Measuring Forecasting Accuracy: Problems and Recommendations (by the Example of SKU-Level Judgmental Adjustments)

Chapter · October 2013

DOI: 10.1007/978-3-642-39869-8_4

CITATIONS
4

READS
1,007

2 authors:

Andrey Davydenko
JSC "CSBI"
6 PUBLICATIONS **74** CITATIONS

SEE PROFILE

Robert Fildes
Lancaster University
156 PUBLICATIONS **5,004** CITATIONS

SEE PROFILE

Please cite this paper as:
Davydenko, A., & Fildes, R. (2014). Measuring Forecasting Accuracy: Problems and Recommendations (by the Example of SKU-Level Judgmental Adjustments). In *Intelligent Fashion Forecasting Systems: Models and Applications* (pp. 43-70). Springer Berlin Heidelberg.

Measuring Forecasting Accuracy: Problems and Recommendations (by the example of SKU-level judgmental adjustments)[1]

Andrey Davydenko[2], Robert Fildes

Department of Management Science, Lancaster University, Lancaster, LA1 4YX, UK

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. The general accuracy evaluation scheme recommended in the paper is applicable in a wide range of settings including forecasting for fashion industry.

Keywords: *judgmental adjustments, forecasting support systems, forecast accuracy, forecast evaluation, forecast error measures*.

[1] This paper is an extended version of Davydenko and Fildes (2013) which appeared in the International Journal of Forecasting.

[2] Corresponding author. E-mail: davydenkoa@yandex.ru.

Please cite this paper as:
Davydenko, A., & Fildes, R. (2014). Measuring Forecasting Accuracy: Problems and Recommendations (by the Example of SKU-Level Judgmental Adjustments). In *Intelligent Fashion Forecasting Systems: Models and Applications* (pp. 43-70). Springer Berlin Heidelberg.

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 (Sanders & Ritzman, 2004; Fildes & Goodwin, 2007). 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 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

Please cite this paper as:

Davydenko, A., & Fildes, R. (2014). Measuring Forecasting Accuracy: Problems and Recommendations (by the Example of SKU-Level Judgmental Adjustments). In *Intelligent Fashion Forecasting Systems: Models and Applications* (pp. 43-70). Springer Berlin Heidelberg.

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 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, of 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 is justified by 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 vari-

Please cite this paper as:

Davydenko, A., & Fildes, R. (2014). Measuring Forecasting Accuracy: Problems and Recommendations (by the Example of SKU-Level Judgmental Adjustments). In *Intelligent Fashion Forecasting Systems: Models and Applications* (pp. 43-70). Springer Berlin Heidelberg.

ous 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 the fourth section, the proposed accuracy measure is introduced. The fifth section 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) which are fashionable in nature. 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 contains 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.

Table 1: Source data summary.

Total number of cases	6882
Total number of time series (SKUs)	412
Period of observations	Mar 2004 to Jul 2007
Total number of adjusted statistical forecasts	4779 (69%)

Please cite this paper as:
Davydenko, A., & Fildes, R. (2014). Measuring Forecasting Accuracy: Problems and Recommendations (by the Example of SKU-Level Judgmental Adjustments). In *Intelligent Fashion Forecasting Systems: Models and Applications* (pp. 43-70). Springer Berlin Heidelberg.

(% of total number of cases)	
Number of zero actual demand periods (% of total number of cases)	271 (4%)
Number of zero-error statistical forecasts (% of total number of cases)	47 (<1%)
Number of zero-error judgmentally adjusted forecasts (% of total number of adjusted forecasts)	61 (1%)
Number of positive adjustments (% of total number of adjusted forecasts)	3394 (71%)
Number of negative adjustments (% of total number of adjusted forecasts)	1385 (29 %)

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 also consider possible situations when the data involve small counts, and zero observations occur more frequently (as is common with intermittent demand data).

As Table 1 shows, for this particular data set, adjustments of positive sign occur more frequently than adjustments of negative sign. However, in order to characterise the average magnitude of the adjustments, an additional analysis is required. In their study of judgmental adjustments, Fildes et al. (2009) analysed the size of judgmental adjustments using the measure of relative adjustments that is defined as $100 \times (\text{Final forecast} - \text{System forecast})/\text{System forecast}$.

As the values of the relative adjustments are scale-independent, they can be compared across time series. However, the above measure is asymmetrical. For example, if an expert doubles a statistical forecast (say from 10 units to 20 units), he/she increases it by 100%, but if he/she halves a statistical forecast (say from 20 units to 10 units), he/she decreases it by 50% (not 100%). The sampling distribution of the relative adjustment is bounded by –100% on the left side and unbounded on the right side (see Fig. 1). Generally, these effects mean that the distribution of the relative adjustment may become non-informative about the size of the adjustment as measured on the original scale. When defining a 'symmetric measure', Mathews and Diamantopoulos (1987) argued for a measure where the adjustment size is measured relative to an average of the system and final forecasts. The same principle is used in the symmetric MAPE (sMAPE) measure proposed by Makridakis (1993). However, Goodwin and Lawton (1999) later showed that such approaches still do not lead to the desirable property of symmetry.

Please cite this paper as:
Davydenko, A., & Fildes, R. (2014). Measuring Forecasting Accuracy: Problems and Recommendations (by the Example of SKU-Level Judgmental Adjustments). In *Intelligent Fashion Forecasting Systems: Models and Applications* (pp. 43-70). Springer Berlin Heidelberg.

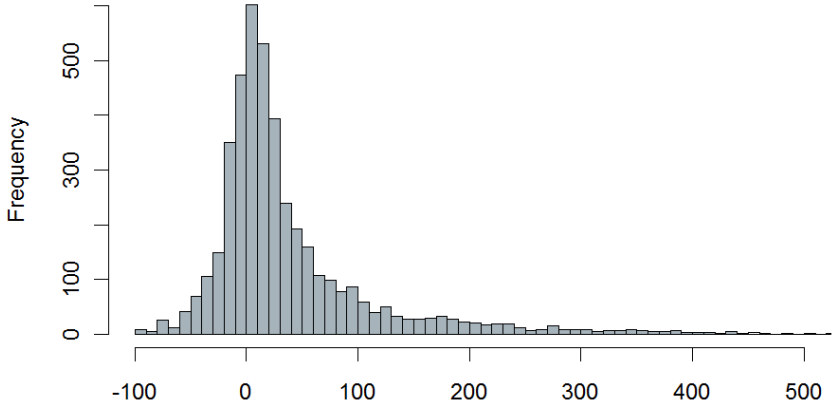

Fig. 1. Histogram of the relative adjustment, measured in percentages.

In this paper, in order to avoid the problem of the non-symmetrical scale of the relative adjustment, we carry out the analysis of the magnitude of adjustments using the natural logarithm of the $(\text{Final forecast}/\text{System forecast})$ ratio. Log-transformation is a common approach to restore symmetry with ratio data since $\ln(A/B) = -\ln(B/A)$ for any positive numbers A and B.

From Fig. 2, it can be seen that the log-transformed relative adjustment follows a leptokurtic distribution and this distribution is still non-symmetrical (although not as severely as for the original data shown on Fig. 1). As is well known, the sample mean is not an efficient measure of location under departures from normality (Wilcox, 2005). We therefore used the trimmed mean as a more robust summary measure of location. The optimal trim level that corresponds to the lowest variance of the trimmed mean depends on the distribution, which is unknown in the current case. Some studies have shown that, for symmetrical distributions, a 5% trim generally ensures a high efficiency with a useful degree of robustness (e.g., Hill & Dixon, 1982). However, it is also known that the trimmed mean gives a biased estimate if the distribution is skewed (Marques, Neves, & Sarmento, 2000). We used a 2% trim in order to eliminate the influence of outliers while at the same time avoiding introducing a substantial bias.

The results presented in Table 2 suggest that, on average, for the dataset under consideration, the magnitude of positive adjustments is higher than the magnitude of negative adjustments, measured relative to the system forecast. Even after using a log scale to treat percentages to baseline symmetrically, the magnitude of positive adjustments is pronouncedly higher than the magnitude of negative ones. The average magnitude of a positive relative adjustment is about twice as large as the

Please cite this paper as:
Davydenko, A., & Fildes, R. (2014). Measuring Forecasting Accuracy: Problems and Recommendations (by the Example of SKU-Level Judgmental Adjustments). In *Intelligent Fashion Forecasting Systems: Models and Applications* (pp. 43-70). Springer Berlin Heidelberg.

average magnitude of a negative adjustment. Also, adjustments with positive signs have much higher ranges than negative ones.

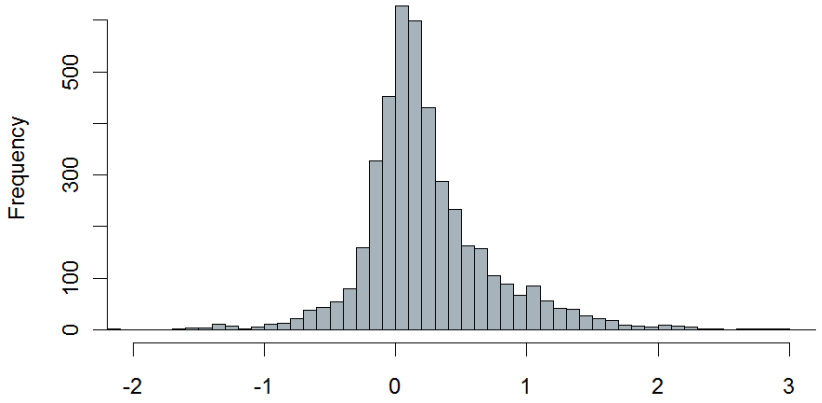

Fig. 2. Histogram of ln(Final forecast/System forecast)**.**

Table 2: Summary statistics for the magnitude of adjustment.

