Judgmental Selection of Forecasting Models

Fotios Petropoulos[a], Nikolaos Kourentzes[b], Konstantinos Nikolopoulos[c], Enno Siemsen[d,*]

[a]*School of Management, University of Bath, UK*
[b]*Lancaster University Management School, Lancaster University, UK*
[c]*Bangor Business School, Bangor University, UK*
[d]*Wisconsin School of Business, University of Wisconsin, USA*

Abstract

In this paper, we explored how judgment can be used to improve the selection of a forecasting model. We compared the performance of judgmental model selection against a standard algorithm based on information criteria. We also examined the efficacy of a judgmental model-build approach, in which experts were asked to decide on the existence of the structural components (trend and seasonality) of the time series instead of directly selecting a model from a choice set. Our behavioral study used data from almost 700 participants, including forecasting practitioners. The results from our experiment suggest that selecting models judgmentally results in performance that is on par, if not better, to that of algorithmic selection. Further, judgmental model selection helps to avoid the worst models more frequently compared to algorithmic selection. Finally, a simple combination of the statistical and judgmental selections and judgmental aggregation significantly outperform both statistical and judgmental selections.

Keywords: model selection, behavioral operations, decomposition, combination

1. Introduction

Planning processes in operations - e.g., capacity, production, inventory, and materials requirement plans - rely on a demand forecast. The quality of these plans depends on the accuracy of this forecast. This relationship is well documented (Gardner, 1990; Ritzman and King, 1993; Sanders and Graman, 2009; Oliva and Watson, 2009). Small improvements

*Correspondence: Enno Siemsen, Operations and Information Management Department, Wisconsin School of Business, University of Wisconsin-Madison, Madison, Wisconsin 53706, USA.

Email addresses: f.petropoulos@bath.ac.uk (Fotios Petropoulos), n.kourentzes@lancaster.ac.uk (Nikolaos Kourentzes), k.nikolopoulos@bangor.ac.uk (Konstantinos Nikolopoulos), esiemsen@wisc.edu (Enno Siemsen)

in forecast accuracy can lead to large reductions in inventory and increases in service levels. There is thus a long history of research in operations management that examines forecasting processes (Seifert et al., 2015; Nenova and May, 2016; van der Laan et al., 2016, are recent examples).

Forecasting model selection has attracted considerable academic and practitioner attention during the last 30 years. There are many models to choose from – different forms of exponential smoothing, autoregressive integrated moving average (ARIMA) models, neural nets, etc. – and forecasters in practice have to select which one to use. Many academic studies have examined different statistical selection methodologies to identify the best model; the holy grail in forecasting research (Petropoulos et al., 2014). If the most appropriate model for each time series can be determined, forecasting accuracy can be significantly improved (Fildes, 2001), typically by as much as 25-30% (Fildes and Petropoulos, 2015).

In general, forecasting software recommends or selects a model based on a statistical algorithm. The performance of candidate models is evaluated either on in-sample data, usually using appropriate information criteria (Burnham and Anderson, 2002), or by withholding a set of data points to create a validation sample (out-of-sample evaluation, Ord et al., 2017, also known as cross-validated error). However, it is easy to devise examples in which statistical model selection (based either on in-sample or out-of-sample evaluation) fails. Such cases are common in real forecasting applications and thus make forecasting model selection a non-trivial task in practice.

Practitioners can apply judgment to different tasks within the forecasting process, namely:

1. definition of a set of candidate models,
2. selection of a model,
3. parametrization of models,
4. production of forecasts, and
5. forecast revisions/adjustments.

Most of the attention in the judgmental forecasting literature focuses on the latter two tasks. Experts are either asked to directly estimate the point forecasts of future values of an event or a time series (see for example Hogarth and Makridakis, 1981; Petropoulos et al., 2017), or they are asked to adjust (or correct) the estimates provided by a statistical method in order to take additional information into account; such information is often called soft data, such as information from the sales team (Fildes et al., 2009).

However, little research has examined the role and importance of human judgment in the other three tasks. In particular, Bunn and Wright (1991) referred to the problem of

judgmental model selection (item 2 in the above list), suggesting that the selection of the most appropriate model(s) can be based on human judgment. They also emphasized the dearth of research in this area. Importantly, the majority of the world-leading forecasting support systems allow human judgment as the final arbiter among a set of possible models.[1] Therefore, the lack of research into how well humans perform this task remains a substantive gap in the literature.

In this study, we examined how well human judgment performs in model selection compared with an algorithm using a large-scale behavioral experiment. We analyzed the efficiency of judgmental model selection of individuals as well as groups of participants. The frequency of selecting the best and worst models provides suggestions on the efficacy of each approach. Moreover, we identified the process that most likely will choose models that lead to improved forecasting performance.

The rest of our paper is organized as follows. The next section provides an overview of the literature concerning model selection for forecasting. The design of the experiment to support the data collection is presented in section 3. Section 4 shows the results of our study. Section 5 discusses the implications for theory, practice, and implementation. Finally, section 6 contains our conclusions.

2. Literature

2.1. Commonly used forecasting models

Business forecasting is commonly based on simple, univariate models. One of the most widely used families of models are exponential smoothing models. Thirty different models fall into this family (Hyndman et al., 2008). Exponential smoothing models are usually abbreviated as ETS, which stands for either ExponenTial Smoothing or Error, Trend, Seasonality (the three terms in such models). More specifically, the error term may be either additive (A) or multiplicative (M), whereas trend and seasonality may be none (N), additive (A), or multiplicative (M). Also, the trend can be linear or damped (d). As an example, ETS(M,Ad,A) refers to an exponential smoothing model with a multiplicative error term, a damped additive trend, and additive seasonality. Maximum likelihood estimation is used to find model parameters that produce optimal one-step-ahead in-sample predictions (Hyndman and Khandakar, 2008).

[1]For example, see the 'Manual Model Selection' feature of SAP Advanced Planning and Optimization (SAP APO), on SAP ERP: https://help.sap.com/viewer/c95f1f0dcd9549628efa8d7d653da63e/7.0.4/en-US/822bc95360267614e10000000a174cb4.html

These models are widely used in practice. In a survey of forecasting practices, the exponential smoothing family of models is the most frequently used (Weller and Crone, 2012). In fact, it is used in almost 1/3 of times (32.1%), with averages coming second (28.1%) and naive methods third (15.4%). More advanced forecasting techniques are only used in 10% of cases. In general, simpler methods are used 3/4 times, a result that is consistent with the relative accuracy of such methods in forecasting competitions. Furthermore, an empirical study that evaluated forecasting practices and judgmental adjustments reveals that "the most common approach to forecasting demand in support of supply chain planning involves the use of a statistical software system which incorporates a simple univariate forecasting method, such as exponential smoothing, to produce an initial forecast" (Fildes et al., 2009, p. 4), while it specifies that three out of four companies examined "use systems that are based on variants of exponential smoothing" (Fildes et al., 2009, p. 7).

There are many alternatives to exponential smoothing for producing business forecasts, such as neural networks and other machine learning methods. Nevertheless, time series extrapolative methods remain very attractive. This is due to their proven track record in practice (Gardner, 2006) as well as their relative performance compared to more complex methods (Makridakis and Hibon, 2000; Armstrong, 2006; Crone et al., 2011). Furthermore, time series methods are fairly intuitive, which makes them easy to specify and use, and enhances their acceptance by the end-users (Dietvorst et al., 2015; Alvarado-Valencia et al., 2017). Complex methods, such as many machine learning algorithms, often appear as black boxes, and provide limited or no insights into how the forecasts are produced and which data elements are important. These attributes of forecasting are often critical for users (Sagaert et al., 2018).

2.2. Algorithmic model selection

Automatic algorithms for model selection are often built on information criteria (Burnham and Anderson, 2002; Hyndman et al., 2002). Models within a certain family (such as exponential smoothing or ARIMA) are fitted to the data, and the model with the minimum value for a specific information criterion is selected as the best. Various information criteria have been considered, such as Akaike's Information Criterion (AIC) or the Bayesian Information Criterion (BIC). The AIC after correction for small sample sizes (AICc) is often recommended as the default option because it is an appropriate criterion for short time series and it differs only minimally from the conventional AIC for longer time series (Burnham and Anderson, 2002). However, research also suggests that if we focus solely on out-of-sample forecasting accuracy, the various information criteria may choose different models

that nonetheless result in almost the same forecast accuracy (Billah et al., 2006).

Information criteria are based on the optimized likelihood function penalized by model complexity. Using a model with optimal likelihood inadvertently assumes that the postulated model is true (Xia and Tong, 2011). In a forecasting context, this assumption manifests itself as follows: The likelihood approach generally optimizes the one-step-ahead errors; for the forecasts to be optimal for multi-step ahead forecasts, the resulting model parameters should be optimal for any longer horizon error distribution as well. This will only occur if the model is true, in which case the model fully describes the structure of the series. Otherwise, the error distributions will vary with the time horizon (Chatfield, 2000). Such time-horizon dependent error distributions are often observed in reality (Barrow and Kourentzes, 2016), providing evidence that any model merely approximates the underlying unknown true process. Not recognizing this can lead to a biased model selection which favors one-step-ahead performance at the expense of longer time horizons that may well be the analyst's real objective.

