Assessing point forecast bias across multiple time series: Measures and visual tools

Andrey Davydenko, Paul Goodwin

International Journal of Statistics and Probability; Vol. 10, No. 5; September 2021
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Assessing Point Forecast Bias Across Multiple Time Series: Measures and Visual Tools

Andrey Davydenko[1], Paul Goodwin[2]

[1] Independent Researcher, Antalya, Turkey

[2] The Management School, University of Bath, Bath, BA2 7AY, United Kingdom

Correspondence: Andrey Davydenko. E-mail: andrey@live.co.uk

Received: May 24, 2021 Accepted: August 5, 2021 Online Published: August 20, 2021

doi:10.5539/ijsp.v10n5p46 URL: https://doi.org/10.5539/ijsp.v10n5p46

Abstract

Measuring bias is important as it helps identify flaws in quantitative forecasting methods or judgmental forecasts. It can, therefore, potentially help improve forecasts. Despite this, bias tends to be under-represented in the literature: many studies focus solely on measuring accuracy. Methods for assessing bias in single series are relatively well-known and well-researched, but for datasets containing thousands of observations for multiple series, the methodology for measuring and reporting bias is less obvious. We compare alternative approaches against a number of criteria when rolling-origin point forecasts are available for different forecasting methods and for multiple horizons over multiple series. We focus on relatively simple, yet interpretable and easy-to-implement metrics and visualization tools that are likely to be applicable in practice. To study the statistical properties of alternative measures we use theoretical concepts and simulation experiments based on artificial data with predetermined features. We describe the difference between *mean* and *median bias*, describe the connection between metrics for accuracy and bias, provide suitable bias measures depending on the loss function used to optimise forecasts, and suggest which measures for accuracy should be used to accompany bias indicators. We propose several new measures and provide our recommendations on how to evaluate forecast bias across multiple series.

Keywords: forecasting, forecast bias, mean bias, median bias, MPE, AvgRel-metrics, AvgRelAME, AvgRelAMdE, RelAME, RelMdE, AvgRelME, AvgRelMdE, OPc

1. Introduction

Generally, bias refers to a systematic error. In a forecasting context, bias is usually measured as a mean forecast error (Hill, 2012, p. 140). This gives an indication of *mean bias* which represents a tendency to produce point forecasts that are typically either too high or low in comparison with the corresponding outcomes, irrespective of their size. Less commonly measured is *regression* (or slope) *bias*, which occurs where the systematic discrepancy between the forecast and outcome depends on the size of the forecast (Goodwin, 2000) so that a unit increase in the point forecast tends not to equate to a unit increase in the outcome. As a result, relatively low forecasts can be systematically too high, while relatively high forecasts tend to be too low, or vice versa. *Regression bias* therefore shows how the mean forecast error depends on the forecast itself.

Additionally, a concept of *median bias* can be introduced (Brown, 1947; van der Vaart, 1961). This type of bias occurs when the median of forecast errors differs from zero. Equivalently, we can say that a forecast is median-biased when the probability of over-prediction is not equal to the probability of under-prediction. As with the mean forecast error, the probabilities of over- and under-estimation may correlate with various factors, including the forecast itself. This would imply *regression bias* with respect to the median.

It is known that optimal forecasts under quadratic loss are mean-unbiased while optimal forecasts under linear loss are median-unbiased (Zellner, 1986). Interestingly, best forecasts under linear loss, while being median-unbiased, may reveal both substantial *mean bias* and substantial *regression bias* (see Davydenko, 2012, pp. 124-129, for a Monte-Carlo experiment illustrating this).

Measuring forecast bias is essential as it may indicate imperfections or flaws in a forecasting procedure and this can be costly. For example, a study by (Sanders and Graman, 2009) reported that the impact on costs resulting from forecasts manifesting *mean bias* in a warehouse environment was significantly greater than the impact resulting from the standard

deviation of forecast errors. However, the nature of the loss function should determine whether *mean* or *median bias* is evaluated. Importantly, if forecast density is believed to be symmetric, the presence of either *median bias* and *mean bias* indicates potential modelling problems. For many time series encountered in practice, however, it is usually the case that forecast density is skewed. In these cases, applying transformations and back-transforming statistical forecasts will lead to forecasts that are optimal under linear loss (Davydenko and Fildes, 2016, p. 240). In such settings therefore the task of detecting *median bias* is more relevant. Despite this, the median-unbiasedness of forecasts has received less attention in the literature compared to mean-unbiasedness. For example, some studies (e.g., Spiliotis et al., 2021) have reported only *mean bias* while employing an accuracy metric that assumed linear loss. Logically, where forecast density is non-symmetric and accuracy is reported with respect to linear loss, the measure for bias should reflect *median, rather than mean, bias*. For judgmental point forecasts, as distinct from statistical forecasts, it is usually difficult to know the exact loss function used, but the presence of forecast bias can give some general indication of potential problems, especially if forecast density is believed to be symmetric.

Here, we focus not only on measures, but also on visual tools that help analysts to gain insights into how forecasting performance may be improved. While bias measurement and visualization has a valuable role in signaling deficiencies in forecasts, it has the additional advantage that it can often indicate what needs to be done to effect improvements. For example, (Petropoulos et al., 2017) found that feeding back information on *mean bias* to judgmental forecasters was more effective in enhancing the accuracy of their forecasts than feedback on accuracy. The detection of bias also allows future forecasts to be corrected. Theil (1966, p. 33) showed that both mean and regression bias can be removed from a set of forecasts ($\boldsymbol{F}$) where the outcomes ($\boldsymbol{Y}$) are known by fitting the OLS (ordinary least squares) regression equation: $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{F}] = a + b\boldsymbol{F}$. Future forecasts can be corrected by substituting the corrected forecast for $\boldsymbol{Y}$ in the equation, assuming that the biases observed in the past will persist. Where the biases are liable to change, (Goodwin, 1997) showed that fitting the regression model using discounted weighted regression (Ameen and Harrison, 1984) could lead to improved forecast accuracy through correction.

The detection and visualization of bias in forecasts of *individual* series is relatively a well-researched area. Tests for rationality are a long-established tool for detecting bias in forecasts as well as inefficiency (e.g., Johnston, 1972, pp. 28-29). In addition, Theil's prediction-realization diagrams enable users to see the extent which biases cause forecasts to depart systematically from a line of perfect forecasts (Theil, 1966, pp. 21-22). However, in some situations it is necessary to assess the bias of a forecasting method over multiple series or to compare its typical level of bias with an alternative method over these series. For example, in forecasting competitions, such as the M4 (Makridakis et al., 2018), researchers may wish to establish which forecasting methods typically exhibit the least bias when they are applied to thousands of time series. Similarly, companies selling large numbers of products may find that it is impractical to assign an individual forecasting method to each product so that they need to identify a single method offering the least bias across their product range. Measuring bias over multiple series poses several challenges, including the need to avoid scale dependence. For example, if some series are measured in millions of units and others in single units, the errors in the larger-scaled series will dominate when a bias measure like the mean error is taken over all the series. Even where multiple series are measured on the same scale, there is also the need to avoid measures being distorted by extreme levels of bias in isolated series. Given these challenges, this paper compares the advantages and disadvantages of using different measures and visualization tools when bias in its different forms needs to be assessed over multiple series, origins and forecast horizons.

2. The Point Forecast Evaluation Setup (PFES)

Bias assessment may be used to answer one of two questions: (i) does a given forecasting method exhibit bias when applied to multiple series and (ii) do alternative forecasting methods differ in the levels of bias they exhibit across multiple series? In relation to the second question, we focus particularly on the use of bias measures as part of forecast-value-added (FVA) analysis (Gilliland, 2008) by examining whether the bias resulting from attempts to improve a set of forecasts is less than that of the original forecasts. For example, judgmentally adjusted forecasts may be compared with unadjusted forecasts to see whether the adjustments are 'adding value' by reducing bias. Alternatively, the bias of a proposed new forecasting method may be compared with that of an existing or simple method such as naïve forecasts.

In this paper we evaluate the measurement and visualization of types of bias under the following conditions, which we refer to as *point forecast evaluation setup* (PFES). This is a particular case of a more general setup defined by (Davydenko et al., 2021, p. 81), where prediction intervals were involved as well. In order to store and access forecast data relating to the PFES, it is possible to use the data formats introduced by (Davydenko et al., 2021, pp. 83-87).

1) We have a set of time series. The set may include from one to thousands of series.

2) Data frequency is the same for each time series (e.g., months, or years), but each series can contain different numbers of observations. Series can also contain missing cases.

3) For each series we have a set of alternative forecasts. Forecasts can be produced from different origins (rolling-origin forecasts) for one or multiple horizons.

4) Forecasts are point estimates produced using statistical or judgmental methods (we do not consider prediction intervals or density forecasts here).

5) Actuals are true outcomes of the quantities being predicted by point forecasts. For example, if we are forecasting the demand for a product, the outcome will be the actual demand for that product, not the level of sales, which may be less than demand where a stock out has been incurred.

For the given forecast data, we assume that we want to measure and compare accuracy and bias with regard to the mean or the median, depending on the distribution of forecast error and the loss function used to optimise forecasts. We may also wish to measure and compare accuracy and bias not only for alternative forecasts, but also for different subsets of the whole dataset. For example, we may wish to compare bias of forecasts obtained as a result of positive and negative judgmental adjustments to statistical forecasts, or to compare bias of forecasts obtained in different seasons or years. So, ideally, it should be possible to slice-and-dice forecast data and to obtain corresponding bias indicators for data subsets (see, e.g., Davydenko et al., 2021, for examples of constructing queries to subset forecast data). A procedure for conducting a formal statistical test for the presence of bias is also desirable.

3. Criteria for a Measure of Forecasting Performance

It is difficult to summarise forecasting performance using just one indicator. For example, as mentioned earlier, bias can depend on various factors, including the size of the forecast itself. Nonetheless, we aim to find the most concise indicators, that still offer a high degree of informativeness. To achieve this, we will assess alternative bias metrics against the following six criteria defined by (Davydenko and Fildes, 2016, p. 240): (i) interpretability, (ii) robustness (i.e., insensitivity to outliers), (iii) applicability in a wide range of settings (e.g., the metric is applicable where errors, forecasts or actuals have values of zero), (iv) informativeness, (v) appropriateness given the loss function that was used for optimisation, and (vi) scale-independence. To these we add (vii) *construct validity*, which reflects the extent to which a metric measures what it is intended to measure. We also consider the extent to which measures meet the criteria of (viii) ease-of-implementation and (ix) ease-of-understanding. The latter criterion is important to ensure that evaluation results are easy-to-communicate to the participants of the forecasting process who may not be technical specialists.

4. Notation

We use a flexible notation where for each time series we, as a general case, have a different number of available observations (an approach previously used by Davydenko, 2012). For simplicity, we assume all forecasts have the same horizon so we do not show the horizon in the equations. We will address the question of averaging bias measures across horizons in later sections.

