

ARTICLES *by* FORECASTERS
for FORECASTERS: Q3:2024

Forecasting Performance Objectives



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



made available to you with permission from the publisher

SPECIAL FEATURE:

FORECASTING PERFORMANCE OBJECTIVES

Scaling-Aware Forecast Rating

MALTE TICHY

PREVIEW *Forecast accuracy goals are sometimes set to meet dubious “industry benchmarks.” But more often, they are set arbitrarily to meet “what management wants to achieve.” So how should we set performance objectives and assess forecast quality? Malte Tichy argues that with traditional KPIs, unavoidable, scale-dependent noise gets misinterpreted as a sign of forecast quality. Instead, no quality conclusion should be drawn without consideration of the sales velocity of the items being evaluated. Three informative commentaries follow presentation of Tichy’s novel approach.*

ORDINARY ACCURACY KPIs SUFFER FROM THE NAÏVE SCALING TRAP

As a provider of software solutions for supply chain applications, project teams at Blue Yonder are constantly asked about and challenged on forecast accuracy. In practice, data scientists on both the customer and provider side have good intuition as to whether achieved values of KPIs are acceptable or a reason for concern. For example, in ABC/XYZ-analysis, one eventually accepts that slowly moving products are harder to predict, and focuses instead on the improvement potential of high-volume, high-fluctuation products.

But making buy/no-buy or go-live/no-go-live decisions that depend upon the subjective judgment of individuals does not feel right. Instead, objective, reproducible forecast accuracy KPIs are sought. One such KPI is the scaled Mean Absolute Error (sMAE), which is the MAE divided by mean sales. (This can also be referred to as the weighted Mean Absolute Percent Error or wMAPE.) But providing bare KPI numbers can raise more questions than it

answers: Is a 25% sMAE an excellent, average, or poor performance? Does a 15% sMAE in “fresh” (e.g. fruit, vegetables, dairy) but a 65% sMAE in “nonperishable” (e.g. canned meals, dry pasta) mean that the model “is doing fine in fresh products, but has a problem in nonperishable”?

While reporting sMAE (or other forecast accuracy KPIs) suggests an air of quantitative, scientific objectivity, accuracy numbers without context were not providing confidence to our customers or to our project teams. Depending on the metrics, the conclusion was either “the slow-movers are doing well, the fast-movers have a problem” or “the fast-movers are doing well, the slow-movers have a problem.” Data scientists’ well-established intuition clashed with the message borne by the KPIs.

From first principles and under mild assumptions, one can establish the level of noise of an ideal forecast. The ideal forecast in retail is one for which the observations are Poisson-distributed around the predicted mean. I have referred to this as a *Poisson-limited prediction* (Tichy,

Key Points

- Demand forecasts of individual items always exhibit *precision scaling*: their forecasting error is velocity-dependent. Even in idealized forecasting situations, slow-movers (low velocity) naturally come with larger relative error than fast-movers (high velocity).
- Accuracy KPIs should not be rated as “good” or “poor” unless the velocity of the forecasted items is known. An sMAE of 70% for slow-movers may be extraordinarily good, but it’s catastrophic for fast-movers.
- Hastily interpreting traditional KPIs leads us into the *naïve scaling trap* – misinterpreting unavoidable, scale-dependent noise as a sign of forecast quality. To avoid this fallacy, a *Scaling-Aware Forecast Rating* puts measured KPIs into a velocity-dependent context.
- The method is exemplified using real-life demand forecasts that Blue Yonder provided to Sainsbury’s UK production forecasting system. Velocity-dependent forecasting errors turn out to be the prime factor for understanding forecasting performance in terms of accuracy KPIs.

2023). Possessing such a Poisson-limited forecast means that one has incorporated all causal variables that influence the demand of products and that one perfectly models all demand characteristics (level, trend, seasonality) – an optimistic assumption. As can be verified in simple numerical experiments, even an ideal Poisson-limited forecast gives a larger relative error for the slow-movers than for the fast-movers. This is known as *precision scaling* – where the forecast error is velocity dependent. Precision scaling is intrinsic to count distributions, as you often find with retail store data where sales are countable quantities 0, 1, 2, 3, ... The effect of precision scaling due to varying sales velocities cannot be ignored. It is a fundamental driver of forecast accuracy KPIs.

Judging assortments based on sMAE alone had made us run into the *naïve scaling trap*: the fallacy of interpreting differences in KPI values (“sMAE of 15% in fresh, 65% in nonperishable”) as representing differences in forecast quality. Rather, these differences reflect the unavoidable scale-dependent width of the underlying distribution.

Even though the Poisson distribution is an idealization, scaling properties of error metrics are shared by most other distributions. For countable events, the meaning of relative errors always depends on whether those counts are many or few.

GROUP BY PREDICTION BUCKET TO PROVIDE VELOCITY CONTEXT

A comparison of different datasets, be it divided by assortments, days, locations, or other criteria, always suffers from the naïve scaling trap. Any two groups of forecasts have slightly different velocities, and any difference in KPIs cannot clearly be attributed to a different forecast quality since precision scaling infects the analysis. Even before practical issues kick in (Kolassa, 2008), the context-free question “Is a 35% sMAE good or bad?” cannot and should not be answered.