An alternative to selecting models via information criteria is to measure the performance of different models in a validation set (Fildes and Petropoulos, 2015; Ord et al., 2017). The available data are divided into fitting 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 measure. The preferred measure can directly match the actual cost function that is used to evaluate the final forecasts.

Forecasts for validation purposes may be produced only once (also known as fixed-origin validation) or multiple times (rolling-origin), which is the cross-validation equivalent for time series data. Evaluating forecasts over multiple origins has several advantages, most importantly their robustness against the peculiarities in data that may appear within a single validation window (Tashman, 2000). Model selection on (cross-)validation 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. Given that product life cycles are shortening, having a validation sample available can be an out of reach luxury for forecasters.

A final category for automatic model selection involves measurement of various time series characteristics (such as trend, seasonality, randomness, skewness, intermittence, vari-

ability, number of available observations) as well as consideration of decision variables (such as the forecast horizon). Appropriate models are selected based on expert rules (Collopy and Armstrong, 1992; Adya et al., 2001) or meta-learning procedures (Wang et al., 2009; Petropoulos et al., 2014). However, such approaches are very sensitive to the selected rules or meta-learning features. No widely accepted set of such rules exists.

Regardless of the approach used for the *automatic* selection of the best model, all processes outlined above (information criteria, validation, and selecting based on rules) are based on statistics or can be implemented through an algorithmic process. A commonality among all algorithmic model selection approaches is that selection is based on historical data. None of these algorithms can evaluate forecasts when the corresponding actual values (for example, actually realized demand) are not yet available. These statistical selection approaches have been adopted from non-time series modeling problems in which the predictive aspect of a model may not be present. Therefore, forecasting is only implicitly accounted for in these algorithmic approaches. Good forecasts, rather than good descriptions of the series, are done with a "leap of faith".

2.3. Model selection and judgment

Despite the fact that the automatic selection of forecasting models has been part of many statistical packages and commercial software, what is often observed in practice is that managers select a model (and in some cases parameters) in a *judgmental* way. Automatic selection procedures are often hard to understand and communicate within companies. In that sense, managers lack trust in automatic statistical forecasting (Alvarado-Valencia et al., 2017). A standard issue is that automatic selection methods tend to change models between successive planning periods, substantially altering the shape of the series of forecasts. This issue reduces the users' trust in the system, especially after the statistical selection makes some poor choices (Dietvorst et al., 2015). Users then eventually resort to either fully overriding the statistical selection or implementing custom ad-hoc judgmental "correction" rules.

Moreover, managers often believe firmly that they better understand the data and the business context that created the data. For example, even if the result of an algorithm suggests that the data lacks any apparent seasonality (and as such a seasonal model would not be appropriate), managers may still manually select a seasonal model because they believe that this better represents the reality of their business. Lastly, the sense of ownership of forecasts can drive experts to override statistical results because more often than not evaluation of their work performance is associated with taking actions about the forecasts

(Önkal and Gönül, 2005; Ord et al., 2017), or is influenced by organizational politics (Kolassa and Siemsen, 2016; Ord et al., 2017).

To the best of our knowledge, the efficacy of judgmental model selection has not been studied. We expect that when forecast models are presented in a graphical environment (actual data versus fitted values plus forecasts, as is the case in the majority of forecasting software), forecasters may pay little attention to the actual fit of the model in the in-sample data (or the respective value of the AIC if provided). However, critical to the judgmental selection will be the matching of the out-of-sample forecasts with the expected reality. Harvey (1995) observed that participants in a laboratory experiment made predictions so that the noise and patterns in the forecasts were representative of the past data. This finding leads us to believe that forecasters perform a mental extrapolation of the available in-sample data, rejecting the models that result in seemingly unreasonable forecasts and accepting the ones that represent a possible reality for them. Thus, in contrast to algorithmic model selection, forecasters will attempt to evaluate the out-of-sample forecasts, even if the future realized values of the forecasted variable are not yet available.

As the amount of information increases, decision makers are unable to process it efficiently and simultaneously (Payne, 1976). Accordingly, research in decision analysis and management judgment has established that decomposition methods, which divide a task into smaller and simpler ones, lead to better judgment. Such methods have also been found to be useful in judgmental forecasting tasks, especially for forecasts that involve trends, seasonality and/or the effect of special events such as promotions. Edmundson (1990) examined the performance of judgmental forecasting under decomposition. Similar to the way exponential smoothing works (Gardner, 2006), forecasters were asked to estimate the structural components of the time series (level, trend, and seasonality) separately. The three estimates were subsequently combined. (Edmundson, 1990) found that estimating the components independently resulted in superior performance compared with producing judgmental forecasts directly. In another study, Webby et al. (2005) observed similar results when the effects of special events were estimated separately. Also, Lee and Siemsen (2017) demonstrated the value of task decomposition on order decisions, especially when coupled with decision support.

We expect that these insights may be applied to judgmental model selection. When judgmentally selecting between forecasting models (through a graphical interface), we expect a model-build approach to outperform the simple choice between different models. In a model-build approach, forecasters are asked to verify the existence (or not) of structural

components (trend and seasonality). This changes the task from identifying the best extrapolation line to determining whether the historical information exhibits specific features that the expert believes will extend into the future.

2.4. Combination and aggregation

Forecast combinations can result in significant improvement in forecast accuracy (Armstrong, 2001). There is also ample evidence that combining the output of algorithms with the output of human judgment can confer benefits. Blattberg and Hoch (1990) used a simple (50-50%) combination and found significant gains in combination methods compared with the separate use of algorithms and judgment. Their results have been repeatedly confirmed in the forecasting literature. Franses and Legerstee (2011) found that a simple combination of forecasts outperformed both statistical and judgmentally adjusted forecasts. Petropoulos et al. (2016) demonstrated that a 50-50 combination of forecasts in the period after a manager's adjustments have resulted in significant losses can indeed increase accuracy by 14%. Wang and Petropoulos (2016) found that a combination is as good, if not better, than selecting between a statistical or an expert forecast. Trapero et al. (2013) demonstrated further gains with more complex combination schemes. We anticipate that a combination will also be beneficial in the context of forecast model selection.

The concept of the wisdom of crowds refers to the aggregation of the judgments of a group of decision makers/stakeholders. Surowiecki (2005) provided several cases in which this concept has been found to increase performance compared with individual judgments. Ferrell (1985) also argued for the importance of combining individual judgments and discussed the significantly improved performance of judgmental aggregation. He suggested that the process of the combination itself is of little significance. However, a later study added that aggregation done mechanically is better than if it is done by one of the forecasters to avoid the possibility of biased weights (Harvey and Harries, 2004). In any case, we expect that the aggregation of judgmental model selections will lead to improved performance compared with selecting a single model, either judgmentally or statistically.

3. Design of the behavioral experiment

3.1. Selecting models judgmentally

Specialized forecasting software lists forecast methods and models. The users of such systems must choose one from the list to extrapolate the data at hand. In some cases, this list of choices is complemented by an option that, based on an algorithm, automatically identifies

and applies the best of the available methods. However, in the context of the current paper, we assumed that forecasters do not have such a recommendation available and instead rely solely on their own judgment. We also assumed that the choice set is constrained to four models able to capture various data patterns (level, trend, and seasonality). This is not an unreasonable setup, with some established systems offering such specific options (such as the well established SAP APO-DP system). We resorted to the exponential smoothing family of models (Hyndman et al., 2008) and focused on the four models presented in Table 1 (mathematical expressions are provided in Appendix A).

Table 1: The four forecasting models considered in this study.

Model description	ETS Model	Trend	Seasonality
Simple exponential smoothing (SES)	A,N,N	✘	✘
SES with additive seasonality	A,N,A	✘	✔
Damped exponential smoothing (DES)	A,Ad,N	✔	✘
DES with additive seasonality	A,Ad,A	✔	✔

To emulate the simple scenario implied by standard forecasting support systems (choose one of the available forecasting models), we used radio buttons to present the different model-choices as a list, as depicted in the left part of Figure 1. A user can navigate across the different choices and examine the forecasts produced by each method. Once the forecasts produced by a method are considered satisfactory, then a manager can submit the choice and move to the next time series. We call this approach "judgmental model selection".

We also considered a second approach where the user builds a model instead of selecting between models. In the "model-build" condition, we ask a user to identify the existence of a trend and/or seasonality in the data; the response can be used to select the respective model from the Table 1. For example, identification of a trend implies damped exponential smoothing; identification of seasonality without a trend implies SES with seasonality. This can be implemented in the software design by including two check-boxes (right panel of Figure 1). Once a change in one of these two check-boxes has been made, the forecasts of the respective model are drawn. To facilitate identification, we provide trend and seasonal plots (with usage instructions) in an attempt to aid the users with the pattern identification task.

