Lancaster University Management School
Working Paper
2010/026

Measuring the Accuracy of Judgmental Adjustments to SKU-level Demand Forecasts

Andrey Davydenko, Robert Fildes and Juan Trapero Arenas

The Department of Management Science
Lancaster University Management School
Lancaster LA1 4YX
UK

Measuring the Accuracy of Judgmental Adjustments to SKU-level Demand Forecasts

Davydenko A., Fildes R., Trapero J.R.

a.davydenko@lancaster.ac.uk

Lancaster University
Department of Management Science
LA1 4YX, UK

June 4, 2010

Abstract

The paper shows that due to the features of SKU (stock-keeping unit) demand data well-known error measures previously used to analyse the accuracy of adjustments are generally not advisable for the task. In particular, percentage errors are affected by outliers and biases arising from a large number of low actual demand values and correlation between forecast errors and actual outcomes. It is also shown that MASE is equivalent to the arithmetic average of relative mean absolute errors (MAEs) and inherently is biased towards overrating the benchmark method. Therefore existing measures cannot deliver easily interpretable and unambiguous results.

To overcome the imperfections of existing schemes a new measure is introduced which indicates average relative improvement of MAE. In contrast to MASE the proposed scheme is based on finding the geometric average of relative MAEs. This allows objective evaluation of relative change in forecasting accuracy yielded by the use of adjustments. Empirical analysis employed a large number of observations collected from a company specialising on manufacturing of fast-moving consumer goods (FMCG). The results suggest that adjustments reduced MAE of baseline statistical forecast on average by approximately 10%. Using a binomial test it was confirmed that adjustments improved the accuracy of forecasts significantly more frequently rather than they reduced it.

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

1. Introduction

Judgmental adjustments to baseline statistical forecasts are widely used for demand forecasting at a level of SKUs (stock-keeping units) (Sanders and Ritzman 2004, Fildes, et al. 2009). At the same time, empirical evidence suggests that judgments under uncertainty are affected by various types of cognitive biases and inherently are non-optimal (Tversky and Kahneman 1974). Therefore it is important to monitor the performance of adjustments in order to ensure the rational use of resources invested in the forecasting process.

This paper shows that due to the features of SKU demand data well-known error measures are generally not advisable for the evaluation of adjustments and can even give misleading results. In particular, percentage measures cannot be efficiently used because of a large number of extremely high percentage errors arising from a relatively low actual demand values. Moreover, it was found that percentage errors penalise the errors of positive and negative adjustments differently due to the correlation between demand values, forecast errors, and the adjustment sign.

MASE (mean absolute scaled error) measure proposed in (Hyndman and Koehler 2006) to overcome the disadvantages of percentage measures was also found to be unsuitable for the adjustments data. The paper shows that MASE is equivalent to the weighted arithmetic average of relative mean absolute errors (MAEs). One of the disadvantages of this scheme is that it introduces a bias towards overrating the performance of a benchmark forecast. This happens because when using the arithmetic average the reward for improving MAE of benchmark forecast does not compensate the penalty given for reducing benchmark MAE by the same quantity. Another disadvantage of MASE scheme in the given context is that it is influenced by outliers arising as a result of dividing by small benchmark MAE values.

To ensure a more reliable evaluation of the effectiveness of adjustments this paper recommends using an enhanced scheme that shows average relative improvement in MAE. In contrast to MASE it is proposed to use the weighted geometric average to find average relative MAE. By taking the statistical forecast as a benchmark it becomes possible to objectively evaluate the relative change in forecasting accuracy yielded by the use of judgmental adjustments. Therefore the proposed statistic can be used to provide a more robust and easily interpretable indicator of changes in accuracy.

Previously the analysis of the accuracy of adjustments was done in a number of empirical studies (Fildes, et al. 2009, Nikolopoulos 2008, Franses and Legerstee 2010). However, different measures were applied to different datasets and suggested different conclusions. The analysis of adjustments was mainly performed with the use of percentage errors. This paper considers the appropriateness of previously used measures and demonstrates the use of the proposed enhanced accuracy measurement scheme based on a real dataset.

The current research employed data collected from a company specialising on manufacturing of fast-moving consumer goods (FMCG). The data contains observed monthly values of actual SKU-level demand, corresponding one-step-ahead statistical

forecasts, and judgmentally adjusted forecasts relating to 413 SKUs. In total, 7544 cases of forecasts and corresponding actual outcomes pertaining to a period of three years have been analysed.

