JUDGMENTAL MODEL SELECTION

Fotios Petropoulos
School of Management, University of Bath, UK

PREVIEW Although judgment plays a significant role in the production and acceptance of forecasts, its performance in model selection has not been tested. In this article, Fotios demonstrates that the application of judgment to the selection of a forecasting model can improve forecast accuracy and presents the conditions under which this is likely to be the case.

KEY POINTS

- Despite the rich literature on forecasting with judgment, one area that has attracted little attention is that of using judgment to choose between different statistical models, which is surprising, since modern forecasting support systems allow the user to either press the magic button or select a model manually.
- The application of judgment to the model selection process has intrinsic appeal: it allows forecasters to make a mental extrapolation of what a reasonable forecast should look like and hence reject models that produce unreasonable forecasts; and by participating in the model selection process, forecasters are better able to "own" the forecasts, thus limiting unnecessary judgmental adjustments to the statistical forecasts.
- In order to see how effectively judgment could be applied to the selection of forecasting models, a sample of 700 users was presented with a pair of *user interfaces:* one in which users were shown a list of model options and told to go through these options and select one model and another in which the users were asked whether or not trend and/or seasonal patterns exist in the data. Results based on the MAPE suggest that both groups of participants perform much better on average than the automatic statistical algorithm.
- A 50-50 combination of judgmental selection and the statistical algorithm and judgmental aggregation of the selections of multiple users are strategies that result in superior performance to both statistical and judgmental selection.

FORECASTING WITH JUDGMENT

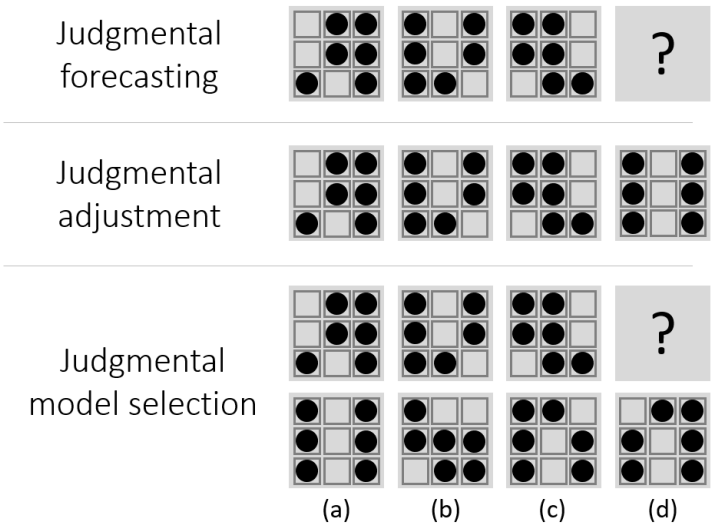
In his featured talk during ISF2013 in Seoul, "Forecasting Without Forecasters", Rob Hyndman was demonstrating how significant has been the progress in statistical forecasting algorithms. These algorithms can automatically identify, through information criteria, the best model within the exponential smoothing or ARIMA families. Even if I agree that advancements in automatic statistical forecasting are significant and allow for batch production of a large number of forecasts, I disagree with the statement that we should now forecast without forecasters.

As the highly-cited review paper from Lawrence and colleagues (2006) has shown, judgment plays a very important role in forecasting and is integrated into numerous stages of the forecasting process. Judgment may be used directly to produce point forecasts or prediction intervals or to adjust a statistical baseline forecast in the light of information that is not captured in the model (see, for example, the Foresight article by Fildes & Goodwin, 2007). When judgment is used, there is evidence that its performance can be improved by timely and salient feedback as well as by aggregation (combining the judgments of different forecasters). In some cases, decomposition of a complex task into smaller, manageable subtasks can also help.

But despite the rich literature on forecasting with judgment, one area that has attracted little attention is that of using judgment to choose between different statistical models. This is surprising, since modern forecasting support systems allow the user to either press the magic button or select a model manually. For instance, the commercial software, *ForecastPro* offers an automatic model selection option called "expert selection", but also provides a user interface for selecting a specific model, such as single exponential smoothing or Holt-Winters. Additionally, SAP's ERP offers a manual model selection feature in its Advanced Planning and Optimization functionalities. Until recently, however, the empirical performance of judgmental model selection had not been investigated.

