ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/284947381

Forecast Error Measures: Critical Review and Practical Recommendations

Chapter · January 2016

DOI: 10.13140/RG.2.1.4539.5281

CITATIONS
5

READS
2,884

2 authors:

Andrey Davydenko
JSC "CSBI"
6 PUBLICATIONS **73** CITATIONS
SEE PROFILE

Robert Fildes
Lancaster University
156 PUBLICATIONS **4,999** CITATIONS
SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Evaluation and development of forecasting approaches applicable to Internet-based telecommunications View project

Building forecasting models of judgment and observed data View project

All content following this page was uploaded by Andrey Davydenko on 29 November 2015.

The user has requested enhancement of the downloaded file.

Please cite this paper as:
Davydenko, A., & Fildes, R. (2016). Forecast Error Measures: Critical Review and Practical Recommendations. In *Business Forecasting: Practical Problems and Solutions*. John Wiley & Sons Inc.

Forecast Error Measures: Critical Review and Practical Recommendations

By Andrey Davydenko and Robert Fildes

1. Introduction

The choice of a measure to assess the accuracy of forecasts across time series is of wide practical importance, since the forecasting function is often evaluated using inappropriate measures distorting the link to economic performance (Armstrong & Fildes, 1995). Despite the continuing interest in the topic, the choice of the most suitable measure still remains controversial. Due to their statistical properties, popular measures do not always ensure easily interpretable results when applied in practice (Hyndman & Koehler, 2006). Surveys show that the proportion of firms tracking the aggregated accuracy is surprisingly small (55% as reported by McCarthy et al., 2006). One apparent reason for this is the inability to agree on appropriate accuracy metrics (Hoover, 2006).

We look at the behaviours of commonly used measures when measuring accuracy across many series (e.g., when dealing with SKU-level data). After identifying most desirable properties of an error measure (including robustness and ease of interpretation), we illustrate that traditional measures are prone to obtaining confusing and even misleading results. Some measures (such as MAPE) are extremely vulnerable to outliers. Limitations of popular error measures have been widely discussed (e.g., see Hyndman & Koehler, 2006). Here we systemize well-known problems and identify a number of additional important limitations of existing measures that have not yet been given enough attention.

To overcome the shortcomings of percentage errors, Hyndman & Koehler (2006) suggest scaling forecasting errors by MAE (mean absolute error) from naïve forecast. This measure is known as MASE (mean absolute scaled error). We show that MASE (i) introduces a bias towards overrating the performance of the benchmark as a result of arithmetic averaging and (ii) is vulnerable to outliers as a result of dividing by small benchmark MAEs.

We propose an enhanced measure that shows an average relative improvement under linear loss. In contrast to MASE, our measure averages relative MAEs using the weighted geometric mean.

Our empirical analysis uses SKU-level data containing statistical forecasts and corresponding judgmental adjustments. We look at the task of measuring the accuracy of such adjustments. Some studies of accuracy of judgmental adjustments have produced conflicting results (e.g., Fildes et al.,

2009; Franses & Legertee, 2010). Different measures were applied to different data and arrived at different conclusions. Several studies reported an interesting picture where adjustments improved MdAPE, while harming MAPE (Fildes et al., 2009; Trapero et al., 2011). These confusing results require better understanding about what lies behind different error measures. We discuss the appropriateness of various measures used and demonstrate the use of the measure we recommend.

Next section describes the data employed for empirical illustrations. Section 3 illustrates the limitations of well-known measures. Section 4 introduces the enhanced measure. Section 5 contains the results of applying different measures. The concluding section summarises our findings and offers recommendations as to which of the different measures can be employed safely.

2. Data

We employ data from a fast-moving consumer goods (FMCG) manufacturer. For each SKU and each month we have:

(i) one-step-ahead forecast computed automatically by a software system (system forecast);

(ii) corresponding judgmentally adjusted forecast obtained from experts after their revision of the statistical forecast (Fildes et al., 2009) (final forecast); and

(iii) actual sales (actuals).

In total, our data contains 412 series and 6882 observations. The data is representative for companies dealing with many series of different lengths relating to different SKUs. The frequency of zero demand and zero error observations for our data was not high. However, our further discussion will also consider situations when small counts and zeroes occur frequently, as is common with intermittent demand.

3. Critical Review of Existing Measures

3.1. Desirable Properties

What are the properties of an ideal error measure? There have been different attempts in literature to identify the most important properties by which an adequacy of an error measure should be judged. In particular, (Fildes, 1992) justifies the properties of interpretability and sensitivity to outliers (robustness). Some authors (e.g., Zellner, 1986) argue that the criterion by which we evaluate forecasts should correspond to the criterion by which we optimise forecasts. In other words, if we optimise

estimates using some given loss function, we must use the same loss function for empirical evaluation in order to find out which model is better.

