FORESIGHT

The International Journal of Applied Forecasting

ARTICLES *by* FORECASTERS *for* FORECASTERS: Q3:2024

Point Forecast Evaluation: State of the Art

Join the *Foresight* readership by becoming a member of the International Institute of Forecasters
forecasters.org/foresight/

Point Forecast Evaluation: State of the Art

IVAN SVETUNKOV

PREVIEW *Of the dozens of error metrics available to evaluate forecasting performance, mean absolute percentage error (MAPE) is the most commonly used – but also the most maligned. MAPE's many flaws are well recognized, but it remains appealing because it is easy to understand by business management. In this article, Ivan Svetunkov looks at alternative approaches to point forecast evaluation that are more appropriate than MAPE.*

There are many articles covering forecast evaluation in general, and issues with the mean absolute percentage error (MAPE) in particular. But repetition is the mother of learning, so I will use this article to consider major issues in forecast evaluation, and to summarize the state of the art in evaluating point forecasts.

WHY EVALUATE FORECASTS?

The main motivation for evaluating forecasts is to monitor and improve our forecasting process. Practitioners are familiar with evaluating the historical performance of their forecasts *after* the actual values are known. Organizations commonly measure and report "forecast accuracy" after each period (e.g., month, week, or day) of their planning cycle. But forecast evaluation also has a key role in selecting an appropriate method or model with which to generate the forecasts.

When choosing from among alternative methods or models, evaluation begins by splitting the available data (e.g., a time series of historical demand) into two parts. First is the *Training Set* – where we fit our candidate models and estimate their parameters. For example, if we have four years of monthly historical demand, we might select the first 36 months as the training set, upon which we build our candidate models. The most recent 12 months then becomes the *Test Set*, over which we evaluate the performance of the candidate models.

Proper evaluation provides the information needed for decision making: whether to change the model, manually adjust forecasts after they have been generated, or amend the overall forecasting process. It is clear that forecast evaluation is important, but we need to be aware of what specifically we are measuring.

WHAT ARE WE MEASURING?

One way to evaluate forecasts is by measuring their accuracy, i.e., understanding how close the forecasts are to the actual value that was forecasted. There are, of course, other ways to evaluate forecasts, such as by measuring bias, but they lie outside of the scope of this article.

When it comes to forecast accuracy, there are dozens of error measures readily available in both open source and commercial software. Common practice has been to leave selection of the error measure to individual preference, based on aspects like simplicity and robustness. Several academic papers claimed over the years that there is no such thing as a single best metric (for example, Koutsandreas and colleagues, 2022). But the modern approach is to start by understanding what specifically we want to measure.

For example, we now realize that if the point forecast generated by our model corresponds to the mean, we should use an error measure based on the root mean squared error (RMSE). On the other hand,

if we produce median forecasts (which is rarely the case for point forecasts), we should use the mean absolute error (MAE). This is because RMSE is minimized by mean, while MAE is minimized by median (Kolassa, 2020).

Why is using the right measure so important? Because if we use a wrong error measure (e.g., use MAE when the model produces mean forecasts), we might end up using a forecasting approach that is inferior to alternatives and looks much better than it should. A classic example is evaluating intermittent demand forecasts, which I illustrate in **Figure 1**.

Here, there are two forecasts: the global mean (blue line), and the zero forecast (red line). Arguably, the mean forecast is more suitable for decision making, because it gives us an idea of how much we can sell on average in the future. For this example, I calculated RMSE and MAE, which are shown in **Table 1**.

Table 1. Error Measures for the Mean and Zero Forecasts for Our Example

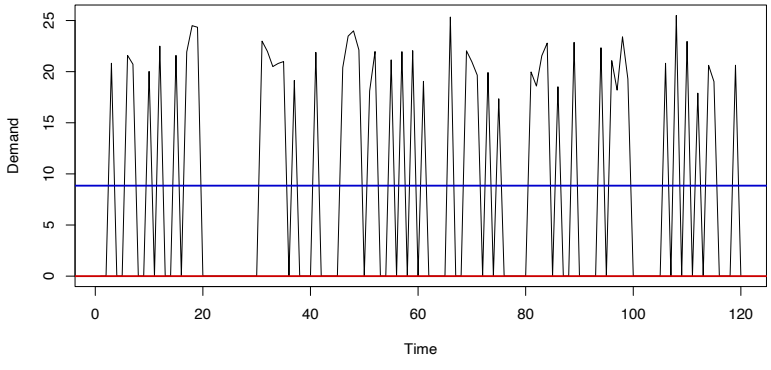
	MAE	RMSE
Mean	10.32	10.54
Zero	8.85	13.76