2. Appropriateness of Existing Measures

2.1. Percentage errors

A traditional way to compare the accuracy of forecasts across multiple time series is based on using absolute percentage errors (Hyndman and Koehler 2006).

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, $F_{i,t}$ is the forecast of $Y_{i,t}$.

The percentage error (PE) is calculated as

$$p_{i,t} = 100 \times e_{i,t}/Y_{i,t}.$$

The most popular PE-based measures are MAPE and MdAPE which are defined as follows:

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

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

These measures served as a main tool for the analysis of judgmental adjustments in some recent empirical studies (Fildes, et al. 2009, Nikolopoulos 2008). In order to determine a change in forecasting accuracy MAPE and MdAPE values were calculated and compared for statistical baseline forecasts and for final judgmentally adjusted forecasts. The significance in the change of accuracy was assessed based on the distribution of the differences between absolute percentage errors (APEs) of forecasts. The difference between APEs is defined as

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

where $\left|p_{i,t}^{\mathrm{f}}\right|$ and $\left|p_{i,t}^{\mathrm{s}}\right|$ denote APEs for the same SKU i and same period t for final and baseline statistical forecasts respectively. In (Nikolopoulos 2008) a paired *t*-test was used to detect if the mean of $d_{i,t}^{\mathrm{APE}}$ was significantly different from zero, while in (Fildes, et al. 2009) it was suggested testing whether the median of $d_{i,t}^{\mathrm{APE}}$ significantly differs from zero using two-sample paired (Wilcoxon) sign rank test.

It can be shown that the sample mean of $d_{i,t}^{\mathrm{APE}}$ is the difference between MAPE values corresponding to statistical and final forecasts:

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

Therefore testing the mean or median of $d_{i,t}^{\mathrm{APE}}$ against zero using the above-mentioned tests means finding out if $\mathrm{MAPE}^{\mathrm{f}}$ significantly differs from $\mathrm{MAPE}^{\mathrm{s}}$.

The reported results suggest that overall values of MAPE and MdAPE were improved by the use of adjustments, but the accuracy of positive and negative adjustments differed substantially. Based on MAPE measure it was found that positive adjustments did not significantly change forecasting accuracy, while negative adjustments lead to significant improvements.

However, the current research has shown that percentage errors have a number of disadvantages when applying to the adjustments data.

One well-known disadvantage of percentage errors is that when the actual value $Y_{i,t}$ in the denominator is relatively small compared to forecast error $e_{i,t}$ the resulting percentage error $p_{i,t}$ becomes extremely large, which distorts the results of further analysis (Hyndman and Koehler 2006). Such high values can be treated as outliers since they do not allow for a meaningful interpretation. However, identifying outliers in a skewed distribution is a non-trivial problem where it is needed to determine an appropriate trimming level in order to achieve adequacy of distribution characteristics, while at the same time not to lose too much information. Usually authors choose the trimming level according to their intuition and experience as 1% or 2% (Fildes, et al. 2009, Nikolopoulos 2008), but this decision still remains highly subjective.

At the same time, SKU-level demand time series typically exhibit a high degree of variation among actual values due to seasonal effects and different stages of a product life cycle. Therefore adjustments data can contain a high proportion of low demand values, which makes PE-based measures inadvisable. In addition, all cases with zero actual values should be excluded from analysis since the percentage error cannot be computed when $Y_{i,t} = 0$ due to its definition.

Obtaining extreme percentage errors can be illustrated using scaled values of errors and actual demand values (Fig. 1). The variables shown were scaled by the standard deviation of actual values in each series in order to eliminate differences between time series. It can be seen that final forecast errors have a truncated and skewed distribution, correlate both with actual values and the sign of adjustments, and a substantial proportion of errors is comparable to 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 in (Fildes, et al. 2009)) still cannot sufficiently improve the properties of percentage errors since a large number of observations remains in the area where the actual demand value is less than the absolute error. This results in extremely high (>100%) and hardly interpretable percentage errors as well as high variance of the difference in absolute percentage errors $d_{i,t}^{\mathrm{APE}}$.