In both cases, once a participant submits his or her decisions, we use the selected forecasting method to produce one-year-ahead (12 months) forecasts for that time series. The

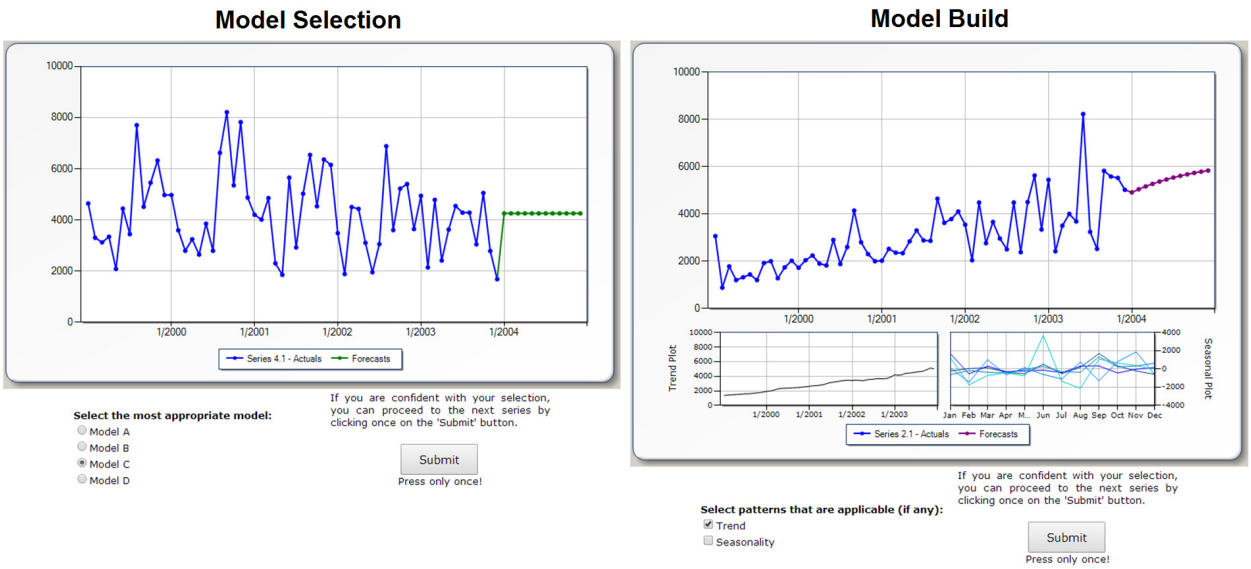

Figure 1: Screens of the Web-based environment of the behavioral experiment. The left panel shows the implementation of model selection; the right panel presents the model-build.

forecasts are compared with the actual future values, which were withheld, to find the forecast accuracy of the submitted choice.

3.2. Data

To compare the performance of statistical versus judgmental model selection, we used a subset of time series from the M3-Competition dataset (Makridakis and Hibon, 2000). This dataset consists of 3,003 real time series of various frequencies and types. It has been used many times in empirical evaluations of new forecasting models or processes (Hyndman et al., 2002; Taylor, 2003; Hibon and Evgeniou, 2005; Crone et al., 2011; Athanasopoulos et al., 2017; Petropoulos et al., 2018). We did not disclose the data source to the participants.

We focused on series with a monthly frequency and handpicked 32 time series. We selected series so that in half of them, the statistical model selection based on minimizing the value of the AIC succeeds in identifying the best model as evaluated in the hold-out sample (out-of-sample observations). For the other half, this minimum-AIC model fails to produce the best out-of-sample forecast. Moreover, the 32 time series were selected so that all four exponential smoothing models considered in this paper (Table 1) are identified as best in some time series according to the AIC criterion.

This success rate of 50% for the statistical algorithm to pick the correct model probably overestimated (but not by much) the true success rate. When the four models presented in

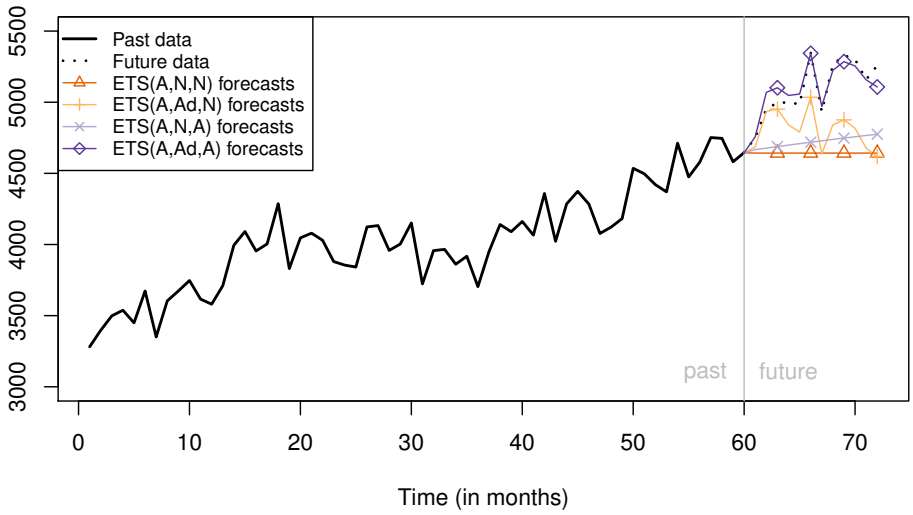

Figure 2: A typical time series used in this research, along with the forecasts from the four models.

section 3.1 were applied on the 1,428 monthly series of the M3-competition, selection based on AIC is accurate in 36% of the cases. As such, if any bias is introduced by our time series selection, we favor the statistical algorithm by giving the algorithm a higher chance of picking the correct model. Consequently, the true effect size with which human judgment improves upon performance may be underestimated in our analysis.

Because time series from the M3-Competition are of various lengths, we truncated all selected series to a history of 72 months (six years). The first five years of data (60 months) were treated as the in-sample data, on which the models were fitted. The last year (12 months) was used for out-of-sample evaluation. Figure 2 depicts a typical time series used in this behavioral experiment. Along with the historical data that cover five years, we also draw the (unobserved) future of the series. Moreover, we show the statistical point forecasts of the four exponential smoothing models considered in different colors. For this example, the AIC method identifies the ETS(A,Ad,A) model as best and, in fact, produced the best out-of-sample forecasts.

3.3. Participants

The behavioral experiment was introduced as an elective exercise to groups of undergraduate and postgraduate students studying at various universities (Institutions with at least 20 participants include Bangor University, Cardiff University, Lancaster University, the National Technical University of Athens, and Universidad de Castilla-La Mancha). Details

regarding the modules where the experiment was introduced as an elective exercise are provided in table B.6 of Appendix B. We ran the exercise as a seminar (workshop) session during the respective modules. The experiment was also posted to several relevant groups on LinkedIn and three major forecasting blogs. As an incentive, the participants were told they would receive £50 if their performance ranked within the top-20 across all the participants (see Lacetera et al., 2014, for some positive effects in rewarding volunteers).

We recruited more than 900 participants; 693 of them completed the task. Upon commencing the task, the participants were asked to self-describe themselves as undergraduate/postgraduate students, researchers, practitioners, or other. At the same time, each participant was randomly assigned to either the "model selection" or "model-build" condition. Table 2 presents the distribution of participants across roles (rows) and experimental conditions (columns).

Table 2: Participants per role and experiment.

Role	Model Selection	Model-Build	Total
UG students	139	137	276
PG students	103	108	211
Researchers	13	31	44
Practitioners	46	44	90
Other	40	32	72
Total	341	352	**693**

Most previous behavioral studies of judgmental forecasting were limited to students' participation (Lee et al., 2007; Thomson et al., 2013), which is common in behavioral experiments (Deck and Smith, 2013). In this study, our sample of student participants was complemented by a sample of practitioners (90 forecasting experts). Practitioner participants come from a variety of industries, as depicted in table B.7 of Appendix B. Our analysis also considered this sub-sample separately to check for differences and similarities between the practitioners and the students.

The completion rate was high for student participants (at 83%). This was expected, as the experiment was conducted in a lab for several cohorts of this population. The rate was lower for the other groups (practitioners 67%, researchers 57%, and other participants 56%). These observed completion rates are slightly lower than other professional web-based surveys (78.6%[2]); this difference may be explained by the duration of this behavioral experiment

[2]http://fluidsurveys.com/university/response-rate-statistics-online-surveys-aiming/

(around 30 minutes, as opposed to the recommended 15 minutes[3]).

3.4. The process of the experiment

After being randomly assigned to experimental conditions, participants were given a short description of the experimental task. The participants assigned to the model-build condition were given brief descriptions of the trend and seasonal plots.

The 32 time series we selected from the M3 Competition were divided into four groups of 8 time series each (the same for all participants), and the actual experiment consisted of 4 rounds. In each round, the participants were provided with different information regarding the forecasts and the fit derived from each model. The purpose was to investigate how different designs and information affect judgmental model selection or model-build. The information provided in each round is as follows:

- Only the out-of-sample forecasts (point forecasts for the next 12 months) were provided.
- The out-of-sample forecasts and the in-sample forecasts (model fit) were provided.
- The out-of-sample forecasts and the value of the AIC, which refers to the fit of the model penalized by the number of parameters, were provided.
- The out-of-sample forecasts, the in-sample forecasts, and the value of the AIC were provided.