So as to better grasp the difference in the mechanics between judgmental point forecasting, judgmental adjustment of a statistical baseline and judgmental model selection, we can use a standard IQ question as an analogy **(Figure 1)**, in which the requirement is to logically define how the last of the four objects in a sequence should look like. In the first instance (judgmental forecasting), we would be given a sequence of three boxes showing different patterns of dots and we would be asked to fill in the blank fourth box. In the second instance (judgmental adjustment), we would be given the same sequence of three boxes together with a pre-filled fourth box and we would be asked to make changes on the fourth box (if we felt that changes are needed) so that it better fits the sequence. In the third instance (judgmental model selection), we would be given four distinct options (a, b, c and d) and we would be asked which one of these should be the fourth box in the sequence.

Figure 1. Forecasting with Judgment: an IQ Test Analogy.



There are two reasons we would expect that judgmental selection of statistical models will work well. The first and most important one is that forecasters, when visually exploring the forecasts from a pool of candidate models, can assess the quality of these forecast by comparing them against a mental extrapolation of what a reasonable forecast should look like. This makes it possible for forecasters to reject models that produce unreasonable forecasts. In contrast, automatic statistical model selection approaches can only assess the performance of the candidate models on historical data, by either measuring the penalized-for-overfitting goodness-of-fit or by evaluating the performance on a validation set.

Secondly, by allowing forecasters to participate in the model selection phase of the forecasting process, such participation might fulfil their wishes of "owning" the forecasts, thus limiting unnecessary judgmental adjustments to the statistical forecasts (Fildes & Goodwin, 2007).

Of course, biases associated with judgment, such as over-optimism and inconsistency, may still be present, but we expect that their negative effect will decrease as the humans' input will be limited to a discrete number of (statistical) choices.

EXPLORING THE PERFORMANCE OF JUDGMENTAL MODEL SELECTION

Along with three colleagues (Petropoulos and colleagues, 2018) I ran an experiment to explored the performance of judgmental model selection for what we believe is the first such research. Our objectives were threefold:

- First, we wished to compare the performance of algorithmic/automatic versus human judgment selection of forecasting models and to determine if humans and algorithms select models differently.
- Second, we sought to understand when judgmental selection performs better.
- Third, we also explored the performance of models that combine statistical and judgmentally selected models. Previous research has shown that both simple combinations of forecasts (50% statistics + 50% judgment) and judgmental aggregation (wisdom of crowds) have worked well in many situations: We felt it would be interesting to see if these strategies also work well for model selection.

Statistical Algorithm vs Human Judgment in Model Selection.

We assumed that automatic statistical selection is done on the basis of information criteria, metrics such as Akaike's Information Criterion (AIC) or the Bayesian Information Criterion (BIC) that measure goodness of fit to past data adjusted to penalize model complexity. Overly complex models can see patterns in what are really random movements in data histories and mistakenly extrapolate these patterns when they produce their forecasts. Use of information criteria is implemented in the very popular *forecast* package for R statistical software.

However, you could instead make an automatic model selection based on the out-of-sample performance of different models. Here the available data are divided into fit and validation sets. Models are fitted using the first set, and their performance is evaluated in the second set. The model with the best performance in the validation set is put forward to produce forecasts for the future. The decision maker can choose the appropriate accuracy metric. The preferred measure can reflect the costs of any errors associated with the forecasts.

Model selection on out of sample performance has two advantages over selection based on information criteria. First, the performance of multiple step-ahead forecasts can be used to inform selection. Second, the validation approach is able to evaluate forecasts derived from any process (including combinations of forecasts from various models). The disadvantage of this approach is that it requires setting aside a validation set, which may not always be feasible, especially for short time series. Given that product life cycles are shortening, having a validation sample available can be a real luxury for forecasters.

Under what conditions does judgmental selection perform better?

To this end, we compared two user interfaces which presented users with a graph of a monthly time series over five years and the forecasts of a particular model for the following twelve months. The first, which we call *model selection*, simply provided candidate models as a list of