Sign of adjustment	ln(Final forecast/System forecast)				
	1st quartile	Median	3rd quartile	$Mean_{(2\% trim)}$	$exp[Mean_{(2\% trim)}]$
Positive	0.123	0.273	0.592	0.412	1.510
Negative	–0.339	–0.153	–0.071	–0.290	0.749
Both	–0.043	0.144	0.425	0.218	1.243

Please cite this paper as:
Davydenko, A., & Fildes, R. (2014). Measuring Forecasting Accuracy: Problems and Recommendations (by the Example of SKU-Level Judgmental Adjustments). In *Intelligent Fashion Forecasting Systems: Models and Applications* (pp. 43-70). Springer Berlin Heidelberg.

3. Appropriateness of existing measures for SKU-level demand data

3.1. Percentage errors

Let the forecasting error for a given time period t and SKU i be

$$e_{i,t} = Y_{i,t} - F_{i,t},$$

where $Y_{i,t}$ is a demand value for SKU i observed at time t, and $F_{i,t}$ is the forecast of $Y_{i,t}$.

A traditional way to compare the accuracy of forecasts across multiple time series is based on using absolute percentage errors (Hyndman & Koehler, 2006). Let us define the percentage error (PE) as $p_{i,t} = 100 \times e_{i,t}/Y_{i,t}$. Hence, the absolute percentage error (APE) is $|p_{i,t}|$. The most popular PE-based measures are MAPE and MdAPE, which are defined as follows:

$$\text{MAPE} = \text{mean}(|p_{i,t}|),$$
$$\text{MdAPE} = \text{median}(|p_{i,t}|),$$

where $\text{mean}(|p_{i,t}|)$ denotes the sample mean of $|p_{i,t}|$ over all available values of *i* and *t*, and $\text{median}(|p_{i,t}|)$ is the sample median.

In the study by Fildes et al. (2009), these measures served as the main tool for the analysis of the accuracy of judgmental adjustments. In order to determine the change in forecasting accuracy, MAPE and MdAPE values of the statistical baseline forecasts and final judgmentally adjusted forecasts were calculated and compared. The significance of the change in accuracy was assessed based on the distribution of the differences between the absolute percentage errors (APEs) of forecasts. The difference between APEs is defined as

$$d_{i,t}^{\text{APE}} = |p_{i,t}^{\text{f}}| - |p_{i,t}^{\text{s}}|,$$

where $|p_{i,t}^{\text{f}}|$ and $|p_{i,t}^{\text{s}}|$ denote APEs for the final and system forecasts, respectively, for a given SKU i and period t. Fildes et al. (2009) used the Wilcoxon's two-sample paired signed rank test to compare the APEs of the final and system forecasts. This is equivalent to performing a one-sample signed rank test to test the median of $d_{i,t}^{\text{APE}}$ against zero.

Please cite this paper as:
Davydenko, A., & Fildes, R. (2014). Measuring Forecasting Accuracy: Problems and Recommendations (by the Example of SKU-Level Judgmental Adjustments). In *Intelligent Fashion Forecasting Systems: Models and Applications* (pp. 43-70). Springer Berlin Heidelberg.

The sample mean of $d_{i,t}^{\mathrm{APE}}$ is the difference between the MAPE values corresponding to the final and system forecasts:

$$\mathrm{mean}\left(d_{i,t}^{\mathrm{APE}}\right) = \mathrm{mean}\left(\left|p_{i,t}^{\mathrm{f}}\right|\right) - \mathrm{mean}\left(\left|p_{i,t}^{\mathrm{s}}\right|\right) = \mathrm{MAPE}^{\mathrm{f}} - \mathrm{MAPE}^{\mathrm{s}}. \quad (1)$$

Therefore, testing the median of $d_{i,t}^{\mathrm{APE}}$ against zero using the above tests leads to establishing whether $\mathrm{MAPE}^{\mathrm{f}}$ differs significantly from $\mathrm{MAPE}^{\mathrm{s}}$ (in case if the distribution of $d_{i,t}^{\mathrm{APE}}$ is symmetric, which is one of the assumptions of the above tests). In fact, the distribution of $d_{i,t}^{\mathrm{APE}}$ is inherently skewed, which complicates matters and may result in drawing erroneous conclusions, but we will now not focus on this particular problem.

The results reported suggest that, overall, the value of MAPE was improved by the use of adjustments, but the accuracy of positive and negative adjustments differed substantially. Based on the MAPE measure, it was found that positive adjustments did not change the forecasting accuracy significantly, while negative adjustments led to significant improvements. However, percentage error measures have a number of disadvantages when applied to the adjustments data, as we explain below.

One well-known disadvantage of percentage errors is that when the actual value $Y_{i,t}$ in the denominator is relatively small compared to the forecast error $e_{i,t}$, the resulting percentage error $p_{i,t}$ becomes extremely large, which distorts the results of further analyses (Hyndman & Koehler, 2006). Such high values can be treated as outliers, since they often do not allow for a meaningful interpretation (large percentage errors are not necessarily harmful or damaging, as they can arise merely from relatively low actual values). However, identifying outliers in a skewed distribution is a non-trivial problem, where it is necessary to determine an appropriate trimming level in order to find robust estimates, while at the same time avoiding losing too much information. Usually authors choose the trimming level for MAPE based on their experience after experimentation (for example, Fildes et al., 2009, used a 2% trim), but this decision still remains subjective. Moreover, the trimmed mean gives a biased estimate of location for highly skewed distributions (Marques et al., 2000), which complicates the interpretation of the trimmed MAPE. In particular, for a random variable that follows a highly skewed distribution, the expected value of the trimmed mean differs from the expected value of the random variable itself. This bias depends on both the trim level and the number of observations used to calculate the trimmed mean. Therefore, it is difficult to compare the measurement results based on the trimmed means for samples that contain different numbers of observations, even when the trim level remains the same.

Please cite this paper as:
Davydenko, A., & Fildes, R. (2014). Measuring Forecasting Accuracy: Problems and Recommendations (by the Example of SKU-Level Judgmental Adjustments). In *Intelligent Fashion Forecasting Systems: Models and Applications* (pp. 43-70). Springer Berlin Heidelberg.

SKU-level demand time series typically exhibit a high degree of variation in actual values, due to seasonal effects and the changing stages of a product's life cycle. Therefore, data on adjustments can contain a high proportion of low demand values, which makes PE-based measures particularly inadvisable in this context. Considering extremes, a common occurrence in the situation of intermittent demand is for many observations (and forecasts) to be zero (see the discussion by Syntetos & Boylan, 2005). All cases with zero actual values must be excluded from the analysis, since the percentage error cannot be computed when $Y_{i,t} = 0$, due to its definition.

The extreme percentage errors that can be obtained can be shown using scaled values of errors and actual demand values (Fig. 3). The variables shown were scaled by the standard deviation of actual values in each series in order to eliminate the differences between time series. It can be seen that the final forecast errors have a skewed distribution and are correlated with both the actual values and the signs of adjustments; it is also clear that a substantial number of the errors are comparable to the actual demand values. Excluding observations with relatively low values on the original scale (here, all observations less than 10 were excluded from the analysis, as was done by Fildes et al., 2009) still cannot improve the properties of percentage errors sufficiently, since a large number of observations still remain in the area where the actual demand value is less than the absolute error. This results in extremely high APEs (>100%), which are all too easy to misinterpret (since very large APEs do not necessarily correspond to very damaging errors, and arise primarily because of low actual demand values). In Fig. 3, the area below the dashed line shows cases in which the errors were higher than the actual demand values. These cases result in extreme percentage errors, as shown in Fig. 4. Due to the presence of extreme percentages, the distribution of APEs becomes highly skewed and heavy-tailed, which makes MAPE-based estimates highly unstable.

A widely used robust alternative to MAPE is MdAPE. However, MdAPE is neither easily interpretable nor sufficiently indicative of changes in accuracy when forecasting methods have different shaped error distributions. The sample median of the APEs is resistant to the influence of extreme cases, but is also insensitive to large errors, even if they are not outliers or extreme percentages. Comparing the accuracy using the MdAPE shows the changes in accuracy that relate to the lowest 50% of APEs. However, MdAPE's improvement can be accompanied by remaining more damaging errors lying above the median if the shapes of the error distributions differ. In Section 5, it will be shown that, while the MdAPE indicates that judgmental adjustments improve the accuracy for a given dataset, the trimmed MAPE suggests the opposite to be the case. Moreover, the task of assessing the statistical significance of changes for MdAPE can be problematic due to the non-symmetric distributions of APEs. Therefore, additional indicators are required in

Please cite this paper as:
Davydenko, A., & Fildes, R. (2014). Measuring Forecasting Accuracy: Problems and Recommendations (by the Example of SKU-Level Judgmental Adjustments). In *Intelligent Fashion Forecasting Systems: Models and Applications* (pp. 43-70). Springer Berlin Heidelberg.

order to be able to draw better-substantiated conclusions with regard to the forecasting accuracy.

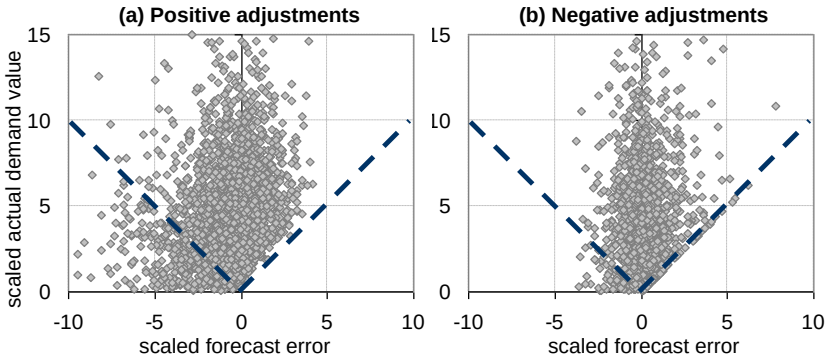

Fig. 3. Dependencies between forecast error, actual value, and the sign of adjustment (based on scaled data).