The order of the rounds, as well as the order of the time series within each round, was randomized for each participant. Attention checks (Abbey and Meloy, 2017) were not performed. To maximize participant attention, round-specific instructions were given at the beginning of each round so that the participants were able to identify and potentially use the information provided.

Our experiment has both a between-subjects factor (model selection vs. model-build) as well as a within-subjects factor (information provided). Since the latter produced little meaningful variation, our analysis focuses on the former. We chose this design since we believed that the difference between model selection vs. model-build could introduce significant sequence effects (making this more suited for between-subjects analysis), but we did not believe that differences in information provided would lead to sequence effects (making this more suited for within-subjects analysis).

[3]http://fluidsurveys.com/university/finding-the-correct-survey-length/

3.5. Measuring forecasting performance

The performance of both algorithmic (based on AIC) and judgmental model selection is measured on the out-of-sample data (12 monthly observations) that were kept hidden during the process of fitting the models and calculating the AIC values. Four metrics were used to this end. The first metric was a *percentage score* based on the ranking of the selections. This was calculated as follows: A participant receives 3 points for the best choice (the model that leads to the best forecasts) for a time series, 2 points for the second best choice, and 1 point for the third best choice. Zero points were awarded for the worst (out of four) choices. The same point scheme can be applied to both judgmental forecasting approaches (model selection and model-build) once the identified patterns are translated to the respective model. The mean absolute error (MAE) was used as the cost function for evaluation. The range of points that anyone could collect is (given the number of time series) 0-96, which was then standardized to the more intuitive scale of 0-100. The *percentage score* of each participant, along with a pie chart presenting the distribution of best, second best, third best, and worst selections, was presented at the very last page of the experiment.

Apart from the percentage score based on the selections, we also use three formal measures of forecasting performance: (1) Mean Percentage Error (MPE) is a measure suitable for measuring any systematic bias in the forecasts, (2) Mean Absolute Percentage Error ($MAPE$) and (3) Mean Absolute Scaled Error ($MASE$; Hyndman and Koehler, 2006) are suitable for measuring the accuracy of the forecasts. Although the $MAPE$ suffers from several drawbacks (Goodwin and Lawton, 1999), it is intuitive, easy to interpret, and widely used in practice. $MASE$ is the Mean Absolute Error scaled by the in-sample Mean Absolute Error of the naive method that uses the last observed value as a forecast. The intuition behind this scaling factor is that it can always be defined and only requires the assumption that the time series has no more than one unit root, which is almost generally true for real time series. Other scaling factors, such as the historical mean, impose additional assumptions, such as stationarity. $MASE$ has desirable statistical properties and is popular in the literature. In particular, $MASE$ is scale independent without having the computational issues of $MAPE$. It is always defined and finite, with the only exception being the extreme case where all historical data would be equal. Note that MAE and $MASE$ would both give the same rankings of the models within a series and, as a result, the same *percentage scores*. However, MAE is a scale-dependent error measure and not suitable for summarizing across series. For all three measures, MPE, $MAPE$, and $MASE$, values closer to zero are better. Moreover, whereas MPE can take both positive and negative values, the values for $MAPE$

and $MASE$ are always non-negative.

The values of MPE, $MAPE$, and $MASE$ for a single time series across forecast horizons are calculated as

$$\begin{aligned} MPE &= \frac{100}{H}\sum_{i=1}^{H}\frac{y_{n+i}-f_{n+i}}{y_{n+i}}, \\ MAPE &= \frac{100}{H}\sum_{i=1}^{H}\frac{\mid y_{n+i}-f_{n+i}\mid}{\mid y_{n+i}\mid}, \\ MASE &= \frac{(n-1)\sum_{i=1}^{H}\mid y_{n+i}-f_{n+i}\mid}{H\sum_{j=2}^{n}\mid y_j-y_{j-1}\mid}, \end{aligned}$$

where y_t and f_t refer to the actual and the forecasted value at period t, n is the size of the training sample, and H is the forecast horizon.

4. Analysis

4.1. Individuals' performance

We next examined the performance of the judgmental model selection and judgmental model-build approaches. We contrasted their performance with the algorithmic model selection by AIC (Hyndman and Khandakar, 2008).

How do judgmental model selection and judgmental model-build perform based on the percentage score? The left panel of Figure 3 presents the percentage scores of the participants under the two approaches. The performance of each participant is depicted with a dot marker (blue for the practitioner participants, gray for all other participants), and the respective box-plots are also drawn. The square symbol represents the arithmetic mean of the percentage score for each approach. The horizontal (red) dashed line refers to the statistical benchmark (performance of automatic model selection based on AIC). Generally, participants performed better under the model-build approach than the model selection approach. In essence, the average participant in model-build performs as well as the participant in the 75t$^{\text{th}}$ percentile of the model selection approach. More importantly, participants under the model-build approach perform on average as well as the statistical selection. However, the differences in scores between individuals are large, with the range spanning between 32% and 83%.

Do humans select similarly to the algorithm? The middle and right panels of Figure 3 present, respectively, how many times the participants selected the best, second best,

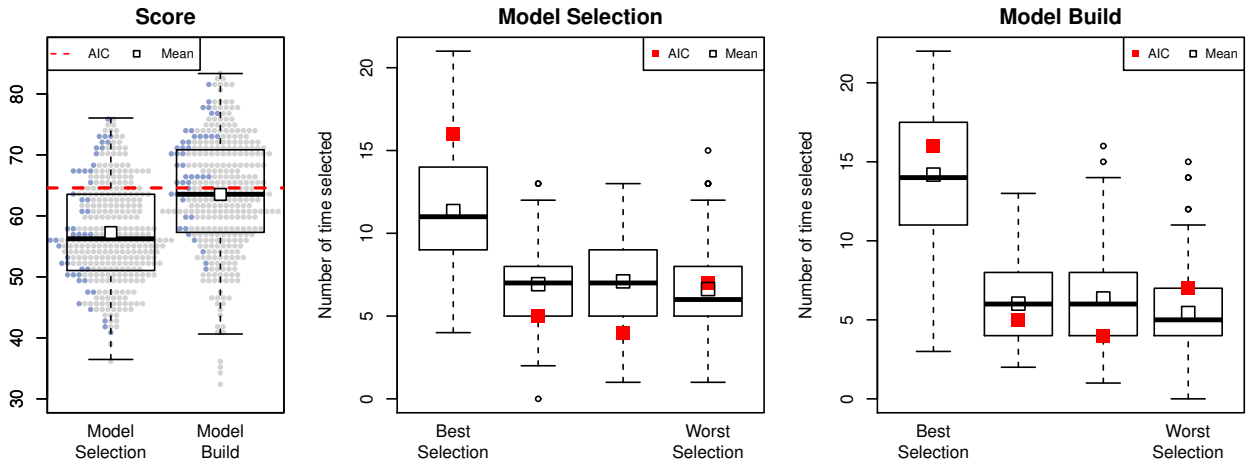

Figure 3: Performance of model selection and model-build in terms of scores and distributions of best/worst selections.

third best, and worst models under the model selection and model-build approaches. The differences in performance (in terms of percentage scores) between model selection and model-build derives from the fact that model-build participants were able to identify the best model more frequently. By comparing how frequently algorithms (red squares) and humans make the best and worst model selection, we can observe that humans are superior to algorithms in avoiding the worst model, especially in the model-build case. The differences are statistically significant according to t-tests for both best and worst selections and both strategies, model selection and model-build ($p < 0.01$). The frequencies that algorithms and humans selected each model are presented in table 3, along with how many times each model performs best in the out-of-sample data. We observe that algorithms generally select the level models (SES and SES with seasonality) more often compared to their trended counterparts. AIC, as with other information criteria, attempts to balance the goodness-of-fit of the model and its complexity as captured by the number of parameters. More uniform distributions are observed for human selection and out-of-sample performance.

Table 3: Frequencies of selected models.