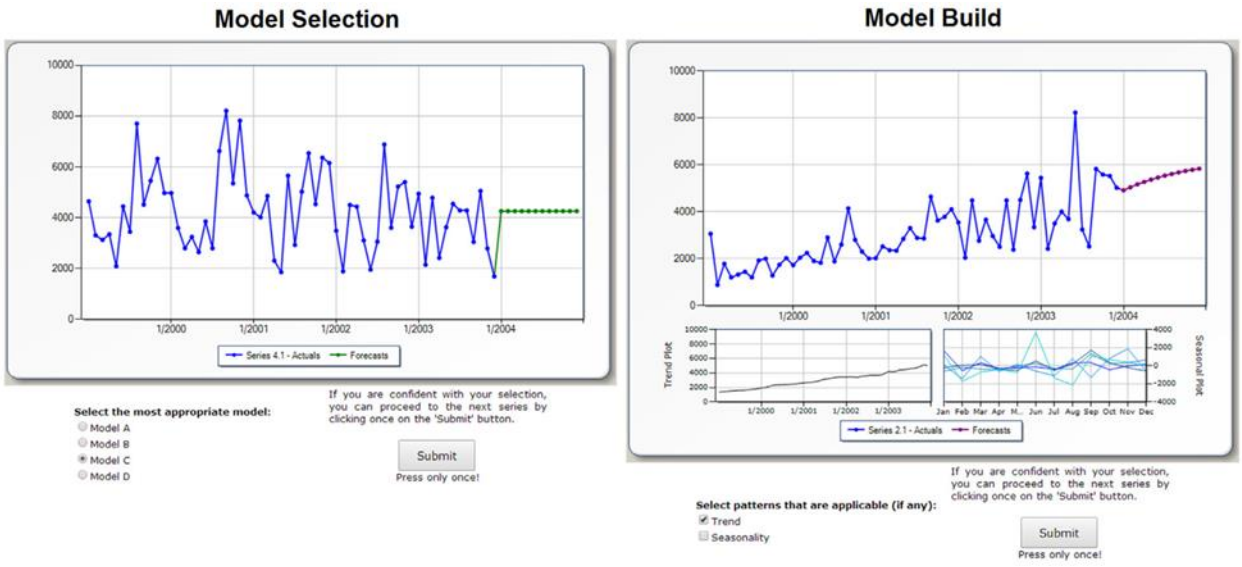
options and users could go through these options and select one. This is very similar to the default interface in the majority of modern forecasting support systems.

The second interface prompted the user to decide if the graphically displayed series exhibited a trend and/or seasonality and on the basis of this decision, a model was selected in the background. This interface we called *model-build* and is based on the principles of decomposition. The model-build interface also provides a trend and a seasonal plot which are built on the in-sample data and support the user in choosing the relevant components.

In both interfaces, a change in the selected model (directly in the case of the model selection interface by clicking another radio button; indirectly in the case of the model-build interface by changes the selected components) will automatically refresh the graph and the forecasts of the new model are displayed.

Figure 2 illustrates the two user interfaces. Both encompass the same four exponential smoothing models (level only, trended only, seasonal only, trended and seasonal) to allow for a direct comparison.

Figure 2. Two Interfaces (Adopted from Petropoulos and colleagues, 2018).



THE BEHAVIORAL EXPERIMENT

We designed a behavioral experiment which was distributed via the Web and completed by almost 700 participants (students studying operations and/or forecasting, forecasting researchers and forecasting practitioners) around the globe. Each of the participants was randomly assigned

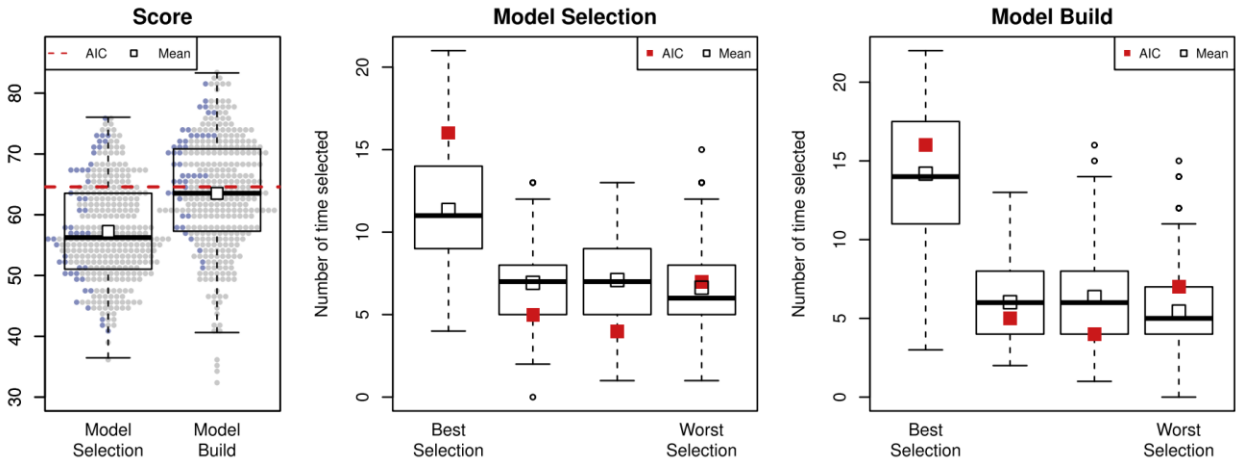
to one user interface, model selection or model-build, and was asked to apply judgment to select a model for each of 32 time series which exhibited a variety of characteristics (trend and seasonality). Subsequently, their judgments were compared against the automatic statistical selections.