The following notation will be used:

N – number of time series,

T_i – the set containing time periods (relating to time series i) for which forecasts from all methods are available,

n_i – number of elements in T_i,

$Y_{i,t}$ – actual for series i, period t,

$F_{i,t,j}$ – out-of-sample forecast produced by method j for period t of time series i,

$e_{i,t,j}$ – forecast error from method j for series i, period t, defined as $e_{i,t,j} = Y_{i,t} - F_{i,t,j}$,

$ME_{i,j}$ – mean error for method j for series i,

$MdE_{i,j}$ – median error for method j for series i.

In formulae, we will use the following indices: i to denote a time series, j to denote a method, and t to denote a time period. $\mathbb{E}[\cdot]$ will denote mathematical expectation, $\mathbb{E}[\cdot|X]$ will denote conditional expectation given X.

5. Types of Bias in Individual Time Series

Before addressing the problem of detecting bias across multiple series, this section considers different types of bias that can be found within a single time series. Later we will evaluate the extent to which a range of measures are able to reflect these biases when multiple series are involved.

5.1 Mean Bias

Assuming that we confine our analysis to one series (say, i) and one method (say, j). Then the mean error, $ME_{i,j}$, indicates *mean bias.* This is given as:

$$ME_{i,j} = \frac{1}{n_i}\sum_{t\in T_i} e_{i,t,j}.$$

Having $ME_{i,j}$ statistically different from 0 suggests that the method is likely to be non-optimal under quadratic loss (DeGroot, 1970). Note that, counter intuitively, positive values for the mean indicate a tendency to forecast too low, while negative values indicate a tendency to forecast too high.

5.2 Regression Bias With Regard to the Mean

We can expand the concept of *mean bias* to see if mean error depends on the forecast itself by using the following regression: $\mathbb{E}[Y_{i,t} \mid F_{i,t,j}] = a_i + b_i F_{i,t,j}$. Obtaining $a_i \neq 0$ or $b_i \neq 1$ provides evidence for non-optimal forecasts under quadratic loss (Johnston, 1972). This relationship may be non-linear and may change in time, see (Davydenko, 2012, pp. 99-129 and 155-158) for examples of non-linear models and models with time-varying coefficients. For the case of many series a number of approaches are available including panel data models, which, in particular, can be effectively estimated using Bayesian models (see Davydenko, 2012, pp. 156-158).

5.3 Median Bias

We can use $MdE_{i,j}$ to indicate *median bias*:

$$MdE_{i,j} = Median\,(e_{i,t,j}),$$

where $Median(e_{i,t,j})$ is the sample median over all $e_{i,t,j}$ belonging to series i and method j.

As with the mean error, a positive value indicates a tendency to forecast too low, and vice versa. Obtaining $MdE_{i,i}$ significantly different from 0 indicates that the method is likely to be non-optimal under linear loss.

Another approach to reporting *median bias* is to use the *Overestimation Percentage* (OP), i.e., the percentage of cases when $Y_{i,t} < F_{i,t,j}$:

$$OP_{i,j} = 100\% \times \frac{1}{n_i}\sum_{t\in T_i} 1\{Y_{i,t} < F_{i,t,j}\}.$$

One potential problem with the OP is that if zero errors occur, even for median-unbiased forecasts we will obtain $OP_{i,j} < 50\%$, which is confusing (the same problem was identified for the "Percent Better" metric when evaluating accuracy, see Davydenko and Fildes, 2016, pp. 243-244). The issue is especially relevant for count data, especially for so-called intermittent demand series.

We therefore propose the following statistic (the *Overestimation Percentage corrected*, OPc) to rectify the problem:

$$OPc_{i,j} = OP_{i,j} + \frac{ZP_{i,j}}{2},$$

where $ZP_{i,j}$ is the percentage of zero errors:

$$ZP_{i,j} = 100\% \times \frac{1}{n_i}\sum_{t\in T_i} 1\{e_{i,t,j} = 0\}.$$

For a median-unbiased forecast the OPc should not significantly differ from 50%, ensuring that the probability of overestimation is equal to the probability of underestimation.

Alternatively, for software implementation, this formula (based on the sign function) may be more suitable:

$$OPc_{i,j} = 100\% \times \frac{1}{n_i}\sum_{t\in T_i} 0.5(1 - \mathrm{sgn}(e_{i,t,j})).$$

For negative errors (cases of overestimation) the expression $0.5(1 - \mathrm{sgn}(e_{i,t,j}))$ is 1, for positive errors (cases of underestimation) it is 0, for zero errors it is 0.5.

5.4 Regression Bias With Respect to Median

As with the mean error, the OPc and the median error may depend on the forecast itself. When median error depends on the size of forecasts, they become non-optimal under linear loss. Possible models to detect and correct *regression bias* with respect to the median can be found, for example, in (Davydenko, 2012, pp. 99-105).

6. Illustrative Datasets

To demonstrate the performance of alternative bias metrics across multiple series, we generated two illustrative datasets using simple rules.

6.1 'Dataset1'

The first dataset was generated using a normal distribution. It contains 1000 series, each series includes 36 actuals. All series were generated using the same equation:

$$Y_{i,t} = 5 + \varepsilon_{i,t},$$

where $\varepsilon_{i,t} \sim N(0,1)$, i.i.d., $i = 1,\ldots,1000$, $t = 1,\ldots,36$.

Fig. 1 shows an example of a series generated using the above equation. The data generated resembles series (with relatively low observations) that can be met in practice.

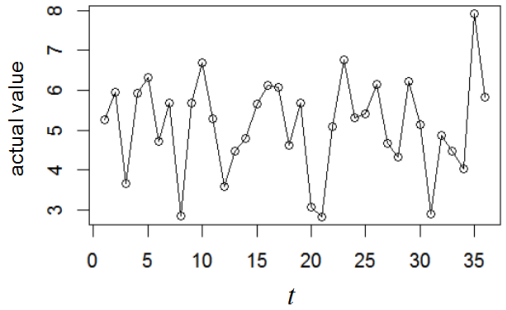

Figure 1. An example of a time series from 'Dataset1'

To generate forecasts ($F_{i.t.i}$) for any period t and any series i in 'Dataset1', we used the equations shown in Table 1. Values for $\mathbb{E}[ME_{i,i}]$ and $\mathbb{E}[MdE_{i,i}]$ given in Table 1 indicate *mean* and *median bias*, respectively. These values were obtained analytically based on the formulae for $Y_{i,t}$ and $F_{i,t,j}$.

Table 1. Equations used to generate 'Dataset1'

Method, j	Equation	Description	$\mathbb{E}[ME_{i,j}]$	$\mathbb{E}[MdE_{i,j}]$
Method 1	$F_{i,t,1} = 5$	Gives optimal estimates under both linear and quadratic loss. $F_{i,t,1}$ is a mean- and median-unbiased forecast since the distribution of $\varepsilon_{i,t}$ is symmetric and $F_{i,t,1} = \mathbb{E}[Y_{i,t}]$.	0	0
Method 2	$F_{i,t,2} = 6$	Both mean- and median-biased: $F_{i,t,2} = \mathbb{E}[Y_{i,t}] + 1$.	-1	-1
Method 3	$F_{i,t,3} = 4$	Both mean- and median-biased: $F_{i,t,3} = \mathbb{E}[Y_{i,t}] - 1$.	1	1
Method 4	$F_{i,t,4} = 7$	Both mean- and median-biased: $F_{i,t,4} = \mathbb{E}[Y_{i,t}] + 2$.	-2	-2
Method 5	$F_{i,t,5} = 5 + \varepsilon_{i,t}$, $\varepsilon_{i,t} \sim N(0, 0.1^2)$, i.i.d.	This is both a mean- and median-unbiased forecast, but it has some noise added. Therefore, the accuracy of this forecast is lower than that of Method 1.	0	0

6.2 'Dataset2'

The second dataset was generated using a non-symmetric distribution in order to better replicate real-world data. More specifically, we assume each actual to follow a log-normal distribution (i.i.d.) with the following parameters:

$$Y_{i,t} \sim Lognormal(\mu = 5, \sigma^2 = 0.5^2).$$

Given the above equation, for any series, and for any period, the expected outcome remains the same and can be found using the well-known formula for the mean of the log-normal distribution:

$$\mathbb{E}[Y_{i,t}] = \exp(\mu + \frac{\sigma^2}{2}) = 168.1741.$$

At the same time, for the log-normal distribution we expect the median to be

$$Median[Y_{i,t}] = \exp(\mu) = 148.4132.$$

When trying to predict $Y_{i.t}$, the optimal forecast under linear loss is therefore 148.4132, whereas the optimal forecast under quadratic loss is 168.1741. Fig. 2 shows a series from 'Dataset2' and the difference between optimal predictions depending on the loss function used to optimise the predictions.

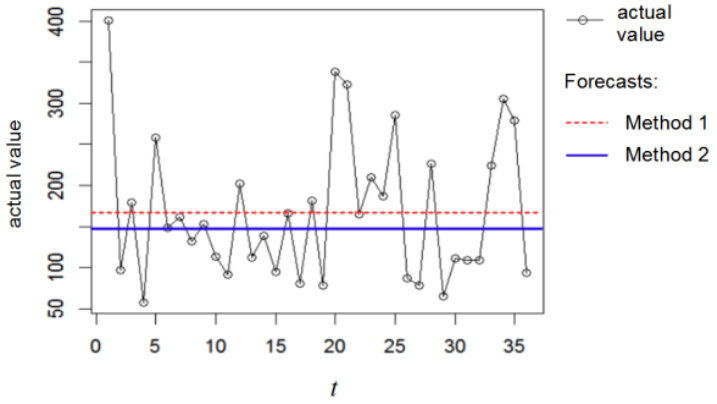

Figure 2. An example of a time series from 'Dataset2'. Methods 1 and 2 (defined in Table 2 below) give optimal predictions under quadratic and linear loss, respectively.

To generate forecasts ($F_{i.t.i}$) for any period t and any series i in 'Dataset2', we used the equations shown in Table 2. Values for $\mathbb{E}[ME_{i.i}]$ and $\mathbb{E}[MdE_{i.i}]$ give indication of *mean* and *median bias*, respectively. These values were found analytically. As shown in Table 2, forecasts from Method 1 correspond to the mean of forecast density and are mean-unbiased ($\mathbb{E}[ME_{i.1}] = 0$), whereas forecasts from Method 2 correspond to the median of forecast density and are median-unbiased ($\mathbb{E}[MdE_{i.2}] = 0$).

We next applied a range of bias measures to the two data sets and evaluated them against the criteria we outlined earlier.

