FORESIGHT

The International Journal of Applied Forecasting

ARTICLES *by* FORECASTERS *for* FORECASTERS: Q2:23

Special Feature: Is It Time to Retire the MAPE?

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

SPECIAL FEATURE:
IS IT TIME TO RETIRE THE MAPE?

Time to Retire the MAPE

MALTE TICHY

PREVIEW *Although forecasting researchers – and even most practitioners – have recognized the flaws in MAPE for many years, it remains a popular metric for reporting forecasting performance. This is largely because of its interpretability: MAPE is easy to understand by the business managers and executives who make decisions based on forecasts. But as more forecasting is done with small, intermittent quantities on a disaggregated level (such as product/location/day in retail), MAPE has become more and more problematic. First-time* Foresight *contributor Malte Tichy believes it's time to present MAPE with a plaque, a gold watch, and a retirement party, as the situations in which MAPE is a suitable metric have become increasingly rare.*

This article is an adaptation of a blog post published on
https://blog.blueyonder.com/mean-absolute-percentage-error-mape-has-served-its-duty-and-should-now-retire/

Introduction: The Gap between Academia and Practice

That the Mean Absolute Percentage Error, or MAPE, is unsuitable for evaluating intermittent or low-volume forecasts meets broad consensus among statisticians. MAPE's failings have been the subject of many previous *Foresight* articles, including Kolassa and Schütz (2007), Valentin (2007), Kolassa and Martin (2011), and Morlidge (2015). Still, expressing an aversion to MAPE induces skepticism and pushback from many practitioners – and, according to Gartner (2018), MAPE remains the most popular evaluation metric in business forecasting.

As forecasts on a product/location/day level in retail are becoming the standard, MAPE is becoming less and less suitable, expanding the gap between statistical theory and forecasting practice to much more than an academic issue. In this article I will walk you over a bridge across that gap, through the application of MAPE in a retail forecasting setting. I will not only tell you that it's a bad metric to use (I already have), but show you and let you convince yourself of its unexpected but serious pitfalls. The examples are aimed at helping you communicate the importance of proper forecast evaluation metrics and explain why there were good reasons for using MAPE 40 years ago in the first M Competition (Makridakis and colleagues, 1982), but there are even better reasons to move away from it now.

Key Points

- Despite consensus in the statistics community that MAPE (Mean Absolute Percentage Error) is a flawed metric and unsuitable in many common circumstances, it is still popular in practice, due to its simplicity and ostensible interpretability.
- While MAPE might be suitable when forecasting large quantities, it is harmful when used for intermittent or low-volume quantities. MAPE can deeply mislead decision makers, exaggerating some problems and disguising others, nudging them to choose forecasts with systematic bias.
- Model evaluation practice has not kept up with progress in forecasting. MAPE values – alone and without context – are not useful as an indicator of forecast quality.
- We simulate a supermarket that relies on a MAPE-optimizing forecast value fed into replenishment. The under- and overstocks in the fast- and slow-sellers quickly push the store out of business.

When Absolute and Relative Errors Contradict – Who Do You Trust?

You predicted a demand of 8 apples and 10 were eventually sold. You predicted 92 bottles of water and 108 were sold. You predicted 2 cans of tuna and 1 was sold. How do you judge these forecasting errors? A straightforward approach is to compute the absolute deviation of the prediction to the actual and divide by that actual. This can be expressed as a percentage value and is called the *absolute percentage error* (APE).

Coming up with APE as a first shot for "forecast quality evaluation" is quite typical. For the three examples, you obtain APEs of seemingly moderate 20% (=|8-10|/10), modest 15% (=|92-108|/108), and alarming 100% (=|2-1|/1), respectively. MAPE is the arithmetic mean of these three percentages, and amounts to 45%. These error percentages convey that the forecast on tuna is worse than the one on apples, and the forecast on water outperforms the others. But does this truly reflect forecast *quality*?

Look again at the beginning of this section – the large *absolute* difference between forecasted and actual bottles of water is worrisome, and its small percentage error cannot really reassure you. On the other hand, the 100% error on tuna could be due to random (bad) luck – it amounts to only a single item. Should you keep your intuition quiet, and blindly rely on the APEs? Consequently, should you revise the tuna forecast and leave the water bottle forecast as it is? If another set of forecasts is issued, with an overall MAPE of only 30%, is that new forecast necessarily better?

The inconvenient truth is that MAPE is unsuitable for forecasts on a granular level (i.e., intermittent or low-volume quantities), due to several intolerable and unsolvable problems. A forecast's MAPE doesn't tell us how good that forecast is, but how oddly APE behaves.

Consciously Ignoring Scale: When Percentage Errors Can Make Sense

Before diving into granular forecasting in retail (on product/location/day level), let's suppose to predict a much larger quantity: the yearly gross domestic product (GDP) of countries, measured in US$. Such a forecast might be used to define policies for entire countries, and these policies should be equally applicable to countries of different sizes. Therefore, it is fair to weight each country equally in this use case: a 5% error on the U.S. GPD (about 25 trillion US$) hurts just as much as a 5% error on the Tuvalu GPD (about 66 million US$, 380,000 times smaller than the U.S. GDP). Here, absolute percentage error (APE) makes sense: the actual GDP is never close to 0 (which would cause a terrible headache when dividing by it; I'll come to that below). And the forecast aim is not to get the overall GDP of the planet right, but to be as close as possible for each individual country, across scales ranging from millions to trillions.