Algorithms vs Judgment

To rate the performance of individuals in judgmentally selecting forecasting models, we assigned a score of 3, 2, 1 or 0 points each time they respectively selected the best, second best, third best or worst of the four models (the rank of each model was decided after comparing their forecasts to the actual future data). The points for each participant were summed to get a total for the 32 time series and then were standardized into a 0-100% scale. The results are illustrated in **Figure 3.**

The standardized performance score for each participant is represented by a point in the left panel of Figure 3. Scores for forecasting practitioners are depicted with blue dots and all other participants with grey dots. The red dashed line is the comparable score of the automatic statistical selection (based on the AIC). The middle and right panels show the percentage of time the best, second best, third best and worst models were selected for each user interface. In each case, we summarize the results using boxplots, where the horizontal boundaries of the boxes are the quartiles and the horizontal bold line within the boxes is the median.

Figure 3. Performance of individual judgmental model selection. (Adopted from Petropoulos and colleagues, 2018).



We offer these key conclusions:

- The performance of different participants significantly varies, even within the same user interface. The range for the model-build scores spans between 32% and 83%. This means

that some participants significantly outperform the automatic statistical selection (which scored 65%), while others fall far short.

- On average, participants using the *model-build interface* are as good as the automatic model selection. The performance of participants using the *model selection interface* was inferior.
- If we compare the blue dots (practitioners) with the grey dots (all other participants), we cannot observe any significant difference in performance. This is supported by formal statistical tests. This result is supportive of the recruitment of students for behavioural experiments on time series forecasting.
- From the middle and left panel of Figure 3, we see that humans select the best model less frequently that the algorithm does. This is particularly true for participants assigned to the model selection interface. At the same time, however, humans succeed in avoiding the worst model more often than the algorithm did. Apparently, when humans visually explore the forecasts from different models they are able to reject the most unreasonable models, performing the mental evaluation of what constitutes a reasonable extrapolation. This is an exciting result and should stimulate more research into how automatic model selection can incorporated mental extrapolations.
- We also compared the forecast accuracy of the judgmental model selections versus that from the automatic model selection. The accuracy metrics used were the mean absolute percentage error (MAPE) and the mean absolute scaled error (MASE). Results based on the MAPE suggest that both participants using either interface, model selection or model-build, perform much better than the automatic statistical selection on the average. Results based on the MASE match the observations of the percentage score presented above (for a description of the MASE, see Hyndman, 2006). Even if humans are not always able to pick the best model, the avoidance of bad models is crucial with regards to average forecast accuracy.

Combined Judgement and Statistical Models

We tested the *combined performance* of automatic model selection and judgmental model selection in this way:

> *If a participant selected a model that differed from that based on the automatic selection, then the forecasts of the two models were combined with equal weights.*

This 50-50% combination resulted in forecasting performance superior to both automatic statistical selection and individual judgmental selection. In fact, when the judgmental selection of each individual is combined with the statistical selection, the performance for 90% of the participants is better (in terms of MASE) than simply using the automatic statistical selection.

Additionally, a 50%-50% combination brings robustness in the sense that the variance in the performance between subjects is halved.