We see that the zero forecast minimizes MAE in this case, because more than 50% of the observations are zeroes (i.e., the median of the data is zero). At the same time the RMSE is minimized by the mean. So, if we were to use MAE for the forecast evaluation in this situation, we would conclude that the best forecast is a zero forecast, implying that nobody will purchase our product in the future (and therefore, we should not replenish our inventory). This is neither true nor useful for decision making. Thus, it is important to select an appropriate metric for forecast evaluation that measures what you expect it to measure. In the majority

Key Points

- Evaluation is important for tracking forecast process performance and understanding whether changes (to forecasts, models, or the overall process) are needed.
- Understand what kind of forecast our models produce, and measure it properly. Most likely, our approach produces the mean (rather than the median) as a point forecast, so root mean squared error (RMSE) should be used to evaluate it.
- To aggregate the error measure across several products, you need to scale it. A reliable way of scaling is to divide the selected error measure by the mean absolute differences of the training data. This way we get rid of the scale and units of the original measure and make sure that its value does not change substantially if we have trend in the data.
- Avoid MAPE!
- To make decisions based on your error measure, consider using the FVA framework, directly comparing performance of your forecasting approach with the performance of some simple benchmark method.

Figure 1. Intermittent Demand Pattern and Two Forecasts: Global Mean and Zero (Median)



of cases it should be RMSE, because the typical forecasting approach produces mean as a point forecast.

FORECAST EVALUATION ACROSS PRODUCTS

RMSE can only be applied to individual time series, for example, on the SKU level. If we calculate it across several SKUs, the aggregation may not provide a good indication of performance, because the sum of apples, oranges, and beer crates is meaningless. Even for a group of similar products, a simple average of the error measures is not a good representation of the overall group performance. Forecasts for products sold in thousands of units will have forecast errors of thousands of units as well. Errors for the low-volume products will be in the scale of tens of units. Averaging errors across high- and low-volume products will mask the performance of models on the low-volume ones.

The most common approach to more appropriately evaluate forecasts across aggregations is to *weight* the individual SKU evaluations by a common unit of measure, such as unit volume, or monetary value. The familiar mean absolute percentage error simply averages the APE across all SKUs. But a "weighted MAPE" (WMAPE) is influenced by the relative size or importance of the individual SKUs. WMAPE should thereby provide a slightly better indicator of overall group performance than the severely flawed MAPE (whose issues are discussed in the next section).

The modern best practice for forecast evaluation across aggregations is to *scale* the error measure to get rid of units and bring everything to a similar level. The main idea is that for a forecast produced for a test set, we scale the error measure by using (in the denominator) something either from the training set or from a different model. This way the error measure will not be deflated by individually very high actual values or inflated by values close to zero.

One candidate for the denominator is a simple arithmetic training set mean, which works fine for level data, especially if we deal with intermittent demand. If we divide MAE (also known as MAD) by the in-sample mean, we will end up with something called "scaled MAE" (sMAE) by Petropoulos and Kourentzes (2015), or the "MAD/MEAN ratio" by Kolassa and Schütz (2007). This is also known to practitioners as "weighted MAPE." It has an interpretation roughly similar to the classical MAPE, showing by how many percent the forecast deviates from the actual values on average. But as I mentioned in the previous section, in the majority of cases we should use RMSE instead of MAE, which can be easily fixed by changing the numerator. We will end up with the scaled RMSE (sRMSE). This error measure also has a relatively straightforward interpretation, also showing roughly the mean percentage error relative to the average sales.

The modern best practice for forecast evaluation across aggregations is to scale the error measure to get rid of units and bring everything to a similar level.