Selection method	SES	Seasonal SES	DES	Seasonal DES
Selection based on AIC	46.88%	31.25%	6.25%	15.62%
Judgmental model selection	17.27%	33.68%	17.64%	31.41%
Judgmental model-build	18.44%	27.85%	24.48%	29.23%
Best out-of-sample performance	34.38%	21.88%	21.88%	21.88%

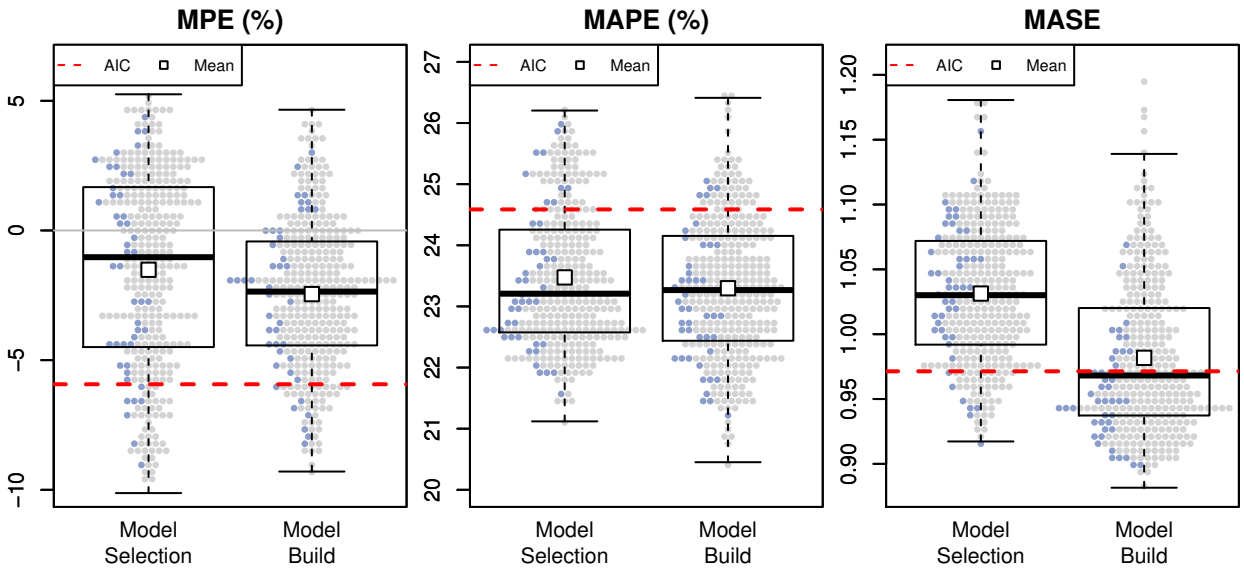

Figure 4: Performance of model selection and model-build in terms of error measures when all participants are considered.

How do the judgmental approaches perform based on error measures? Figure 4 presents the performance of all 693 participants for MPE, $MAPE$, and $MASE$, the three error measures considered in this study. In both the model selection and the model-build treatments, human judgment is significantly better (less biased and more accurate) than statistical selection in terms of MPE and $MAPE$. At the same time, although the judgmental model-build performs on a par with statistical selection for $MASE$, the judgmental model selection performs worse than the statistical selection. Differences in the insights provided by the different error measures can be attributed to their statistical properties.

It is noteworthy that statistical selection is, on average, positively biased (negative values for MPE). However, this is not the case for all participants. In fact, slightly more than 30% of all participants (42% of those that were assigned in the model selection experiment) are, on average, negatively biased. The positive bias of statistical methods is consistent with the results of Kourentzes et al. (2014) who investigated the performance of statistical methods on all M3-competition data.

Do the results differ if only the practitioners' subgroup is analyzed? Independent sample t-tests were performed to compare the performance of practitioners and students. The results showed that differences are not statistically significant, apart from the case of model-build and $MASE$ where practitioner participants perform significantly better. The similarities in the results between student and practitioner participants are relevant for the discussion

of the external validity of behavioral experiments using students (Deck and Smith, 2013). Similar to Kremer et al. (2015), our results suggest that student samples may be used for (at least) forecasting behavioral experiments.

4.2. Effects of individuals' skill and time series properties

Going beyond the descriptives presented thus far, we constructed a linear mixed effects model to account for the variability in the skills of individual participants, as well as for the properties of each time series that was used. We model values of $MASE$ (containing the performance of the participants on each individual response: 693 participants $\times$ 32 time series) considering the following fixed effects: (i) experimental condition (model selection or model-build); (ii) interface information (out-of-sample only, in- and out-of-sample and these options supplemented with fit statistics; see section 3.4); and (iii) the role of the participants (see Table 2 in which both under- and postgraduate students were grouped together). We accounted for the variation between participants as a random effect that reflects any variability in skill. Similarly, we consider the variability between time series as a second random effect, given that they have varying properties.

To conduct the analysis we use the `lme4` package (Bates et al., 2015) for R statistical computing language (R Core Team, 2016). We evaluated the contribution of each variable in model by using two information criteria (AIC and BIC). To facilitate the comparison of the alternative model specification using information criteria, we estimated the model using maximum likelihood. We found only marginal differences between the models recommended by the two criteria, a result suggesting that the most parsimonious option was the better choice.

We concluded that only the effect of the experimental condition was important, and that the interface information and role of participant did not explain enough variability to justify the increased model complexity. Consequently, these two effects were removed from the model. Both random effects were deemed useful. We investigated for random slopes as well, but found that such slopes did not explain any additional variability in $MASE$. Therefore, they were not considered further.

The resulting fully crossed random intercept model is reported in Table 4. The estimated model indicates that model-build improves $MASE$ by 0.0496 over model select, which is consistent with the analysis so far. The standard deviations of the participant and series effects, respectively, are 0.0314 and 0.5551 (the intraclass correlations are 0.0026 and 0.8178). This shows that the skills of the participants account for only a small degree of the variability of $MASE$. This helps explain the insignificant role of the participant background. To put

these values into perspective, the standard deviation of $MASE$ on individual time series responses is 0.6144.

Table 4: Linear mixed effects model output

Fixed effects			
	Estimate	Standard Error	
Intercept	1.0313	0.0982	
Experiment setup	-0.0496	0.0042	
Random effects			
	Standard Deviation		
Participant (intercept)	0.0314		
Time series (intercept)	0.5551		
Residual	0.2601		
Model statistics			
	AIC	BIC	Log Likelihood
	3742.3	3782.3	-1866.1

4.3. 50% statistics + 50% judgment

The seminal work by Blattberg and Hoch (1990) suggested that combining the outputs of statistical models with managerial judgment will provide more accurate outputs than each single-source approach (model *or* manager), while a 50-50% combination is "a nonoptimal but pragmatic solution" (Blattberg and Hoch, 1990, p. 898). Their result has been confirmed in many subsequent studies. In this study, we considered the simple average of the two predictions, i.e., the equal-weight combination of the forecasts produced by the model selected by the statistics (AIC) and the model chosen by each participant.

How does a 50-50 combination of statistical and judgmental selection perform? Figure 5 presents the performance of the simple combination of statistical and judgmental selection for the three error measures considered in this study and categorized by the two judgmental approaches. We observed that performance for both judgmental model selection and model-build was improved significantly compared with using statistical selection alone (horizontal dashed red line). In fact, the combination of the statistical + judgmental selection is less biased than statistical selection in 86% of the cases and produces lower values for $MAPE$ and $MASE$ for 99% and 90% of the cases, respectively. Moreover, the differences in the performance of the two approaches are now minimized.

Does a 50-50 combination bring robustness? On top of the improvements in performance, an equal-weight combination also reduces the between-subject variation in performance. Fo-

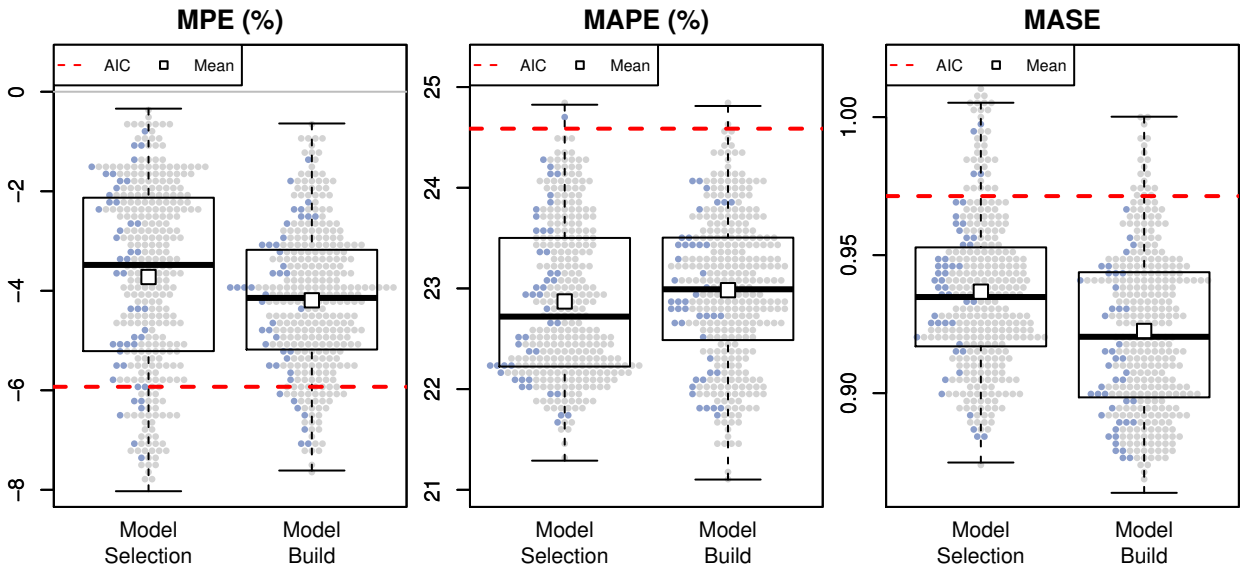

Figure 5: Performance of 50-50% combination of statistical and judgmental selection.

cusing, for instance, on $MASE$ and the judgmental model-build approach, Figure 4 suggests a range of 0.314 (between 0.882 and 1.196) between best and worst performers. The comparable range according to Figure 5 is 0.136 (between 0.864 and 1). Therefore, a 50-50 combination renders the judgmental selection approaches more robust.

