# The Myth of the MAPE . . . and how to avoid it

Hans Levenbach, PhD, Executive Director – CPDF Training and Certification Program; URL: www.cpdftraining.org

## In the Land of the APEs, Is the MAPE a King or a Myth?

Demand planners in supply chain organizations are accustomed to using the Mean Absolute Percentage Error (MAPE) as their best answer to measuring forecast accuracy. It is so ubiquitous that it is hardly questioned. I do not even find a consensus on the definition of the underlying forecast error among supply chain practitioners participating in the demand forecasting workshops I conduct worldwide.

For some, Actual (A) minus Forecast (F) is the forecast error, for others just the opposite. If bias is the difference, what is a positive versus a negative bias? Who is right and why? Among practitioners, it is a jungle out there trying to understand the role of the APEs in the measurement of accuracy. Bias is one component of accuracy, and precision measured with Absolute Percentage Errors (APEs) is the other. I find that accuracy is commonly measured just by the MAPE.

Outliers in forecast errors and other sources of unusual data values should never be ignored in the accuracy measurement process. With the measurement of bias, for example, the calculation of the mean forecast error ME (the arithmetic mean of Actual (A) minus Forecast (F)) will drive the estimate towards the outlier. An otherwise unbiased pattern of performance can be distorted by just a single unusual value.

When we deal with forecast accuracy in practice, a demand forecaster typically reports averages of quantities based on forecast errors (squared errors, absolute errors, percentage errors, etc.). To properly interpret a measure of forecast accuracy, we must also be sensitive to the role of unusual values in these calculations. For example, consider the following set of numbers, representing 12 absolute percentage forecast errors of monthly forecasts made over 1 year:

{1.1%, 1.6%, 4.7%, 2.1%, 3.1%, 32.7%, 5.8%, 2.6%, 4.8%, 1.9%, 3.7%. 2.6%}

By ranking the data from smallest to largest, we obtain the ordered set:

{1.1%, 1.6% 1.9%, 2.1%, 2.6%, 2.6%, 3.1%, 3.7%, 4.7%, 4.8%, 5.8%, 32.7%}

Over the twelve monthly periods, the mean absolute percentage error (MAPE = 5.6%) can be viewed as a typical percentage error for a month. The median absolute percentage error (MdAPE = [2.6 + 3.1]/2 = 2.9%), on the other hand, is an outlier-resistant measure and gives quite a different answer. Note that the arithmetic mean has been severely distorted by the outlying value 32.7%. The arithmetic mean of the numbers when we exclude the outlier is 3.1% and, like the MdAPE, appears to be much more typical of the underlying data. Overall, an average absolute percentage error for the 12 months is more likely to be around 3% per month than around 6% per month, in this example. Among practitioners, the MAPE is routinely reported for business planning purposes to summarize forecast performance.

What should a demand planner and forecaster do in practice? Although the arithmetic mean is the conventional estimator of central tendency, forecasters and business planners should not accept it uncritically. As our example illustrates, one outlier can have an undue effect on the arithmetic mean and pull an estimate of the bulk of data away from a representative or typical value. It is always best to calculate and compare *multiple* measures for the same quantity to be estimated, just to be assured that you are not misled by the arithmetic mean. If the measures are

practically close, you report the conventional measure. If not, you check out the APEs for anything that appears unusual. Then work with domain experts to find a credible rationale (stockouts, weather, strikes, etc.)

### Item-Level versus Aggregate Performance

Forecast evaluations are also useful in multi-series comparisons. Production and inventory managers typically need demand or shipment forecasts for hundreds to tens of thousands of items (SKUs) based on historical data for each item. Financial forecasters need to issue forecasts for dozens of budget categories in a strategic plan on the basis of past values of each source of revenue. In a multi-series comparison, the forecaster should appraise the method based not only on its performance for the individual item but also on the basis of its overall accuracy when tested over various summaries of the data. How to do that most effectively will depend on the context of the forecasting problem and can, in general, not be determined a priori. In the next section, we discuss various measures of forecast accuracy.

## Are There Better Accuracy Measures Than the MAPE?

There are certainly better measures than the MAPE and they have been around for at least half a century. Nowadays with cheap and fast computing, there is little reason not to explore and analyze data. Data-driven approaches (in contrast to theory-driven techniques) can be found in some textbooks, like *Forecasting: Practice and Process for Demand Management* (2005) by Levenbach and Cleary**.** However, with a few exceptions, forecast researchers have not considered data-driven metrics in the context of forecast accuracy measurement. Even the notion of forecast accuracy is not well understood and frequently misinterpreted by supply chain planners and managers.