Minimizing the total absolute error of the model (i.e., error in US$, not in percentages) puts the largest economies into the spotlight and disregards the small ones. It does not weight each country equally, but by its economic power. A model with a nice 3% error on the U.S. GDP and an unacceptable 200% error on the Tuvalu GDP would appear to be "better" than a model with 4% error on the U.S. GDP and 10% error on Tuvalu GDP in absolute US$ terms. MAPE, on the other hand, points toward using the latter forecast, which sacrifices a lot of absolute GDP accuracy on the U.S. (1% of 25 trillion US$) for a modest absolute improvement of the accuracy on Tuvalu (190% of 66 million US$). The U.S. GPD is much larger than Tuvalu's, but one would be consciously deciding to treat them equally when using MAPE. (Note that both the U.S. and Tuvalu can be considered "large" in the sense that one can't expect statistical fluctuations or "bad luck" to be responsible for forecast error. Deviations will typically be statistically significant and point toward model improvement potential.)

In summary, whenever single instances of a forecast of different values should be treated in an equal way, i.e., whenever we are fine with comparing enormous apples to minuscule oranges, MAPE can make sense. But is an equal treatment always fair?

The Impact of Scale on Achievable Forecast Accuracy

Let's return to our previous example in grocery and talk about apples, tuna cans, and bottles. Here, comparing APEs makes little sense, for two reasons.

By definition, a slow-seller sells less often than a fast-seller. Assuming equivalent financial characteristics of the two products, the business impact of an unreliable slow-seller forecast is much less severe than for an equally unreliable fast-seller forecast. A 5% sales loss due to stock-outs in some marginal slow-seller is merely inconvenient for the vendor, while a 5% sales loss on the best-selling item can be quite dramatic. At the end of the day, absolute numbers count for a business. You overpredict the total demand of your main product in the U.S. by 20%? You probably have a problem and need to deal with lots of unsold stock, which might put your entire business in jeopardy. You overpredict the total demand of that same product by 20% in Tuvalu? That error won't sink your business. You can tolerate much larger relative error in petty assortments or markets than in your high-volume/high-revenue categories, yet MAPE treats them all equally.

Adding to this obvious difference (small is small and large is large), there is a subtle but important statistical effect: scale-dependence of achievable forecast accuracy. Being 10% off for a product that sells 10 times a day is unavoidable sometimes (being off just one unit can give you that for such a low-volume item). But being 10% off on a product that sells 10,000 times a day clearly points towards a problem. Not only is the slow-seller less important businesswise than the fast-seller, but it naturally comes with larger relative errors (Tichy, 2022a, 2022b).

For the grocery forecasts above, you probably have just been unlucky regarding the tuna on that day. The 16 additional bottles of water seem less excusable. Therefore, absolute percentage error (APE) does not catch achievable forecast quality well, neither in business terms (it weights unequal things equally) nor in statistical terms (its achievable value needs the context of the forecasted value itself).

Governing Replenishment by MAPE Leads to Catastrophic Stock Levels

We have shown that, by itself and with no context, MAPE is not a good indicator of forecast quality. Whether 20%, 70%, or 90%, MAPE has no immediate interpretable meaning, and one cannot jump to any conclusion. Even accepting that a MAPE value, by itself, tells you little to nothing about your overall model quality, you might nevertheless expect that, for a given forecasting situation, the MAPE-winning forecast should be the best one. But as we'll see now, you need to also give up that weaker expectation:

the MAPE-optimal forecast answers an abstract mathematical question that you probably didn't ask.

Consider a supermarket that offers many different products – from slow-sellers that sell about once per quarter up to fast-sellers that sell 100 times per day.

To keep things simple, we'll focus on five exemplary products: apples, bananas, cashews, dragon fruits, and eggplants, with true mean daily selling rates of 0.01, 0.1, 1, 10, and 100. All items are ultra-perishable: either they are sold on their first day on the shelves or go to waste in the evening (so each day's closing inventory is always zero). The actual sales are distributed according to the Poisson distribution with the respective mean selling rate. The slowest, apples, sells about once per quarter; the fastest, eggplants, sells 100 times per day. (You are right if you suspect that numbers were not made up for real-world plausibility, but rather mathematical clarity and simplicity.) In this thought experiment, we know these selling rates, and they are the best possible estimate for the mean sales for each product by construction. Using the Poisson distribution to simulate actual sales, we can determine what is the forecast value with the best MAPE. Since there is no inventory carried over from the previous day (it has either been sold or gone to waste), let the replenishment of items be done by a system that picks the daily MAPE-optimal forecast and preorders according to it. That is, it chooses the forecast value for which the MAPE is lowest. How would that supermarket perform?

Poisson distribution. Let **n** customers visit a store; each considers buying a given item and does so with probability **p**. The number of sales of that item then follows a ***binomial*** distribution, with expected mean **p * n**. The ***Poisson*** distribution is the limit in which the number of customers **n** becomes large, while the expected number of sold items **p * n** is kept constant. For product-based forecasts in retail, a probabilistic Poisson forecast is an appropriate representation of sales.

For each product, **Table 1** shows the true selling rate (which is the unbiased best daily forecast), its simulated MAPE, and the optimized MAPE-winning forecast (along with simulated MAPE and resulting bias):

What happens if replenishment uses the MAPE-winning forecast? The supermarket overstocks on the slow-movers: for every day, one apple, one banana, and one cashew are replenished – but apples only sell once every 100 days and bananas once every 10 days! Apples and bananas go to waste almost every evening, cashews do so from time to time, while the demand of dragon fruits is not met: on average, more customers want to buy dragon fruits than are available. For the fast-moving eggplants, the 1% error might be excusable – nevertheless, it is striking that the "best" forecast is always biased, unless the true selling rate equals one.