4.4. Wisdom of crowds

An alternative to combining the statistical and judgmental selections is judgmental aggregation – a combination of the judgmental selections of multiple participants. The concept of the "wisdom of crowds" is not new (Surowiecki, 2005) and has repeatedly been shown to improve forecast accuracy as well as the quality of judgments in general.

For example, consider that a group of 10 experts is randomly selected. Given their selections regarding the best model, we can derive the frequencies that show how many times each model is identified as best. In other words, experts preferences are equally considered (each expert has exactly one vote, and all votes carry the same weight). This procedure leads to a weighted combination of the four models for which the performance can be measured. We consider groups of 1 to 25 experts randomly re-sampled 1000 times.

Does judgmental aggregation improve forecasting performance? Figure 6 presents the results of judgmental aggregation. The light blue area describes the range of the performance of judgmental aggregation for various group sizes when the model-build approach is considered. The middle-shaded blue area refers to the 50% range of performances, and the

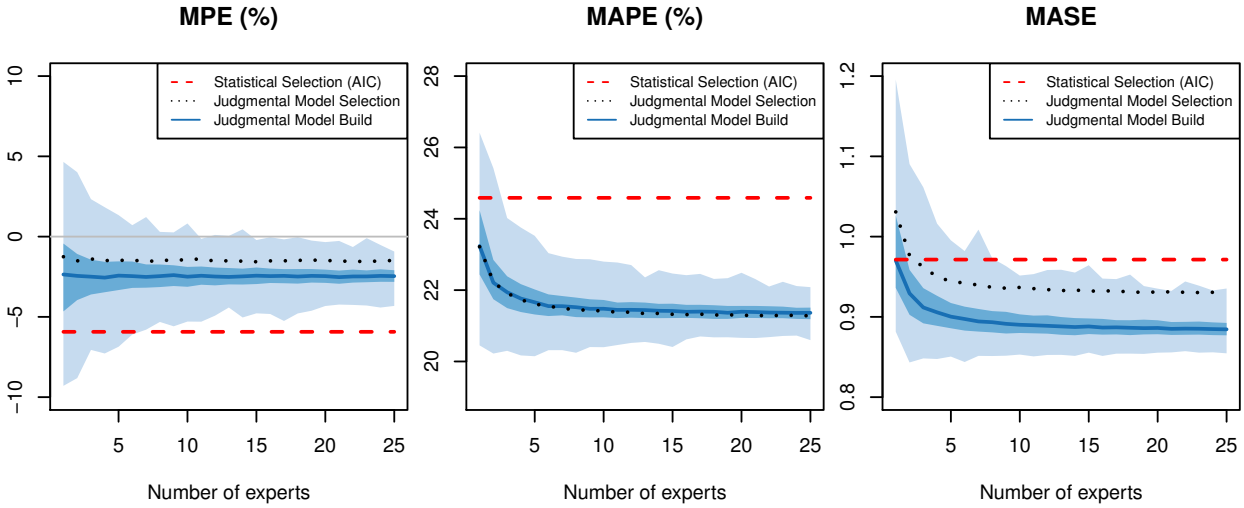

Figure 6: Wisdom of crowds' performance for different numbers of experts.

dark blue line refers to the median performance. In other words, if one considers a vertical line (i.e., for particular group size), the points at which this line intersects provide the minimum, first quartile, median, third quartile, and maximum descriptive summary of the performance. For the judgmental model selection approach, only the median is drawn in the black dotted line, and the performance of statistical selection is represented by a red dashed horizontal line. We observed significant gains in performance as the group size increased, coupled with lower variance in the performance of different equally sized groups. We also observed the convergence of performance, meaning that no further gains were noticed in the average performance for group sizes higher than 20. Judgmental aggregation outperforms both statistical and individual selection.

How many experts are enough? A careful examination of Figure 6 reveals, on top of the improvements if aggregation is to be used, the critical thresholds for deciding on the optimal number of experts in groups. We observed that if groups of five participants are considered, then their forecasting performance was almost always better than that of the statistical selection on all three measures, regardless of their role (undergraduate/postgraduate students, practitioners, researchers, or other). Of even more interest, the third quartile of groups of size two always outperforms the statistical benchmark. It is not the first time that the thresholds of two and five appear in the literature. These results confirm previous findings: "only two to five individuals' forecasts must be included to achieve much of the total improvement" (Ashton and Ashton, 1985, p. 1499). This result holds even if we only consider

specific sub-populations of our sample, e.g. practitioners only. Once judgmental aggregation is used, the results from practitioners are all-but identical to results from students and other participants.

Note that in this study we consider only unweighted combinations of the judgmental model selections. If the same exercise were carried dynamically (over time), then one could also consider, based on past performance, weighted combinations that have been proven to enhance performance in other forecasting tasks (Tetlock and Gardner, 2015).

How does the model with the most votes perform? Instead of considering a weighted-combination forecast based on votes across the four models, we have also examined a case in which the aggregate selection is the model with the most votes. In case two (or more) models are tied for first place in votes, then an equal-weight combination among them was calculated. The performance of this strategy is worse, in terms of the accuracy measures considered, than the wisdom of crowds with weighted combinations; moreover, the quality of the performance depends heavily on the sample selected (the variance is high even for groups with a large number of experts). However, on average, merely choosing the most-popular model still outperforms statistical selection.

4.5. Evaluation summary and discussion

Table 5 summarizes the results presented in the above sections. As a sanity check, we also provide the performance of four additional statistical benchmarks:

- *Random selection* refers to a performance obtained by randomly selecting one of the four available choices. We have reported the arithmetic mean performance when such a procedure is repeated 1000 times for each series. This benchmark is included to validate the choice of the time series and to demonstrate that non-random selection is indeed meaningful (in either a statistical or a judgmental manner).
- *Equal-weight combination* refers to the simple average of the forecasts across all four models. In other words, each model was assigned a weight of one-quarter. It is included as a benchmark of the performance of 50-50 combinations and the wisdom of crowds.
- *Weighted combinations based on AIC* were proposed by Burnham and Anderson (2002) and evaluated by Kolassa (2011). This approach showed improved performance over selecting the model with the lowest AIC. It is used in this study as a more sophisticated benchmark for the wisdom of crowds.
- *Combination of best two based on AIC* refers to the equal-weight combination of the best and second-best models based on the AIC values.

Table 5: Summary of the results; top-method is underlined; top-three methods are in boldface.

Method	MPE (%)	MAPE (%)	MASE
Individual selection			
Random selection	-2.91	24.52	1.104
Selection based on AIC	-5.93	24.59	0.971
Judgmental model selection	**-1.52**	23.48	1.031
Judgmental model-build	**-2.45**	23.30	0.982
Combination			
Equal-weight combination	-2.90	**21.96**	0.985
Weighted combination based on AIC	-4.84	23.39	0.931
Combination of best two based on AIC	-4.65	23.12	**0.921**
50-50% combination of AIC and judgment	-3.96	**22.93**	**0.930**
Wisdom of crowds: 5 humans (model-build)	**-2.43**	**21.68**	**0.903**

Focusing on the first four rows of Table 5 that refer to selecting a single model with different approaches, random selection performs poorly compared with all other approaches. This is especially true in terms of $MASE$. Moreover, although it seems to be less biased than statistical selection, random selection's absolute value of MPE is larger than any of the two judgmental approaches.

The last five rows of Table 5 present the performance of the various combination approaches. First, we observe that the equal-weight combination performs very well according to all metrics, apart from $MASE$. A weighted combination based on AIC improves the performance of the statistical benchmark, confirming the results by Kolassa (2011); however, it is always outperformed by both 50-50 combinations of statistical and judgmental selection and by the wisdom of crowds. Combination of the best two models based on AIC performs slightly better than weighted combination based on AIC.

We also present in boldface the top three performers for each metric. The top performer is underlined. We observe that the wisdom of crowds (which is based on model-build) is always within the top three and ranked first for two of the metrics. The wisdom of crowds based on model selection also performs on par. We believe that this is an exciting result because it demonstrates that using experts to select the appropriate method performs best against state-of-the-art benchmarks.

5. Implications for theory, practice, and implementation

This work provides a framework for judgmental forecasting model selection, and highlights the conditions for achieving maximal gains. We now discuss the implications of our

work for theory and practice as well as issues of implementation.

