International Journal of Forecasting 29 (2013) 510–522

Contents lists available at SciVerse ScienceDirect

International Journal of Forecasting

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

# Measuring forecasting accuracy: The case of judgmental adjustments to SKU-level demand forecasts

Andrey Davydenko *, Robert Fildes

*Department of Management Science, Lancaster University, Lancaster, Lancashire, LA1 4YX, United Kingdom*

ARTICLE INFO

*Keywords:*
Judgmental adjustments
Forecasting support systems
Forecast accuracy
Forecast evaluation
Forecast error measures

ABSTRACT

Forecast adjustment commonly occurs when organizational forecasters adjust a statistical forecast of demand to take into account factors which are excluded from the statistical calculation. This paper addresses the question of how to measure the accuracy of such adjustments. We show that many existing error measures are generally not suited to the task, due to specific features of the demand data. Alongside the well-known weaknesses of existing measures, a number of additional effects are demonstrated that complicate the interpretation of measurement results and can even lead to false conclusions being drawn. In order to ensure an interpretable and unambiguous evaluation, we recommend the use of a metric based on aggregating performance ratios across time series using the weighted geometric mean. We illustrate that this measure has the advantage of treating over- and under-forecasting even-handedly, has a more symmetric distribution, and is robust.

Empirical analysis using the recommended metric showed that, on average, adjustments yielded improvements under symmetric linear loss, while harming accuracy in terms of some traditional measures. This provides further support to the critical importance of selecting appropriate error measures when evaluating the forecasting accuracy.


## 1. Introduction

The most well-established approach to forecasting within supply chain companies starts with a statistical time series forecast, which is then adjusted by managers in the company based on their expert knowledge. This process is usually carried out at a highly disaggregated level of SKUs (stock-keeping units), where there are often hundreds if not thousands of series to consider (Fildes & Goodwin, 2007; Sanders & Ritzman, 2004). At the same time, the empirical evidence suggests that judgments under uncertainty are affected by various types of cognitive biases and are inherently non-optimal (Tversky & Kahneman, 1974). Such biases and inefficiencies have been shown to apply specifically to judgmental adjustments (Fildes, Goodwin, Lawrence, & Nikolopoulos, 2009). Therefore, it is important to monitor the accuracy of judgmental adjustments in order to ensure the rational use of the organisation's resources which are invested in the forecasting process.

The task of measuring the accuracy of judgmental adjustments is inseparably linked with the need to choose an appropriate error measure. In fact, the choice of an error measure for assessing the accuracy of forecasts across time series is itself an important topic for forecasting research. It has theoretical implications for the comparison of forecasting methods and is of wide practical importance, since the forecasting function is often evaluated using inappropriate measures (see, for example, Armstrong & Collopy, 1992; Armstrong & Fildes, 1995), and therefore the link to economic performance may well be distorted. Despite the continuing interest in the topic, the choice of the most suitable error measure for evaluating companies' forecasts still remains controversial. Due to their statistical properties, popular error measures do not always ensure easily interpretable results when applied to real-world data (Hyndman & Koehler, 2006; Kolassa & Schutz, 2007). In practice, the proportion of firms which track

* Corresponding author. Tel.: +44 1524 593879.
*E-mail address:* a.davydenko@lancaster.ac.uk (A. Davydenko).

0169-2070/$ – see front matter 
doi:10.1016/j.ijforecast.2012.09.002

the aggregated accuracy is surprisingly small, and one apparent reason for this is the inability to agree on appropriate accuracy metrics (Hoover, 2006). As McCarthy, Davis, Golicic, and Mentzer (2006) reported, only 55% of the companies surveyed believed that their forecasting performance was being formally evaluated.

The key issue when evaluating a forecasting process is the improvements achieved in supply chain performance. While this has only an indirect link to the forecasting accuracy, organisations rely on accuracy improvements as a suitable proxy measure, not least because of their ease of calculation. This paper examines the behaviours of various well-known error measures in the particular context of demand forecasting in the supply chain. We show that, due to the features of SKU demand data, well-known error measures are generally not advisable for the evaluation of judgmental adjustments, and can even give misleading results. To be useful in supply chain applications, an error measure usually needs to have the following properties: (i) scale independence—though it is sometimes desirable to weight measures according to some characteristic such as their profitability; (ii) robustness to outliers; and (iii) interpretability (though the focus might occasionally shift to extremes, e.g., where ensuring a minimum level of supply is important).