Table 2. Equations used to generate 'Dataset2'

Method, j	Equation	Description	$\mathbb{E}[ME_{i,j}]$	$\mathbb{E}[MdE_{i,j}]$
Method 1	$F_{i,t,1} = \mathbb{E}[Y_{i,t}] = 168.1741$	Gives optimal estimates under quadratic loss, but not optimal under linear loss. Mean-unbiased, median-biased.	0	-19.7609
Method 2	$F_{i,t,2} = Median[Y_{i,t}] = 148.4132$	Optimal under linear loss, but not optimal under quadratic loss. Median-unbiased, mean-biased.	-19.7609	0
Method 3	$F_{i,t,3} = \mathbb{E}[Y_{i,t}] + 30$	Both mean- and median-biased.	-30	-49.7609
Method 4	$F_{i,t,4} = \mathbb{E}[Y_{i,t}] - 30$	Both mean- and median-biased.	30	10.2391
Method 5	$F_{i,t,5} = \mathbb{E}[Y_{i,t}] + 60$	Both mean- and median-biased.	-60	-79.7609

7. Measuring and Comparing Bias Across Series

7.1 Percentage Errors

Aggregating ME and MdE across series is problematic as they are scale-dependent. One well-known approach is to use percentage errors (PEs) instead of the original errors. A PE is given by:

$$PE_{i,t,j} = e_{i,t,j}/Y_{i,t} \times 100\%.$$

For a given series i and method j, mean percentage error (MPE) calculated within series is:

$$MPE_{i,j} = \frac{1}{n_i}\sum_{t\in T_i} PE_{i,t,j}.$$

MPE for method j across all series is (assuming all series have equal length):

$$MPE_j = \frac{1}{N}\sum_{i=1}^{N} MPE_{i,j}.$$

This approach, however, has disadvantages arising due to the intractable features of PEs (e.g., see Davydenko, 2012; Goodwin, 2018). In particular, PEs are arguably unsuitable for trended or seasonal series. In the former case, for a given error, the PE declines as the level in the series increases. In the latter case, a given error will be associated with a smaller PE at a seasonal peak than at a seasonal trough (Goodwin, 2018). Crucially, such PEs cannot be calculated when an outcome is zero, as is frequently the case with intermittent demand. In addition, very small actual values can be associated with very large PEs even when the forecast is close to the outcome. These can lead to highly skewed distributions of PEs with long tails. For time series containing only positive values, such as those representing product demand, positive PEs will have an upper bound of 100% (this will occur when the forecast equals zero). However, there is no lower band to negative PEs, where the forecast exceeds the outcome. This can lead to the mean PE having unrepresentatively large negative values when they are measured both within and across series. Also, PEs require positive actuals, making them inapplicable for some tasks (e.g., weather forecasts). For example, an actual of 2 units and a forecast of 4 units has the same percentage error (-100%) as an actual of -2 and a forecast of -4, despite the biases being in opposite directions.

As we show below, the use of PEs generally distorts the original properties of errors. Fig. 3 shows side-by-side boxplots of MPEs for the artificial datasets and Table 3 shows corresponding MPE values. As expected, the upper bound for MPE is 100%, whereas there is no lower bound, which makes the distribution skewed. For 'Dataset1' for Method 1 we expect to have zero *mean* and *median bias*, but the MPEs show a significant bias towards overestimation. Moreover, we expect Methods 2 and 3 ('Dataset1') to have equal absolute bias, but the absolute MPE of Method 2 is much higher compared to that of Method 3, demonstrating that the MPE tends to place a heavier weight on overestimation compared to underestimation.

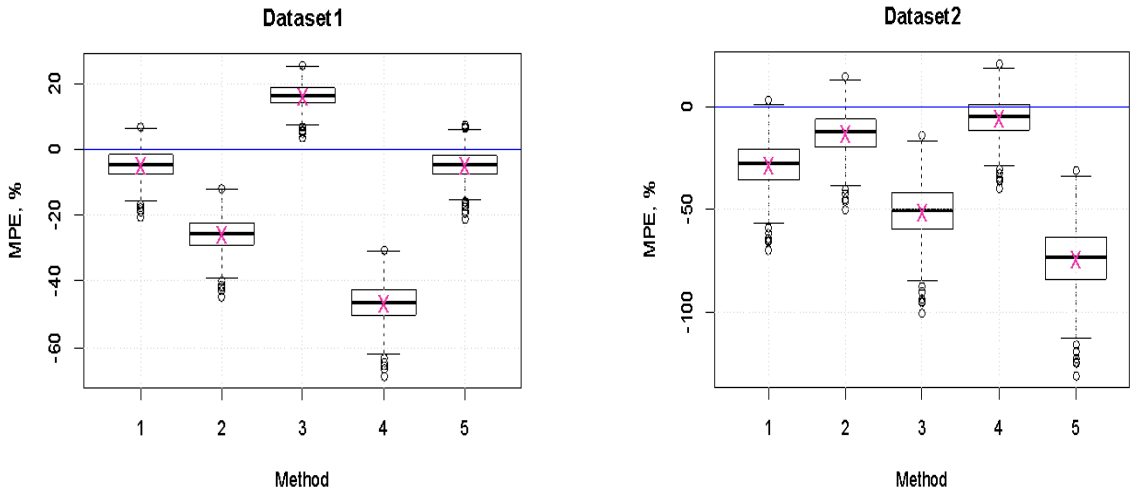

Figure 3. MPE-boxplots showing MPE values for individual series

Note: 'X' denotes sample means (or MPEs across series). Values below the line indicate overestimation, values above the line mean underestimation. Due to the non-symmetric features of PEs, cases of overestimation receive heavier penalties compared to those of underestimation, which complicates the interpretation. Generally, MPEs are not good proxies for MEs.

Table 3. MPEs and true ME/MEAN ratios (desirable values)

Indicator	Dataset1					Dataset2				
	Method, j					Method, j				
	1	2	3	4	5	1	2	3	4	5
MPE_i, %	**-4.65**	-25.58	16.27	-46.51	-4.66	-28.42	-13.33	-51.32	**-5.5**	-74.23
$\mathbb{E}[ME_{i,j}]$	**0**	-1	1	-2	0	**0**	19.76	-30	30	-60
$\mathbb{E}[Y_{i,t}]$	5	5	5	5	5	168.17	168.17	168.17	168.17	168.17
Desirable value: $\frac{\mathbb{E}[ME_{i,j}]}{\mathbb{E}[Y_{i,t}]} \times 100$	**0**	-20	20	-40	**0**	**0**	11.75	-17.84	17.84	-35.68

Note: Bold font shows the best methods in terms of *mean bias* depending on the indicator. Index i denotes any series, index t denotes any time period.

For 'Dataset2', where the distributions of actuals is non-symmetric, the results are even worse. We should expect different signs of bias for Method 3 and Method 4 and no bias for Method 1, but, according to the MPE, all methods overestimate actuals. Table 3 shows that MPEs do not reflect the ME/MEAN ratio for 'Dataset2' and do not even reflect the true direction of bias. The interpretation of MPE results is therefore counterintuitive and can lead to erroneous conclusions.

Our simulations show that the use of MPE as an indicator of *mean bias* and a proxy for ME is not advisable. Similar experiments we conducted showed that the median percentage error (MdPE) is not advisable as an indicator of *median bias* and a proxy for MdE (for brevity we have not presented these results here). Interestingly, (Nikolopoulos et al., 2005) modelled *regression bias* based on PEs where errors were divided by forecasts instead of actuals. But, again, due to the distortions introduced, the results of this approach are also prone to error, as indicated in (Davydenko 2012, p. 160). Overall, it is clear that PE-based metrics fail to meet the criteria of robustness, applicability in settings where actuals are zero and *construct validity* that we set out earlier.

7.2 Scaled Errors

Some disadvantages of PEs can be avoided by dividing errors by the in-sample MAE (mean absolute error) of the naïve forecast, as proposed by (Hyndman and Koehler, 2006). A scaled error is

$$q_{i,t,j} = \frac{e_{i,t,j}}{MAE^{NAIVE}{}_i},$$

where $MAE^{NAIVE}{}_i$ is the in-sample MAE for the naïve method for time series i.

Scaled errors have been used in some studies (e.g., Spilotis et al., 2021) to analyse bias. In particular, (Spilotis et al., 2021) used the following formula for the Absolute Mean Scaled Error (AMScE) for series i and method j:

$$AMScE_{i,j} = \left| \frac{1}{n_i} \sum_{t \in T_i} q_{i,t,j} \right|.$$

Where there are multiple series, the mean AMScE is obtained by averaging AMScEs across series (we assume all series are of equal length, a weighted mean can be used to reflect different lengths of series):

$$AMScE_j = \frac{1}{N} \sum_{i=1}^{N} AMScE_{i,j}.$$

Fig. 4 shows boxplots for AMScE for 'Dataset1' and Table 4 shows corresponding AMScE values. The use of AMScE is problematic for two reasons. Firstly, if we use absolute values of the MScE, even unbiased forecasts will show some

bias and the extent of this erroneous indication will depend on the sample size. Secondly, due to the distribution introduced by the arithmetic mean, AMScE will not represent the desirable true ratio $\left|\mathbb{E}[ME_{i,j}]\right|/\mathbb{E}[MAE^{NAIVE}{}_i]$ (where i denotes any series) reliably. Additional problems may arise when some MAEs appearing in the denominator are small so that the underlying distribution of the metric is highly skewed. Instead of scaling by in-sample MAE of the naïve forecast, errors can be scaled by series means (see Davydenko and Fildes, 2016, p. 245, for the disadvantages of this approach) or series standard deviations, but these alternatives will not eliminate the above problems.

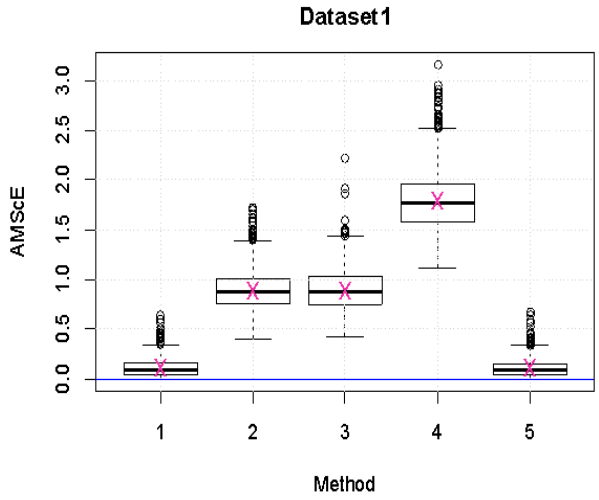

Figure 4. AMScE-boxplots showing distribution of $AMScE_{i,j}$

Note: 'X' denotes AMScE values across series. One evident problem is that unbiased forecasts (Methods 1 and 5) are still shown to be biased.

Table 4. AMScE and true expected ratios for 'Dataset1'

Indicator	Method, j				
	1	2	3	4	5
$AMScE_i$	**0.12**	0.9	0.9	1.8	**0.12**
$MScE_i$	**0**	-0.9	0.9	-1.8	**0**
$\mathbb{E}[ME_{i,j}]$	**0**	-1	1	-2	**0**
$\mathbb{E}[MAE^{NAIVE}{}_i]$	1.13	1.13	1.13	1.13	1.13
Desirable value: $\left\|\mathbb{E}[ME_{i,j}]\right\|/\mathbb{E}[MAE^{NAIVE}{}_i]$	**0**	-0.88	0.88	-1.77	**0**

Note: Bold font shows the best methods in terms of *mean bias* depending on the indicator. Index i denotes any series, $\mathbb{E}[MAE^{NAIVE}{}_i]$ was found using a Monte-Carlo simulation.

