to combine? The inventory performance of model and expert forecasts. *International Journal of Production Research*, *54*(17), 5271–5282.

Wang, X., Smith-Miles, K., & Hyndman, R. (2009). Rule induction for forecasting method selection: Meta-learning the characteristics of univariate time series. *Neurocomputing*, *72*(10–12), 2581–2594.

Webby, R., O'Connor, M., & Edmundson, B. (2005). Forecasting support systems for the incorporation of event information: An empirical investigation. *International Journal of Forecasting*, *21*(3), 411–423.

Wen, R., Torkkola, K., Narayanaswamy, B., & Madeka, D. (2017). *A Multi-Horizon Quantile Recurrent Forecaster*. *arXiv [stat.ML]*. Retrieved from http://arxiv.org/abs/1711.11053

Winkler, R. L. (1972). A Decision-Theoretic Approach to Interval Estimation. *Journal of the American Statistical Association*, *67*(337), 187–191.

Wucker, M. (2016). *The Gray Rhino: How to Recognize and Act on the Obvious Dangers We Ignore*. St. Martin's Press.