The numbers computed for Table 1 reflect a perfect world in which forecasters can assume that the actual sales will follow a Poisson distribution around the predicted mean. For retail sales, the Poisson distribution is the narrowest distribution that can be achieved. For a more realistic model in which some moderate additional uncertainty (technically speaking: overdispersion, set to a value of 0.2) is present, the situation immediately looks worse in **Table 2**:

Table 1. Performance of MAPE-winning forecast

Product	True daily selling rate, unbiased daily forecast	MAE of true selling rate	MAPE- winning daily forecast	MAPE of MAPE-winning forecast	Forecast bias of MAPE- winning forecast
Apples	0.01	99%	1	0.25%	+9,900%
Bananas	0.1	90%	1	2.5%	+900%
Cashews	1	23.3%	1	23.3%	0%
Dragon fruits	10	31%	9	29%	-10%
Eggplants	100	8.11%	99	8.05%	-1%

The gap between the MAPE value computed at the true selling rate and the MAPE value of the MAPE-winning forecast has increased substantially. In other words, the user might think that

Table 2. Performance of MAPE-winning forecast when there is additional uncertainty

Product	True daily selling rate, unbiased daily forecast	MAE of true selling rate	MAPE- winning daily forecast	MAPE of MAPE-winning forecast	Forecast bias of MAPE- winning forecast
Apples	0.01	99%	1	0.3%	+9,900%
Bananas	0.1	90%	1	3%	+900%
Cashews	1	25%	1	25%	0%
Dragon fruits	10	73%	6	53%	-40%
Eggplants	100	49%	72	40%	-28%

the "evidence" that the MAPE-winning forecast is better than the other is even stronger than above. The MAPE-optimal forecast is, however, more strongly biased than in the ideal situation: the underforecasting in dragon fruits and eggplants now amounts to 40% and 28%, respectively – a repeated stock-out situation on many days would be the consequence. Below, we will see why more uncertainty means "we need to play safe" and why that means "we need to bias our forecasts low."

Clearly, a supermarket that runs with the MAPE strategy will not survive for long! The problems with MAPE thus go beyond the business *interpretability* (it's unsuitable to answer the question "how good is the forecast?"), but can potentially lead to severe *operational problems* (choosing a forecast value over another one that would be inarguably better for the application at hand). Let's next explore why, and how we could replenish that supermarket instead.

MAPE Censors Zero-Count-Events, with Catastrophic Consequences

It is well recognized that APE is problematic whenever there are periods where the actual is zero. This is because APE uses the actual in the denominator, and so is undefined. Of the many ways to deal with the problem of zero actuals (see the commentary by Kolassa [2023]), we will use the approach of simply removing the zero-sales periods from the data.

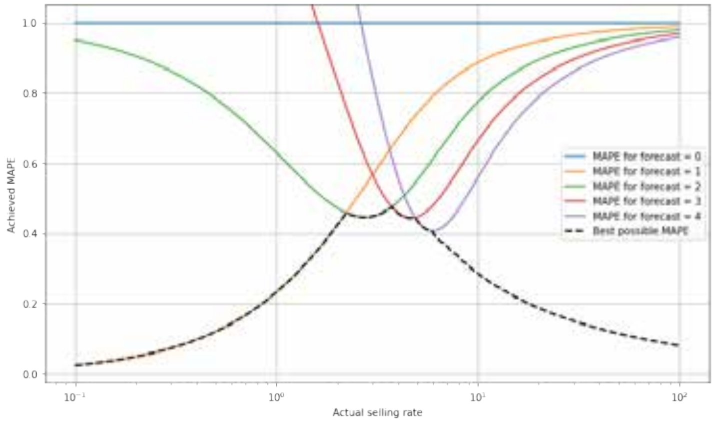
Unfortunately, this data removal is as bad as it feels: it leads to a blatant overprediction bias on super-slow-movers (which sell once or less per time period) in a MAPE-optimal prediction. Since periods with zero sales are ignored, the lowest reasonable prediction for any product, location, and day is 1 – even for a product that sells once per year! Since the MAPE-optimized forecast can safely ignore the outcome "0," playing safe is proposing "1" as lowest forecast value.

Alternatives to removal of periods with zero sales (e.g., assigning 100% error instead of removal) do not solve this problem. A prediction of 1.7 with outcome 0 is clearly less problematic than a prediction of 17,000 with outcome 0, and assigning those two events the same artificial APE of 100% makes no sense. Thus, whenever your data could plausibly contain "0" as actual value for any event, MAPE is extremely problematic. Optimizing it will lead to overpredictions in the super-slow-moving items – as we see in the first two rows of the tables.

MAPE Penalizes Under- and Overforecasts Differently, Leading to Skewed Estimates

Suppose you predict 1 and observe 7: the APE is $|1-7|/7 = 86\%$. Does that seem a lot to you? If so, exchange the numbers and predict 7, observe 1: your APE becomes $|7-1|/1 = 600\%$! APE penalizes an overprediction by a certain amount much more heavily than for an underprediction by the same amount. For underpredictions, the worst possible APE is 100%; for overpredictions, it is unbounded. As a result, since you will never be certain about the outcome, playing safe is biasing your forecast low. Avoid strong overforecasting at (almost) any cost, whereas some massive underforecasting will not break your neck.

Figure 1. "Mount MAPE" – the best achievable MAPE per selling rate for a Poisson-limited forecast



Even under minimal forecast uncertainty, which we assumed in Table 1, the MAPE-optimal forecast is an underprediction for selling rates above 1 (last two rows). Moreover, the larger the variability of the training data is (as in Table 2), the more uncertain the model, and the more will the MAPE-optimal forecast underforecast. The lesson again is that playing safe is forecasting low. The more uncertainty you face, the safer you want to be, and the lower the MAPE-optimal forecast becomes. This hedging against overpredictions leads to the strong bias in the last two rows of Table 2. This asymmetry is (somewhat) addressed by alternative flavors of MAPE, such as symmetric MAPE (sMAPE), where the percentage error is computed with respect to the mean of prediction and actual instead of actual only. However, even sMAPE doesn't fully solve the asymmetry (Goodwin and Lawton, 1999), and the various MAPE alternatives may induce other problems and paradoxes.