Some problems can be mitigated by just using the MScE instead of the AMScE. But, generally, the AMScE is not a good proxy for ME (see Table 4). The AMScE or MScE will exaggerate bias in the same way as the MASE (mean absolute scaled error) will exaggerate the performance of the benchmark forecasting method (see Davydenko and Fildes, 2013). Further experiments (not discussed here for brevity) showed that the use of absolute median scaled errors may not give a reliable indicator of *median bias*. Another problem is that we may want to aggregate the AMScE across horizons and the arithmetic mean will lead to the over-influence of forecasts with greater horizons (Davydenko et al., 2021). There have been attempts to model regression bias based on the use of scaled errors and scaled forecasts (Fildes et al., 2009; Davydenko et al., 2010), but this approach may lead to spurious correlation due to the correlation between the error and the scale (Davydenko, 2012, pp. 150-152).

7.3 The General AvgRel-metric and Its Principles

In order to address the problems identified above and to provide an improved approach for measuring performance across series, Davydenko (2012, Chapter 2) introduced a new class of metrics (which we will refer to as AvgRel-metrics) based on the following principles:

- The forecasting performance of a method is assessed as a relative indicator showing how it compares with a benchmark (for example, the performance of the naïve method). This principle is similar to that of the MASE or the MAD/MEAN ratio (see Hyndman and Koehler, 2006) that are used to assess accuracy.
- The performance of the method and the benchmark indicator are first calculated for each time series individually. This involves the use of rolling-origin forecasts having the same fixed horizon. Importantly, both the performance for the method and the benchmark indicator should relate to the same period of time (the same evaluation sample should be used for the method and for the benchmark). This helps avoid problems arising due to structural breaks and it is where this approach differs from the MASE or MAD/MEAN ratio.
- *Relative performances* are then obtained as ratios of the performance of the method and the benchmark indicator.
- Averaging *relative performances* across series is based on the weighted geometric mean. Based on research by (Fleming and Wallace, 1986), (Davydenko, 2012) identified the following advantages of the geometric mean in the context of averaging relative forecasting performances over multiple series. Most importantly, the geometric mean ensures invariance of rankings (Davydenko, 2012, p. 66). Generally, the median or the arithmetic mean do not ensure this property. This property, in turn, comes from the fact that the geometric mean gives equal weight to reciprocal relative changes (Davydenko, 2012, p. 61).
- Importantly, the function initially used to optimise forecasts should correspond to the function used as a performance indicator (Davydenko, 2012, p. 63). In particular, if forecasts are calculated as means of forecast densities then the MSE (mean squared error) or RMSE (root mean squared error) should be used as a function to measure forecast accuracy. If the median of density forecast was used as point forecasts, then the MAE (mean absolute error) should be used as a function to measure forecast accuracy (see Davydenko, 2012, p. 84).

The combination of the above principles makes the AvgRel-metrics a novel approach compared to exiting methods in the literature. The following general AvgRel-metric was suggested by (Davydenko, 2012, p. 62) to indicate average relative performance across multiple series for a given method j:

$$AvgRelP_j = \left(\prod_{i=1}^{N} RelP_{i,j}^{n_i} \right)^{\frac{1}{\sum_{i=1}^{N} n_i}},$$

where $RelP_{i,j}$ denotes *relative performance* found as

$$RelP_{i,j} = \frac{c_{i,j}}{c_i^B},$$

where $c_{i,j}$ - characteristic of forecasting errors of method j for series i (e.g., $MAE_{i,j}$), c_i^B - characteristic of the benchmark for series i (e.g., the MAE of the naïve method for series i), n_i - number of time periods used to calculate $c_{i,j}$ assuming both $c_{i,j}$ and c_i^B are calculated using the same time periods.

The "AvgRel*" prefix introduced in (Davydenko, 2012, Chapter 2) helps avoid confusion with some well-known measures based on the arithmetic mean and make the metric more recognizable across studies (see Davydenko et al., 2021, p. 96).

Obtaining $AvgRelP_j < 1$ means the performance indicator of method j for individual series is an average lower than the benchmark, $AvgRelP_j > 1$ means the opposite.

Importantly, following the same principle of averaging relative performances using the geometric mean, $AvgRelP_j$ can conveniently be averaged across horizons (Davydenko et al., 2021, pp. 95-96). Suppose $AvgRel_{j,h}$ denotes AvgRelP for horizon h and method j. Then

$$AvgRelP_j = \left(\prod_{h=1}^{H} AvgRelP_{j,h}^{l_h} \right)^{\frac{1}{\sum_{h=1}^{H} l_h}},$$

where H - number of forecast horizons available, l_h - number of forecasts available for horizon h.

Note that the geometric mean is equivalent to the antilog of the arithmetic mean of logarithms, enabling analysts to explore the distribution of $log(RelP_{i,j})$ for potential outliers, and the presence of skew. This can also help them to identify and perform appropriate statistical tests. Davydenko and Fildes (2016) proposed a statistical test to check if AvgRelP significantly differs from 1 and also a robust version of the AvgRelP based on using the concept of the trimmed mean. The underlying distribution for the AvgRelP can be explored using boxplots of $log(RelP_{i,j})$, as was done by (Davydenko and Fildes, 2013). Here we introduce an improved variant of boxplots to explore the distribution of RelPs featuring a double-scale to represent both the log-scale and the original scale to improve readability of plots.

7.4 AvgRel-metrics for Accuracy and Bias and Their Interconnection

For measuring forecasting accuracy in terms of symmetric linear loss the Average Relative Mean Absolute Error (AvgRelMAE) was defined by (Davydenko, 2012, p. 63):

$$AvgRelMAE_j = \left(\prod_{i=1}^{N} RelMAE_{i,j}^{n_i}\right)^{\frac{1}{\sum_{i=1}^{N} n_i}},$$

$$RelMAE_{i,j} = \frac{MAE_{i,j}}{MAE_{i,B}},$$

where $MAE_{i,j}$ is the mean absolute error (MAE) calculated for method j and series i using observations relating to time periods t $(t \in T_i)$: $MAE_{i,j} = {}^{1}/_{n_i} \sum_{t \in T_i} |e_{i,t,j}|$, B is the index of the benchmark method.

Similarly, the Average Relative Mean Squared Error (AvgRelMSE) was defined by (Davydenko, 2012, p. 63) in order to measure forecasting performance when the target loss is quadratic:

$$AvgRelMSE_j = \left(\prod_{i=1}^{N} RelMSE_{i,j}^{n_i}\right)^{\frac{1}{\sum_{i=1}^{N} n_i}},$$

$$RelMSE_{i,j} = \frac{MSE_{i,j}}{MSE_{i,B}},$$

where $MSE_{i,j}$ is the mean square error (MSE) calculated for method j and series i using observations relating to time periods t $(t \in T_i)$: $MSE_{i,j} = {}^{1}/_{n_i} \sum_{t \in T_i} e_{i,t,j}^2$, B is the index of the benchmark method.

For better interpretation, instead of the MSE we can use the root mean square error (RMSE), which is found as a square root of MSE. The relative RMSE (RelRMSE) is the ratio of two RMSEs and represents a relative variance of errors (see Davydenko and Fildes, 2013, p. 517). By analogy to the AvgRelMSE, we can use the AvgRelRMSE abbreviation (introduced by Davydenko, 2012, p. 262). The rankings of methods in terms of these measures will stay the same since the AvgRelRMSE is a square root of the AvgRelMSE:

$$AvgRelRMSE_j = \sqrt[2]{AvgRelMSE_j}.$$

AvgRel-metrics have been successfully used in a number of studies (e.g., Fildes and Goodwin, 2021). However, one limitation of the AvgRelMAE is that it converges to the geometric mean of relative absolute errors (GMRAE) when n_i gets close to 1 (Davydenko and Fildes, 2013). As explained by (Davydenko and Fildes, 2013, p. 518), GMRAE is sometimes not a good proxy for RelMAE. In practice, however, this effect is usually negligible when $n_i > 5$. The same considerations relate to the other AvgRel-metrics for accuracy (such as the AvgRelRMSE).

Regarding the indicators for bias, the following considerations apply when finding the correspondence between the optimisation of forecasts and the metric used for forecast evaluation. When point forecasts are equivalent to the mean of density forecasts they are optimised for minimizing the absolute mean error (AME), and hence are designed to avoid *mean bias*. Such forecasts are optimised under quadratic loss and should therefore be evaluated using the MSE or RMSE for accuracy and the AME for bias. When point forecast are equivalent to the median of density forecasts, they are optimised for minimizing the absolute median error (AMdE), and hence are designed to avoid *median bias*. These forecasts are optimised under linear loss and should be evaluated using the MAE for accuracy and the AMdE for bias.

Fig. 5 demonstrates the application of AvgRelMAE to 'Dataset2' where the distribution of actuals is skewed. Under this condition, we would expect forecasts optimal under quadratic loss to have no *mean bias* and forecasts optimal under linear loss to have no *median bias.* Method 2 that is optimised for linear loss has no *median bias* and, as expected, shows the best performance when measured by AvgRelMAE as the AvgRelMAE is a proxy for MAE. Note that where AvgRel-metrics are being applied, we recommend that boxplots have a double scale, as in Fig. 5, with one scale showing original values of the metric to aid interpretation. We also recommend that the mean of distributions should be plotted to demonstrate their degree of skewness. Fig. 6 shows the application of AvgRelMSE to 'Dataset 2'. In this case, Method 1, optimised under quadratic loss, and therefore having zero *mean bias*, is the best performer rather than Method 2. This demonstrates both that the AvgRel-metrics are reflecting the true underlying performance conditions and that it is important to match the metric for accuracy and the metric for bias. However, this raises the question of how the AvgRel-metrics can be adapted to measure forecast bias.

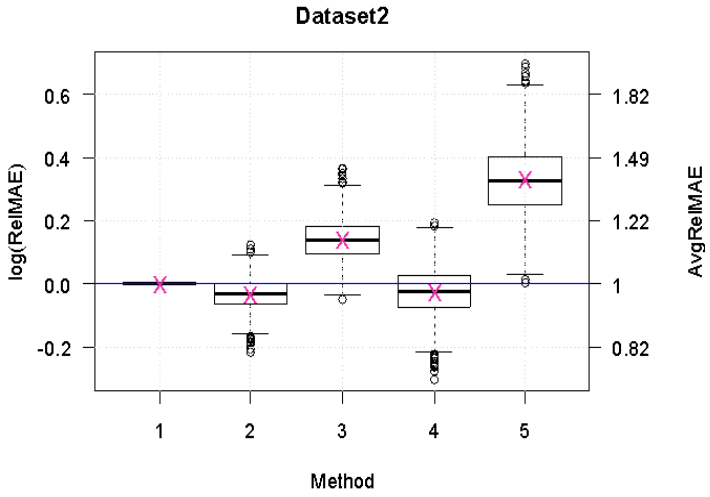

Figure 5. AvgRelMAE-boxplots (benchmark: Method 1). 'X' denotes AvgRelMAEs for each method

Note: AvgRelMAE is a proxy for MAE and Method 2, having zero *median bias*, is the best in terms of MAE, as expected.