So, we start with the basic questions: What is a forecast error? Is it A-F or F-A? What defines forecast precision: A/F or F/A? You can start by checking out a freely downloadable White Paper **Taming Uncertainty: All You Need to Know about Measuring Forecast Accuracy, But Are Afraid to Ask**

### The Mean APE is Not the Best King of the APEs

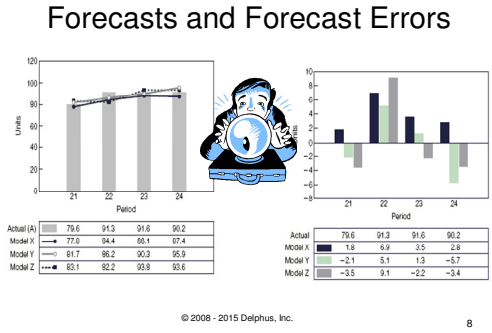
To start on the same page, two important aspects of forecast accuracy measurement are **bias** and **precision**. **Bias** is a problem of direction: Forecasts are typically too low (downward bias) or typically too high (upward bias). **Precision** is an issue of magnitudes: Forecast errors can be too large (in either direction) using a particular forecasting technique. What is a **forecast error**? It is not unusual to find inconsistent definitions and different misinterpretations for concepts surrounding accuracy measurement among analysts, planners and managers in the same company. Each deviation represents a **forecast error** (or sometimes called a forecast *miss)* for the associated period:

$$\text{Forecast error (E)} = \text{Actual (A)} - \text{Forecast (F)}$$

Contrast this with a fitting error (or **residual**) of a forecasting model over a (historical) fit period, which is

$$\text{Fit error (E)} = \text{Actual (A)} - \text{Model fit (}\hat{A}\text{)}$$

If (F – A) is the preferred usage in some organizations but not others, then demand forecasters and planners should give (F – A) a different name, like **forecast variance**, which is the more conventional usage with revenue-oriented planners. The distinction is important because of the interpretation of bias in under- and overforecasting situations.




The exhibit shows (a) bar charts and (b) tables showing Actuals (A), Forecasts (F) and forecast errors for three forecasting techniques. The most common averages of absolute values are mean absolute error (MAE), mean absolute percentage error (MAPE), and median absolute percentage error (MdAPE).

The interpretations of the averages of absolute error measures are straightforward. From the exhibit, we calculate that the MAE is 4.6 for technique Z, from which we can conclude that the forecast errors from this technique average 4.6 per period. The MAPE is 5.2%, which tells us that the period forecast errors vary around 5.2%.
The MdAPE for technique Z is approximately 4.1% (the average of the two middle values: 4.4 and 3.8). Thus, half the time the forecast errors exceeded 4.1%, and half the time they were smaller than 4.1%. When there is a serious outlier among the forecast errors, as with technique Z, it is useful to know the MdAPE in addition to the MAPE because medians are less sensitive (more resistant) than mean values to distortion from outliers. This is why the MdAPE is a full percentage point below the MAPE for technique Z. Sometimes, as with technique Y, the MdAPE and the MAPE are virtually identical. In this case, we can report the MAPE because it is the far more common measure.

## Introducing M-Estimators: The Need for Nonconventional Methods

The need for a nonconventional approach to estimation is motivated by two problems. First, a forecaster never has an accurate knowledge of the true underlying distribution of the random errors in a model. Second, even slight deviations from a strict parametric model can give rise to poor statistical performance for classical (i.e., associated with the ordinary least-squares method) estimation techniques. Estimators that are less sensitive to these factors are referred to as *robust* by statisticians.

The M-estimation method can be used to reduce automatically the effect of outliers by giving them a reduced weight when we compute the typical value for a dataset. The method is based on an estimator that makes repeated use of the underlying data in an iterative procedure. One example is the Huber distribution (original source paper: *Robust estimation of a location parameter* by Peter Huber, 1964), which behaves like a normal distribution in the middle range and like an exponential distribution in the tails. Thus, the bulk of the data appears normally distributed but there is a greater chance of having extreme observations in the sample.

In the case of the MAPE, a family of robust estimators, called M-estimators, is obtained by minimizing a specified function of the absolute percentage errors (APE). Alternate forms of the function produce the various M-estimators. Generally, the estimates are computed by iterated weighted least squares.