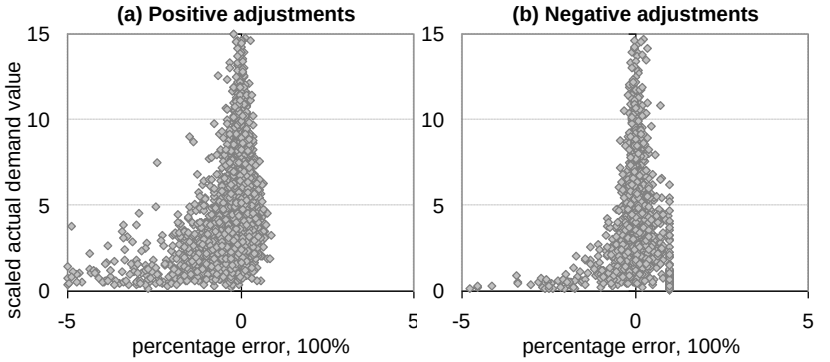

Fig. 4. Percentage errors, depending on the actual demand value and adjustment sign.

Apart from the presence of extreme APEs, another problem with using PE-based measures is that they can bias the comparison in favour of methods that issue low forecasts (Armstrong, 1985; Armstrong & Collopy, 1992; Kolassa & Schutz, 2007). This happens because, under certain conditions, percentage errors put a heavier penalty on positive errors than on negative errors. In particular, we can observe it when the forecast is taken as fixed. To illustrate this phenomenon, Kolassa and Schutz (2007) provide the following example. Assume that we have a time series that contains values distributed uniformly between 10 and 50. If we are

Please cite this paper as:
Davydenko, A., & Fildes, R. (2014). Measuring Forecasting Accuracy: Problems and Recommendations (by the Example of SKU-Level Judgmental Adjustments). In *Intelligent Fashion Forecasting Systems: Models and Applications* (pp. 43-70). Springer Berlin Heidelberg.

using a symmetrical loss function, the best forecast for this time series would be 30. However, a forecast of 22 produces a better accuracy in terms of MAPE. As a result, if the aim is to choose a method that is better in terms of a linear loss, then the values of PE-based measures can be misleading. The way in which the use of MAPE can bias the comparison of the performances of judgmental adjustments of different signs will be illustrated below.

One important effect which arises from the presence of cognitive biases and the non-negative nature of demand values is the fact that the most damaging positive adjustments (producing the largest absolute errors) typically correspond to relatively low actuals (left corner of Fig. 3(a)), while the worst negative adjustments (producing the largest absolute errors) correspond to higher actuals (centre section, Fig. 3(b)). More specifically, the following general dependency can be found within most time series. The difference between the absolute final forecast error $\left|e_{i,t}^{\mathrm{f}}\right|$ and the absolute statistical forecast error $\left|e_{i,t}^{\mathrm{s}}\right|$ is positively correlated with the actual value $Y_{i,t}$ for positive adjustments, while there is a negative correlation for negative adjustments. To reveal this effect, distribution-free measures of the association between variables were used. For each SKU i, Spearman's ρ coefficients were calculated, representing the correlation between the improvement in terms of absolute errors $\left(\left|e_{i,t}^{\mathrm{f}}\right| - \left|e_{i,t}^{\mathrm{s}}\right|\right)$ and the actual value $Y_{i,t}$. Fig. 5 shows the distributions of the coefficients ρ_i^+, calculated for positive adjustments, and ρ_i^-, corresponding to negative adjustments (the coefficients can take values 1 and –1 when only a few observations are present in a series). For the given dataset, $\mathrm{mean}(\rho_i^+) \approx 0.47$ and $\mathrm{mean}(\rho_i^-) \approx -0.44$, indicating that the improvement in forecasting is markedly correlated with the actual demand values. This illustrates the fact that positive adjustments are most effective for larger values of demand, and least effective (or even damaging) for smaller values of demand. Actually, efficient averaging of correlation coefficients requires applying Fisher's z transformation to them and then transforming back the result (see, e.g., Mudhoklar, 1983). But here we used raw coefficients because we only wanted to show that the ρ value clearly correlates with the adjustment sign.

Please cite this paper as:

Davydenko, A., & Fildes, R. (2014). Measuring Forecasting Accuracy: Problems and Recommendations (by the Example of SKU-Level Judgmental Adjustments). In *Intelligent Fashion Forecasting Systems: Models and Applications* (pp. 43-70). Springer Berlin Heidelberg.

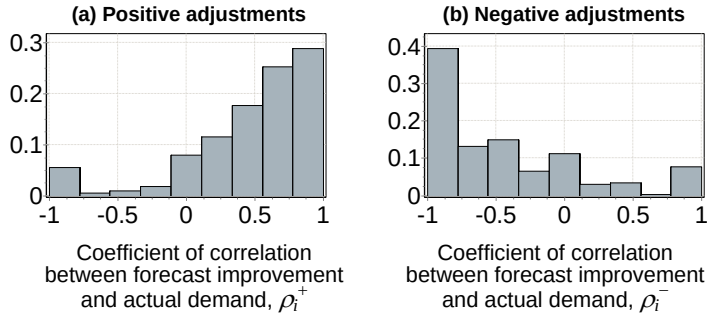

Fig. 5. Spearman's ρ coefficients showing the correlation between the improvement in accuracy and the actual demand value for each time series (relative frequency histograms).

Because of the division by the scale factor that is correlated with the numerator, the difference of APEs (which is calculated as $d_{i,t}^{\mathrm{APE}} = 100 \times \left(\left|e_{i,t}^{\mathrm{f}}\right| - \left|e_{i,t}^{s}\right|\right)/Y_{i,t}$) will not reflect changes in forecasting accuracy in terms of a symmetric loss function. More specifically, for positive adjustments, $d_{i,t}^{\mathrm{APE}}$ will systematically downgrade improvements in accuracy and exaggerate degradations of accuracy (on the percentage scale). In contrast, for negative adjustments, the improvements will be exaggerated, while the errors from harmful forecasts will receive smaller weights. Since the difference in MAPEs is calculated as the sample mean of $d_{i,t}^{\mathrm{APE}}$ (in accordance with equation (1)), the comparison of forecasts using MAPE will also give a result which is biased towards underrating positive adjustments and overrating negative adjustments. Consequently, since the forecast errors arising from adjustments of different signs are penalised differently, the MAPE measure is flawed when comparing the performances of adjustments of different signs. One of the aims of the present research has therefore been to reinterpret the results of previous studies through the use of alternative measures.

A second measure based on percentage errors was also used by Franses and Legerstee (2010). In order to evaluate the accuracy of improvements, the RMSPE (root mean square percentage error) was calculated for the statistical and judgmentally adjusted forecasts, and the resulting values were then compared. Based on this measure, it was concluded that the expert adjusted forecasts were no better than the model forecasts. However, the RMSPE is also based on percentage errors, and is affected by the outliers and biases described above even more strongly.

Please cite this paper as:
Davydenko, A., & Fildes, R. (2014). Measuring Forecasting Accuracy: Problems and Recommendations (by the Example of SKU-Level Judgmental Adjustments). In *Intelligent Fashion Forecasting Systems: Models and Applications* (pp. 43-70). Springer Berlin Heidelberg.

3.2. Relative errors

Another approach to obtaining scale-independent measures is based on using relative errors. The relative error (RE) is defined as

$$RE_{i,t} = e_{i,t}/e_{i,t}^{\mathrm{b}},$$

where $e_{i,t}^{\mathrm{b}}$ is the forecast error obtained from a benchmark method. Usually a naïve forecast is taken as the benchmark method.

Well-known measures based on relative errors include Mean Relative Absolute Error (MRAE), Median Relative Absolute Error (MdRAE), and Geometric Mean Relative Absolute Error (GMRAE):

$$\begin{aligned}\mathrm{MRAE} &= \mathrm{mean}(|RE_{i,t}|),\\ \mathrm{MdRAE} &= \mathrm{median}(|RE_{i,t}|),\\ \mathrm{GMRAE} &= \mathrm{gmean}(|RE_{i,t}|),\end{aligned}$$

where mean, median, and gmean respectively denote the sample mean, sample median, and the sample geometric mean over all possible values of i and t.

Averaging the ratios of absolute errors across individual observations overcomes the problems related to dividing by actual values. In particular, the RE-based measures are not affected by the presence of low actual values, or by the correlation between errors and actual outcomes. However, REs also have a number of limitations.

The calculation of $RE_{i,t}$ requires division by the non-zero error of the benchmark forecast $e_{i,t}^{\mathrm{b}}$. In the case of calculating GMRAE, it is also required that $e_{i,t} \neq 0$. The actual and forecasted demands are usually count data, which means that the forecasting errors are count data as well. With count data, the probability of a zero error of the benchmark forecast can be non-zero. Such cases must be excluded from the analysis when using relative errors. When using intermittent demand data, the use of relative errors becomes impossible due to the frequent occurrences of zero errors (Hyndman, 2006; Syntetos & Boylan, 2005).

As was pointed out by Hyndman & Koehler (2006), in the case of continuous distributions, the benchmark forecast error $e_{i,t}^{\mathrm{b}}$ can have a positive probability density at zero, and therefore the use of MRAE can be problematic. In particular, $RE_{i,t}$ can follow a heavy-tailed distribution for which the sample mean becomes a highly inefficient estimate that is vulnerable to outliers. In addition, the distribution of $|RE_{i,t}|$ is highly skewed. At the same time, while MdRAE is highly robust, it cannot be sufficiently informative, as it is insensitive to large REs which lie in

Please cite this paper as:
Davydenko, A., & Fildes, R. (2014). Measuring Forecasting Accuracy: Problems and Recommendations (by the Example of SKU-Level Judgmental Adjustments). In *Intelligent Fashion Forecasting Systems: Models and Applications* (pp. 43-70). Springer Berlin Heidelberg.

the tails of the distribution. Thus, even if the large REs are not outliers which arise from the division by relatively small benchmark errors, they still will not be taken into account when using MdRAE. Averaging the absolute REs using GMRAE is preferable to using either MRAE or MdRAE, as it provides a reliable and robust estimate and at the same time takes into account the values of REs which lie in the tails of the distribution. Also, when averaging the benchmark ratios, the geometric mean has the advantage that it produces rankings which are invariant to the choice of the benchmark (see Fleming & Wallace, 1986).