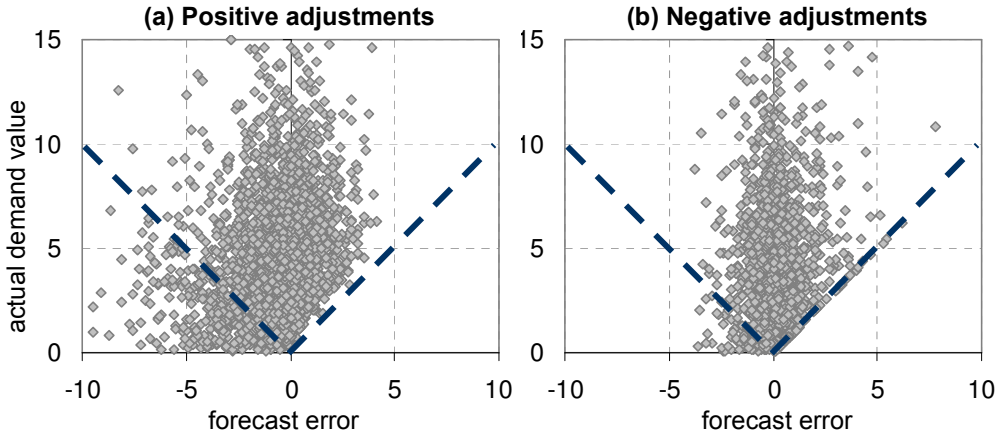

Figure 1. Dependencies between forecast error, actual value, and the sign of adjustment (based on scaled data). Absolute errors in the area below the dashed line are higher than corresponding actual demand values and therefore result in extreme percentage errors (Fig. 2).

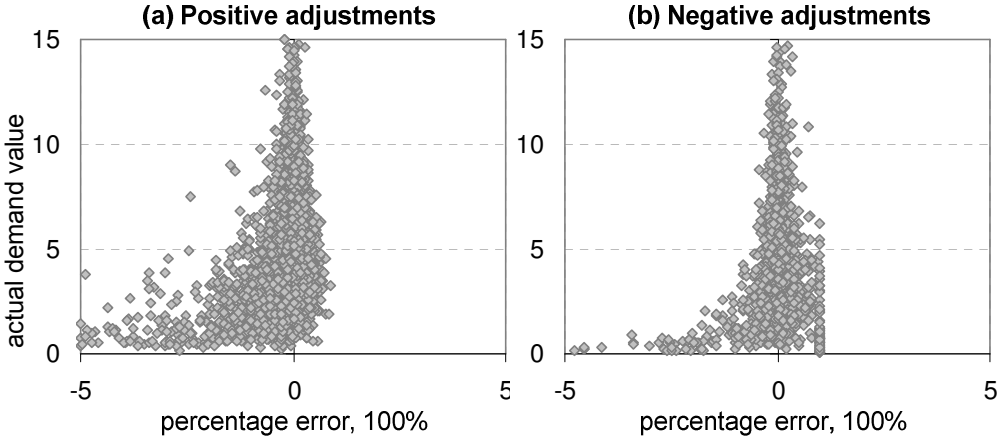

Figure 2. Percentage errors depending on actual demand value and adjustment sign.

Fig. 1(a,b) illustrates that errors arising from positive adjustments are on average negative and correspond to low actual values of demand, while negative adjustments on average lead to positive errors and relate to higher actuals. Transition to percentage measures magnifies the errors of positively adjusted forecasts (Fig. 1(a), Fig. 2(a)) due to low values of actual demand. The opposite transformation happens with the errors of negative adjustments (Fig. 1(b), Fig. 2(b)). Moreover, the distribution of $d_{i,t}^{\mathrm{APE}}$ for positive adjustments becomes highly diffuse, which does not allow a proper estimation of its characteristics.

One of the important effects arising from the presence of cognitive biases and non-negative nature of demand values is that most damaging positive adjustments typically correspond to low actuals, while worst negative adjustments correspond to high actuals. More specifically, the following general dependency can be found within most time series.

The difference between absolute final forecast error $\left|e_{i,t}^{\mathrm{f}}\right|$ and absolute statistical forecast error $\left|e_{i,t}^{\mathrm{s}}\right|$ is positively correlated with actual value $Y_{i,t}$ for positive adjustments, while for negative adjustments there is a negative correlation. To reveal this effect distribution-free measures of association between variables were used. For each time series 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 actual value $Y_{i,t}$. Fig. 3 shows the distribution of coefficients ρ_i^+ calculated for positive adjustments and ρ_i^- that correspond to negative adjustments. For the given dataset $\mathrm{mean}(\rho_i^+) \approx 0.47$ and $\mathrm{mean}(\rho_i^-) \approx -0.44$, which indicates that the improvement in forecasting markedly correlates with actual demand values and this relationship is inversely different for positive and negative adjustments.