Now, for the technical details if you want to implement this, you can use such function $\xi$ (.), which takes the form:

$$\xi(e) = 0.5\, e^2 \quad \text{if } |e| \le c = Ks$$

$$= c\,|e| - 0.5\, e^2 \qquad \text{if } |e| > c = Ks,$$

where $s$ is a scale estimate, such as UMdAD, the (unbiased) median absolute deviation from the median, divided by 0.6745.

If the APE data are normally distributed and the number of observations is large, the divisor 0.6745 is used because then s approximates the standard deviation of the normal distribution.

Usually the sample standard deviation is **no**t used as a value because it itself is influenced too much by outliers and thus is not resistant against them. The constant K is chosen to obtain a desired level of efficiency (another desirable statistical property), and it is often set to between 1 and 2. If K is sufficiently large, the M-estimate will be equivalent to ordinary least-squares estimates.

Minimizing $\xi(e)$ yields an estimate of location; the minimization requires *Huber weights*, defined by

$$W_i = 1 \qquad \text{if } |e_i| \le Ks$$
$$W_i = Ks / |e_i| \qquad \text{if } |e_i| > Ks$$

The statistic *s* approximates the standard deviation of a normal distribution, and the constant K is chosen to some number close to 1.5 (based on empirical evidence).

The iterative procedure involves the following steps:

1. Obtain an initial estimate of the typical APE. The initial estimates can be obtained in a number of ways such as the mean or median APE.
2. Compute $e_i$ and calculate Ks = K (UMdAD), where UMdAD is defined above.
3. Compute the weights and perform the weighted average.
4. Repeat steps 1 through 3 until convergence or a reasonable number of iterations.

A second *M*-estimator, called the *bisquare estimator*, gives zero weight to data whose residuals are quite far from zero (original source book: Mosteller and Tukey, **Data Analysis and Regression**, 1977). The *Bisquare weighting function* is defined by

$$W_i = 0 \qquad \text{if } |e_i| > Ks$$
$$W_i = [1 - (e_i / Ks)^2]^2 \qquad \text{if } |e_i| \le Ks.$$

It is worth noting that the bisquare-weighting scheme is more severe than the Huber scheme. In the bisquare scheme, all data for which $|e_i| \le Ks$ *will* have a weight less than 1. Data having weights greater than 0.9 are not considered extreme. Data with weights less than 0.5 are regarded as extreme, and data with zero weight are, of course, ignored.

*To counteract the impact of outliers, the bisquare estimator gives zero weight to data whose residuals are quite far from zero*

## A Sample Calculation

The table shows the calculations for the Huber M-estimator of location and its corresponding standard error for a data set {-67, -48, 6, 8, 14, 16, 23, 24, 28, 29, 41, 49, 56, 60, 75}. The median of the data is 24. Column 2 is the difference between the forecast errors and the median of the forecast errors. The median absolute deviation MdAD is 17; it is the median of the absolute value of this column (the eighth largest absolute value in column 2). An approximate unbiased scale statistic is the UMdAD = MdAD/0.6745 = 25.2.

The calculations of the Huber M-estimator of Location (K = 2, and $\theta_0 = 24$, median) are shown below

| $Y_{(i)}$ | $Y_{(i)} - \theta_0$ | $w^2_{I1}$ | $Y_{(i)} - \theta_1$ | $w^2_{I2}$ |
|---|---|---|---|---|
| -67 | -91 | 50.4/91 | -92.13 | 50.8/92.13 |
| -48 | -72 | 50.4/72 | -73.13 | 50.8/73.13 |
| 6 | -18 | 1 | -19.13 | 1 |
| 8 | -16 | 1 | -17.13 | 1 |
| 14 | -10 | 1 | -11.13 | 1 |
| 16 | -8 | 1 | -9.13 | 1 |
| 23 | -1 | 1 | -2.13 | 1 |
| 24 | 0 | 1 | -1.13 | 1 |
| 28 | 4 | 1 | 2.87 | 1 |
| 29 | 5 | 1 | 3.87 | 1 |
| 41 | 17 | 1 | 15.87 | 1 |
| 49 | 25 | 1 | 23.87 | 1 |
| 56 | 32 | 1 | 30.87 | 1 |
| 60 | 36 | 1 | 34.87 | 1 |
| 75 | 51 | 50.4/51 | 49.87 | 1 |