Fildes (1992) recommends the use of the Relative Geometric Root Mean Square Error (RelGRMSE). The RelGRMSE for a particular time series i is defined as

$$\mathrm{RelGRMSE}_i = \left(\frac{\prod_{t \in T_i} \left(e_{i,t}\right)^2}{\prod_{t \in T_i} \left(e_{i,t}^{\mathrm{b}}\right)^2} \right)^{\frac{1}{2n_i}},$$

where T_i is a set containing time periods for which non-zero errors $e_{i,t}$ and $e_{i,t}^{\mathrm{b}}$ are available, and n_i is the number of elements in T_i.

After obtaining the RelGRMSE for each series, Fildes (1992) recommends finding the geometric mean of the RelGRMSEs across all time series, thus obtaining $\mathrm{gmean}(\mathrm{RelGRMSE}_i)$. As Hyndman (2006) pointed out, the Geometric Root Mean Square Error (GRMSE) and the Geometric Mean Absolute Error (GMAE) are identical because the square roots cancel each other in a geometric mean. Similarly, it can be shown that

$$\mathrm{gmean}(\mathrm{RelGRMSE}_i) = \mathrm{GMRAE}.$$

An alternative representation of GMRAE is:

$$\mathrm{GMRAE} = \exp\left[\frac{1}{\sum_{i=1}^{m} n_i} \sum_{i=1}^{m} \sum_{t \in T_i} \ln\left|RE_{i,t}\right| \right],$$

where m is the total number of time series, and other variables retain their previous meaning.

For the adjustments data set under consideration, only a small proportion of observations contain zero errors (about 1%). It has been found empirically that for the given data set the log-transformed absolute REs, $\ln\left|RE_{i,t}\right|$, can be approximated adequately using a distribution which has a finite variance. In fact, even if a heavy-tailed distribution of $\ln\left|RE_{i,t}\right|$ arises, the influence of extreme cases can be eliminated based on various robustifying schemes such as trimming or Winsoriz-

Please cite this paper as:
Davydenko, A., & Fildes, R. (2014). Measuring Forecasting Accuracy: Problems and Recommendations (by the Example of SKU-Level Judgmental Adjustments). In *Intelligent Fashion Forecasting Systems: Models and Applications* (pp. 43-70). Springer Berlin Heidelberg.

ing. In contrast to APEs, the use of such schemes for $\ln|RE_{i,t}|$ is unlikely to lead to biased estimates, since the distribution of $\ln|RE_{i,t}|$ is not highly skewed.

Though GMRAE (or, equivalently, $\mathrm{gmean}(\mathrm{RelGRMSE}_i)$) has some desirable statistical properties and can give a reliable aggregated indication of changes in accuracy, its use can be complicated for the following two reasons. Firstly, as was mentioned previously, zero-error forecasts cannot be taken into account directly. Secondly, in a similar way to the median, the geometric mean of absolute errors generally does not reflect changes in accuracy under standard loss functions. For instance, for a particular time series, GMAE (and, hence, GMRAE) favours methods which produce errors with heavier tailed-distributions, while for the same series RMSE (root mean square error) can suggest the opposite ranking.

The latter aspect of using GMRAE can be illustrated using the following example. Suppose that for a particular time series, method A produces errors e_t^{A} that are independent and identically distributed variables following a heavy-tailed distribution. More specifically, let e_t^{A} follow the *t*-distribution with $\nu = 3$ degrees of freedom: $e_t^{\mathrm{A}} \sim t_\nu$. Also, let method B produce independent errors that follow the normal distribution: $e_t^{\mathrm{B}} \sim N(0, 3)$. Let method B be the benchmark method. It can be shown analytically that the variances for e_t^{A} and e_t^{B} are equal: $\mathrm{Var}(e_t^{\mathrm{A}}) = \mathrm{Var}(e_t^{\mathrm{B}}) = 3$. Thus, the relative RMSE (RelRMSE, the ratio of the two RMSEs) for this series is 1. However, the Relative Geometric RMSE (or GMRAE) will show that method A is better than method B: GMRAE ≈ 0.69 (based on 10^6 simulated pairs of e_t^{A} and e_t^{B}). Now if, for example, $e_t^{\mathrm{B}} \sim N(0, 2.5)$, then the RelRMSE and GMRAE will be 1.10 and 0.76, respectively. This means that method B is now preferable in terms of the variance of errors, while method A is still (substantially) better in terms of the GMRAE. However, the geometric mean absolute error is rarely used when optimising predictions with the use of mathematical models. Some authors claim that the comparison based on RelRMSE can be more desirable, as in this case the criterion used for the optimisation of predictions corresponds to the evaluation criteria (Zellner, 1986; Diebold, 1993).

The above example has demonstrated that even for a single time series a statistically significant improvement of GMRAE is not equivalent to a statistically significant improvement in terms of RMSE. Analogously, it can be demonstrated that the GMRAE is not indicative of changes in terms of MAE.

Thus, analogously to what was said with regard to PE-based measures, if the aim of the comparison is to choose a method that is better in terms of a linear or a quadratic loss, then GMRAE may not be sufficiently informative, or may even lead to counterintuitive conclusions.

Please cite this paper as:
Davydenko, A., & Fildes, R. (2014). Measuring Forecasting Accuracy: Problems and Recommendations (by the Example of SKU-Level Judgmental Adjustments). In *Intelligent Fashion Forecasting Systems: Models and Applications* (pp. 43-70). Springer Berlin Heidelberg.

3.3. Percent Better

A simple approach to compare forecasting accuracy of methods A and B is to calculate the percentage of cases when method A was closer to the actual observation than method B. This measure is known as 'Percent Better' (further abbreviated as PB) and was recommended by some authors as a fairly good indicator (see, e.g., Armstrong & Collopy, 1992; Chatfield, 2001). It has the advantage of being immune to outliers and is scale-independent (it can therefore be used to assess accuracy across series). In addition, it can be used for qualitative forecasts (but we will not look at this kind of forecasts in this paper). Although the measure seems to be easy to interpret, the following important limitations should be taken into account.

One problem with PB is that it does not show the magnitude of changes in accuracy (Hyndman & Koehler, 2006). Therefore, it becomes hard to assess the consequences of using one method instead of another. Moreover, as was the case for the GMRAE, we can show that if shapes of error distributions are different for different methods, PB becomes non-indicative of changes in terms of a linear or quadratic loss even for a single series.

Another problem arises when methods A and B frequently produce equal forecasts (e.g., this happens with intermittent demand data). In such situations, obtaining a PB value that is lower than 50% is not necessarily a bad result, but without additional information we cannot draw any conclusions about the changes in accuracy. Suppose absolute errors for methods A and B can be approximated using the Poisson distribution: $\left|e_t^{\mathrm{A}}\right| \sim \mathrm{Pois}(\lambda = 1)$ and $\left|e_t^{\mathrm{B}}\right| \sim \mathrm{Pois}(\lambda = 3)$. Method A is much better than method B in terms of MAE: $E\left[\left|e_t^{\mathrm{A}}\right|\right]/E\left[\left|e_t^{\mathrm{B}}\right|\right] = 1/3$, but $P(\left|e_t^{\mathrm{A}}\right| < \left|e_t^{\mathrm{B}}\right|) \approx 0.077$. Thus, the PB is, approximately, only 7.7 % – a figure that can be misleading. For this example, even looking at 'Percent Worse' and relating it to the PB will also not give us an informative and easily interpretable indication of accuracy.

Thus, in spite of its apparent simplicity, the PB measure is often confusing and does not necessarily show changes in accuracy under linear loss. Moreover, it is not representative of the magnitude of changes and therefore it cannot ensure a complete and reliable analysis of accuracy.

3.4. Scaled errors

In order to overcome the imperfections of PE-based measures, Hyndman and Koehler (2006) proposed the use of the MASE (mean absolute scaled error). For

Please cite this paper as:
Davydenko, A., & Fildes, R. (2014). Measuring Forecasting Accuracy: Problems and Recommendations (by the Example of SKU-Level Judgmental Adjustments). In *Intelligent Fashion Forecasting Systems: Models and Applications* (pp. 43-70). Springer Berlin Heidelberg.

the scenario when forecasts are produced from varying origins but with a constant horizon, the MASE is calculated as follows (see Appendix A):

$$q_{i,t} = \frac{e_{i,t}}{\mathrm{MAE}_i^{\mathrm{b}}}, \qquad \mathrm{MASE} = \mathrm{mean}\left(|q_{i,t}|\right),$$

where $q_{i,t}$ is the scaled error and $\mathrm{MAE}_i^{\mathrm{b}}$ is the mean absolute error (MAE) of the naïve (benchmark) forecast for series i.

Though this was not specified by Hyndman and Koehler (2006), it is possible to show (see Appendix A) that in the given scenario, the MASE is equivalent to the weighted arithmetic mean of relative MAEs, where the number of available values of $e_{i,t}$ is used as the weight:

$$\mathrm{MASE} = \frac{1}{\sum_{i=1}^{m} n_i} \sum_{i=1}^{m} n_i \, r_i, \quad r_i = \frac{\mathrm{MAE}_i}{\mathrm{MAE}_i^{\mathrm{b}}}, \tag{2}$$

where m is the total number of series, n_i is the number of available values of $e_{i,t}$ for series i, $\mathrm{MAE}_i^{\mathrm{b}}$ is the MAE of the benchmark forecast for series i, and MAE_i is the MAE of the forecast being evaluated against the benchmark.