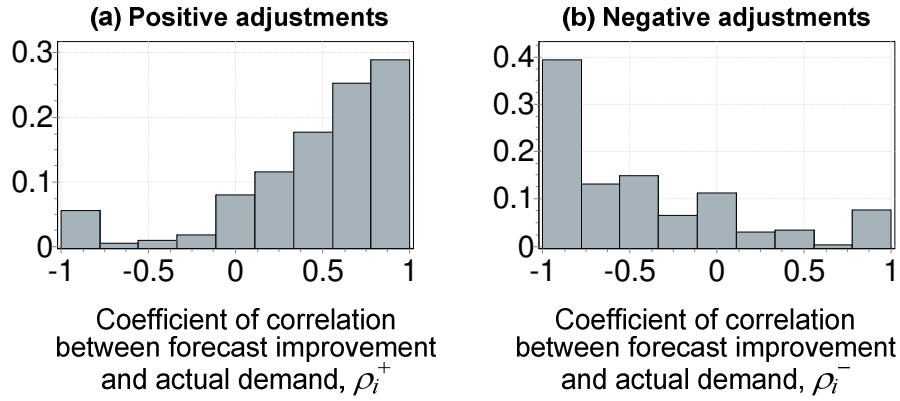

Figure 3. Spearman's ρ coefficients showing correlation between the improvement in accuracy and actual demand value.

This means that the improvement in terms of percentage errors $d_{i,t}^{\mathrm{APE}}$ (which can as well be expressed as $d_{i,t}^{\mathrm{APE}} = 100 \times \left(\left|e_{i,t}^{\mathrm{f}}\right| - \left|e_{i,t}^{\mathrm{s}}\right|\right)/Y_{i,t}$) will underrate the accuracy of positive adjustments as a result of dividing the difference of absolute errors by higher actuals. In the same way it will overrate the accuracy of negative adjustments. Since the difference in MAPEs is calculated as a mean improvement in terms of percentage errors (in accordance with formula (1)), the comparison of forecasts using MAPE will also give a biased result towards overrating positive adjustments and underrating negative adjustments. Consequently, since the forecast errors arising from adjustments of different signs are penalised differently, the MAPE measure is not sufficiently appropriate for the comparison of the performance of adjustments of different signs. One of the aims of the present research therefore has been to reconsider the results of previous studies with the use of alternative measures.

Another measure based on percentage errors was used in (Franses and Legerstee 2010). In order to evaluate the accuracy of improvements RMSPE (root mean square percentage error) was calculated for both statistical and judgmentally adjusted forecast then compared. Based on this measure it was concluded that expert forecasts were not better than the model forecasts. However, RMSPE is also based on percentage errors and is even more affected by the outliers and biases described above.

2.2. MASE (Mean Absolute Scaled Error)

In order to overcome the imperfections of PE-based measures it was proposed in (Hyndman and Koehler 2006) to use MASE (mean absolute scaled error). The MASE is found as follows (see Appendix 1):

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

where $\mathrm{MAE}_i^{\mathrm{b}}$ – mean absolute error (MAE) of the naïve (benchmark) forecast for series i.

The naïve method was chosen as a benchmark to ensure a sufficient number of forecasts for finding a non-zero and stable denominator. In the current case the number of available statistical forecasts is larger than the number of in-sample naïve forecasts. Therefore scaling can be done more efficiently using the MAE of statistical forecast:

$$q_{i,t} = \frac{e_{i,t}^{\mathrm{f}}}{\mathrm{MAE}_i^{\mathrm{s}}}, \qquad \mathrm{MAE}_i^{\mathrm{s}} = \frac{1}{n_i}\sum\nolimits_{j \in T_i} |e_{i,j}^{\mathrm{s}}|\,,$$

where $e_{i,t}^{\mathrm{f}}$ – error of judgmentally adjusted forecast for series i and period t, $e_{i,j}^{\mathrm{s}}$ – error of baseline statistical forecast for series i and period j, T_i – a set containing all time indexes for which the values of $e_{i,t}^{\mathrm{f}}$ for series i are known, n_i – the number of elements in T_i.