The problem of scaling by the training set mean is that in the presence of the trend, the mean might change with the addition of new observations. As a result, if the sales of a product grow, the denominator of such error measure will increase, which can lead to the decline of the sRMSE over time, even if the model starts performing worse than before. To resolve this, Hyndman and Koehler (2006) proposed to divide the error measure by the mean absolute differences of the data (i.e., use the period-to-period changes in sales instead of using the original data). This way we end up with so-called "mean absolute scaled error" (MASE) or "root mean squared scaled error" (RMSSE). This addresses one issue, but at the cost of interpretability. While error expressed as

a percent is easy to understand, it is now harder to explain what, for example, a value like 1.14 means in that error measure. While MASE or RMSSE might not be suitable for a report to management, these values are useful to the forecast modelers. This is because during performance evaluation and model selection, the modelers typically need to compare error measures between models, not to aim at getting values lower than some arbitrary threshold.

An alternative approach for scaling is to calculate the error measure from some benchmark method (such as naive, or an arithmetic mean of the series) and divide our measure by it. The resulting "relative RMSE" (rRMSE) first proposed by Stock and Watson (2005) addresses the issues with scaling and has the added benefit of a straightforward interpretation: it shows by how much one approach is better than the benchmark. For example, if rRMSE is 0.85, this means that our method is better by 15%.

The downside of this approach appears when the error measure either in the numerator or denominator becomes close to zero. In these cases, the rRMSE becomes either close to zero or to infinity, respectively. Besides, the denominator in such measure might change with the change of sample size. And in that situation, an increase of rRMSE over time might imply either the increase of the forecast error of our approach, or a decrease of the forecast error of the benchmark method. It is not possible to tell the difference between the two.

ISSUES WITH MAPE

But what's wrong with the good old mean absolute percentage error? As a reminder, here's how MAPE is calculated across a set of point forecasts and the actual values:

MAPE=mean(|forecast-actual|/actual)

The original rationale behind MAPE is clear: we need to get rid of scale, and we want something that is easy to calculate and interpret. But we also want something that measures accuracy well. Unfortunately, MAPE is fraught with issues:

1. Foremost, it is not clear what minimizes MAPE. Stephan Kolassa (2016) provided an insight for a special case if the data follows Log-Normal distribution. But in reality, we cannot count on data following some specific theoretical distribution, so the minimum of MAPE is a mystery.
2. MAPE is scale sensitive. If you have sales in thousands of units, then the actual value in the denominator will bring the overall measure down and you will have a very low number even if the model is not doing well. Similarly, if you deal with very low volumes, they will inflate the measure, making it easily hundreds of percent, even if the model does a very good job. This also means that the measure breaks on intermittent demand. And it also means that if your sales have strong seasonality, MAPE might tell you that your models are doing a very good job in summer (when the sales are high) and a very bad one in winter (with low volume of sales), even if in reality the performance across the year is consistent.
3. It is well known that MAPE prefers when you underforecast (Fildes, 1992). It is not symmetric, and it can be misleading. A proposed modified version called "symmetric MAPE" was found to be no better and was not symmetric either (Goodwin and Lawton, 1999).
4. Yes, MAPE is easy to calculate and interpret. But the value itself is not a reliable indicator of performance of your model. MAPE can get the value of 1% either because your model is very accurate, or because you are dealing with the high-volume data (see point 2 above). Furthermore, if we change the units of the actual demand (e.g. measure it in thousands of units), the measure would change as well. So, even straightforward interpretation of the measure may be misleading.

Almost all the issues appearing above are because of the division of each individual value in the test set by the specific actual

value. Substituting the denominator by either the in-sample mean (in case of level data) or the mean absolute differences (as discussed above) fixes those issues.

FORECAST VALUE ADDED

In practice, we want to evaluate performance of models to decide which one to use and when to interfere and make some changes in the forecasting process. One of the ways of doing that, which has now become more commonly employed, is the "forecast value added" or FVA approach (Gilliland, 2010, 2023). Following the FVA framework, instead of trying to find an arbitrary threshold for the error measure, we should calculate the direct improvement our forecasting approach brings in comparison to a benchmark. The straightforward formula for FVA aligns with the rRMSE that we discussed above:

$$FVA=1-(Error_a/Error_b)$$

where $Error_a$ is the error measure (e.g., RMSSE) of the approach under consideration, while $Error_b$ is the error measure of a benchmark approach. Typically, some simple method is used for the benchmark, such as a naive, a global average, a simple moving average, or a method already used by the company. FVA tells us by how many percent the accuracy improves (or worsens) relative to the benchmark when an alternative approach is used.

