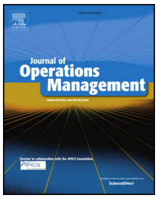
Contents lists available at ScienceDirect

Journal of Operations Management

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



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

ARTICLE INFO

Keywords:
Model selection
Behavioral operations
Decomposition
Combination

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.

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 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;

[*] Corresponding author.
E-mail address: esiemsen@wisc.edu (E. Siemsen).

https://doi.org/10.1016/j.jom.2018.05.005
Received 3 October 2017; Received in revised form 22 May 2018; Accepted 23 May 2018

0272-6963/ © 2018 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/BY/4.0/).

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).

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

[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.

Download English Version:

https://daneshyari.com/en/article/7436139

Download Persian Version:

https://daneshyari.com/article/7436139

Daneshyari.com