MAPE Exhibits Particularly Complex Scaling Behavior, Leaving Us Ignorant on How Good a Forecast Really Is

Admittedly, the lack of interpretability (is 50% MAPE good or bad?) is not an exclusive feature of MAPE. Every metric has scale-dependence and assumes different values for slow- and fast-movers. Nevertheless, the scaling of MAPE is especially troublesome due to the combination of the two aforementioned effects:

- A MAPE-optimal forecast will never output a number smaller than 1 when we just remove the 0-sale-outcomes;
- Relative errors decrease for large selling rates.

In **Figure 1**, we show "Mount MAPE," the best possible achievable MAPE as a function of selling rate, assuming the actual sales follow a Poisson distribution.

Let me explain what you see. The x-scale is logarithmic so we can observe small selling rates well – the scale goes from 0.1 to 100, super-slow to fast. For small selling rates below around 2, a forecast of 1 is the best possible; it yields the MAPE value given by the orange line that goes from the lower left (where it's overlaid by the black dashed line) to the upper right. The forecast 2 would lead to large MAPE in the slow-movers (green line), close to 95% for a selling rate of 0.1. The forecast 0 always leads to a constant MAPE of 100% (blue line): for any outcome that is not 0 (and those zero actuals have been removed from evaluation), we have APE=|actual-0|/actual=100%. At a selling rate of around 2.3, the forecast 2 becomes the optimal one, hence the black dashed line (the best possible MAPE) jumps from the orange to the green line. It further takes turns whenever the best forecast jumps from one value to the next (shown for forecast 3 and 4 in red and purple, respectively).

The best possible MAPE decreases when we go to very slowly moving items (to the left). Since 0-sales-events are removed from the data, the "surviving" events are mostly 1-sale-events, and even more so the more slowly the item sells. For a selling rate of 0.1, observing 2 items sold on a single day is already highly unlikely, and the forecast "1" is therefore, in most of the non-0-cases, perfect, and the achieved MAPE quite low. In other words, when you know that "0" will be removed from the data and the item is slow, then "1" is a safe bet for the number of sales that occur. For midsized values around 1 to 5, we see the "turn-taking" of the best possible MAPE. For large forecasts of 10 or higher (to the right-hand side of the plot), the achievable MAPE decreases again: the Poisson distribution becomes relatively narrow in the limit of large rates (Tichy, 2022a, 2022b).

I really did my best to explain the shape of “Mount MAPE.” It took me more than 300 words in two paragraphs, but I fear it might not be entirely successful. Did you understand it in such a way that you’ll be able to intuitively judge MAPEs in the future, in the context of predicted selling rates? If you don’t feel you will – don’t worry: this complexity is yet another argument that, even among professionals, it is unlikely that an intuitive correct judgement of MAPE-values ever becomes widespread.

The **median^(-1)**. Given a probability mass function P(k), the **median** of P is the best point estimate to optimize the forecasted value when evaluating using **Mean Absolute Error** (MAE). The median^(-1) is the best point estimator for MAPE: it’s computed using the artificial distribution P(k)/k, i.e., for which we divide each probability mass P(k) by the actual k, such that we need to exclude observing k=0. In other words, it’s the median of a skewed distribution that results from dividing the event probabilities P(k) by the event values k.

You don’t want to optimize an abstract mathematical function – you want to maximize business value.

MAPE-Optimal Forecasts Are Irrelevant to Business, Jeopardizing Potential Forecast Value

The forecast that wins at MAPE is not the unbiased forecast that you would wish in many applications. But what does it then mean to “optimize for MAPE”? Mathematically, the value that minimizes MAPE minimizes a cumbersome-looking expression, the median^(-1) (Gneiting, 2011). But that expression has no meaningful business interpretation. Whatever you want to achieve with your forecast – ensure availability, reduce waste, plan promotions and markdowns, replenish items, plan workforce, etc. – the business cost of a wrong forecast in your application is certainly not reflected by MAPE! Ideally, you should utilize an evaluation metric that reflects the actual financial cost of “being off.” You don’t want to optimize an abstract mathematical function – you want to maximize business value.

The Alternative: Let the Metric Directly Reflect Business

Apart from situations involving large numbers (like predicting GDPs country-wise) and under strong assumptions, MAPE is neither suitable to indicate how good a forecasting model is (due to scaling), nor a suitable decision driver to choose among two competing models (MAPE-winning forecasts are biased). What is the alternative? Optimally, the metric that is used directly reflects the business value. Mean Absolute Error (MAE) quantifies situations in which the cost of one overstocked item is the same as the cost of one missing item – a strong assumption, but certainly closer to reality than MAPE.

Let’s re-simulate the replenishment strategy of our supermarket using now the MAE-optimal value, the median of the

Table 3. MAE-optimal forecast value for Tables 1 and 2

Product	True daily selling rate, unbiased daily forecast	MAE-winning daily forecast (Poisson uncertainty [Table1])	MAE-winning daily forecast (larger-than-Poisson uncertainty [Table 2])
Apples	0.01	0	0
Bananas	0.1	0	0
Cashews	1	1	1
Dragon fruits	10	10	9
Eggplants	100	100	99

distribution (**Table 3**). For a minimum-uncertainty Poisson distribution (Table 1) and medium- to fast-sellers, the median matches the mean. For slow-sellers, the median automatically manages our categories by assuming the value zero: under the assumed cost function, it is clearly unprofitable to offer any ultra-perishable super-slow-seller at all, since we would let about 99 apples or 9 bananas go to waste on average before finally selling one piece! For the high-uncertainty scenario of Table 2, the median also plays slightly low – the distribution is long-tailed, with the median being shifted to lower values. Setting more realistic values for the cost of over- and under-stocks, possibly on item-level, will make us select different percentiles of the distribution as optimal values for replenishment.

MAE carries the same dimension as the prediction itself ("number of items"), and thereby strongly depends on scale. By dividing MAE by the mean sales, we obtain the dimensionless weighted MAPE, wMAPE, for which we do not need to remove 0-sales-events from the data, which is a great advantage (Kolassa and Schütz, 2007). Due to the scaling property of the Poisson distribution, wMAPE is not scale-independent either. Scale-dependence therefore always needs to be addressed explicitly.