Though it was not specified in (Hyndman and Koehler 2006), it is possible to show (Appendix 1) that MASE is equivalent to the weighted arithmetic mean of relative MAEs:

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

where m – total number of series, $\mathrm{MAE}_i^{\mathrm{b}}$ – MAE for a benchmark forecast for series i, MAE_i – MAE for the forecast being evaluated against the benchmark, n_i – the number of errors used to calculate MAE_i.

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 and Weinstein 2008). As a result of using the arithmetic mean of MAE ratios formula (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 MASE rewards and penalises forecasts it can be represented as

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

The reward for improving benchmark MAE from A to B $(A > B)$ in a series i is found as $R_i = n_i(1 - B/A)$, while the penalty for reducing benchmark MAE from B to A

in the same series is $P_i = n_i(B/A - 1)$. Since $R_i < P_i$ the reward given for improving benchmark MAE cannot compensate the penalty given for reducing benchmark MAE by the same quantity. As a result, for some datasets it can be the case that $\text{MASE} > 1$ regardless of the choice of the benchmark method, which cannot ensure unambiguity of the comparison of the accuracy of forecasts.

For example, suppose a comparison of accuracy of two forecasting methods is performed across two series ($m = 2$). For the first series the MAE ratio is $r_1 = 1/2$ and for the second series the MAE ratio is the opposite: $r_2 = 2/1$. Averaging the ratios gives $\text{MASE} = ½\ (r_1 + r_2) = 1.25$, which indicates that two methods have different accuracy. Moreover, the benchmark method is superior regardless of which method is chosen as a benchmark.

With regard to the available data the bias introduced by MASE was found to be substantial, especially in case of short series and large differences in accuracies. In addition, using MASE (in the same way as MAPE) results in unstable estimates as the arithmetic mean is severely influenced by extreme cases arising from dividing by relatively small values. In this case outliers occur when dividing by relatively small MAEs of benchmark forecast which can appear in short series.

The next section presents an improved statistic which is more suitable for comparing accuracy of SKU-level forecasts.

3. Recommended Accuracy Evaluation Scheme

By changing the arithmetic mean to the geometric mean in formula (2) it is possible to define an unbiased measure of average relative MAE (ARMAE):

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

where MAE_i^{s} is MAE for baseline statistical forecast for series i, MAE_i^{f} is MAE for judgmentally adjusted final forecast, other variables have their previous meaning.

This measure is immediately interpretable as it adequately represents the average relative value of MAE and directly shows how adjustments improve/reduce MAE compared to baseline statistical forecast. Obtaining $\text{ARMAE} < 1$ means that on average $\text{MAE}_i^{\text{f}} < \text{MAE}_i^{\text{s}}$ and adjustments improve accuracy, while $\text{ARMAE} > 1$ indicates the opposite. The average percentage improvement in MAE of forecasts is found as $(1 - \text{ARMAE}) \times 100$.

It also can be used to answer the question 'if for a given series MAE of statistical forecast is x, what will be the MAE of final forecast for the same series?' by computing $x \times \text{ARMAE}$. Therefore based on ARMAE it is possible to find a rough estimate of the magnitude of final forecast error on a real scale.

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

$$\text{ARMAE} = \exp\left[\frac{1}{\sum_{i=1}^{m} n_i} \sum_{i=1}^{m} n_i \ln r_i\right].$$

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

Since no scaling by actual values is required, this scheme can applied 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 (3) should be greater than zero for all series.

Thus, the advantages of the recommended accuracy evaluation scheme are that it i) can be easily interpreted, ii) objectively represents the performance of the adjustments (without introduction of additional biases or outliers), iii) is informative and efficiently uses all available information, and iv) is applicable in a wide range of settings with minimal assumptions about the features of the data.

4. Results of Empirical Evaluation

The results of applying the described above measures are shown in Table 1.

Table 1: Accuracy of adjustments according to different error measures

	Positive adjustments		**Negative adjustments**		**All nonzero adjustments**	
	Statistical forecast	*Adjusted forecast*	*Statistical forecast*	*Adjusted forecast*	*Statistical forecast*	*Adjusted forecast*
MAPE, % (2% trim)	**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**
MASE	**1.00**	1.12	1.00	**0.86**	**1.00**	1.02
ARMAE	1.00	**0.96**	1.00	**0.71**	1.00	**0.90**
Avg. improve-ment in MAE (1-ARMAE)		**0.04**		**0.29**		**0.10**

For the given dataset a large number of percentage errors have extreme values (>100%) arising from low actual demand values (Fig. 4). Though 2% trimmed MAPE values were used, it is difficult to determine the trim level since there is no indication of what proportion of data will represent the percentage error adequately. As a result, the difference in APEs has a very high dispersion and cannot be used efficiently to assess the improvements in accuracy. It can also be seen that the distribution of APEs is highly skewed.