Iteration 0 $\theta_0$ = Median = 24

Iteration 1 $s$ = MdAD / 0.6745 = 17/ 0.6745 = 25.20;
$Ks$ = 50.40

$\theta_1 = [\sum W^2_i y_i / \sum W^2_i] = 357.38 / 14.24 = 25.13$

Iteration 2 $s$ = MdAD / 0.6745 = 17.13 / 0.6745 = 25.40
$Ks$ = 50.80

$\theta_2 = [\sum W^2_i y_i / \sum W^2_i] = 358.72 / 14.25 = 25.17$

$V(\theta) = [\sum W^4_i (y_i - \theta_2)^2 / (n^*)^2]^{1/2} = [\sum W^4_i (y_i - 25.17)^2 / (13)^2]^{1/2} = 8.28$
where $n^*$ = number of observations receiving full weight

## Introducing the LHBB **T**ypical **A**bsolute **P**ercentage **E**rror (TAPE) for Measuring Precision

The normality assumption is widely used among practitioners to justify the use of the arithmetic mean. This justification is primarily made for the following reason. In the words of the original data scientist John W. Tukey (1915-2000): "Practice dictates a choice between what can be done and should be done. In the absence of anything better, normality usually implies what can be done. If we find something better, then we want to be more interested in what should be done."

What we need, for best practices, are robust/resistant procedures that are resistant to outlying values and robust against non-normal characteristics in the data distribution, so that they give rise to estimates that are more reliable and credible than those based on normality assumptions.

The Princeton Robustness Study (1972) led by John Tukey was an early effort to analyze systematically this concept for estimates of location. Estimates that have robust efficiency are often very resistant to outliers. This can be a very valuable consideration because real-life data are frequently non-normal and possess hard-to-detect outlying observations. This is becoming the challenge for the data scientist in the world of Big Data.

Taking a data-driven approach to using APE data to measure precision, we can create many TAPE measures. However, we recommend that you start with the MAPE or MdAPE for the first iteration. Then use the Huber scheme for the next iteration and finish with one or two more iterations of the Bisquare scheme. The LHBB (Levenbach Huber Bisquare Bisquare) TAPE measure has worked quite well for me in practice and can be readily automated in a spreadsheet.

### Detecting Unusual Values with Outlier-Resistant Measures

How can we detect and identify outliers automatically with outlier-resistant methods? In the APE data set {1.1%, 1.6%, 4.7%, 2.1%. 3.1%, 32.7%, 5.8%. 2.6%. 4.8%, 1.9%, 3.7%. 2.6%}, the standard deviation is 8.6, the MdAD is 1.1, and the IQD is 2.75. (The IQD is the Interquartile difference, or the difference between the 75$^{th}$ and 25$^{th}$ percentile in the distribution.) For normally distributed data, the standard deviation can be approximated by dividing the MdAD by 0.6745 (= 1.63) or by dividing the IQD by 1.35 (= 2.04).

We determine conventional outlier boundaries by calculating the mean plus and minus three standard deviations, which almost encompasses the suspected outlier (= 32.7%). This is because the calculation of the standard deviation is itself distorted by the outlier as it gives equal weight to all observations. However, if we consider the alternative means of determining outlier boundaries, the median plus and minus three times UMdAD, we get 8.3 (=3.4 + 3 * 1.63) for the upper limit. Now 32.7% clearly shows to be very unusual and to be far away from the bulk of the data!

## Try it on your own data and tell us if the Mean APE is still King in your jungle!

## Final Take-away

*"Practice dictates a choice between what can be done and should be done. In the absence of anything better, normality usually implies what can be done. If we find something better, then we want to be more interested in what should be done."*

JOHN W. TUKEY (1915-2000) – The Original Data Scientist

- When underlying data in a forecast accuracy measurement process suggest nonnormal distributions with possible outliers, robust/resistant methods should always be considered:
- Robust methods are a way of dealing with estimation and modeling problems in the presence of outliers and nonnormality.
- Resistant procedures are recommended as complements to the conventional procedures. When they are in agreement, they should be reported together. When substantial differences exist in the two analyses, the data should be examined thoroughly for outliers or unusual values. Even if you use robust/resistant techniques, you should still make graphical displays for a thorough examination of it.

What we desire, in practice, are robust procedures that are resistant to nonnormal tails in the data distribution, so that they give rise to estimates that are much better than those based on normality.