Statistical model selection has been dominated by goodness-of-fit derived approaches, such as information criteria, and by others based on cross-validated errors (Fildes and Petropoulos, 2015). The findings of our research challenge these approaches and suggest that an alternative approach based on human judgment is feasible and performs well. Eliciting how experts perform the selection of forecasts may yield still more novel approaches to statistical model selection. Our research provides evidence that a model-build approach works better for humans. We postulate that model-build implies a suitable structure (and a set of restrictions) that aid selection. Statistical procedures could potentially benefit from a similar framing.

Our findings are aligned with the literature on judgmental forecasting as well as research in behavioral operations. The good performance of the 50-50 combination (of judgment and algorithm) and judgmental aggregation resonates with findings in forecasting and cognitive science (Blattberg and Hoch, 1990; Surowiecki, 2005).

Our research looks at an everyday problem that experts face in practice. Planners and managers are regularly tasked with the responsibility of choosing the best method to produce the various forecasts needed in their organizations. We not only benchmark human judgment against a state-of-the-art statistical selection but also provide insights into how to aid experts. Another exciting aspect of this research is that it demonstrates that expert systems that rely on algorithms to select the right model, such as Forecast Pro, Autobox, SAP Advanced Planning and Optimization - Demand Planning (APO-DP), IBM SPSS Forecasting, SAS, etc., may be outperformed by human experts, if these experts are supported appropriately in their decision making. This has substantial implications for the design of algorithms for both expert systems algorithms and the user interfaces of forecasting support systems.

Judgmental model selection is used in practice because it has some endearing properties. It is intuitive: A problem that necessitates human intervention is always more meaningful and intellectually and intuitively appealing for users. It is interpretable: Practitioners understand how this process works. The version of model-build that is based on judgmental decomposition is easy to explain and adapt to real-life setups. This simplicity is a welcome property (Zellner et al., 2002). In fact, the configuration used in our experiment is already offered in a similar format in popular software packages. For example, SAP APO-DP provides a manual (judgmental) forecasting model selection process, providing clear guidance of a judgmental selection to be driven from prevailing components (most notably trend and

seasonality) as perceived by the user (manager). Specialized off-the-shelf forecasting support systems like Forecast Pro also allow their optimal algorithmic selection to be overridden by the user.

At its most basic form, implementing judgmental model selection requires no investment. Nonetheless, to obtain maximum gains, existing interfaces will need some redesign to allow incorporation of the model-build approach. However, a crucial limitation is the cost of using human experts. Having an expert going through all items that need to be forecasted may not be feasible for many organizations, such as large retailers that often require millions of forecasts. Of course, using the judgmental aggregation approach requires even more experts.

In a standard forecasting and inventory setting, ABC analysis is often used to classify the different stock keeping units (SKUs) into importance classes. In this approach, the 20% most important and the 50% least important items are classified as A and C items, respectively. Additionally, XYZ analysis is used in conjunction with ABC analysis to further classify the products into easy-to-forecast (X items) to hard-to-forecast (Z items). As such, nine classes are considered, as depicted in Figure 7 (Ord et al., 2017). We propose the application of the wisdom of crowds for judgmental model selection/build on the AZ items (those of high importance but difficult to forecast). This class is shaded with gray color in Figure 7. In many cases, these items represent only a small fraction of the total number of SKUs. Thus, judgmentally weighted selections across the available models of a forecasting support system should be deduced by the individual choices, either in terms of models or patterns, of a small group of managers; our analysis showed that selections from five managers would suffice.

A potential limitation of our current study, especially in the big data era, is the number of series and respective contexts examined. As a future direction, we suggest the extension of our study by increasing the amount of time series represented in each context and the number of contexts to allow the evaluation of the robustness of the superior performance of the judgmental model selection in each context. This could include more and higher frequencies and exogenous information. Such extensions would also benefit the ongoing limitation of controlled experimental studies towards more generalizable results.

6. Conclusions

Model selection of appropriate forecasting models is an open problem. Optimal ex-ante identification of the best ex-post model can bring significant benefits regarding forecasting performance. The literature has so far focused on automatic/statistical approaches for model selection. However, demand managers and forecasting practitioners often tend to

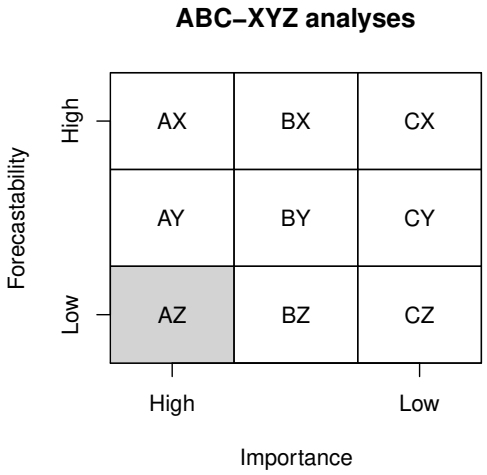

Figure 7: A visual representation of ABC-XYZ analyses; resources for judgmental model selection should be allocated towards the important and low-forecastable items.

ignore system recommendations and apply judgment when selecting a forecasting model. This study is the first, to the best of our knowledge, to investigate the performance of judgmental model selection.

We devised a behavioral experiment and tested the efficacy of two judgmental approaches to select models, namely simple model selection and model-build. The latter one was based on the judgmental identification of time series features (trend and seasonality). We compared the performance of these methods against that of a statistical benchmark based on information criteria. Judgmental model-build outperformed both judgmental and statistical model selection. Significant performance improvements over statistical selection were recorded for the equal-weight combination of statistical and judgmental selection. Judgmental aggregation (weighted combinations of models based on the selections of multiple experts) resulted in the best performance of any approaches we considered. Finally, an exciting result is that humans are better, compared to statistics, in avoiding the worst model.

The results of this study suggest that companies should consider judgmental forecasting selection as a complementary tool to statistical model selection. Moreover, we believe that applying the judgmental aggregation of a handful of experts to the most important items is a trade-off between resources and performance improvement that companies should be willing to consider. However, forecasting support systems that incorporate simple graphical interfaces and judgmental identification of time series features are a prerequisite to the

successful implementation of do-it-yourself (DIY) forecasting. This does not seem too much to ask for software in the big data era.

Given the good performance of judgment in model selection forecasting tasks, the emulation of human selection processes through artificial intelligence approaches seems a natural way forward toward eventually deriving an alternative statistical approach. We leave this for future research. Furthermore, we expect to further investigate the reasons behind the difference in the performance of judgmental model selection and judgmental model-build. To this end, we plan to run a simplified version of the experiment of this study that will be coupled with the use of an electroencephalogram (EEG) to record electrical brain activity. Future research could also focus on the conditions (in terms of time series characteristics, data availability, and forecasting horizon) under which judgmental model selection brings more benefits. Finally, field experiments would provide further external validity for our findings.

Acknowledgments

FP and NK would like to acknowledge the support for conducting this research provided by the Lancaster University Management School Early Career Research Grant MTA7690.

References

Abbey, J. D., Meloy, M. G., 2017. Attention by design: Using attention checks to detect inattentive respondents and improve data quality. Journal of Operations Management 53-56, 63–70.

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.

Alvarado-Valencia, J., Barrero, L. H., Önkal, D., Dennerlein, J. T., 2017. Expertise, credibility of system forecasts and integration methods in judgmental demand forecasting. International Journal of Forecasting 33 (1), 298–313.

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

Armstrong, S. J., 2001. Combining forecasts. In: Principles of Forecasting. International Series in Operations Research & Management Science. Springer, Boston, MA, pp. 417–439.

Ashton, A. H., Ashton, R. H., 1985. Aggregating subjective forecasts: Some empirical results. Management Science 31 (12), 1499–1508.

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

Barrow, D. K., Kourentzes, N., 2016. Distributions of forecasting errors of forecast combinations: Implications for inventory management. International Journal of Production Economics 177, 24–33.

Bates, D., Mächler, M., Bolker, B., Walker, S., 2015. Fitting linear mixed-effects models using lme4. Journal of Statistical Software 67 (1), 1–48.

Billah, B., King, M. L., Snyder, R., Koehler, A. B., 2006. Exponential smoothing model selection for forecasting. International Journal of Forecasting 22 (2), 239–247.

Blattberg, R. C., Hoch, S. J., 1990. Database models and managerial intuition: 50% model + 50% manager. Management Science 36 (8), 887–899.

Bunn, D., Wright, G., 1991. Interaction of judgemental and statistical forecasting methods: Issues & analysis. Management Science 37 (5), 501–518.

Burnham, K. P., Anderson, D. R., 2002. Model selection and multi-model inference: a practical information-theoretic approach. Springer, New York.

Chatfield, C., 2000. Time-series forecasting. CRC Press.

Collopy, F., Armstrong, J. S., 1992. Rule-based forecasting: Development and validation of an expert systems approach to combining time series extrapolations. Management Science 38 (10), 1394–1414.

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.

Deck, C., Smith, V., 2013. Using laboratory experiments in logistics and supply chain research. Journal of Business Logistics 34 (1), 6–14.

Dietvorst, B. J., Simmons, J. P., Massey, C., 2015. Algorithm aversion: People erroneously avoid algorithms after seeing them err. Journal of Experimental Psychology: General 144 (1), 114.