Fitting a statistical model usually delivers forecasts optimal under quadratic loss. This, e.g., happens when we fit a linear regression. If our density forecast from statistical modelling is symmetric, then forecasts optimal under quadratic loss are also optimal under linear loss. But, if we stabilise the variance by log-transformations and then transform back forecasts by exponentiation, we get forecasts optimal only under linear loss. If we use another loss, we must first obtain the density forecast using a statistical model, and then adjust our estimate given our specific loss function (see examples of doing this in Goodwin, 2000).

Let's assume we want to empirically compare two methods and find out which method is better in terms of a symmetric linear loss (since this type of loss is commonly used in modelling). If we have only one time series, it seems natural to use a mean absolute error (MAE). Also, MAE is attractive as it is simple to understand and calculate (Hyndman, 2006). Potentially, MAE has the following limitation: absolute errors follow a highly skewed distribution with a heavy right tail, which means that MAE is not robust (in other words, it is a highly inefficient estimate). But there is a more important problem: when comparing accuracy across series, MAE becomes unsuitable as it is not scale-independent.

In this paper we address the question of how to adequately represent forecasting performance under symmetric linear loss when measuring accuracy across series. We aim for the following properties: i) interpretability, ii) robustness, iii) applicability in a wide range of settings (e.g. allows zero errors or forecasts/actuals, etc.), iv) informativeness, v) the use of the same loss function that was used for optimisation, and vi) scale-independence.

3.2. Percentage Errors

Although MAPE is very popular, it has many problems.

Problem 1: zero and negative actuals cannot be used. MAPE is therefore unsuitable for intermittent demand data.

Problem 2: extreme percentages. The sample mean of APEs due to the skewed and diffuse distribution gives a highly inefficient estimate and is severely affected by extreme cases. The distribution of APEs for our data is illustrated by Figure 1. APEs are often larger than 100%. Such extremes do not allow for a meaningful interpretation since corresponding forecasting errors are not necessarily very harmful or damaging. Large percentages often arise merely due to the relatively low actual values. Due to the large influence of outliers the sample mean in a highly skewed distribution becomes inefficient. In other words, for highly skewed distributions it can take a very big sample size before the most likely value of the sample mean approaches the true population mean (Fleming, 2008).

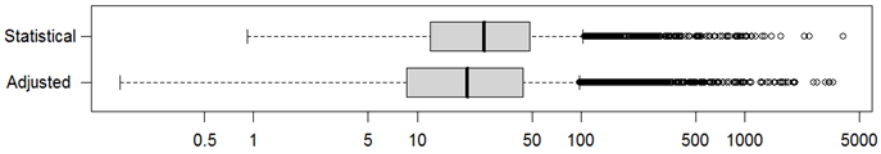

Figure 1. Box-plot for APEs (log scale)

Problem 3: non-symmetric loss. Percentage errors put a heavier penalty on positive errors than on negative errors when the forecast is taken as fixed. This leads to a serious bias when trying to average APEs using the arithmetic mean. Kolassa and Schutz (2007) provide the following example. Assume that we have a series containing values distributed uniformly between 10 and 50. If we are using a symmetrical loss, the best forecast would be 30. However, a forecast of 22 produces better MAPE. Thus, MAPE is not indicative of accuracy in terms of a symmetric loss even for a single series.

Problem 4: misleading when errors correlate with actuals. The comparison of forecasting performance based on percentage errors can give misleading results when the improvement in accuracy correlates with actual value on the original scale (Davydenko and Fildes, 2013).

Various improvements have been proposed in literature, but none of them solves the problems:

Trimmed/Winsorized MAPE	This approach aims to improve robustness, but introduces another problem. Since the distribution of APEs is non-symmetric, the use of trimmed or Winsorized means makes the resulting estimates biased. Moreover, this does not solve problems 1, 3, and 4.
Symmetric MAPE, SMAPE	As shown in literature, SMAPE does not solve problem 2 at all (Goodwin & Lawton, 1999). In fact, this approach does not solve any of the above problems whatsoever. The only correct approach to average ratios is through the use of logarithms (Fleming & Wallace, 1986), but see the next comments.
Geometric mean APE, GMAPE	This is equivalent to the mean of log-transformed APEs as suggested by (Swanson et al., 2000). This solves problems 2 and 3, but does not solve problems 1 and 4. Also, zero errors are not allowed.
MdAPE	MdAPE is not easily interpretable and is not sufficiently indicative of changes in accuracy when forecasting methods have different shapes of error distributions. The sample median of APEs is resistant to the influence of extreme cases, but at the same time it is insensitive to large errors even if they are not outliers or extreme percentages. Also, difficult to assess statistical significance of difference in accuracy. Problem 4 still persists.