It is known that the arithmetic mean is not strictly appropriate for averaging observations representing relative quantities, and in such situations the geometric mean should be used instead (Spizman & Weinstein, 2008). As a result of using the arithmetic mean of MAE ratios, equation (2) introduces a bias towards overrating the accuracy of a benchmark forecasting method. In other words, the penalty for bad forecasting becomes larger than the reward for good forecasting.

To show how the MASE rewards and penalises forecasts, it can be represented as

$$\mathrm{MASE} = 1 + \frac{1}{\sum_{i=1}^{m} n_i} \sum_{i=1}^{m} n_i \, (r_i - 1).$$

The reward for improving the benchmark MAE from A to B $(A > B)$ in a series i is $R_i = n_i(1 - B/A)$, while the penalty for harming MAE by changing it from B to A is $P_i = n_i(A/B - 1)$. Since $R_i < P_i$, the reward given for improving the benchmark MAE cannot balance the penalty given for reducing the benchmark MAE by the same quantity. As a result, obtaining $\mathrm{MASE} > 1$ does not necessarily indicate that the accuracy of the benchmark method was better on average. This leads to ambiguity in the comparison of the accuracy of forecasts.

For example, suppose that the performance of some forecasting method is compared with the performance of the naïve method across two series $(m = 2)$

Please cite this paper as:
Davydenko, A., & Fildes, R. (2014). Measuring Forecasting Accuracy: Problems and Recommendations (by the Example of SKU-Level Judgmental Adjustments). In *Intelligent Fashion Forecasting Systems: Models and Applications* (pp. 43-70). Springer Berlin Heidelberg.

which contain equal numbers of forecasts and observations. For the first series, the MAE ratio is $r_1 = 1/2$, and for the second series, the MAE ratio is the opposite: $r_2 = 2/1$. The improvement in accuracy for the first series obtained using the forecasting method is the same as the reduction for the second series. However, averaging the ratios gives MASE $= ½ (r_1 + r_2) = 1.25$, which indicates that the benchmark method is better. While this is a well-known point, its implications for error measures, with the potential for misleading conclusions, are widely ignored.

In addition to the above effect, the use of MASE (as for MAPE) may result in unstable estimates, as the arithmetic mean is severely influenced by extreme cases which arise from dividing by relatively small values. In this case, outliers occur when dividing by the relatively small MAEs of benchmark forecast which can appear in short series.

Some authors (e.g., Hoover, 2006) recommend the use of the MAD/MEAN ratio. In contrast to the MASE, the MAD/MEAN ratio approach assumes that the forecasting errors are scaled by the mean of time series elements, instead of by the in-sample MAE of the naïve forecast. The advantage of this scheme is that it reduces the risk of dividing by a small denominator (see Kolassa & Schutz, 2007). However, Hyndman (2006) notes that the MAD/MEAN ratio assumes that the mean is stable over time, which may make it unreliable when the data exhibit trends or seasonal patterns. In Section 5, we show that both the MASE and the MAD/MEAN are prone to outliers for the data set we consider in this paper. Generally, the use of these schemes has the risk of producing unreliable estimates that are based on highly skewed left-bounded distributions.

Thus, while the use of the standard MAPE has long been known to be flawed, the newly proposed MASE also suffers from some of the same limitations, and may also lead to an unreliable interpretation of the empirical results. We therefore need a measure that does not suffer from these problems. The next section presents an improved statistic which is more suitable for comparing the accuracies of SKU-level forecasts.

4. Recommended accuracy evaluation scheme

4.1. Measuring the accuracy of judgmental adjustments

The recommended forecast evaluation scheme is based on averaging the relative efficiencies of adjustments across time series. The geometric mean is the correct average to use for averaging benchmark ratio results, since it gives equal weight to reciprocal relative changes (Fleming & Wallace, 1986). Using the geo-

Please cite this paper as:
Davydenko, A., & Fildes, R. (2014). Measuring Forecasting Accuracy: Problems and Recommendations (by the Example of SKU-Level Judgmental Adjustments). In *Intelligent Fashion Forecasting Systems: Models and Applications* (pp. 43-70). Springer Berlin Heidelberg.

metric mean of MAE ratios, it is possible to define an appropriate measure of the average relative MAE (AvgRelMAE). If the baseline statistical forecast is taken as the benchmark, then the AvgRelMAE showing how the judgmentally adjusted forecasts improve/reduce the accuracy can be found as

$$\text{AvgRelMAE} = \left(\prod_{i=1}^{m} r_i^{n_i}\right)^{1/\sum_{i=1}^{m} n_i}, \quad r_i = \frac{\text{MAE}_i^{\text{f}}}{\text{MAE}_i^{\text{s}}}, \tag{3}$$

where MAE_i^{s} is the MAE of the baseline statistical forecast for series i, MAE_i^{f} is the MAE of the judgmentally adjusted forecast for series i, n_i is the number of available errors of judgmentally adjusted forecasts for series i, and m is the total number of time series. This differs from the proposals of Fildes (1992), who examined the behaviour of the GRMSEs of the individual relative errors.

The MAEs in equation (3) are found as

$$\text{MAE}_i^{\text{f}} = \frac{1}{n_i}\sum_{t\in T_i}\left|e_{i,t}^{\text{f}}\right|, \qquad \text{MAE}_i^{\text{s}} = \frac{1}{n_i}\sum_{t\in T_i}\left|e_{i,t}^{\text{s}}\right|,$$

where $e_{i,t}^{\text{f}}$ is the error of the judgmentally adjusted forecast for period t and series i, T_i is a set containing the time periods for which $e_{i,t}^{\text{f}}$ are available, and $e_{i,t}^{\text{s}}$ is the error of the baseline statistical forecast for period t and series i.

AvgRelMAE is immediately interpretable, as it represents the average relative value of MAE adequately, and directly shows how the adjustments improve/reduce the MAE compared to the baseline statistical forecast. Obtaining $\text{AvgRelMAE} < 1$ means that on average $\text{MAE}_i^{\text{f}} < \text{MAE}_i^{\text{s}}$, and therefore adjustments improve the accuracy, while $\text{AvgRelMAE} > 1$ indicates the opposite. The average percentage improvement in MAE of forecasts is found as $(1 - \text{AvgRelMAE}) \times 100$. If required, equation (3) can also be extended to other measures of dispersion or loss functions. For example, instead of MAE one might use the MSE (mean square error), interquartile range, or mean prediction interval length. The choice of the measure depends on the purposes of analysis. In this study, we use MAE, assuming that the penalty is proportional to the absolute error.

Equivalently, the geometric mean of MAE ratios can be found as

$$\text{AvgRelMAE} = \exp\left(\frac{1}{\sum_{i=1}^{m} n_i}\sum_{i=1}^{m} n_i \ln r_i\right).$$

Please cite this paper as:
Davydenko, A., & Fildes, R. (2014). Measuring Forecasting Accuracy: Problems and Recommendations (by the Example of SKU-Level Judgmental Adjustments). In *Intelligent Fashion Forecasting Systems: Models and Applications* (pp. 43-70). Springer Berlin Heidelberg.

Therefore, obtaining $\sum_{i=1}^{m} n_i \ln r_i < 0$ means an average improvement of accuracy, and $\sum_{i=1}^{m} n_i \ln r_i > 0$ means the opposite.

In theory, the following effect may complicate the interpretation of the AvgRelMAE value. If the distributions of errors $e_{i,t}^{\mathrm{f}}$ and $e_{i,t}^{\mathrm{s}}$ within a given series i have different levels of the kurtosis, then $\ln r_i$ is a biased estimate of $\ln\left(E\left|e_{i,t}^{\mathrm{f}}\right| / E\left|e_{i,t}^{\mathrm{s}}\right|\right)$. Thus, the indication of an improvement under linear loss given by the AvgRelMAE may be biased. In fact, if $n_i = 1$ for each i, then the AvgRelMAE becomes equivalent to the GMRAE, which has the limitations described in Section 3.2. However, our experiments have shown that the bias of $\ln r_i$ diminishes rapidly as n_i increases, becoming negligible for $n_i > 4$.

To eliminate the influence of outliers and extreme cases, the trimmed mean can be used in order to define a measure of location for the relative MAE. The trimmed AvgRelMAE for a given threshold t ($0 \le t \le 0.5$) is calculated by excluding the $[tm]$ lowest and $[tm]$ highest values of $n_i \ln r_i$ from the calculations (square brackets indicate the integer part of tm). As was mentioned in Section 2, the optimal trim level depends on the distribution. In practice, the choice of the trim level usually remains subjective, since the distribution is unknown. Wilcox (1996) wrote that 'Currently there is no way of being certain how much trimming should be done in a given situation, but the important point is that some trimming often gives substantially better results, compared to no trimming' (p. 16). Our experiments show that a 5% level can be recommended for the AvgRelMAE measure. This level ensures high efficiency, because the underlying distribution usually does not exhibit very large departures from the normal distribution. A manual screening for outliers could also be performed in order to exclude time series with non-typical properties from the analysis.

The results described in the next section show that the robust estimates obtained using a 5% trimming level are very close to the estimates based on the whole sample. The distribution of $n_i \ln r_i$ is more symmetrical than the distribution of either the APEs or absolute scaled errors. Therefore, the analysis of the outliers in relative MAEs can be performed more efficiently than the analysis of outliers when using the measures considered previously. Besides, we can assess the statistical significance of changes in accuracy by testing the mean of $n_i \ln r_i$ against zero.

Since the AvgRelMAE does not require scaling by actual values, it can be used in cases of low or zero actuals, as well as in cases of zero forecasting errors. Consequently, it is suitable for intermittent demand forecasts. The only limitation is that the MAEs in equation (3) should be greater than zero for all series. If zero MAEs do occur, they can be handled by the procedure that we describe below.