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

Ferrell, W. R., 1985. Combining individual judgments. In: Wright, G. (Ed.), Behavioral Decision Making. Springer US, pp. 111–145.

Fildes, R., 2001. Beyond forecasting competitions. International Journal of Forecasting 17, 556–560.

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.

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

Gardner, E. S., 1990. Evaluating forecast performance in an inventory control system. Management Science 36 (4), 490–499.

Gardner, E. S., 2006. Exponential smoothing: The state of the art - Part II. International Journal of Forecasting 22 (4), 637–666.

Goodwin, P., Lawton, R., 1999. On the asymmetry of the symmetric MAPE. International Journal of Forecasting 15 (4), 405–408.

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., Harries, C., 2004. Effects of judges' forecasting on their later combination of forecasts for the

same outcomes. International Journal of Forecasting 20 (3), 391–409.

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

Hogarth, R. M., Makridakis, S., 1981. Forecasting and planning: An evaluation. Management Science 27 (2), 115–138.

Hyndman, R., Koehler, A. B., Ord, J. K., Snyder, R. D., 2008. Forecasting with exponential smoothing: The state space approach. Springer Science & Business Media.

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.

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

Kolassa, S., Siemsen, E., 2016. Demand Forecasting for Managers. Business Expert Press.

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.

Kremer, M., Siemsen, E., Thomas, D. J., 2015. The sum and its parts: Judgmental hierarchical forecasting. Management Science 62 (9), 2745–2764.

Lacetera, N., Macis, M., Slonim, R., 2014. Rewarding volunteers: A field experiment. Management Science 60 (5), 1107–1129.

Lee, W. Y., Goodwin, P., Fildes, R., Nikolopoulos, K., Lawrence, M., 2007. Providing support for the use of analogies in demand forecasting tasks. International Journal of Forecasting 23 (3), 377–390.

Lee, Y. S., Siemsen, E., 2017. Task decomposition and newsvendor decision making. Management Science 63 (10), 3226–3245.

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

Nenova, Z., May, J. H., 2016. Determining an optimal hierarchical forecasting model based on the characteristics of the dataset: Technical note. Journal of Operations Management 44 (5), 62–88.

Oliva, R., Watson, N., 2009. Managing functional biases in organizational forecasts: A case study of consensus forecasting in supply chain planning. Production and Operations Management 18 (2), 138–151.

Önkal, D., Gönül, M. S., 2005. Judgmental adjustment: A challenge for providers and users of forecasts. Foresight: The International Journal of Applied Forecasting 1, 13–17.

Ord, J. K., Fildes, R., Kourentzes, N., 2017. Principles of Business Forecasting, 2nd Edition. Wessex Press Publishing Co.

Payne, J. W., 1976. Task complexity and contingent processing in decision making: An information search and protocol analysis. Organizational Behavior and Human Performance 16 (2), 366–387.

Petropoulos, F., Fildes, R., Goodwin, P., 2016. Do 'big losses' in judgmental adjustments to statistical forecasts affect experts' behaviour? European Journal of Operational Research 249 (3), 842–852.

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., 2018. Exploring the sources of uncertainty: Why does bagging for time series forecasting work? European Journal of Operational Research.

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

R Core Team, 2016. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
URL `https://www.R-project.org/`

Ritzman, L. P., King, B. E., 1993. The relative significance of forecast errors in multistage manufacturing. Journal of Operations Management 11 (1), 51–65.

Sagaert, Y. R., Aghezzaf, E.-H., Kourentzes, N., Desmet, B., 2018. Tactical sales forecasting using a very large set of macroeconomic indicators. European Journal of Operational Research 264 (2), 558–569.

Sanders, N. R., Graman, G. A., 2009. Quantifying costs of forecast errors: A case study of the warehouse environment. Omega 37 (1), 116–125.

Seifert, M., Siemsen, E., Hadida, A., Eisingerich, A., 2015. Effective judgmental forecasting in the context of fashion products. Journal of Operations Management 36 (1), 33–45.

Surowiecki, J., 2005. The Wisdom of Crowds: Why the Many Are Smarter Than the Few. Abacus.

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

Taylor, J. W., 2003. Exponential smoothing with a damped multiplicative trend. International Journal of Forecasting 19 (4), 715–725.

Tetlock, P., Gardner, D., 2015. Superforecasting: The Art and Science of Prediction. Crown.

Thomson, M. E., Pollock, A. C., Gönül, M. S., Önkal, D., 2013. Effects of trend strength and direction on performance and consistency in judgmental exchange rate forecasting. International Journal of Forecasting 29 (2), 337–353.

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.

van der Laan, E., van Dalen, J., Rohrmoser, M., Simpson, R., 2016. Demand forecasting and order planning for humanitarian logistics: An empirical assessment. Journal of Operations Management 45 (7), 114–122.

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.

Weller, M., Crone, S., 2012. Supply Chain Forecasting: Best Practices & Benchmarking Study. Lancaster Centre For Forecasting, Technical Report.

Xia, Y., Tong, H., 2011. Feature matching in time series modeling. Statistical Science, 21–46.

Zellner, A., Keuzenkamp, H. A., McAleer, M. (Eds.), 2002. Simplicity, Inference and Modelling: Keeping it Sophisticatedly Simple. Cambridge University Press.

Appendix A. Forecasting models

We denote:
α: smoothing parameter for the level ($0 \leq \alpha \leq 1$).
β: smoothing parameter for the trend ($0 \leq \beta \leq 1$).
γ: smoothing parameter for the seasonal indices ($0 \leq \gamma \leq 1$).
ϕ: damping parameter (usually $0.8 \leq \phi \leq 1$).
y_t: actual (observed) value at period t.
l_t: smoothed level at the end of period t.
b_t: smoothed trend at the end of period t.
s_t: smoothed seasonal index at the end of period t.
m: Number of periods within a seasonal cycle (e.g., 4 for quarterly, 12 for monthly).
h: forecast horizon.
$\hat{y}_{t+h}$: forecast for h periods ahead from origin t.

SES, or ETS(A,N,N), is expressed as:

$$l_t = \alpha y_t + (1-\alpha) l_{t-1}, \tag{A.1}$$
$$\hat{y}_{t+h} = l_t. \tag{A.2}$$

SES with additive seasonality, or ETS(A,N,A), is expressed as:

$$l_t = \alpha(y_t - s_{t-m}) + (1-\alpha) l_{t-1}, \tag{A.3}$$
$$s_t = \gamma(y_t - l_t) + (1-\gamma) s_{t-m}, \tag{A.4}$$
$$\hat{y}_{t+h} = l_t + s_{t+h-m}. \tag{A.5}$$

DES, or ETS(A,Ad,N), is expressed as:

$$l_t = \alpha y_t + (1-\alpha)(l_{t-1} + \phi b_{t-1}), \tag{A.6}$$
$$b_t = \beta(l_t - l_{t-1}) + (1-\beta)\phi b_{t-1}, \tag{A.7}$$
$$\hat{y}_{t+h} = l_t + \sum_{i=1}^{h} \phi^i b_t. \tag{A.8}$$

DES with additive seasonality, or ETS(A,Ad,A), is expressed as:

$$l_t = \alpha(y_t - s_{t-m}) + (1-\alpha)(l_{t-1} + \phi b_{t-1}), \tag{A.9}$$

$$b_t = \beta(l_t - l_{t-1}) + (1-\beta)\phi b_{t-1}, \tag{A.10}$$

$$s_t = \gamma(y_t - l_t) + (1-\gamma)s_{t-m}, \tag{A.11}$$

$$\hat{y}_{t+h} = l_t + \sum_{i=1}^{h} \phi^i b_t + s_{t+h-m}. \tag{A.12}$$

Appendix B. Participants details

Table B.6: University modules with at least 20 participants where the behavioral experiment was introduced as an elective exercise.

University	Module (and keywords)	Level
Bangor University	**Applied Business Projects: Operations Management** operations, strategy, competitiveness, supply chain, capacity, planning, inventory, forecasting	PG
Cardiff University	**Logistics Modelling** business statistics, forecasting, stock control, system dynamics, bull-whip effect, queuing analysis, simulation	PG
Lancaster University	**Business Forecasting** time series, forecasting, regression, evaluation, model selection, judgment	UG
Lancaster University	**Forecasting** time series, forecasting, univariate and causal models, evaluation, model selection, judgment	PG
National Technical University of Athens	**Forecasting Techniques** time series, forecasting, decomposition, univariate and causal models, evaluation, support systems, judgment	UG
Universidad de Castilla-La Mancha	**Manufacturing planning and control** planning, forecasting, manufacturing, just-in-time, stock control, inventory models	UG

Table B.7: Industries associated with the practitioner participants.

Industry	Participants
Consulting (including analytics)	14
Banking & Finance	11
Software (including forecasting software)	9
Advertising & Marketing	9
Retail	8
Health	8
Government	6
Manufacturing	5
Food & Beverage	4
Energy	3
Logistics	3
Telecommunications	3
Automotive	2
Other	5