Just ignoring that optimal MAPE-estimates are biased, however, is not an option. Important strategic decisions hinge on reliable, meaningful, business-relevant forecast evaluation! Shall we go with software vendor A, with software vendor B, or with our in-house solution? On what assortments should we focus our model improvement efforts? Is the forecast in that new category "good enough" for taking an automated system live? Forecast evaluation should provide clear, interpretable, business-reflecting evidence to answer these and many other questions.

MAPE can't help us with that. It's time to retire the MAPE.

Malte Tichy has a research background in quantum physics, with a PhD from University of Freiburg. He is a Fellow and Senior Staff Data Scientist at Blue Yonder, working at the interface between customer project implementation, research, and product development to ensure that probabilistic forecasts can unleash their full potential.

malte.tichy@blueyonder.com

REFERENCES

Gartner (2018). Sales & Operations Planning Success Survey.

Gneiting, T. (2011). Making and Evaluating Point Forecasts, *Journal of the American Statistical Association*, 106, 746-762.

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

Kolassa, S. & Martin, R. (2011). Percentage Errors Can Ruin your Day (and Rolling the Dice Shows How), *Foresight*, Issue 23, 21-27.

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

Kolassa, S. (2023). How We Deal with Zero Actuals Has a Huge Impact on the MAPE and Optimal Forecasts, *Foresight*, Issue 69, 13-16.

Makridakis, S. et al. (1982). The Accuracy of Extrapolation (Time Series) Methods: Results of a Forecasting Competition, *Journal of Forecasting*, 1, 111-153.

Morlidge, S. (2015). Measuring the Quality of Intermittent Demand Forecasts: It's Worse Than We've Thought! *Foresight*, Issue 37, 37-42.

Tichy, M. (2022a). Forecasting Few Is Different (Part 1), ***https://blog.blueyonder.com/forecasting-few-is-different-part-1/***

Tichy, M. (2022b). Forecasting Few Is Different (Part 2), ***https://blog.blueyonder.com/forecasting-few-is-different-part-2/***

Valentin, L. (2007). Use Scaled Errors Instead of Percentage Errors in Forecast Evaluations, *Foresight*, Issue 7, 17-22.

Commentary: How We Deal with Zero Actuals Has a Huge Impact on the MAPE and Optimal Forecasts

STEPHAN KOLASSA

The Mean Absolute Percentage Error (MAPE) has been discussed many times in *Foresight* and elsewhere and, as Malte Tichy writes, it is time for it to retire. I have to admit, though, that I differ as to whether it has ever served any duty – from my perspective, it has been problematic from day one.

Although the shortcomings of the MAPE have been known for a long time, the *consequences* of these shortcomings are becoming more and more serious because of the way forecasting is evolving. Decades ago, we would be forecasting on monthly granularity for a small number of products, or even on category level. Nowadays, the data are available on much finer granularity, such as days in the time dimension, and quite naturally we have moved to forecasting on this finer granularity, too. The bias induced by the MAPE, which was small in relative terms for high volumes, becomes larger and larger as the volumes to be forecasted become smaller.

The continuing trend towards ever finer granularity, and therefore towards intermittent demand forecasting, uncovers yet another serious issue with the MAPE beyond the problems it already has when forecasting nonzero data (Kolassa, 2017). When I first read Malte's excellent paper, his A-E fruit example had me scratching my head. For slowly moving items, such as the ones he considers, the Absolute Percentage Error (APE) is not defined whenever the actual is zero, and what we do in such a situation is absolutely crucial to MAPE calculations. We can't just divide by zero and then take the mean of undefined numbers! Not noting what convention he used in this case looked like a glaring oversight to me, until I read on and discovered that he was using the "if the actual is zero, this data point is removed from MAPE calculation" convention.

That is good to know, and now the A-E fruit example does make sense, but there is nothing set in stone about this particular convention. The APE has been undefined when the actual is zero since the dawn of time, and various people have proposed various ways of "dealing" with this issue. And therein lies another problem! Each convention can be defended and argued for, but how we deal with zero actuals has a major impact on the final MAPE, and on what the optimal forecast for intermittent demands is.

Although the shortcomings of the MAPE have been known for a long time, the consequences of these shortcomings are becoming more and more serious because of the way forecasting is evolving.

Approaches for Data with Zeros

Let's take a look at a few things we can do for data that has zeros.

- We can replace each zero in the denominator with a one. I am indebted to Flavio von Rickenbach for this proposal.
- We can simply discard all observations where the actual is zero, which is the convention Malte is using. This amounts to completely ignoring what the forecast is for such data points, which is *not* a good idea – whether we forecast 1 or 10 for an actual of 0 does make a difference and should not be ignored (Hoover, 2006).
- We can always use the average of the forecast and the actual in the denominator of the APE – this has been called