Thus, the advantages of the recommended accuracy evaluation scheme are that it (i) can be interpreted easily, (ii) represents the performance of the adjustments objectively (without the introduction of substantial biases or outliers), (iii) is in-

Please cite this paper as:
Davydenko, A., & Fildes, R. (2014). Measuring Forecasting Accuracy: Problems and Recommendations (by the Example of SKU-Level Judgmental Adjustments). In *Intelligent Fashion Forecasting Systems: Models and Applications* (pp. 43-70). Springer Berlin Heidelberg.

formative and uses available information efficiently, (iv) is applicable in a wide range of settings, with minimal assumptions about the features of the data, and (v) gives rankings and indicates relative improvements that are invariant to the choice of the benchmark. Importantly, the last property can be ensured only through the use of the geometric mean. If we used a sample median or sample mean instead, this could lead to different rankings depending on the choice of the benchmark.

4.2. Generalized scheme for measuring the accuracy of point forecasts

In general, in order to ensure a reliable evaluation of forecasting accuracy under a symmetric linear loss, we recommend using the following scheme. Suppose we want to measure the accuracy of h-step-ahead forecasts produced with some forecasting method A across m time series. Firstly, we need to select a benchmark method. This, in particular, can be the naïve method. Let n_i denote the number of periods for which both the h-step-ahead forecasts and actual observations are available for series i. Then the accuracy measurement procedure is as follows:

1. For each i in $1 \dots m$
   a. Calculate the relative MAE as $r_i = \frac{\mathrm{MAE}_i^{\mathrm{A}}}{\mathrm{MAE}_i^{\mathrm{B}}}$,
      where $\mathrm{MAE}_i^{\mathrm{A}}$ and $\mathrm{MAE}_i^{\mathrm{B}}$ denote out-of-sample h-step-ahead MAEs for method A and for the benchmark, respectively.
   b. Calculate the weighted log relative MAE as $l_i = n_i \ln r_i$.
2. Calculate the Average Relative MAE as

$$\mathrm{AvgRelMAE} = \exp\left(\frac{1}{\sum_{i=1}^{m} n_i} \sum_{i=1}^{m} l_i\right).$$

   If there is an evidence for a non-normal distribution of l_i, use the following procedure to ensure more efficient estimates:
   a. Find the indices of l_i that correspond to the 5% of largest and 5% of lowest values. Let R be a set that contains the remaining indices.

Please cite this paper as:
Davydenko, A., & Fildes, R. (2014). Measuring Forecasting Accuracy: Problems and Recommendations (by the Example of SKU-Level Judgmental Adjustments). In *Intelligent Fashion Forecasting Systems: Models and Applications* (pp. 43-70). Springer Berlin Heidelberg.

b. Calculate the trimmed version of the AvgRelMAE:

$$\text{AvgRelMAE}^{\text{trimmed}} = \exp\left(\frac{1}{\sum_{i\in R} n_i}\sum_{i\in R} l_i\right).$$

3. Assess the statistical significance of changes by testing the mean of l_i against zero. For this purpose, the Wilcoxon's one-sample signed rank test can be used (assuming that the distribution of l_i is symmetric, but not necessarily normal). If the distribution of l_i is non-symmetric, the binomial test can be used to test the median of l_i against zero. If the distribution has a negative skew then it is likely that the negative median will indicate negative mean as well.

Notes: (a) For low volume data it can be the case that $\text{MAE}_i^{\text{A}} = 0$ or $\text{MAE}_i^{\text{B}} = 0$ (or both). Essentially, MAE represents our estimate of the expected value of absolute error. But our prior knowledge suggests that the expected value of absolute error is larger than zero because for any forecasting task we assume that some level of uncertainty is present. Therefore, obtaining a zero MAE is an inadequate result and we may use some sufficiently small number instead (say MAE=0.001). The extreme r_i values corresponding to such cases should then be excluded from the analysis on step 2 by setting a sufficiently large trim level. If the frequency of obtaining zero MAEs is too high (say larger than 30%), a reliable estimation of the average relative MAE becomes unavailable, and we then have to resort to simply estimating the success rate for the MAE improvement. This can be done by calculating the number of cases when $\text{MAE}_i^{\text{A}} < \text{MAE}_i^{\text{B}}, i = 1 \dots, m$, and then dividing this number by the total number of time series, m. Importantly, as mentioned in Section 3.3, getting a success rate that is statistically lower than 0.5 does not necessarily indicate that method A is worse than method B for count data (because of the possibility of equal MAEs); therefore the sum of ranks should be reported as well. But it is also important to keep in mind that neither the success rate nor the sum of ranks will be indicative of improvements under linear loss if sampling distribution for l_i is heavily skewed.

(b) If distribution of absolute errors is heavily skewed, the MAE, as any sample mean, becomes a very inefficient estimate of the expected value of absolute error. One simple method to improve the efficiency of the estimates while not introducing substantial bias is to use asymmetric trimming algorithms, such as those described by (Alkhazeleh and

Please cite this paper as:
Davydenko, A., & Fildes, R. (2014). Measuring Forecasting Accuracy: Problems and Recommendations (by the Example of SKU-Level Judgmental Adjustments). In *Intelligent Fashion Forecasting Systems: Models and Applications* (pp. 43-70). Springer Berlin Heidelberg.

Razali, 2010). However, further discussions on this topic are outside the scope of our paper.

(c) If a suitable benchmark method is unavailable, we can use the sample mean of time series values instead of the benchmark MAE. The procedure then becomes similar to the MAD/MEAN ratio approach described in Section 3.4, but here the use of the geometric mean i) ensures the correct averaging of ratios (i.e., deviations from the mean will be treated symmetrically) and ii) gives more robust measurement results in cases when mean time series values are relatively small compared to absolute forecasting errors.

(d) In step 2, the optimal trim level depends on the shape of the distribution of l_i. Our experiments suggest that, for the distributions that are likely to be obtained, the efficiency of the trimmed mean is not highly sensitive to the choice of the trim level and any value between 2% and 10% gives reasonably good results. Generally, as was shown by (Andrews et al., 1972), when the underlying distribution is symmetrical and heavy-tailed relative to the Gaussian, the variance of the trimmed mean is quite a lot smaller than the variance of the sample mean. Therefore, the use of the trimmed means for symmetrical distributions can be highly recommended.

5. Results of empirical evaluation

The results of applying the measures described above are shown in Table 3.

For the given dataset, a large number of APEs have extreme values (>100%) which arise from low actual demand values (Fig. 6). Following Fildes et al. (2009), we used a 2% trim level for MAPE values. However, as noted, it is difficult to determine an appropriate trim level. As a result, the difference in APEs between the system and final forecasts has a very high dispersion and cannot be used efficiently to assess improvements in accuracy. It can also be seen that the distribution of APEs is highly skewed, which means that the trimmed means cannot be considered as unbiased estimates of the location. Albeit the distribution of the APEs has a very high kurtosis, our experiments show that increasing the trim level (say from 2% to 5%) would substantially bias the estimates of the location of the APEs due to the extremely high skewness of the distribution. We therefore use the 2% trimmed MAPE in this study. Also, the use of this trim level makes the measurement results comparable to the results of Fildes et al. (2009).

Please cite this paper as:
Davydenko, A., & Fildes, R. (2014). Measuring Forecasting Accuracy: Problems and Recommendations (by the Example of SKU-Level Judgmental Adjustments). In *Intelligent Fashion Forecasting Systems: Models and Applications* (pp. 43-70). Springer Berlin Heidelberg.

Table 3: Accuracy of adjustments according to different error measures

Error measure	**Positive adjustments**		**Negative adjustments**		**All nonzero adjustments**	
	Statistical forecast	*Adjusted forecast*	*Statistical forecast*	*Adjusted forecast*	*Statistical forecast*	*Adjusted forecast*
MAPE, % (untrimmed)	**38.85**	61.54	70.45	**45.13**	**47.88**	56.85
MAPE, % (2 % trimmed)	**30.98**	40.56	48.71	**30.12**	**34.51**	37.22
MdAPE, %	25.48	**20.65**	23.90	**17.27**	24.98	**19.98**
GMRAE	1.00	**0.93**	1.00	**0.70**	1.00	**0.86**
GMRAE (5 % trimmed)	1.00	**0.94**	1.00	**0.71**	1.00	**0.87**
MASE	**0.97**	0.97	0.95	**0.70**	0.96	**0.90**
Mean (MAD/Mean)	**0.37**	0.42	0.33	**0.24**	**0.36**	0.37
Mean (MAD/Mean) (5 % trimmed)	**0.34**	0.35	0.29	**0.21**	0.33	**0.31**
AvgRelMAE	1.00	**0.96**	1.00	**0.71**	1.00	**0.90**
AvgRelMAE (5 % trimmed)	1.00	**0.96**	1.00	**0.73**	1.00	**0.89**
Avg. improvement based on AvgRelMAE	0.00	**0.04**	0.00	**0.29**	0.00	**0.10**

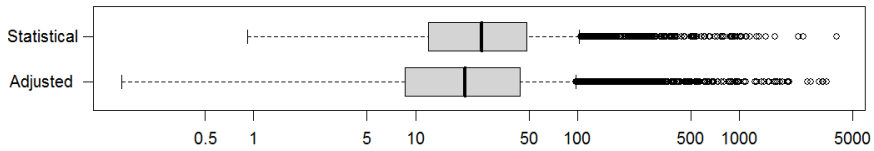

Fig. 6. Box-and-whisker plot for absolute percentage errors (log scale, zero-error forecasts excluded).