The most popular measure used in practice is the mean absolute percentage error, MAPE (Fildes & Goodwin, 2007), which has long been being criticised (see, for example, Fildes, 1992, Hyndman & Koehler, 2006, Kolassa & Schutz, 2007). In particular, the use of percentage errors is often inadvisable, due to the large number of extremely high percentages which arise from relatively low actual demand values.

To overcome the disadvantages of percentage measures, the MASE (mean absolute scaled error) measure was proposed by Hyndman and Koehler (2006). The MASE is a relative error measure which uses the MAE (mean absolute error) of a benchmark forecast (specifically, the random walk) as its denominator. In this paper we analyse the MASE and show that, like the MAPE, it also has a number of disadvantages. Most importantly: (i) it introduces a bias towards overrating the performance of a benchmark forecast as a result of arithmetic averaging; and (ii) it is vulnerable to outliers, as a result of dividing by small benchmark MAE values.

To ensure a more reliable evaluation of the effectiveness of adjustments, this paper proposes the use of an enhanced measure that shows the average relative improvement in MAE. In contrast to MASE, it is proposed that the weighted geometric average be used to find the average relative MAE. By taking the statistical forecast as a benchmark, it becomes possible to evaluate the relative change in forecasting accuracy yielded by the use of judgmental adjustments, without experiencing the limitations of other standard measures. Therefore, the proposed statistic can be used to provide a more robust and easily interpretable indicator of changes in accuracy, meeting the criteria laid down earlier.

The importance of the choice of an appropriate error measure can be seen from the fact that previous studies of the gains in accuracy from the judgmental adjustment process have produced conflicting results (e.g., Fildes et al., 2009, Franses & Legerstee, 2010). In these studies, different measures were applied to different datasets and arrived at different conclusions. Some studies where a set of measures was employed reported an interesting picture, where adjustments improved the accuracy in certain settings according to MdAPE (median absolute percentage error), while harming the accuracy in the same settings according to MAPE (Fildes et al., 2009; Trapero, Pedregal, Fildes, & Weller, 2011). In practice, such results may be damaging for forecasters and forecast users, since they do not give a clear indication of the changes in accuracy that correspond to some well-known loss function. Using real-world data, this paper considers the appropriateness of various previously used measures, and demonstrates the use of the proposed enhanced accuracy measurement scheme.

The next section describes the data employed for the analysis in this paper. Section 3 illustrates the disadvantages and limitations of various well-known error measures when they are applied to SKU-level data. In Section 4, the proposed accuracy measure is introduced. Section 5 contains the results from measuring the accuracy of judgmental adjustments with real-world data using the alternative measures and explains the differences in the results, demonstrating the benefits of the proposed enhanced accuracy measure. The concluding section summarises the results of the empirical evaluation and offers practical recommendations as to which of the different error measures can be employed safely.

## 2. Descriptive analysis of the source data

The current research employed data collected from a company specialising in the manufacture of fast-moving consumer goods (FMCG). This is an extended data set from one of the companies considered by Fildes et al. (2009). The company concerned is a leading European provider of household and personal care products to a wide range of major retailers. Table 1 summarises the data set and indicates the number of cases used for the analysis. Each case includes (i) the one-step-ahead monthly forecast prepared using some statistical method (this will be called the system forecast); (ii) the corresponding judgmentally adjusted forecast (this will be called the final forecast); and (iii) the corresponding actual demand value. The system forecast was obtained using an enterprise software package, and the final forecast was obtained as a result of a revision of the statistical forecast by experts (Fildes et al., 2009). The two forecasts coincide when the experts had no extra information to add. The data set is representative of most FMCG manufacturing or distribution companies which deal with large numbers of time series of different lengths relating to different products, and is similar to the other manufacturing data sets considered by Fildes et al. (2009), in terms of the total number of time series, the proportion of judgmentally adjusted forecasts and the frequencies of occurrence of zero errors and zero actuals.

Since the data relate to FMCG, the numbers of cases of zero demand periods and zero errors are not large (see Table 1). However, the further investigation of the properties of error measures presented in Section 3 will