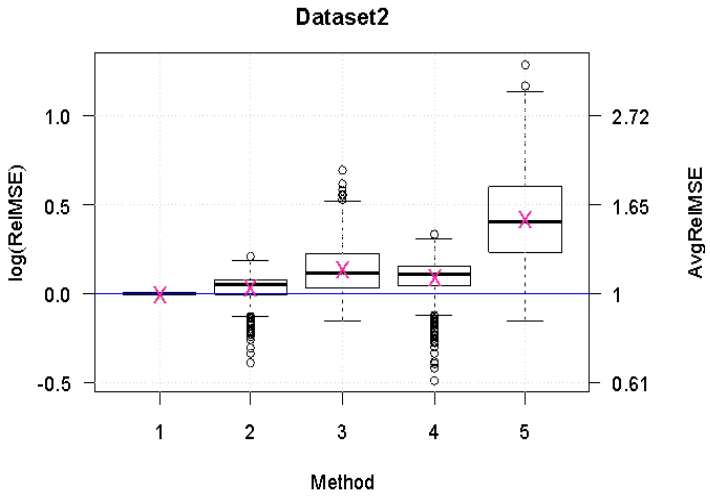

Figure 6. AvgRelMSE-boxplots (benchmark: Method 1). 'X' denotes AvgRelMSEs for each method

Note: AvgRelMSE is a proxy for MSE and Method 1, having zero *mean bias*, is the best in terms of MSE, as expected.

Using ME directly in the equation for the AvgRelP (see Section 7.3) is not possible, as it can take negative values. Thus, deriving a proxy for ME (and MdE) is not straightforward. In order to overcome this problem, where multiple series are involved (Davydenko, 2012, p. 64) proposed the Relative Absolute Mean Error (RelAME) to indicate relative bias. The RelAME for method j and time series i is defined as

$$RelAME_{i,j} = \left|\frac{ME_{i,j}}{ME_{i,B}}\right|,$$

where B denotes the index of the benchmark method.

To average RelAMEs across series, (Davydenko, 2012, p. 64) proposed the Average Relative Absolute Mean Error (AvgRelAME):

$$AvgRelAME_j = \left(\prod_{i=1}^{N} RelAME_{i,j}{}^{n_i}\right)^{1/\sum_{i=1}^{N} n_i}.$$

Finding the AvgRelAME therefore involves calculating the ratio of the absolute mean errors for each series and then finding the geometric mean of these ratios across the series. By analogy, for *median bias*, we propose the Average Relative Absolute Median Error (AvgRelAMdE):

$$AvgRelAMdE_j = \left(\prod_{i=1}^{N} RelAMdE_{i,j}{}^{n_i}\right)^{1/\sum_{i=1}^{N} n_i},$$

$$RelAMdE_{i,j} = \left|\frac{MdE_{i,j}}{MdE_{i,B}}\right|,$$

where B denotes the index of the benchmark method.

Fig. 7 shows side-by-side boxplots for RelAMEs (benchmark: Method 1) for 'Dataset1'. The problem is that since Method 1 is unbiased, the relative performances of alternative (biased) methods appear to be very poor (see Table 5). The interpretation of the AvgRelAME becomes problematic in this case. Nonetheless, the AvgRelAME identified the ranks correctly.

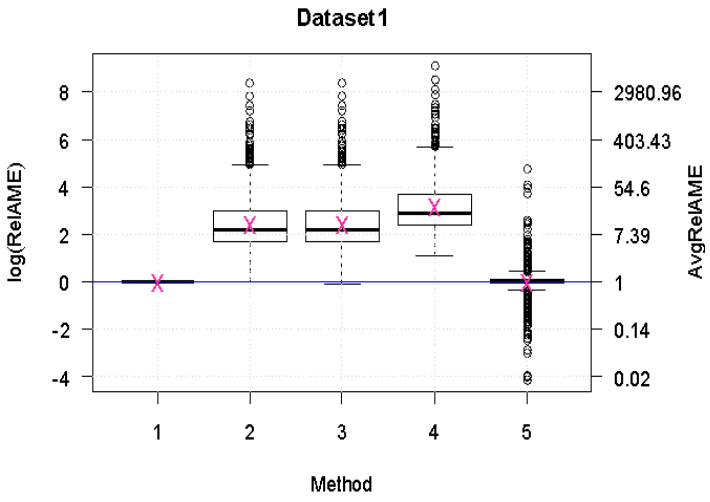

Figure 7. AvgRelAME-boxplot with double-scale (benchmark: Method 1)

Note: Method 1 is unbiased, the relative performance of alternative (biased) methods is therefore extremely high.

Table 5. AvgRelAMEs and true expected ratios for 'Dataset1'

Indicator	Method, j				
	1	2	3	4	5
$AvgRelAME_i$, benchmark: ***Method 1***	**1.00**	11.53	11.51	23.29	**1.00**
Desirable value: $\lvert\mathbb{E}[ME_{i,j}]\rvert/\lvert\mathbb{E}[ME_{i,1}]\rvert$	**1**	∞	∞	∞	**1**
$AvgRelAME_i$, benchmark: ***Method 2***	**0.09**	1.00	1.00	2.02	**0.09**
Desirable value: $\lvert\mathbb{E}[ME_{i,j}]\rvert/\lvert\mathbb{E}[ME_{i,2}]\rvert$	**0**	1	1	2	**0**

Note: Bold font shows the best methods in terms of *mean bias* depending on the indicator. Index i denotes any series.

To avoid dividing by near-zero AMEs, one quick solution is to use another (biased) method as the benchmark or to use the MSE of the benchmark method as the benchmark. Nonetheless, if some methods in a dataset are unbiased, it is difficult to use the AvgRelAME or AvgRelAMdE as proxies for ME and MdE. Additionally, these AvgRel-metrics do not show the direction of bias, only the magnitude. Moreover, if n_i approaches 1, AvgRelAME and AvgRelAMdE will converge to the GMRAE and will no longer be good proxies for ME and MdE.

Table 6 shows the results of applying the AvgRel-metrics described above to 'Dataset2'. Here we used Method 3 as the benchmark to avoid the extreme AvgRelAMEs resulting from the unbiasedness of the benchmark. The results correspond exactly to what should be expected. Method 2 (median-unbiased) delivers the best MAE and MdE, whereas Method 1 (mean-unbiased) delivers the best MSE and ME.

In the subsections that follow we continue our search for improved metrics that provide an indication of both the direction and the magnitude of bias and avoid the problem of extreme values when benchmark forecasts are unbiased.

Table 6. Results of applying the AvgRel-metrics to 'Dataset2'

Target loss function	Optimised forecast should be found as	What is measured?	By representing a proxy for	AvgRel-metric, benchmark: ***Method 3***	Method				
					1	2	3	4	5
Linear	Median of forecast density	Accuracy under linear loss	MAE	AvgRelMAE	0.87	**0.84**	1.00	0.85	1.21
		Median bias	MdE	AvgRelAMdE	0.34	**0.17**	1.00	0.21	1.68
Quadratic	Mean of forecast density	Accuracy under quadratic loss	MSE	AvgRelMSE	**0.87**	0.90	1.00	0.95	1.33
			RMSE	AvgRelRMSE	**0.93**	0.95	1.00	0.98	1.15
		Mean bias	ME	AvgRelAME	**0.31**	0.60	1.00	1.01	2.26

Note: the best methods in terms of the corresponding AvgRel-metric are indicated in bold.

7.5 The Relative ME (RelME)

One well-known approach to make errors scale-independent is to divide them by the time series mean (see Hyndman and Koehler, 2006). Some studies (e.g., Medina and Tian, 2020, p. 1015) have used the Relative Mean Error (RelME) metric which adopts this approach. We will use the following formula for the RelME for series i and method j:

$$RelME_{i,j} = \frac{ME_{i,j}}{\overline{Y_i}},$$

where $\overline{Y_i} = \frac{1}{n_i}\sum_{t\in T_i} Y_{i,t}$, assuming $Y_{i,t} \geq 0$ and $F_{i,t,j} \geq 0$.

One problem of the RelME is the risk of obtaining extreme cases and skewed underlying distributions due to dividing by a small denominator (see Davydenko and Fildes, 2016, p. 245).

Another problem is that averaging the RelME using the arithmetic mean is prone to biases. In particular, if $n_i = 1$, the RelME becomes the PE and therefore has all the disadvantages described for MPE. Generally, even if forecasts are unbiased on the original scale, the arithmetic mean of RelMEs may still indicate bias (as was the case for PEs), making the results counter-intuitive.

7.6 The LnQ-Metric

Tofallis (2014, p. 2) advocated the use of the LnQ metric defined as the logarithm of the ratio of the predicted value to the actual value. In our notation, the LnQ is:

$$\mathrm{Ln}Q_{i,t,j} = log\frac{F_{i,t,j}}{Y_{i,t}}.$$

Tofallis (2014) notes the following properties of Q. Firstly, "Q is the complement of the relative error: 1—(relative error), and so apart from the shift of one unit, will have the same distribution as the relative error". Secondly, Q is asymmetric because its value is bounded from below by zero, whereas it is unbounded from above. To overcome this asymmetry problem the logarithm of Q is used, obtaining the LnQ. LnQ can be viewed as *log*[1-(relative error)], where relative error is the percentage error divided by 100. When the geometric mean of Q is 1, this indicates that predictions are unbiased "in relative terms" (Tofallis, 2014, p. 4). However, a value of Q=1 does not directly correspond to the case when ME=0. In other words, the geometric mean of Q is not always a good indicator of (1-RelME). For example for 'Dataset2' we know that Method 1 is mean-unbiased (see Table 2), but the mean of its LnQ is 0.12 (the geometric mean of Q is 1.13), which indicates the presence of bias. Another limitation of the LnQ is that it assumes only positive values.

7.7 The Average Relative Mean Error (AvgRelME)

In the approach below we replace the 'relative error' in LnQ with the RelME (in order to obtain a good proxy for ME), then calculate the weighted arithmetic mean of *log*(1-RelME), and then transform the variable back so that we obtain an estimate for the RelME. Equivalently, the geometric mean can be used. We also apply the principles we defined for the AvgRel-metrics, such as using the same sample for the numerator and denominator and using forecasts optimised under quadratic loss. The metric proposed is the Average Relative Mean Error (AvgRelME) defined (for method j) as

$$AvgRelME_j = 1 - \left(\prod_{i=1}^{N}(1 - RelME_{i,j})^{n_i}\right)^{1/\sum_{i=1}^{N} n_i}.$$

A value of zero for the AvgRelME indicates no *mean bias*. A negative value indicates positive *mean bias*, that is a tendency to forecast too high by an amount equal to |AvgRelME| × [time series mean], while a positive value indicates a tendency to forecast too low by AvgRelME × [time series mean]. Our experiments show that the AvgRelME can serve as a good representation of the RelME (assuming at least one non-zero forecast within each series to ensure that RelME<1, which is needed to use the geometric mean). Fig. 8 shows boxplots of $log(1 - RelME_{i,j})$. The underlying distribution has the desirable property of symmetry. Table 7 shows that the AvgRelME works as expected.