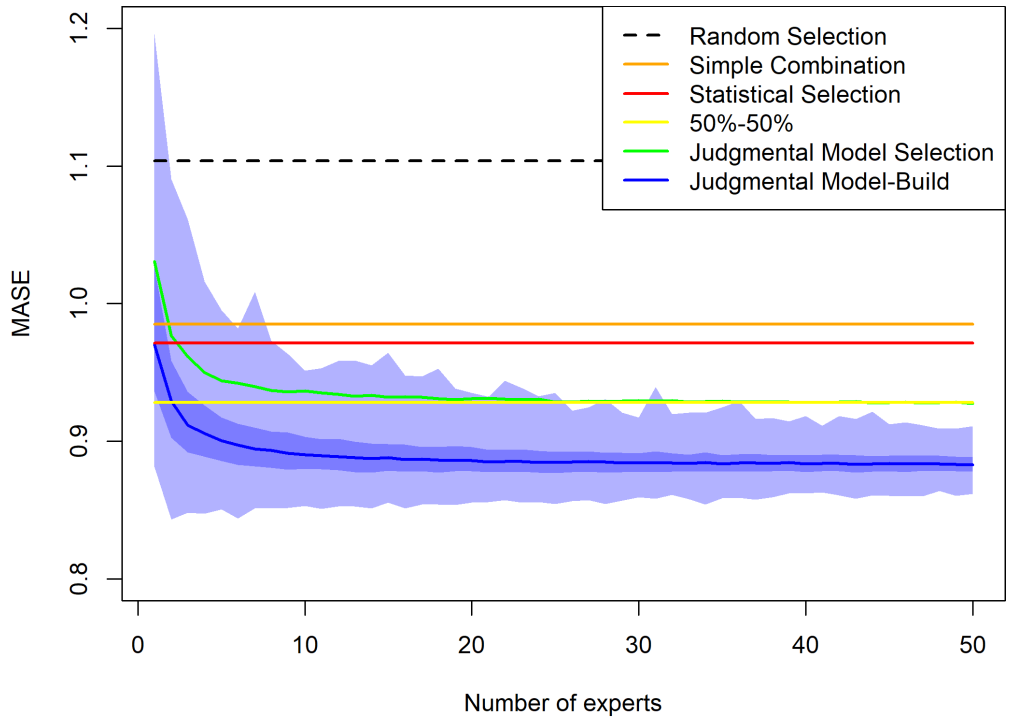
The Wisdom of Crowds

We also examined the performance of groups of participants (wisdom of crowds). We aggregated participants' model selections. For example, assume that we randomly select 5 participants from our sample of almost 700 participants. As these participants might have selected different models for each series, we calculated the weighted forecast combinations of their selected models by assigning linear weights to the forecasting models: the more often a particular model was selected the higher the weight. Consequently, we can calculate the MASE across series. We repeated this for a total 1000 groups of 5 participants each, with resampling. We also assessed the sensitivity of the results to the size of the groups, considering groups from 1 to 50 participants.

The results for the model-build interface are presented in Figure 4 in the form of "boxplots". The horizontal axis shows the size of the group (number of experts). For a given group-size, the blue line represents the median performance, the dark blue area represents the box of the boxplot (with its limits being the upper and lower quartiles) and the limits of the light blue area shows the maximum and minimum values of the MASE. The green line presents the average performance for the wisdom of crowds using the model selection interface. In the same plot, we also present the statistical benchmark (red), the average performance of the 50%-50% combination (yellow) as well as two additional simple benchmarks: random selection (black dashed line) and simple combination of all candidate models (amber). The results suggest that:

- Two experts are enough to outperform the statistical selection in 75% of the cases.
- Five experts will suffice to almost always outperform the statistical selection.
- The average performance and the variance in the performance across randomly sampled groups of experts is not noticeably improved when the selections of more than 20 experts are aggregated.

Figure 4. Summary of the Results



WHY MODEL-BUILD WORKS BETTER

I participated in another research project (Han and colleagues, 2018) that investigated the reasons behind the good performance of *model-build* compared to *model selection*. We designed a process in which users were exposed to trended or non-trended series and then were asked to perform:

- Model selection: select the forecasts from a non-trended (single exponential smoothing) or a trended (Holt's linear trend) exponential smoothing model.
- Model-build: decide if the presented series has a trend or not.

An innovative aspect of this experiment is that we managed to measure the cognitive effort that was required to perform each of the above tasks. This was achieved by applying a psycho-physiological process – an electroencephalogram (EEG) – to capture the brain activity. The EEG records the changes in the electric potential in different regions of the scalp through a number of electrodes (small metal discs attached on wires). We focused on three particular changes in the electrical potential that the neuroscience literature has connected with attention and cognitive load and working memory.

The results from this research confirmed the empirical superiority of model-build and, additionally, showed that:

- Users require more time to submit their choices under the model selection interface compared to the model-build.
- Model-build requires significantly less mental effort.
- Model selection is associated with increased working memory storage and retrieval.