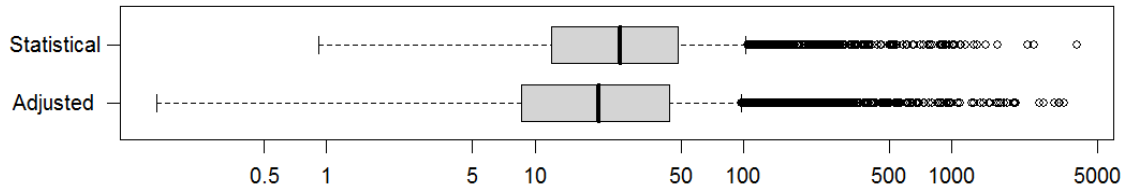

Figure 4. Box-and-whisker plot for absolute percentage errors (log scale). For the given dataset the range and the variation of percentage errors is extremely high, which makes them difficult to interpret and to analyse.

Table 1 shows that MdAPE and MAPE values are different and suggest different conclusions about the effectiveness of adjustments. While MdAPE is resistant to outliers, it is not sufficiently informative as it is insensitive to APEs lying above the median. In addition, the improvement in terms of percentage errors is biased since the improvement on a real scale within each series markedly correlates with the actual value (as was described in Subsection 2.1). Therefore applying percentage errors in the current settings leads to ambiguous results and brings confusion in their interpretation.

Scaled errors found according to the MASE scheme (calculated as described in Subsection 2.2) are also affected by extreme values and have a non-symmetrical distribution (Fig. 5). Outliers commonly occur in short series where MAE of statistical forecast is smaller than the error of judgmental forecast. For adjustments data the lengths of series vary substantially, which makes MASE seriously affected by outliers.

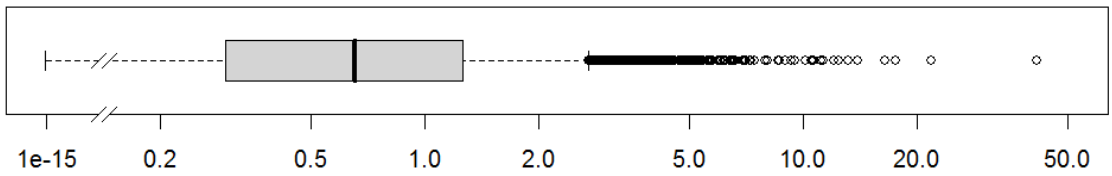

Figure 5. Box-and-whisker plot for scaled errors found according to the MASE scheme (log scale). Extreme cases arise due to dividing by small values of statistical forecast MAE.

Average relative MAE represents the effectiveness of adjustments more adequately. This measure gives a directly interpretable meaning and is not affected by extreme cases arising when using the alternative schemes. The sample mean of the log-transformed ratios is not severely influenced by outliers and can by reliably estimated based on the sample data (Fig. 6). Therefore ARMAE can serve as a robust indicator of changes in accuracy.

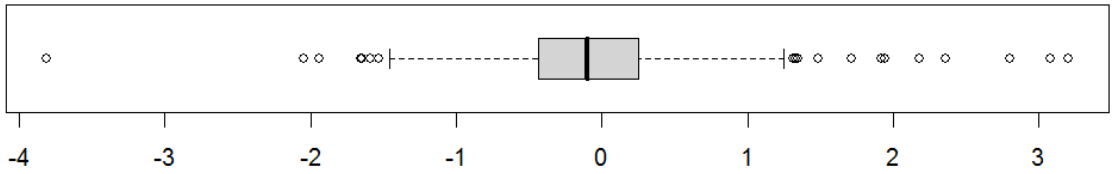

Figure 6. Box-and-whisker plot for logarithms of MAE ratios.