In contrast with the rRMSE, FVA can be applied on the aggregate level after scaling the error measures to provide a clear indication that the new approach improves (or worsens) forecasting performance on average across all our products.

While FVA has straightforward practical implications, it also has some potential issues. These are similar to the ones discussed in the case of rRMSE, namely (1) the issue with the zero-error measure, and (2) a change in the FVA can be due to a change in performance of either the benchmark or the alternative approach. In contrast with the rRMSE, FVA can be applied on the aggregate level after scaling the error measures to provide a clear indication that the new approach improves (or worsens) forecasting performance on average across all our products.

CONCLUSIONS

I would like to repeat the main points from this short article, because repetition is the mother of learning:

1. Forecast evaluation is important to track the forecasting process and understand whether or not some changes are needed.
2. When evaluating forecasts, we need to understand what specifically our models produce and measure it properly. Most probably, your approaches produce mean as a point forecast, so you should use root mean squared error (RMSE).
3. If you want to aggregate the error measure across several products, you need to scale it. And one of the reliable ways of scaling is to divide the selected error measure by the mean absolute differences of the training data. This way we will get rid of the scale and units of the original measure and make sure that its value does not change substantially if we have trend in the data.
4. Avoid MAPE!
5. If you want to make decisions based on your error measure, consider using the FVA methodology, directly comparing performance of your forecasting approach with the performance of some simple benchmark method.

Following these steps will ensure that you are using the state-of-the-art methodology for forecast evaluation, developed over the years through research and practice in the forecasting community.

REFERENCES

Fildes, R. (1992). The Evaluation of Extrapolative Forecasting Methods, *International Journal of Forecasting*, 8(1), 81-98.

Gilliland, M. (2023). 20 Years of FVA: A Critical Retrospective, *Foresight*, Issue 71, 10-17.

Gilliland, M. (2010). *The Business Forecasting Deal*, Wiley.

Goodwin, P. & Lawton, R. (1999). On the Asymmetry of the Symmetric MAPE, *International Journal of Forecasting*, 15(4), 405-408.

Hyndman, R.J. & Koehler, A.B. (2006). Another Look at Measures of Forecast Accuracy, *International Journal of Forecasting*, 22(4), 679-688.

Kolassa, S. (2020). Why the "Best" Point Forecast Depends on the Error or Accuracy Measure, *International Journal of Forecasting*, 36(1), 208-211.

Kolassa, S. (2016). Evaluating Predictive Count Data Distributions in Retail Sales Forecasting, *International Journal of Forecasting*, 32(3), 788-803.

Kolassa, S. & Schütz, W. (2007). Advantages of the MAD/MEAN Ratio Over the MAPE, *Foresight*, Issue 6, 40-43.

Koutsandreas, D., Spiliotis, E., Petropoulos, F., & Assimakopoulos, V. (2022). On the Selection of Forecasting Accuracy Measures, *Journal of the Operational Research Society*, 73(5), 937-954.

Petropoulos, F. & Kourentzes, N. (2015). Forecast Combinations for Intermittent Demand, *Journal of the Operational Research Society*, 66(6), 914-924.

Stock, J.H. & Watson, M.W. (2004). Combination Forecasts of Output Growth in a Seven-Country Data Set, *Journal of Forecasting*, 23(6), 405-430.

Ivan Svetunkov is an Associate Professor of Marketing Analytics at Lancaster University (UK), where he earned a PhD in management science. His main area of interest is statistical learning for demand forecasting. He is a creator and a maintainer of several forecasting- and analytics-related R packages and the author of many papers and a monograph "Forecasting and Analytics with the Augmented Dynamic Adaptive Model" (***openforecast.org/adam/***).

i.svetunkov@lancaster.ac.uk

Recent Advances in Supply Chain Forecasting: A workshop in memory of Professor John E. Boylan

Lancaster University, UK on June 13-14, 2024

The workshop was a great opportunity to share wonderful memories of John and to reflect on the profound impact he has made on forecasting, inventory management, supply chain and other areas. With many talks from academics and practitioners as well as a tutorial on stochastic lead times, the workshop followed John's footsteps to stimulate new research ideas and inspire Early Career Researchers.

This article originally appeared in *Foresight*, Issue 74 (forecasters.org/foresight) and is made available with permission from *Foresight* and the International Institute of Forecasters.