In addition, we found that regardless of the user interface used, the superiority of judgmental model selection grows as the noisiness in the series decreases, the strength of trend increases, or the trend has a positive direction.

FINAL COMMENTS

Judgmental model selection has been and underexplored concept despite the imposing fact that it is explicitly offered by major forecasting providers and widely exercised. This paper provided a summary of the results of the two published papers on this area. On balance, we can say that there is much potential for a model selection process that is informed by judgment, either in terms of combining judgmental and statistical selections or even by aggregating multiple judgmental selections.

And new avenues are opening for the designers of forecasting software:

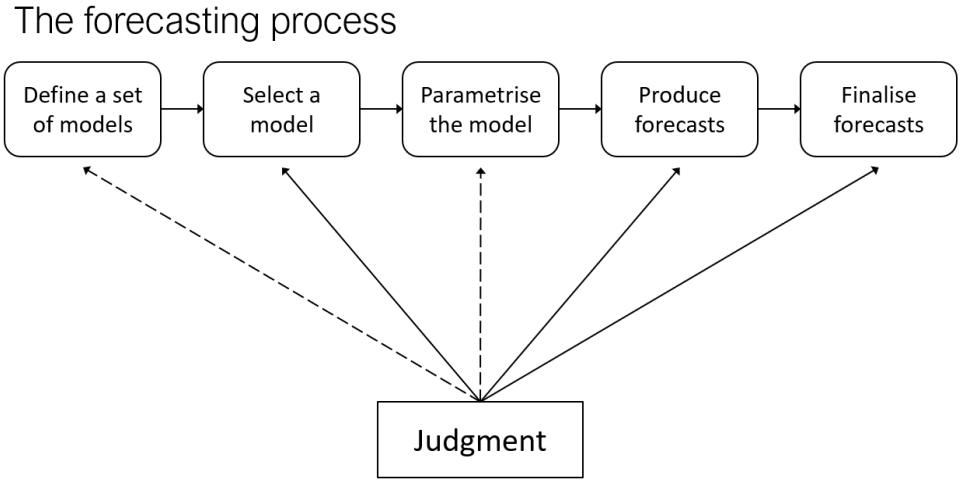
- Forecasting software should not only allow judgmentally selecting models but also should encourage and support its application.
- Judgmental selections could be used independently, but we suggest that these should be combined with the automatic selections of the forecasting software.
- The superior performance of the model-build interface over the simpler model selection suggests that some changes need to be made in the way that software guide users to manually select models. The adoption of the new model-build interface will also lead to a substantial decrease in the cognitive load of the users.

A big question is whether such an approach can be scaled, especially given that modern inventory settings may involve hundreds of thousands of stock-keeping units (SKUs). The simple answer is that we may not need to scale it if we simply focus on what matters: the most important and hard-to-forecast items. These would be the AZ items in a standard ABC/XYZ classification, where ABC is the Pareto classification for importance (20% of the SKUs are classified as the most important, A; 50% as the least important, C) and XYZ the respective classification for forecastability (50% of the SKUs are classified as the most forecastable, X; 20% as the least forecastable, Z).

Finally, there are other unexplored opportunities for injecting judgment into the model selection as summarized in **Figure 5**. Apart from selecting models, producing forecasts and finalizing

forecasts through adjustments, judgment could be potentially used to define a set of contender models (from the long-list to the short-list) and, (why not?) to select the parameter values of the models, such as the smoothing parameters in the exponential smoothing models.

Figure 5. Judgment within the forecasting process



References

1. Fildes R. & Goodwin P. (2007) "Good and bad judgment in forecasting: lessons from four companies", *Foresight: The International Journal of Applied Forecasting*, Issue 8: 5-10
2. Han W., Wang X., Petropoulos F. & Wang J. (2018) "Brain imaging and forecasting: insights from judgmental model selection", *Omega: The International Journal of Management Science*, https://doi.org/10.1016/j.omega.2018.11.015
3. Hyndman, R. J. (2006) "Another look at forecast-accuracy metrics for intermittent demand", *Foresight: The International Journal of Applied Forecasting*, Issue 4: 43-46
4. Lawrence M., Goodwin P., O'Connor M. & Onkal D. (2006) "Judgmental forecasting: A review of progress over the last 25 years", *International Journal of Forecasting*, 22(3): 493-518
5. Petropoulos F., Kourentzes N., Nikolopoulos K. & Siemsen E. (2018) "Judgmental Selection of Forecasting Models", *Journal of Operations Management*, 60: 34-46