Figure 1. Simulated Poisson distributed demand

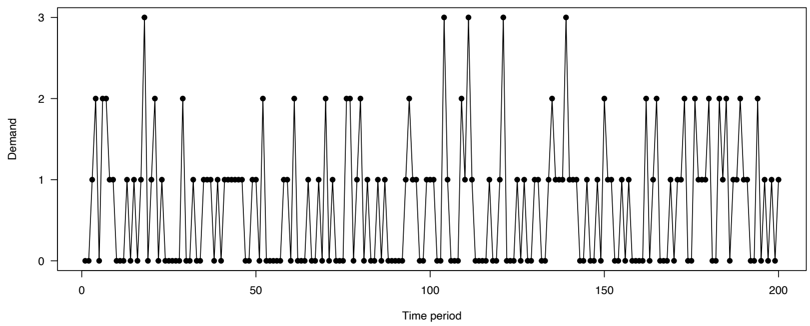
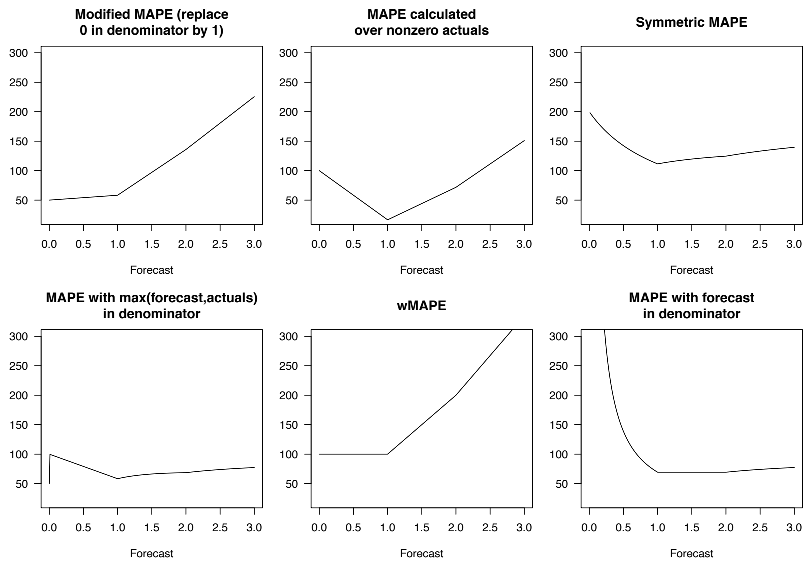


Figure 2. Expected MAPEs against forecasts for six different ways of dealing with zero actuals (for Poisson distributed actuals with expectation log 2)



a "symmetric MAPE." It avoids the division-by-zero problem as long as we always forecast higher than zero, but comes with its own kind of asymmetry (Goodwin and Lawton, 1999). As a matter of fact, the symmetric APE always contributes 200% to the MAPE calculation whenever the actual is zero, regardless of the actuals (Boylan and Syntetos, 2006).

- We can always use the maximum of the forecast and the actual in the denominator of the MAPE, with the additional convention that the APE is set to zero whenever both the forecast and the actual are zero (and we would divide zero by zero, which is mathematically undefined).
- We can divide the sum of absolute errors by the sum of actuals, which does not divide by zero unless all actuals are zero and can be interpreted as a "weighted MAPE" (Kolassa and Schütz, 2007).
- Lastly, we can always use the forecast instead of the actual in the denominator of the MAPE, which again does not divide by zero as long as we don't forecast zero (Green and Tashman, 2009).

Simulating MAPE Calculations with a Poisson Distribution

Each of these possibilities has advantages and can be argued for. So, let's assume we have chosen one convention and are looking to optimize our forecast for some intermittent demand time series. Let's assume that our demand just happens to be Poisson distributed with a mean of log 2≈0.69. This is a rather fast-moving intermittent time series with half of the actuals being zero (see **Figure 1** for a simulated example). Working with an assumption about the statistical distribution of demand allows us to focus on the properties of the MAPE, rather than get bogged down in possible ways of capturing demand dynamics in the forecasts – because there simply aren't any.

We will consider any forecast between 0 and 3. **Figure 2** shows what MAPE we can expect for each forecast for our five possible conventions, and we can immediately read off what forecast will minimize "this" MAPE (and thus conceivably maximize our bonus). For instance, suppose we want to calculate the MAPE using the first possible way of dealing with a zero actual: by simply replacing a zero by a one in this case. We note that the probability of a demand y in a Poisson distribution with a mean of of λ=log2 is

$$P(y) = \frac{\lambda^y e^{-\lambda}}{y!}, \text{ or}$$

$P(0) = 0.5, P(1) \approx 0.347,\ P(2) \approx 0.120, P(3) \approx 0.028,\ P(4) \approx 0.005,\ P(5) \approx 0.001 \ldots$

For a forecast of ŷ, the expected MAPE under this convention then is

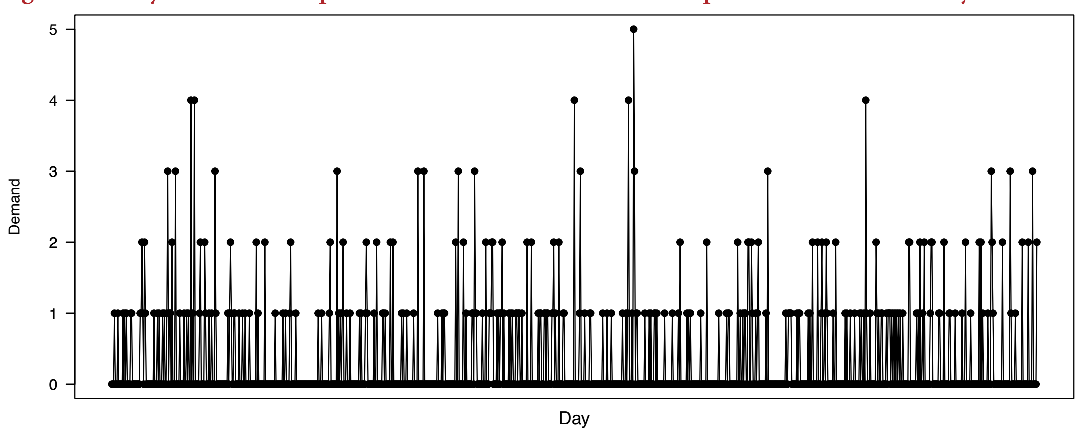
$$\text{MAPE} = \sum_{y=0}^{\infty} P(y)\frac{|\hat{y}-y|}{\max\{y,1\}} = P(0)\hat{y} + P(1)|\hat{y}-1| + P(2)\frac{|\hat{y}-2|}{2} + P(3)\frac{|\hat{y}-3|}{3} + \cdots$$