Table 3 shows that the rankings based on the trimmed MAPE and MdAPE differ, suggesting different conclusions about the effectiveness of adjustments. As was explained in Section 3.1, the interpretation of PE-based measures is not straightforward. While MdAPE is resistant to outliers, it is not sufficiently in-

Please cite this paper as:
Davydenko, A., & Fildes, R. (2014). Measuring Forecasting Accuracy: Problems and Recommendations (by the Example of SKU-Level Judgmental Adjustments). In *Intelligent Fashion Forecasting Systems: Models and Applications* (pp. 43-70). Springer Berlin Heidelberg.

formative, as it is insensitive to APEs which lie above the median. Also, PE-measures produce a biased comparison, since the improvement on the real scale within each series is correlated markedly with the actual value. Therefore, applying percentage errors in the current setting leads to ambiguous results and to confusion in their interpretation. For example, for positive adjustments, the trimmed MAPE and MdAPE suggest the opposite rankings: while the trimmed MAPE shows a substantial worsening of the final forecast due to the judgmental adjustments, the MdAPE value points in the opposite direction.

The absolute scaled errors found using the MASE scheme (as described in Section 3.4) also follow a non-symmetrical distribution and can take extremely large values (Fig. 7) in short series where the MAE of the naïve forecast is smaller than the error of judgmental forecast. For the adjustments data, the lengths of the series vary substantially, so the MASE is affected seriously by outliers. Fig. 8 shows that the use of the MAD/MEAN scheme instead of the MASE does not improve the properties of the distribution of the scaled errors. Table 3 shows that a trimmed version of the MAD/MEAN scheme gives the opposite rankings with regard to the overall accuracy of adjustments, which indicates that this scheme is highly unstable. Moreover, with such distributions, the use of trimming for either MASE or MAD/MEAN leads to biased estimates, as was the case with MAPE.

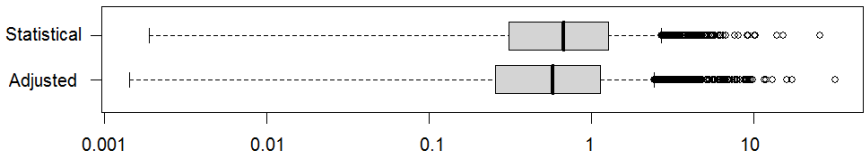

Fig. 7. Box-and-whisker plot for the absolute scaled errors found by the MASE scheme (log scale, zero-error forecasts excluded).

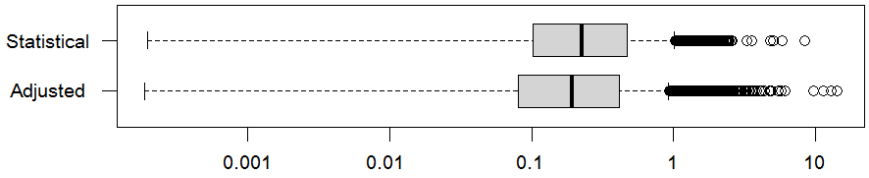

Fig. 8. Box-and-whisker plot for absolute scaled errors found by the MAD/MEAN scheme (log scale, zero-error forecasts excluded).

Please cite this paper as:
Davydenko, A., & Fildes, R. (2014). Measuring Forecasting Accuracy: Problems and Recommendations (by the Example of SKU-Level Judgmental Adjustments). In *Intelligent Fashion Forecasting Systems: Models and Applications* (pp. 43-70). Springer Berlin Heidelberg.

Fig. 9 shows that the log-transformed relative absolute errors follow a symmetric distribution and contain outliers that are easier to detect and to eliminate. Based on the shape of the underlying distribution, it seems that using a 5% trimmed GMRAE would give a location estimate with a reasonable level of efficiency. Although the GMRAE measure is not vulnerable to outliers, its interpretation can present difficulties, for the reasons explained in Section 3.2.

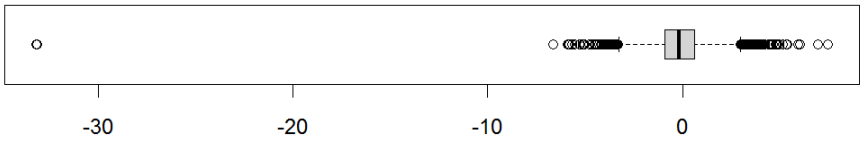

Fig. 9. Box-and-whisker plot for the log-transformed relative absolute errors (using the statistical forecast as the benchmark).

Compared to the APEs and the absolute scaled errors, the log-transformed relative MAEs are not affected severely by outliers and have a more symmetrical distribution (Fig. 10). The AvgRelMAE can therefore serve as a more reliable indicator of changes in accuracy. At the same time, in terms of a linear loss function the AvgRelMAE scheme represents the effectiveness of adjustments adequately and gives a directly interpretable meaning.

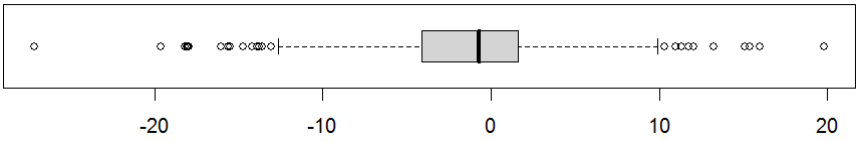

Fig. 10. Box-and-whisker plot for the weighted log-transformed relative MAEs $(n_i \ln r_i)$**.**

The AvgRelMAE result shows improvements from both positive and negative adjustments, whereas according to MAPE and MASE, only negative adjustments improve the accuracy. For the whole sample, adjustments improve the MAE of statistical forecast by 10%, on average. Positive adjustments are less accurate than negative adjustments and provide only minor improvements. To assess the significance of changes in accuracy in terms of MAE, we applied the two-sided Wilcoxon test to test the mean of the weighted relative log-transformed MAEs against zero. The *p*-value was less than 0.01 for the set containing the adjustments of both

Please cite this paper as:

Davydenko, A., & Fildes, R. (2014). Measuring Forecasting Accuracy: Problems and Recommendations (by the Example of SKU-Level Judgmental Adjustments). In *Intelligent Fashion Forecasting Systems: Models and Applications* (pp. 43-70). Springer Berlin Heidelberg.

signs, less than 0.05 for the set containing only positive adjustments, and less than $2.2 \cdot 10^{-16}$ for the set containing only negative adjustments.

To determine whether the probability of a successful adjustment is higher than 0.5, the two-sided binomial test was applied. The results are shown in Table 4.

Table 4: Results of using the binomial test to analyse the frequency of a successful adjustment.

Adjustment sign	**Total number of adjustments**	**Number of adjustments that improved forecast**	***p*-value**	**Probability of a successful adjustment**	**95% confidence interval for the probability of a successful adjustment**	
Positive	3394	1815	<0.001	0.535	0.518	0.552
Negative	1385	915	<0.001	0.661	0.635	0.686
Both	4779	2730	<0.001	0.571	0.557	0.585

Based on the *p*-values obtained for each sample, it can be concluded that adjustments improved the accuracy of forecasts more frequently than they reduced it. However, the probability of a successful intervention was rather low for positive adjustments.

6. Conclusions

The appropriate measurement of forecasting accuracy is important in many organizational settings, and is not of merely academic interest. Where an inappropriate error measure is used the consequences can be the adoption of a poor forecasting process. In addition forecasters can be rewarded or penalized (appropriately or not) for their performance and this is evaluated through the organization's choice of error measure. Due to the specific features of SKU-level demand data, many well-known error measures are not appropriate for use in evaluating the effectiveness of adjustments. This is especially true for fast-moving fashionable products. In particular, the use of percentage errors is not advisable because of the considerable proportion of low actual values, which lead to high percentage errors with no direct interpretation for practical use. Moreover, the errors corresponding to adjustments of different signs are penalised differently when using percentage errors, because the forecasting errors are correlated with both the actual demand values

and the adjustment sign. As a result, measures such as MAPE and MdAPE do not provide sufficient indication of the effectiveness of adjustments, in terms of a linear loss function. Similar arguments were also found to apply to the calculation of MASE, which can also induce biases and outliers as a result of using the arithmetic mean to average relative quantities. Thus, an organization which determines its forecast improvement strategy based on an inadequate measure will misallocate its resources, and will therefore fail in its objective of improving the accuracy at the SKU level.

In order to overcome the disadvantages of existing measures, it is recommended that an average relative MAE be used which is calculated as the geometric mean of relative MAE values. This scheme allows for the objective comparison of forecasts, and is more reliable for the analysis of adjustments.

For the empirical dataset, the analysis has shown that adjustments improved accuracy in terms of the average relative MAE (AvgRelMAE) by approximately 10%. For the same dataset, a range of well-known error measures, including MAPE, MdAPE, GMRAE, MASE, and the MAD/MEAN ratio, indicated conflicting results. The MAPE-based results suggested that, on the whole, adjustments did not improve the accuracy, while the MdAPE results showed a substantial improvement (dropping from 25% to 20%, approximately). The analysis using MASE and the MAD/MEAN ratio was complicated, due to a highly skewed underlying distribution, and did not allow any firm conclusions to be reached. The GMRAE showed that adjustments improved the accuracy by 13%, a result that is close to that one obtained using the AvgRelMAE. Since analyses based on different measures can lead to different conclusions, it is important to have a clear understanding of the statistical properties of any error measure used. We have described various undesirable effects that complicate the interpretation of the well-known error measures. As an improved scheme which is appropriate for evaluating changes in accuracy under linear loss, we recommend using the AvgRelMAE. The generalisation of this scheme can be obtained straightforwardly for other loss functions as well.

One question that arises after the analysis of the accuracy of judgmental adjustments is whether or not these adjustments are systematically biased and can we improve them using some statistical calibration. A number of studies have been now conducted to address these questions (see, e.g., Davydenko and Fildes, 2008; Fildes et al., 2009; Franses and Legerstee, 2010; Trapero, Fildes, and Davydenko, 2011) and it has been found that judgmental adjustments do contain persistent systematic errors. Albeit this topic is outside the scope of the current paper, but, of course, we think that any study of procedures for the correction of judgmental forecasts should contain thorough analysis of accuracy based on appropriate and well-justified error measures.