3.3. Relative Errors (REs)

Well-known RE-measures include Mean Relative Absolute Error (MRAE), Median Relative Absolute Error (MdRAE), and Geometric Mean Relative Absolute Error (GMRAE).

When averaging benchmark ratios the geometric mean has the advantage over the arithmetic mean (Fleming & Wallace, 1986). The geometric mean produces rankings that are invariant to the choice of the benchmark. Suppose method A is compared with method B. Let method A is used as the benchmark and the arithmetic mean of absolute REs indicates that method B is superior. Then, if method B is used as a benchmark instead of method A, the arithmetic mean can indicate that now method A is superior. Such results are ambiguous and can lead to confusion in their interpretation. Of the measures based on REs, GMRAE is the only measure that has the property of not changing the ranking depending on what method is used as the benchmark. But GMRAE has its limitations.

Problem 1. Zero errors are not allowed. When using intermittent demand data, the use of relative errors becomes impossible due to the frequent occurrences of zero errors (Hyndman, 2006).

Problem 2. GMRAE generally does not reflect changes in accuracy under linear or quadratic loss. For instance, for a particular time series GMAE can compare methods in favour of a method producing errors with a heavier tailed-distribution, while for the same series MAE or MSE can suggest the opposite ranking.

Consider 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 (methods have the same performance under quadratic loss): $\mathrm{Var}\left(e_t^{\mathrm{A}}\right) = \mathrm{Var}\left(e_t^{\mathrm{B}}\right) = 3$. However, GMRAE shows method A being better than method B: $\mathrm{GMRAE} \approx 0.69$.

Thus, even for a single series, a statistically significant improvement of GMRAE is not equivalent to a statistically significant improvement under quadratic or linear loss.

3.4 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 actual than method B. This measure, known as 'Percent Better' (PB), was recommended by some authors as a fairly good indicator (e.g., Chatfield, 2001). It has the advantage of being immune to outliers and scale-independent. Although PB seems to be easy to interpret, the following important limitations should be taken into account.

Problem 1: PB does not show the magnitude of changes in accuracy (Hyndman & Koehler, 2006). Thus, it becomes hard to assess the consequences of using one method instead of another.

Problem 2: PB does not reflect changes under linear loss. 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 under linear loss even for a single series.

Problem 3: many equal forecasts lead to confusing results. When methods A and B frequently produce equal forecasts (this often happens with intermittent demand data), obtaining PB<50% is not necessarily a bad result. But, without additional information, we cannot draw any conclusions about the changes in accuracy.

3.5. Scaled Errors

When forecasts are produced from varying origins but with a constant horizon, the MASE is calculated as

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

where $e_{i,t}$ is forecasting error for period t for time series i, $q_{i,t}$ is the scaled error, and $\mathrm{MAE}_i^{\mathrm{b}}$ is the in-sample MAE of naïve forecast for series i.

It is possible to show that, in this scenario, 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:

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

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.

Problem 1: Bias towards overrating the benchmark.

As noted previously, the arithmetic mean is not appropriate for averaging observations representing relative quantities. In such situations the geometric mean should be used instead. As a result of using the arithmetic mean of r_i, equation (1) 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.

For example, suppose that the performance of some forecasting method is compared with the performance of the naïve method across two series ($m = 2$) 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 $\text{MASE} = ½\ (r_1 + r_2) = 1.25$, 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.

Problem 2: Skewed, heavy-tailed, left-bounded distribution.

In addition to the above effect, the use of MASE (like MAPE) may result in unstable estimates, as the arithmetic mean is severely influenced by extreme cases arising from dividing by relatively small values (see Figure 2). In case of MASE outliers occur when dividing by relatively small benchmark MAEs. Such MAEs are likely to appear in short series. At the same time, attempts to trim or Winsorize MASE lead to biased results.

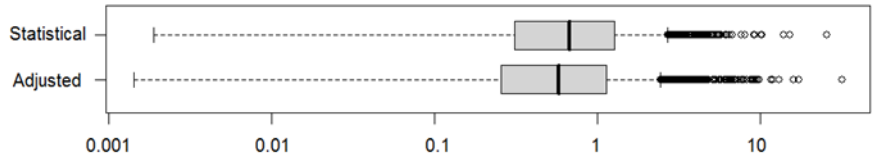

Figure 2. Box-plot for absolute scaled errors (log scale)

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.