While ARMAE shows improvements that correspond to both positive and negative adjustments, according to MAPE and MASE only negative adjustments improve accuracy. ARMAE value indicates that adjustments improve MAE of statistical forecast on average by 10% . Positive adjustments are less accurate than negative adjustments and bring only minor improvements.

To determine if the probability of a successful adjustment is higher than 0.5 a two-sided binomial test was applied. The results are shown in Table 2.

Table 2: Results of a binomial test

Adjust-ment sign	Total number of adjust-ments	Number of adjustments that improved forecast	p-value	Probability of a successful adjustment	95% confidence interval for the probability of a successful adjustment	
Positive	3161	1662	0.004	0.526	0.508	0.543
Negative	1504	1034	0.000	0.688	0.663	0.711
Both	4665	2696	0.000	0.578	0.564	0.592

According to the obtained p-values in all cases it can be concluded that adjustments improved accuracy of forecasts more frequently rather than reduced it. However, for positive adjustments the probability of success was rather low.

5. Conclusions

Due to the features of SKU-level demand data many well-known error measures cannot be efficiently used to evaluate the effectiveness of adjustments. In particular, the use of percentage errors is not advisable because of a considerable proportion of low actual values which lead to high percentage errors with no direct interpretation for practical use. Moreover, errors corresponding to adjustments of different signs are penalised differently when using percentage errors because forecasting errors correlate both with actual demand values and with the adjustment sign. As a result measures such as MAPE or MdAPE do not give enough information to justify drawing conclusions about the improvements yielded by the use of adjustments. At the same, time it was found that MASE can also induce biases and outliers as a result of using the arithmetic mean to average relative quantities.

In order to overcome the disadvantages of existing measures it is recommended to use average relative MAE which is calculated as the geometric mean of relative MAE values. This scheme allows for objective comparison of forecasts and is more reliable for the analysis of adjustments. For the given dataset the analysis has shown that adjustments improved average relative MAE by approximately 10%.

Appendix A

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

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

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

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

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

where n_i – number of errors $e_{i,t}$ in series i used to calculate MASE, m – total number of time series.

Then

$$\begin{aligned}
\mathrm{MASE} &= \frac{1}{\sum_{i=1}^{m} n_i}\sum_{i=1}^{m}\sum_{t=1}^{n_i}\frac{|e_{i,t}|}{\mathrm{MAE}_i^{\mathrm{b}}} \\
&= \frac{1}{\sum_{i=1}^{m} n_i}\sum_{i=1}^{m}\frac{\sum_{t=1}^{n_i}|e_{i,t}|}{\mathrm{MAE}_i^{\mathrm{b}}} \\
&= \frac{1}{\sum_{i=1}^{m} n_i}\sum_{i=1}^{m} n_i\frac{\frac{1}{n_i}\sum_{t=1}^{n_i}|e_{i,t}|}{\mathrm{MAE}_i^{\mathrm{b}}} \\
&= \frac{1}{\sum_{i=1}^{m} n_i}\sum_{i=1}^{m} n_i r_i, \quad r_i = \frac{\mathrm{MAE}_i}{\mathrm{MAE}_i^{\mathrm{b}}},
\end{aligned}$$

where MAE_i – MAE for the forecast being evaluated against the benchmark, n_i – number of errors used to calculate MAE_i.

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

References

Fildes, R., P. Goodwin, M. Lawrence, and K. Nikolopoulos. "Effective forecasting and judgmental adjustments: an empirical evaluation and strategies for improvement in supply-chain planning." *International Journal of Forecasting* 25, no. 1 (2009): 3-23.

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

Hyndman, R.J., and Y. Khandakar. "Automatic Time Series Forecasting: The forecast Package for R." *Journal of Statistical Software* (American Statistical Association) 27, no. 03 (July 2008).

Hyndman, R.J., and A.B. Koehler. "Another look at measures of forecast accuracy." *International Journal of Forecasting* 22, no. 4 (2006): 679-688.

Nikolopoulos, K. *On the accuracy of judgmental interventions on Statistical Forecasts.* Working Paper 0021, University of Peloponnese, Department of Economics, 2008.

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

Spizman, L., and M.A. Weinstein. "A Note on Utilizing the Geometric Mean: When, Why and How the Forensic Economist Should Employ the Geometric Mean." *Journal of Legal Economics* 15, no. 1 (2008): 43-55.

Tversky, A., and D. Kahneman. "Judgment under uncertainty: Heuristics and biases." *Science*, 1974: 1124-1131.