We can calculate this for various forecasts ŷ by summing for 0≤y≤30, after which P(y) gets so small as to not make a difference. The different conventions for calculating the MAPE only make a difference for the very first term in this summation: for the first convention we are looking at here, we have P(0) $\hat{y}$, but if we just remove any zero actuals, this term is simply dropped and the entire MAPE rescaled. However, as Figure 2 shows, this difference has a major impact on how the MAPE depends on the forecast ŷ:

- If we replace actuals of 0 by 1 in the APE denominator, our best forecast is 0.
- If we calculate the MAPE over nonzero actuals only, the best forecast is 1 (as in Malte's paper).
- The best forecast for the symmetric MAPE is again 1.
- Something surprising appears if we use the maximum of the forecast and the actual in the denominator: a forecast of zero yields the lowest MAPE, and the MAPE then immediately jumps to 100%, decreases and increases again, with a second local optimum for a forecast of 1.
- Things also get funky for the wMAPE: any forecast between 0 and 1 is optimal.
- Similarly, any forecast between 1 and 2 is optimal for the "MAPE with actuals in the denominator."

Observe that the "different MAPEs" yield very different values: replacing zeros with ones, or calculating the MAPE over nonzero actuals only, or using forecasts in the denominator has a good chance of giving us a MAPE less than 100% (no certainty of this, though), whereas the symmetric MAPE and the wMAPE do give us certainty – of having MAPEs at least 100%. If you calculate

Figure 3. Daily demand for a particular white door in a home improvement store over 3 years



"accuracy" as 1-MAPE, such numbers give you negative accuracy, which can give rise either to very uncomfortable discussions with your colleagues, or to enlightening teaching moments. Lastly, using the maximum of the forecast and the actual in the denominator will give a MAPE that will always be no larger than 100%, at the cost of having two locally optimal forecasts, one of them the rather useless zero forecast.

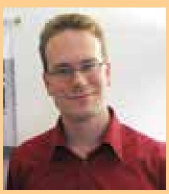
MAPE Calculation for a Real Demand Series

Let me reassure you that there is nothing particularly artificial about learning from abstract Poisson distributions above. For instance, Figure 3 gives a real demand time series with 34% zero demands. There is a little signal in the series, with slightly higher average sales on Saturday, but we can get much better MAPEs by gaming how we calculate the MAPE in the presence of zeros than by accounting for this signal. We get almost the same result as for the abstract Poisson example above: the best forecast is zero if we replace zero demands by one, if we use the maximum of the forecast or the actual, or if we use the wMAPE; the best forecast is one if we only calculate the MAPE over nonzero actuals or use the symmetric MAPE, and any forecast between 1 and 2 is optimal if we use the forecast in the denominator. And the range of the MAPEs is also very similar to Figure 2.

Conclusion

Should an error measure, and the optimal forecast, depend on the specific technical way used to address a division-by-zero issue? I believe not. This dependence alone on what is basically a matter of taste should disqualify the MAPE from serious discussions about forecasting on fine granularity. In any case, if your bonus depends on getting a low MAPE, do your best to have zero actuals ignored in the calculations. In the examples above, this convention gives you a good chance of getting the MAPE down to 20%, whereas other conventions may doom you to MAPEs no less than 100%. And don't forget to game the error metric in determining the optimal forecast (Kolassa, 2020).

Let's retire the MAPE.



Stephan Kolassa is Deputy Editor of *Foresight*, a Data Science Expert at SAP Switzerland AG, and an Honorary Researcher at the Centre for Marketing Analytics and Forecasting, Lancaster University Management School.

stephan.kolassa@gmx.de

REFERENCES

Boylan, J. & Syntetos, A. (2006). Accuracy and Accuracy-Implication Metrics for Intermittent Demand, *Foresight*, 4, 39-42.

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

Green, K. & Tashman, L. (2009). Percentage Error: What Denominator, *Foresight*, 12, 36-40.

Hoover, J. (2006). Measuring Forecast Accuracy: Omissions in Today's Forecasting Engines and Demand-Planning Software, *Foresight*, 4, 32-35.

Kolassa, S. (2020). Why the "Best" Point Forecast Depends on the Error or Accuracy Measure (Invited commentary on the M4 forecasting competition), *International Journal of Forecasting*, 36(1), 208-211.

Kolassa, S. (2017). What Are the Shortcomings of the Mean Absolute Percentage Error (MAPE)? CrossValidated thread, ***https://stats.stackexchange.com/q/299712/1352***.

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

Commentary: MAPE, What Else?

FLAVIO VON RICKENBACH

When I joined the retail forecasting game not that long ago, I had (and still have) to do a lot of reading. One important part was to learn how forecast accuracy is measured in the retail business. I guess to no one's surprise, I found Mean Absolute Percent Error (MAPE) advocated as an accuracy metric quite early and often. But I also encountered various discussions and blogs with an opposing viewpoint, where experts in the field like Stephan Kolassa (2017) and Nicolas Vandeput (2019) show you why you should not use MAPE to measure how well your forecast is doing or to compare models with each other. Dr. Malte Tichy now offers further insights and good examples of why MAPE should not be used – at least not for granular retail forecasting.