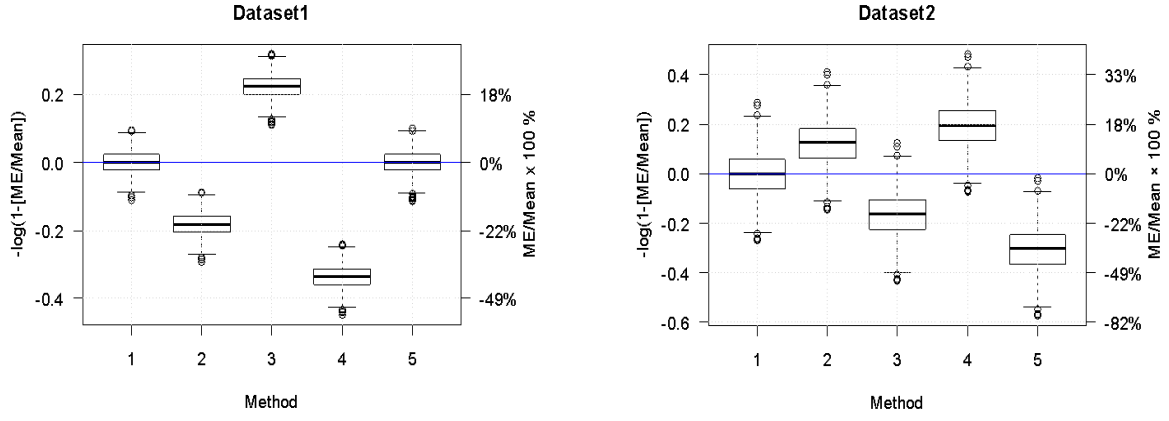

Figure 8. AvgRelME-boxplots. The AvgRelME correctly identified the magnitude and the direction of *mean bias*

Table 7. AvgRelMEs and the desirable ratios

Indicator	Dataset1					Dataset2				
	Method, j					Method, j				
	1	2	3	4	5	1	2	3	4	5
$AvgRelME_i \times 100$	**0.00**	-20.01	19.94	-40.10	**0.00**	**0.00**	11.50	-18.18	17.60	-36.07
$\mathbb{E}[ME_{i,j}]/\mathbb{E}[Y_{i,t}] \times 100$	**0**	-20	20	-40	**0**	**0**	11.75	-17.84	17.84	-35.68

Note: Bold font shows the best methods in terms of *mean bias* depending on the indicator. Index i denotes any series, index t denotes any time period.

7.8 The Average Relative Median Error (AvgRelMdE)

By analogy to the AvgRelME, we propose the following proxy for MdE in order to measure *median bias* (the Average Relative Median Error, AvgRelMdE):

$$AvgRelMdE_j = 1 - \left(\prod_{i=1}^{N}(1 - RelMdE_{i,j})^{n_i}\right)^{1/\sum_{i=1}^{N} n_i},$$

$$RelMdE_{i,j} = \frac{MdE_{i,j}}{MdY_i},$$

where MdY_i - sample median for $Y_{i,t}$, $t \in T_i$, assuming non-negative actuals and forecasts and $MdY_i > 0$. Also, in order to use the geometric mean, we need to assume that the Relative Median Error (RelMdE) is less than 1. Equivalently, this means the following: $MdE_{i,j} < MdY_i$. These assumptions may limit the application of the AvgRelMdE to intermittent demand data as there may be time series where $MdY_i = MdE_{i,j} = 0$. For such series, however, we can assume that $RelMdE_{i,j} = 0$, which then allows the calculation of the AvgRelMdE.

Fig. 9 shows boxplots of $log(1 - RelMdE_{i,j})$. The underlying distributions have the desirable property of symmetry. Table 8 shows that the AvgRelMdE works as expected, confirming its *construct validity*.

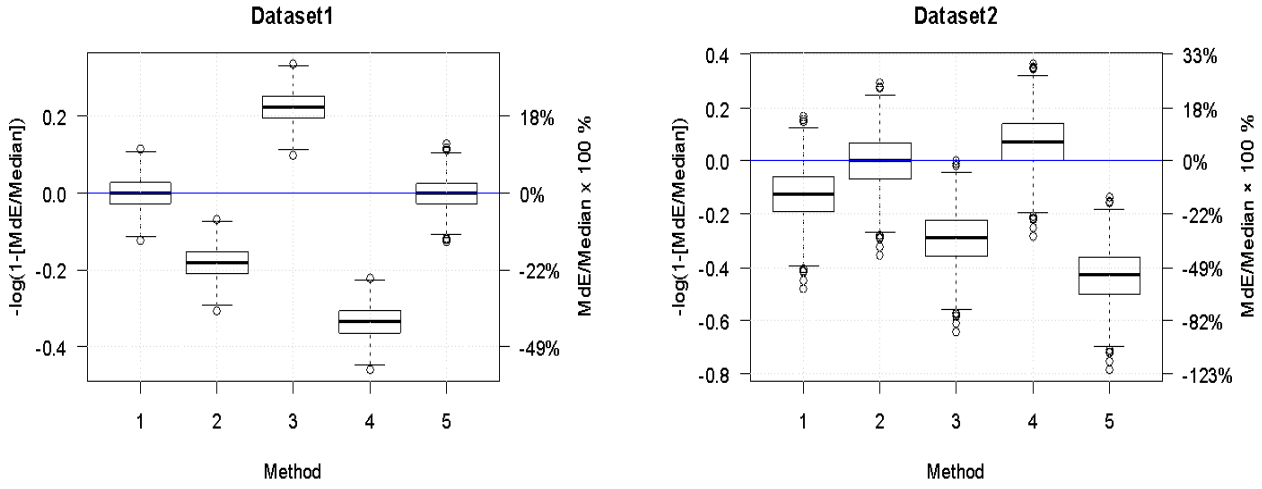

Figure 9. AvgRelMdE-boxplots

Note: The AvgRelMdE correctly identified the magnitude and the direction of *median bias*.

Table 8. AvgRelMdEs and the desirable ratios

Indicator	Dataset1					Dataset2				
	Method, j					Method, j				
	1	2	3	4	5	1	2	3	4	5
$AvgRelMdE_i \times 100$	**0.00**	-20.03	19.98	-40.00	**0.00**	-13.39	**-0.01**	-33.62	6.84	-53.85
$\mathbb{E}[MdE_{i,j}]/$ $Median[Y_{i,t}] \times 100$	**0**	-20	20	-40	**0**	-13.31	**0**	-33.53	6.9	-53.74

Note: Bold font shows the best methods in terms of *median bias* depending on the indicator. Index i denotes any series, index t denotes any time period.

7.9 The Overestimation Percentage Corrected (OPc) Calculated Across Multiple Series and Horizons

One potential problem with the AvgRelMdE is that it can be found only when series contain non-negative values. Also, the calculation of the AvgRelMdE is relatively complex.

In order to obtain a simpler metric for *median bias*, we propose to apply the *Overestimation Percentage corrected* (OPc), which we introduced earlier, to multiple series. The OPc for a given method j calculated across all series is:

$$OPc_j = OP_j + ZP_j/2,$$

where OP_j - the number of overestimates for method j over the whole dataset divided by total number of cases and expressed in percentages, ZP_j - the same for zero errors. If evaluation is made across multiple horizons, forecasts relating to all horizons are used.

We can also construct the following formula using the sign function:

$$OPc_j = 100\% \times \frac{1}{\sum_{i=1}^{N} n_i} \sum_{i=1}^{N} \sum_{t \in T_i} 0.5(1 - sgn(e_{i,t,j})).$$

If method j is median-unbiased, we should expect OPc_j to be close to 50%.

The OPc is immune to outliers, scale independent, easy to interpret, and shows the direction of bias. The OPc can be applied in different settings including negative observations, count data and intermittent demand data. The major limitation of the OPc, however, is that it will only tell about the frequency of overestimates or underestimates, but not about their magnitude. This is why additional metrics, such as the AvgRelMdE or the AvgRelAMdE may still be needed. Nonetheless, the OPc is a simple yet powerful tool to detect the presence of *median bias*.

To perform a statistical test to see if the OPc differs from 50%, we can use the binomial test (to compare the probabilities of over- and underestimation), assuming independent forecast errors. If this assumption does not hold (we can, for example, use the runs test to check this), alternative approaches may be possible. For example, if errors within series are not independent, but those in different series are, tests could be carried out on the hypothesis that mean the OPc is 50%, with the series, rather than individual cases, treated as the sampling units.

We propose the following visual tools to indicate the OPc where multiple series are involved. Firstly, we can use boxplots to explore the distribution of OPc across series, as shown on Fig. 10. The graph on Fig. 10 features the line representing median-unbiasedness. For both datasets we can see that OPc worked as expected showing values very close to 50% for median-unbiased methods. The corresponding results are presented in Table 9.

Alternatively, we can use the OPc barchart-diagram shown Fig. 11 featuring error bars and the line indicating the OPc of a median-unbiased forecast. The error bars indicate confidence intervals (CIs) for the probability of overestimation given non-zero error. To approximate the CIs for a population proportion one can use the well-known z-score formula (see, e.g., Illowsky and Dean, 2014).

The y-axis for the OPc-based graphs should have the lower limit of 0% and upper limit of 100%, as shown on Fig. 10 and Fig. 11, in order to allow a better readability.

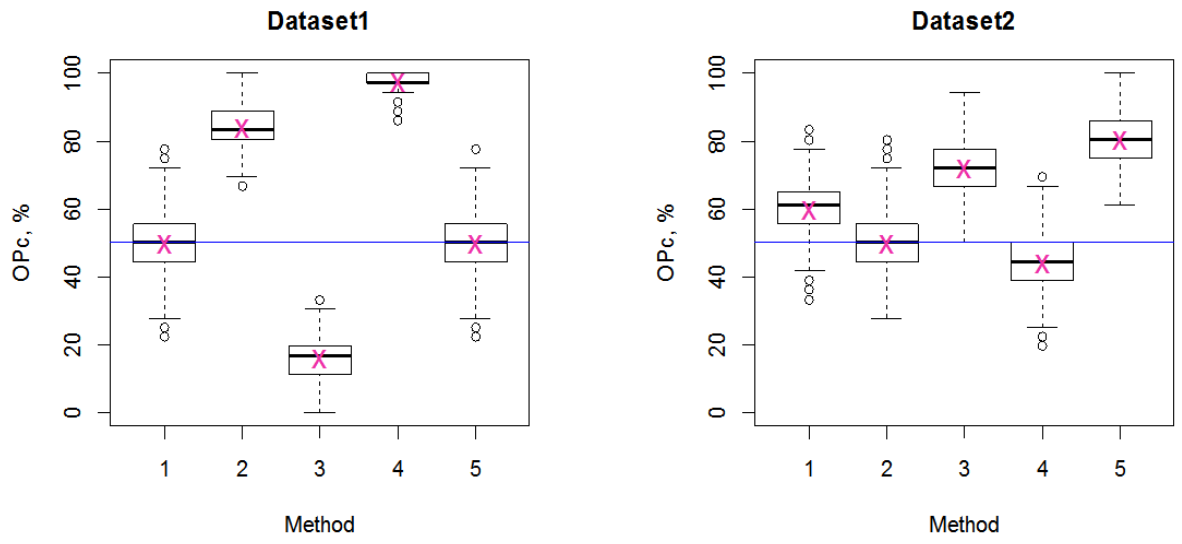

Figure 10. OPc-boxplots. The OPc correctly identified the presence and the direction of *median bias*

Table 9. OPc and true overestimation rates

Indicator	Dataset1					Dataset2				
	Method, j					Method, j				
	1	2	3	4	5	1	2	3	4	5
OPc_j, %	**50.00**	84.07	15.91	97.71	50.12	59.89	**50.00**	71.98	44.31	80.40
True overestimation rate × 100	**50.00**	84.13	15.86	97.72	**50.00**	59.87	**50.00**	71.85	44.31	80.52

Note: The "true overestimation rate" is the frequency of overestimated actuals for a very long series (we used 10^8 actuals). Bold font indicates the best method in terms of *median bias* depending on the indicator.