Please cite this paper as:

Davydenko, A., & Fildes, R. (2014). Measuring Forecasting Accuracy: Problems and Recommendations (by the Example of SKU-Level Judgmental Adjustments). In *Intelligent Fashion Forecasting Systems: Models and Applications* (pp. 43-70). Springer Berlin Heidelberg.

The process by which a new error measure is developed and accepted by an organisation has not received any research attention. A case in point is intermittent demand, where service improvements can be achieved, but only by abandoning the standard error metrics and replacing them with service-level objectives (Syntetos & Boylan, 2005). When an organisation and those to whom the forecasting function reports insist on retaining MAPE or similar (as will mostly be the case), the forecaster's objective must then shift to delivering to the organisation's chosen performance measure, whilst using a more appropriate measure, such as the AvgRelMAE, to interpret what is really going on with the data. In essence, the forecaster cannot reasonably resort to using the organisation's measure and expect to achieve a cost-effective result.

Please cite this paper as:
Davydenko, A., & Fildes, R. (2014). Measuring Forecasting Accuracy: Problems and Recommendations (by the Example of SKU-Level Judgmental Adjustments). In *Intelligent Fashion Forecasting Systems: Models and Applications* (pp. 43-70). Springer Berlin Heidelberg.

Appendix A. Alternative representation of MASE

According to Hyndman and Koehler (2006), for the scenario when forecasts are made from varying origins but with a constant horizon (here taken as 1), the scaled error is defined as[3]

$$q_{i,t} = \frac{e_{i,t}}{\mathrm{MAE}_i^{\mathrm{b}}}, \qquad \mathrm{MAE}_i^{\mathrm{b}} = \frac{1}{l_i - 1}\sum_{j=2}^{l_i} |Y_{i,j} - Y_{i,j-1}|,$$

where $\mathrm{MAE}_i^{\mathrm{b}}$ is the MAE from the benchmark (naïve) method for series i, $e_{i,t}$ is the error of a forecast being evaluated against the benchmark for series i and period t, l_i is the number of elements in series i, and $Y_{i,j}$ is the actual value observed at time j for series i.

Let the mean absolute scaled error (MASE) be calculated by averaging the absolute scaled errors across time periods and time series:

$$\mathrm{MASE} = \frac{1}{\sum_{i=1}^{m} n_i}\sum_{i=1}^{m}\sum_{t\in T_i}\frac{|e_{i,t}|}{\mathrm{MAE}_i^{\mathrm{b}}},$$

where n_i is the number of available values of $e_{i,t}$ for series i, m is the total number of series, and T_i is a set containing time periods for which the errors $e_{i,t}$ are available for series i.

Then,

$$\mathrm{MASE} = \frac{1}{\sum_{i=1}^{m} n_i}\sum_{i=1}^{m}\sum_{t\in T_i}\frac{|e_{i,t}|}{\mathrm{MAE}_i^{\mathrm{b}}}$$

$$= \frac{1}{\sum_{i=1}^{m} n_i}\sum_{i=1}^{m}\frac{\sum_{t\in T_i}|e_{i,t}|}{\mathrm{MAE}_i^{\mathrm{b}}}$$

$$= \frac{1}{\sum_{i=1}^{m} n_i}\sum_{i=1}^{m} n_i\frac{\frac{1}{n_i}\sum_{t\in T_i}|e_{i,t}|}{\mathrm{MAE}_i^{\mathrm{b}}}$$

[3] The formula corresponds to the software implementation described by Hyndman and Khandakar (2008).

Please cite this paper as:
Davydenko, A., & Fildes, R. (2014). Measuring Forecasting Accuracy: Problems and Recommendations (by the Example of SKU-Level Judgmental Adjustments). In *Intelligent Fashion Forecasting Systems: Models and Applications* (pp. 43-70). Springer Berlin Heidelberg.

$$= \frac{1}{\sum_{i=1}^{m} n_i} \sum_{i=1}^{m} n_i r_i, \quad r_i = \frac{\mathrm{MAE}_i}{\mathrm{MAE}_i^{\mathrm{b}}},$$

where MAE_i is the MAE for series i for the forecast being evaluated against the benchmark.

Please cite this paper as:

Davydenko, A., & Fildes, R. (2014). Measuring Forecasting Accuracy: Problems and Recommendations (by the Example of SKU-Level Judgmental Adjustments). In *Intelligent Fashion Forecasting Systems: Models and Applications* (pp. 43-70). Springer Berlin Heidelberg.

References

Alkhazaleh, A. M. H., & Razali, A. M. (2010). New technique to estimate the asymmetric trimming mean. *Journal of Probability and Statistics, vol. 2010*.

Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H. & Tuckey, J. W. (1972). *Robust Estimates of Location*. Princeton University Press, Princeton, NJ

Armstrong, S. (1985). *Long-range forecasting: from crystal ball to computer*. New York: John Wiley.

Armstrong, J. S., & Collopy, F. (1992). Error measures for generalizing about forecasting methods: Empirical comparisons. *International Journal of Forecasting*, *8*, 69-80.

Armstrong, J. S., & Fildes, R. (1995). Correspondence on the selection of error measures for comparisons among forecasting methods. *Journal of Forecasting*, *14*(1), 67-71.

Chatfield, C. (2001) *Time-series Forecasting*. Chapman & Hall.

Davydenko, A., & Fildes, R. (2008, June 22-25). *Models for product demand forecasting with the use of judgmental adjustments to statistical forecasts*. Paper presented at the 28th International Symposium on Forecasting (ISF2008), Nice. Retrieved from http://www.forecasters.org/submissions08/DAVYDENKOANDREYISF2008.pdf

Davydenko, A., & Fildes, R. (2013). Measuring forecasting accuracy: The case of judgmental adjustments to SKU-level demand forecasts. *International Journal of Forecasting*, *29*(3), 510-522.

Diebold, F. X. (1993). On the limitations of comparing mean square forecast errors: Comment. *Journal of Forecasting*, *12*, 641-642.

Fildes, R. (1992). The evaluation of extrapolative forecasting methods. *International Journal of Forecasting*, *8*(1), 81-98.

Fildes, R., & Goodwin, P. (2007). Against your better judgment? How organizations can improve their use of management judgment in forecasting. *Interfaces*, *37*, 570-576.

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.

Fleming, P. J., & Wallace, J. J. (1986). How not to lie with statistics: the correct way to summarize benchmark results. *Communications of the ACM*, *29*(3), 218-221.

Franses, P. H., & Legerstee, R. (2010). Do experts' adjustments on model-based SKU-level forecasts improve forecast quality? *Journal of Forecasting*, *29*, 331-340.

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

Hill, M., & Dixon, W. J. (1982). Robustness in real life: A study of clinical laboratory data. *Biometrics*, *38*, 377-396.

Hoover, J. (2006). Measuring forecast accuracy: Omissions in today's forecasting engines and demand-planning software. *Foresight: The International Journal of Applied Forecasting*, *4*, 32-35.

Hyndman, R. J. (2006). Another look at forecast-accuracy metrics for intermittent demand. *Foresight: The International Journal of Applied Forecasting*, *4*(4), 43-46.

Hyndman, R. J., & Khandakar, Y. (2008). Automatic time series forecasting: the forecast package for R. *Journal of Statistical Software*, *27*(3).

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

Kolassa, S., & Schutz, W. (2007). Advantages of the MAD/MEAN ratio over the MAPE. *Foresight: The International Journal of Applied Forecasting*, *6*, 40-43.

Makridakis, S. (1993). Accuracy measures: theoretical and practical concerns. *International Journal of Forecasting*, *9*, 527–529

Marques, C. R., Neves, P. D., & Sarmento, L. M. (2000). *Evaluating core inflation indicators*. Working Paper 3-00, Economics Research Department, Banco de Portugal.

Mathews, B., & Diamantopoulos, A. (1987). Alternative indicators of forecast revision and improvement. *Marketing Intelligence*, *5*(2), 20-23.

McCarthy, T. M., Davis, D. F., Golicic, S. L., & Mentzer, J. T. (2006). The evolution of sales forecasting management: a 20-year longitudinal study of forecasting practice. *Journal of Forecasting*, *25*, 303-324.

Mudholkar, G. S. (1983). Fisher's *z*-transformation, *Encyclopedia of Statistical Sciences, 3*, 130–135.

Sanders, N., & Ritzman, L. (2004). Integrating judgmental and quantitative forecasts: methodologies for pooling marketing and operations information. *International Journal of Operations and Production Management*, *24*, 514-529.

Spizman, L., & Weinstein, M. (2008). A note on utilizing the geometric mean: when, why and how the forensic economist should employ the geometric mean. *Journal of Legal Economics*, *15*(1), 43-55.

Syntetos, A. A., & Boylan, J. E. (2005). The accuracy of intermittent demand estimates. *International Journal of Forecasting*, *21*(2), 303-314.

Trapero, J.R., Fildes, R.A., & Davydenko A. (2011). Non-linear identification of judgmental forecasts at SKU-level. *Journal of Forecasting, 30*(5), 490-508.

Trapero, J. R., Pedregal, D. J., Fildes, R., & Weller, M. (2011). *Analysis of judgmental adjustments in presence of promotions.* Paper presented at the 31th International Symposium on Forecasting (ISF2011), Prague.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*, 1124-1130.

Wilcox, R. R. (1996). *Statistics for the social sciences*. San Diego, CA: Academic Press.

Wilcox, R. R. (2005). Trimmed means. *Encyclopedia of Statistics in Behavioral Science*, *4*, 2066-2067.

Zellner, A. (1986). A tale of forecasting 1001 series: The Bayesian knight strikes again. *International Journal of Forecasting*, *2*, 491-494.