So why is the MAPE still used in practice? Based on my brief experience in the field, I think these are the reasons:

1. **MAPE is simple.** If you tell someone with basic mathematical knowledge to create a percentage error measure, I wouldn't be surprised if they came up with the idea of MAPE. Its simplicity allows you to show and explain MAPE to customers who want to understand how you measure model and forecasting performance. The MAPE is also easy to calculate in Excel/R/Python without any special libraries, and the risk of getting the formula wrong is low.
2. **You can summarize MAPE.** When you forecast products and locations by days (or even weeks) you can quickly amass huge volumes of forecasts, making it impossible to sufficiently evaluate your forecasting performance at each product location. Since MAPE is a percentage, this makes it possible to summarize the MAPEs of different product locations with different sales rates. Of course, doing so does not always make sense (even though management may want to see it!). For example, when you mix slow- and fast-selling products, small absolute errors on intermittent goods can lead to a very high MAPE. (Note that taking the median of all MAPEs instead of the mean when summarizing can reduce this effect, as can using a weighted MAPE.)
3. **Old habits die hard.** Some of the arguments why MAPE should not be used are straightforward (such as MAPE being undefined when actuals are zero). Other arguments against MAPE are not obvious enough to reveal themselves in daily business (such as how MAPE induces bias). I suspect that lots of practitioners who use MAPE have found workarounds to deal with straightforward problems. (For example, Stephan Kolassa's commentary [2023] compares the different options to deal with division by zero actuals.) But these practitioners may not be aware of or concerned with the other flaws.

MAPE, or…?

Tichy shows some alternatives to MAPE in his article, but from my viewpoint, this part of the article is too short. He mentions the Mean Absolute Error (MAE), but then as a forecaster my first thought is why this and not the Mean Squared Error (MSE) or the Root Mean Square Error (RMSE)? All three are simple, but they are not percentages and summarizing them doesn't make sense when you have different scales.

I'm aware that the goal of Tichy was to show why we should not use MAPE. But I also believe that without offering data analysts, consultants, demand planners, and business decision makers a simple alternative – that fits the retail business (different scales), but without having any of the issues the MAPE has – the MAPE will not be retiring.

Maybe there is no single accuracy metric that's simple to understand, can be summarized, and is unbiased? Maybe we need to use more than one formula to measure how well our forecasts are doing?

Maybe there is no single accuracy metric that's simple to understand, can be summarized, and is unbiased? Maybe we need to use more than one formula to measure how well our forecasts are doing?

Vandeput (2019) argued to measure accuracy and bias to evaluate retail forecasts. Morlidge (2015) also proposed to measure the bias, combining it with accuracy into a new metric called the Bias-Adjusted Mean Absolute Error (BAMAE). BAMAE is computed as follows:

- *First, calculate bias by the Mean Net Error (MNE).*
- *Second, calculate the magnitude of variation of error around the MNE.*
- *Third, add the MNE (expressed in absolute terms) and dispersion measurement.*

But disparaging MAPE is not enough – there needs to be a solution for how we can replace it.

I like the approach to measure the bias in addition to the accuracy but have some concerns about how to decide on the weighting of the two aspects. Further, I can think of retail scenarios where the way both authors calculated the bias can lead to misleading results. For example, if we overforecast a product location most of the days but then there is an unforeseen special event leading to huge higher demand compared to our forecast, we get a bias that tells us that we underforecast this product location, which can lead to wrong actions. An alternative would be to compare how many times, e.g., how many days we overforecast vs. how many days we underforecast.

Vandeput mentions RMSE or MAE to measure accuracy, but both of them have their disadvantages for certain demand patterns. As he points out, RMSE is highly influenced by outliers and the MAE is not suited for intermittent demand. And neither method is in percentages. I had not heard of BAMAE before seeing the Morlidge article. A quick Web search for BAMAE does not bring up many results, and this lack of research about BAMAE leaves me a bit unsure if it is fitted for all retail scenarios.

MAPE, and what else? Apparently, the answer is "it depends" – but on what? Hewamalage and colleagues (2022) offer a comprehensive overview of different error metrics and flow charts for different data characteristics. But it is a dense and difficult read, and it would be easy for a practitioner to get lost in it. Davydenko and Fildes (2015) provide a broad critical review of existing error metrics, landing on the Average Relative MAE as their recommended scheme for evaluating point forecasts across many series. Observing that different measures can lead to different conclusions, they emphasize the importance of understanding the statistical properties of any error measure used – but this may be asking a lot of practitioners who lack advanced statistical training.

What I would love is a simplified guide like these, but specific for practitioners in the retail field. Maybe there is already such a guide/paper out there?

Conclusion

Malte Tichy did a fantastic job in showing why we should not use MAPE as an accuracy measure for retail point forecasts. But disparaging MAPE is not enough – there needs to be a solution for how we can replace it. Otherwise, I fear that this is not the last article about the MAPE in retail.

REFERENCES

Davydenko, A. & Fildes, R. (2015). Forecast Error Measures: Critical Review and Practical Recommendations, in Gilliland, M., Tashman, L. & Sglavo, U. (eds.), *Business Forecasting: Practical Problems and Solutions*, Hoboken, NJ: John Wiley & Sons.

Hewamalage, H., Ackermann, K. & Bergmeir, C. (2022). Forecast Evaluation for Data Scientists: Common Pitfalls and Best Practices, *Data Mining and Knowledge Discovery*, **https://doi.org/10.1007/s10618-022-00894-5.**

Kolassa, S. (2023). How We Deal with Zero Actuals Has a Huge Impact on MAPE and Optimal Forecasts, *Foresight*, Issue 69, 13-16.

Kolassa, S. (2017). What Are the Shortcomings of the Mean Absolute Percentage Error (MAPE)? CrossValidated Thread, **https://stats.stackexchange.com/q/299712/1352** (last accessed December 21, 2022).

Morlidge, S. (2015). Measuring the Quality of Intermittent-Demand Forecasts: It's Worse than We've Thought!, *Foresight*, Issue 37, 37-42.

Vandeput, N. (2019). Forecast KPIs: RMSE, MAE, MAPE & Bias, **https://towardsdatascience.com/forecast-kpi-rmse-mae-mape-bias-cdc5703d242d** (accessed December 21, 2022).

Flavio von Rickenbach works as Data Analyst and SAP Consultant at retailsolutions Schweiz AG. He always trusts weather forecasts when they predict good weather; if they predict bad weather, he assumes they are wrong.

flavio.rickenbach@retailsolutions.ch