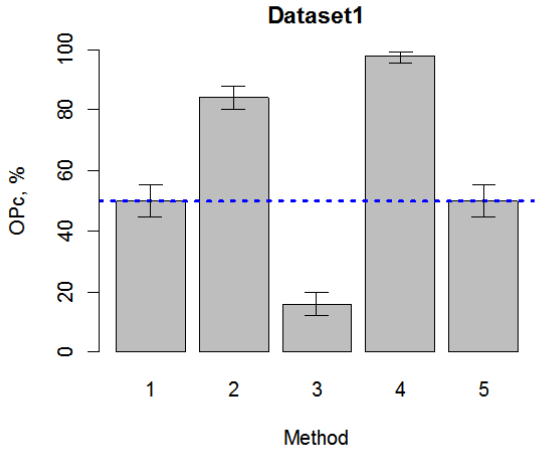

Fig. 11. OPc-diagram for 'Dataset1'

Note: As expected, since Methods 1 and 5 are median-unbiased, they have OPc near 50%. A reduced sample (the first 10 elements from each series) was used to construct this diagram in order to make the error bars clearer. The error bars indicate 90% CIs for the probability of overestimation.

7.10 Pooled Prediction-Realization Diagrams

To further explore the distribution of actuals, forecasts, and errors we can use the prediction-realisation diagram (PRD) proposed by (Theil, 1966). The PRD is a scatterplot with forecasts on the *x*-axis and outcomes on the *y*-axis. A '*Y=X*' line depicts perfect forecasts. The PRD is especially useful when exploring the presence of *regression bias*. When many methods and many horizons are available, we can use a pooled version of the diagram with different colors and marks representing different methods, as used in the variant introduced by (Davydenko et al., 2021, p. 89), which we will refer to as *pooled PRD*. Interestingly, with a good choice of colors and markers, even when many series are shown, this plot is still useful. An example is shown in Fig. 12 where the results of the M3 dataset (Makridakis and Hibon, 2000) are displayed on one graph. Alternatively, we can plot forecast errors against forecasts (see Davydenko, 2012, p. 96).

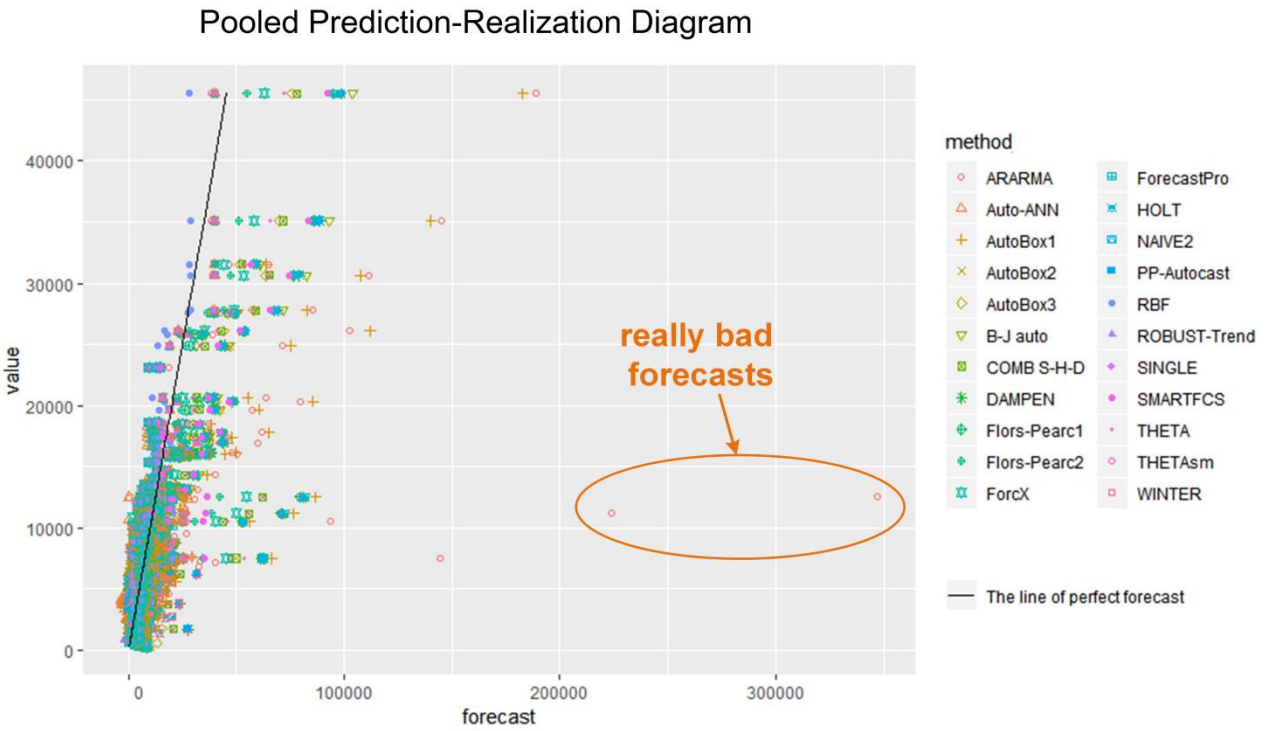

Fig. 12. The *pooled prediction-realization diagram* for M3 competition dataset

Note: Fig. adapted from (Davydenko et al., 2021, p. 89).

8. Forecast Evaluation Workflows (FEWs)

For the *point forecast evaluation setup* we defined earlier, we propose two alternative step-by-step procedures (workflows) for forecast evaluation and comparison depending on the loss function used to optimise and compare forecasts. The term *forecast evaluation workflow* (FEW) will denote a set of sequential activities aiming to ensure a comprehensive, informative, and reliable forecast evaluation. Here we assume non-negative actuals and forecasts allowing the calculation of the AvgRelME and AvgRelMdE. For the OPc, however, these assumptions are not needed. The workflows presented below are largely based on the workflow proposed by (Davydenko et al., 2021, pp. 99-100). One important aim of providing these workflows is to avoid conflicting results obtained using alternative metrics (e.g., AvgRelMAE and AvgRelRMSE).

Table 10 provides a workflow assuming that forecasts were optimised for the symmetric linear loss function. This workflow is further abbreviated as FEW-L1.

Table 11 shows the second workflow that is designed for the evaluation and comparison of forecasts in terms of the symmetric quadratic loss (FEW-L2).

Table 10. Steps for FEW-L1

Evaluation and Comparison of Forecasts in Terms of the Symmetric Linear Loss (FEW-L1)	
Step	Activities
Step 1	If forecasts are obtained using statistical models allowing producing forecast densities, make sure point forecasts are found as medians of these densities. If forecasts are first obtained using a Box-Cox transformation or its special cases (such as log- or sqrt-transformation) and then transformed back, the back-transformed forecasts can be used for this workflow since median-unbiasedness is invariant to such transformations (see Davydenko and Fildes, 2016, p. 240).
Step 2	Use the *pooled prediction-realization diagram* (PPRD, see Fig. 12 above) to explore forecast data, see if data was loaded correctly, and detect potential data flaws. Use individual time series plots (such as the *fixed origin* or *fixed horizon* plots presented in Davydenko et al., 2021) to explore series with pronounced outliers.
Step 3	Use the AvgRelMAE-boxplots (Fig. 5) to explore accuracy across series, and the AvgRelMdE- and OPc-boxplots (Fig. 9 and Fig. 10) to explore *median bias* across series. Use plots for individual series to explore time series with unusual values of RelMAE, RelMdE, or OPc.
Step 4	Report accuracy in terms of the AvgRelMAE, report *median bias* in terms of the AvgRelMdE and OPc. Use statistical tests to compare AvgRelMAE against 1, AvgRelMdE against 0, and OPc against 50%. In order to construct a test to compare the AvgRelMAE against 1, (Davydenko and Fildes, 2016, p. 246) suggested the use of the Wilcoxon one-sample signed rank test applied to *log*(RelMAE). A similar approach can be used to compare the AvgRelMdE against 0. In this case the distribution of *log*(1-RelMdE) should be explored. With regard to possible tests for the OPc, see notes in Section 7.9.
Step 5	Use "accuracy *vs* horizon" or "bias *vs* horizon" plots, if relevant (see Davydenko et al., 2021, pp. 93-100, for illustrative examples and more details).
Step 6	Interpret the results: AvgRelMAE<1 means improvement in comparison with the benchmark. OPc≠50% or AvgRelMdE≠0 means the possibility of improvement using better statistical modelling. Scatterplots showing the *log*(RelMAE) *vs log*(time series mean) and *log*(1-RelMdE) *vs log*(time series mean) dependencies may also be useful to explore the heterogeneity between series. Plotting $OPc_{i,j}$ *vs log*(time series mean) may also help understand the reasons for obtaining biased estimates.

Table 11. Steps for FEW-L2

Evaluation and Comparison of Forecasts in Terms of the Symmetric Quadratic Loss (FEW-L2)	
Step	Activities
Step 1	If forecasts are obtained using statistical models allowing producing forecast densities, make sure point forecasts are obtained as means of these densities. If forecasts are first obtained on a transformed scale and then transformed back, this workflow will not adequately show the accuracy of alternative methods, use FEW-L1 instead.
Step 2	The same as for FEW-L1.
Step 3	Use the AvgRelMSE-boxplots (by analogy to Fig. 5) and AvgRelME-boxplots (Fig. 8) to explore accuracy and bias across series. Use plots for individual series to examine unusual cases. Alternatively, the AvgRelRMSE may be used instead of the AvgRelMSE. The RelRMSE may be interpreted as an estimate of the relative variance of forecast errors.
Step 4	Report accuracy in terms of the AvgRelMSE, report *mean bias* in terms of the AvgRelME. Use statistical tests to compare AvgRelMSE against 1, AvgRelME against 0. See notes for Step 4 for FEW-L1 on how to construct statistical tests for the AvgRelMSE and AvgRelME.
Step 5	The same as for FEW-L1.
Step 6	Interpret the results: AvgRelMSE<1 means improvement in comparison with the benchmark. AvgRelME≠0 suggests the possibility of improvement through better statistical modelling. Scatterplots showing the *log*(RelMSE) *vs log*(time series mean) and *log*(1-RelME) vs *log*(time series mean) dependencies may also be useful to explore the heterogeneity between series.

9. Conclusions

This paper makes the following contributions to the fields of applied statistical analysis and forecasting.

Firstly, we defined the *point forecast evaluation setup* (PFES) assuming the specific settings where an aggregated set of forecast data is available and it is needed to evaluate accuracy and bias across series.