3.6. MAD/MEAN Ratio

In contrast to the MASE, the MAD/MEAN ratio assumes that the forecasting errors are scaled by the mean of series actuals instead of by the in-sample MAE of naïve forecast. This reduces the risk of dividing by a small denominator (see Kolassa & Schutz, 2007), however:

Problem 1: MAD/MEAN assumes stable mean. Hyndman (2006) notes that the MAD/MEAN ratio assumes the series mean is stable over time, which may make it unreliable when the data exhibit trends or seasonal patterns.

Problem 2: Outliers. Figure 3 shows that the MAD/MEAN scheme is prone to outliers for the dataset we consider in this paper. Again, attempts to trim/Winsorize lead to biases. Generally, MAD/MEAN

ratio introduces the risk of producing unreliable estimates that are based on highly skewed left-bounded distributions.

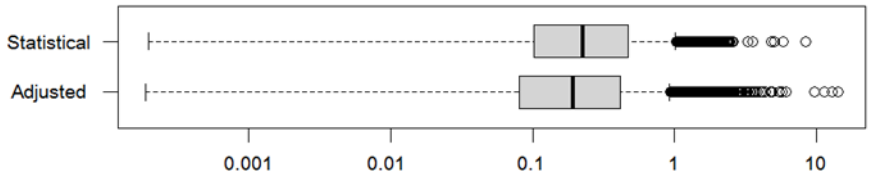

Figure 3. Box-plot for absolute scaled errors found by the MAD/MEAN scheme (log scale)

4. Recommended scheme for measuring the accuracy of point forecasts across many series

To ensure a reliable evaluation of forecasting accuracy under 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 time series 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 Average Relative MAE:

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

If there is 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.

b. Calculate trimmed 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) In theory, the following effect may complicate the interpretation of AvgRelMAE. Since r_i is a ratio estimator, if the distributions of errors from method A, and the benchmark, $e_{i,t}^{\text{A}}$ and $e_{i,t}^{\text{B}}$, within series i have different levels of kurtosis, then $\ln r_i$ is a biased estimate of $\ln\left(E\left|e_{i,t}^{\text{A}}\right|/E\left|e_{i,t}^{\text{B}}\right|\right)$. In fact, if $n_i = 1$ for each i, then the AvgRelMAE becomes equivalent to the GMRAE, which has the limitations described in Section 3.3. However, in practice this effect is usually negligible when $n_i > 5$. But, if necessary, standard correction methods for ratio estimators can be used.

(b) If distribution of absolute errors is heavily skewed, MAE 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 (Alkhazaleh and Razali, 2010). However, further discussions on this topic are outside the scope of our paper.

(c) 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. Any trim between 2% and 10% gives reasonably good results. Generally, when the underlying distribution is symmetrical and heavy-tailed relative to Gaussian, the variance of the trimmed mean is quite a lot smaller than the variance of the sample mean. Therefore, it is highly recommended to use trimmed means for symmetrical distributions.

5. Results of Empirical Evaluation

The results of applying the measures described above are shown in Table 1. When calculating the AvgRelMAE we used statistical forecast as the benchmark.

For the empirical dataset, the analysis has shown that adjustments improved accuracy in terms of the AvgRelMAE, but 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 analysis using MAPE, MASE, and the MAD/MEAN was affected by the highly skewed underlying distribution.

Table 1: 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**

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 <0.01 for the set containing the adjustments of both signs, <0.05 for only positive adjustments, and $<2.2 \cdot 10^{-16}$ for only negative adjustments.

6. Conclusions

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 showed that in practice many well-known error measures become inappropriate.

In order to overcome the disadvantages of existing measures, we recommend the use of the average relative MAE (AvgRelMAE) measure which is calculated as the geometric mean of relative MAE values.

In practice, the adoption of a new error measure may present difficulties due to organisational factors. If organisation insists on using percentages, we recommend using geometric mean APE instead of MAPE because it helps overcome some problems, as described in Section 3.2.

References

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

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

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

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, G., (2008). Yep, we're skewed. Variance 2(2), 179-183.

Fleming, P. J., & Wallace, J. J. (1986). How not to lie with statistics: the correct way to summa-rize 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.

Goodwin, P. (2000). Improving the voluntary integration of statistical forecasts and judgment. International Journal of Forecasting, 16, 85-99.

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

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

Kolassa, S., & Schutz, W. (2007). Advantages of the MAD/MEAN ratio over the MAPE.

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 Fore-casting, 25, 303-324.

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.

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

View publication stats