Secondly, by expanding the previous research done by (Davydenko and Fildes, 2016), we formulated a set of criteria for an ideal error measure in the context of the PFES. Namely, ideally, a measure should: (i) be easy to interpret, (ii) be robust to occasional unusual observations, (iii) be applicable for various data domains (for example, including those with negative actuals or zero errors), (iv) provide useful information for forecast evaluation and comparison, (v) adequately reflect the cost function used to optimise forecasts, (vi) be scale-independent, (vii) have *construct validity*, (viii) be easy to implement, and (ix) be easy to understand, and communicate.

Thirdly, given the setup and the above criteria, we conducted simulation experiments to evaluate the appropriateness of alternative error measures. Importantly, we developed a special simulation design where for each time series we generated actuals and forecasts with identical features of errors for each method. We then compared estimates based on various measures with the true parameters used to generate errors. Our experiments showed that existing measures can be counterintuitive due to imperfections of their design and they do not meet many of the above criteria. In particular, we demonstrated that use of the MPE is generally not advisable and the AvgRelAME, LnQ, and AMScE have their own limitations. In particular, our special emphasis was on *construct validity* and the ease of interpretation.

Fourthly, we proposed improved measures and visual tools for detecting bias: the AvgRelAMdE, AvgRelME, AvgRelMdE, OPc, AvgRel-boxplots, OPc-boxplots, and the OPc-diagram. These tools help analysts to detect problematic series and to compare forecasting performance with regard to *mean* and *median bias* where multiple series are involved.

Among most simple procedures to evaluate the presence of bias we recommend the *Overestimation Percentage corrected* (OPc) as an indicator of *median bias*. The OPc metric is immune to outliers, very easy-to-interpret, meets the criteria of *construct validity*, and also allows for the implementation of a simple statistical test. More complex, but more informative measures involve the AvgRelME and AvgRelMdE, serving as proxies for ME and MdE, respectively. We introduced special variants of boxplots showing the underlying distribution of AvgRel-metrics (AvgRel-boxplots) to ensure a more reliable analysis. We also recommend the use of the *pooled prediction-realization diagram* showing data across series on one plot.

Bias in practice can depend on many factors and should be evaluated with regard to a specific loss function. A general test with regard to many factors can be found in (Davydenko, 2012, p. 85), but here our aim has been to provide simple, concise and easily interpretable assessment procedures. The indicators proposed should help analysts to detect the most serious deviations from the desirable properties of forecasts.

Finally, we defined two detailed workflows (named FEW-L1 and FEW-L2) depending on the loss function of interest and provided a guide to the interpretation of measurement results. The result is a statistical framework (in the sense of the definition proposed in Davydenko and Charith, 2020) including the settings, criteria, methods and tools, and the workflow for the particular task of measuring forecast bias. Software implementation is straightforward and can be based on the flexible data formats proposed in (Davydenko et al., 2021). Almost any software environment (including Microsoft Excel) can be used to implement the simple methods proposed, but we recommend `R` because it allows flexible implementation of visual tools.

The workflows proposed can be directly used in a wide range of settings ranging from inventory control applications to climate forecasting. In particular, these workflows are useful for inventory control settings and the case of intermittent demand data where non-negative actuals and forecasts are assumed.

Although our focus has been on forecasting and time series data the methods are also applicable for regression analysis, especially for panel data and multi-target regression. Our suggested procedures are applicable both to academic researchers who are developing and evaluating new forecasting methods and practitioners wishing to evaluate the current forecasting performance of their organisation. The framework presented allows the preparation of reports in accordance with FVA-principles and methodologies for carrying out data science projects.

Supplementary Materials

Supplementary materials for this paper (including the datasets, implementation scripts, and additional results) can be found in the following GitHub repository: https://github.com/anvdavy/fcBiasData

List of Abbreviations

Abbreviation	Meaning	Section where defined or first used
AMdE	Absolute Median Error	7.4
AME	Absolute Mean Error	7.4
AMScE	Absolute Mean Scaled Error	7.2
AvgRelAMdE	Average Relative Absolute Median Error	7.4
AvgRelAME	Average Relative Absolute Mean Error	7.4
AvgRelMAE, AvgRelMSE, AvgRelRMSE	Average Relative MAE, MSE, RMSE, respectively	7.4
AvgRelMdE	Average Relative Median Error	7.8
AvgRelME	Average Relative Mean Error	7.7
AvgRelP	Average Relative Performance	7.3
FEW	Forecast Evaluation Workflow	8
FVA	Forecast-value-added	2
MAE	Mean Absolute Error	7.2
MASE	Mean Absolute Scaled Error	7.2
MdE	Median Error	5.3
MdPE	Median Percentage Error	7.1
ME	Mean Error	5.1
MPE	Mean Percentage Error	7.1
MScE	Mean Scaled Error	7.2
MSE	Mean Squared Error	7.3
OP	Overestimation Percentage	5.3
OPc	Overestimation Percentage corrected	5.3 (individual time series) and 7.9 (multiple time series)
PE	Percentage Error	7.1
PFES	Point Forecast Evaluation Setup	2
PPRD	Pooled PRD	7.10
PRD	Prediction-Realization Diagram	7.10
RelAMdE	Relative Absolute Median Error	7.4
RelAME	Relative Absolute Mean Error	7.4
RelMdE	Relative Median Error	7.8
RelME	Relative Mean Error	7.5
RelMAE, RelMSE, RelRMSE	Relative MAE, MSE, RMSE, respectively	7.4
RelP	Relative Performance	7.3
RMSE	Root Mean Squared Error	7.3
ZP	Percentage of zero errors	5.3 and 7.9

References

Ameen, J. R. M., & Harrison, P. J. (1984). Discount weighted estimation. *Journal of Forecasting*, *3*(3), 285-296. https://doi.org/10.1002/for.3980030306

Brown, G. W. (1947). On small-sample estimation. *The Annals of Mathematical Statistics, 18*(4), 582–585. https://doi.org/10.1214/aoms/1177730349

Davydenko, A. (2012). *Integration of judgmental and statistical approaches for demand forecasting: Models and methods* (doctoral dissertation). Lancaster University, UK. https://doi.org/10.13140/RG.2.2.31788.62083

Davydenko, A., & Charith, K. (2020, July 29-30). *A visual framework for longitudinal and panel studies (with examples in R)* [ePoster]. IRCUWU-2020 online conference. https://doi.org/10.6084/m9.figshare.12749432

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. https://doi.org/10.1016/j.ijforecast.2012.09.002

Davydenko, A., & Fildes, R. (2016). Forecast error measures: Critical review and practical recommendations. In: M. Gilliland, L. Tashman, & U. Slavo (Eds.), *Business forecasting: Practical problems and solutions* (pp. 238-250). Chichester: Wiley. ISBN: 111922456X, 9781119224563. http://dx.doi.org/10.13140/RG.2.1.4539.5281

Davydenko, A., Fildes, R. A., & Trapero, A. (2010). *Judgmental adjustments to demand forecasts: Accuracy evaluation and bias correction* (Lancaster University Management School Working Paper 2010/03). Lancaster University. https://eprints.lancs.ac.uk/id/eprint/48981/1/Document.pdf

Davydenko, A., Sai, C., & Shcherbakov, M. (2021). Forecast evaluation techniques for I4.0 systems. In A.G. Kravets, A.A. Bolshakov, & M. Shcherbakov (Eds.), *Cyber-physical systems: Modelling and intelligent control* (pp. 79-102). Springer, Cham. https://doi.org/10.1007/978-3-030-66077-2_7

DeGroot, M. (1970). *Optimal statistical decisions*. New York**:** McGraw-Hill.

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. https://doi.org/10.1016/j.ijforecast.2008.11.010

Fildes, R., & Goodwin, P. (2021). Stability in the inefficient use of forecasting systems: A case study in a supply chain company. *International Journal of Forecasting*, *37,* 1031-1046. https://doi.org/10.1016/j.ijforecast.2020.11.004

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. https://doi.org/10.1145/5666.5673

Gilliland, M. (2008). *Forecast value added analysis: Step-by-step* [White paper]. SAS Institute.

Goodwin, P. (1997). Adjusting judgemental extrapolations using Theil's method and discounted weighted regression. *Journal of Forecasting*, *16*(1), 37-46. https://doi.org/10.1002/(SICI)1099-131X(199701)16:1<37::AID-FOR647>3.0.CO;2-T

Goodwin, P. (2000). Correct or combine? Mechanically integrating judgmental forecasts with statistical methods. *International Journal of Forecasting*, *16*(2), 261-275. https://doi.org/10.1016/S0169-2070(00)00038-8

Goodwin, P. (2018). *Profit from your forecasting software: A best practice guide for sales forecasters*. Hoboken, NJ: Wiley. https://doi.org/10.1002/9781119415992

Hill, A. (2012). *The encyclopedia of operations management: A field manual and glossary of operations management terms and concepts*. FT Press Operations Management.

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International journal of forecasting*, *22*(4), 679-688. https://doi.org/10.1016/j.ijforecast.2006.03.001

Illowsky, B., & Dean, S. (2014). *Collaborative Statistics*. Retrieved from https://cnx.org/contents/XgdE-Z55@40.9:qzyOSfZa@20/Confidence-Interval-for-a-Population-Proportion

Johnston, J. (1972). *Econometric methods*. New York: McGraw-Hill.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). The M4 Competition: Results, findings, conclusion and way forward. *International Journal of Forecasting*, *34*(4), 802-808. https://doi.org/10.1016/j.ijforecast.2018.06.001

Medina, H., & Tian, D. (2020). Comparison of probabilistic post-processing approaches for improving numerical weather prediction-based daily and weekly reference evapotranspiration forecasts. *Hydrology and Earth System Sciences*, *24*(2), 1011-1030. https://doi.org/10.5194/hess-24-1011-2020

Nikolopoulos, K., Fildes, R., Goodwin, P., & Lawrence, M. (2005). *On the accuracy of judgmental interventions on forecasting support systems*. Lancaster University Management School Working paper 2005/022.

Petropoulos, F., Goodwin, P., & Fildes, R. (2017). Using a rolling training approach to improve judgmental

extrapolations elicited from forecasters with technical knowledge. *International Journal of Forecasting*, *33*(1), 314-324. https://doi.org/10.1016/j.ijforecast.2015.12.006

Sanders, N. R., & Graman, G. A. (2009). Quantifying costs of forecast errors: A case study of the warehouse environment. *Omega*, *37*(1), 116-125. https://doi.org/10.1016/j.omega.2006.10.004

Spiliotis, E., Doukas, H., Assimakopoulos, V., & Petropoulos, F. (2021). Forecasting week-ahead hourly electricity prices in Belgium with statistical and machine learning methods. In *Mathematical modelling of contemporary electricity markets* (pp. 59–74). https://doi.org/10.1016/b978-0-12-821838-9.00005-0

Theil, H. (1966). *Applied economic forecasting*. Amsterdam: North Holland Publishing Company.

Tofallis, C. (2014). A better measure of relative prediction accuracy for model selection and model estimation. *Journal of the Operational Research Society*. Advance online publication. https://doi.org/10.1057/jors.2014.103

van der Vaart, H. R. (1961). Some extensions of the idea of bias. *The Annals of Mathematical Statistics, 32*(2), 436–447. https://doi.org/10.1214/aoms/1177705051

Zellner, A. (1986). *An introduction to Bayesian inference in econometrics*. Wiley Classics Library. 629, New York: Wiley.