To make groups of forecasted product-location-day combinations comparable, we segregate the forecasts (which we assume to be on a product-location-day level) further, into buckets of similar value. For each group of forecasts that we investigate, we group all forecasts of approximately 1 into one bucket, those of approximately 2 into another bucket, those of approximately 4 in the next, and so on in a logarithmic way. We thereby also split individual products that come with different predictions on different days and locations into different buckets: an apple might be predicted to sell 10 times in London on Monday, and 50 times in Manchester on Tuesday – the two forecasts then enter two different buckets. For each bucket, i.e. for each collection of similar forecasts, we evaluate metrics separately. The buckets are characterized by prediction value and not by

outcome to avoid the hindsight bias effect I previously discussed (Tichy, 2024).

The result of this procedure is shown in **Figure 1** for forecasts that Blue Yonder provided to Sainsbury’s UK production forecasting system (i.e., not a proof-of-concept or pilot project). The Normalized Mean Ranked Probability Score (NMRPS), defined in Tichy and colleagues (2022), is used as forecast accuracy metric. It generalizes sMAE to probabilistic forecasts and behaves similarly. (For the sake of this example, we can take it as synonymous with sMAE.) The x-axis is indexed by forecasted value, ranging from ultra-slow-movers (0.1 per day) to ultra-fast-movers (1,000 per day). The y-axis shows the achieved value of NMRPS. The blue line is the value of NMRPS that would be achieved if the forecast were a perfect Poisson prediction – it does not depend on any parameter or data, but only on the assumption that no customer-individual information is available to model the demand of products (Tichy, 2023). The other lines are simple parameterizations for “Excellent,” “Good,” “OK,” “Fair,” “Insufficient” and “Unacceptable” – based on experience with many retail customers. Predicted slow-moving product-location-day combinations are found to the left, fast-moving ones to the right.

Clearly, one needs to accept different values of NMRPS, depending on the velocities that one deals with (slow-movers or fast-movers). The dark-blue circles represent the values of NMRPS achieved in production, for forecasts provided to Sainsbury’s by Blue Yonder’s Demand Edge for Retail product, based on Wick and colleagues (2021). The circle size reflects the total number of sales in the bucket, assuming here that more sales is tantamount to being more important. The scaling of the circles can be chosen differently, depending on the KPI in focus (e.g. revenue or profit).

The achieved values of NMRPS follow the “perfect” line in shape. Forecasts of about 1,000 come with a 5% NMRPS, forecasts of about 1 come with a 60% NMRPS – but the difference in forecast *quality* is marginal, as both come close to the Poisson limit. The representation allows us to

Figure 1. Normalized Mean Ranked Probability Score (NMRPS) as a Function of Prediction (Tichy and colleagues, 2022).

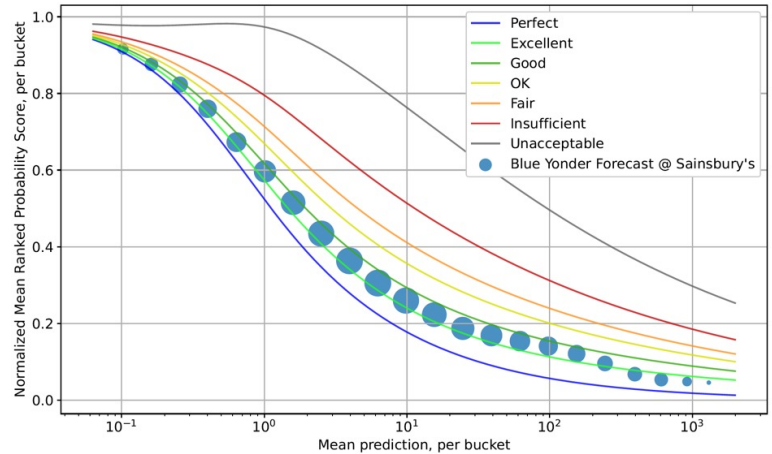
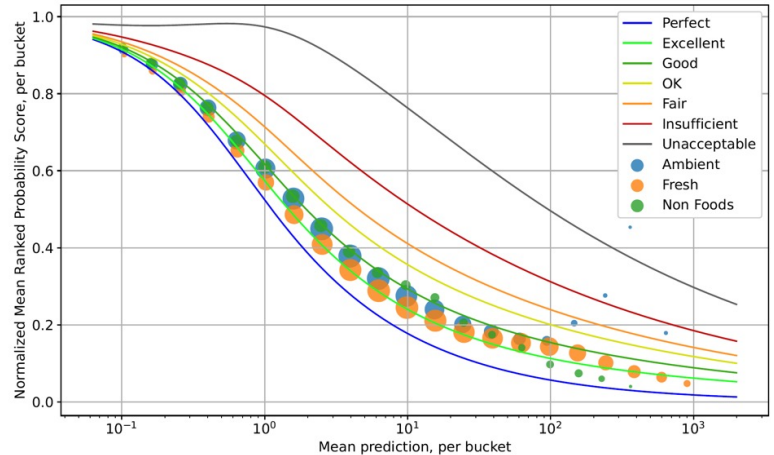


Figure 2. Normalized Mean Ranked Probability Score (NMRPS) as a Function of Prediction (Tichy and colleagues, 2022).



avoid the naïve scaling trap. Knowing that the blue “perfect” line is the limit of what is possible in principle, it is visually intuitive that, despite the quite different NMRPS values, the slow- and fast-movers are performing very similarly.

The true power of the method is unleashed when ratings are applied on a lower level, such as product group. **Figure 2** shows the same representation as Figure 1, but on the highest product group level at Sainsbury’s: “Fresh” (chilled), “Ambient” (i.e. nonperishable, kept at ambient temperatures) and “Non Foods.” Overall accuracy KPIs vary considerably: Fresh comes with an sMAE of 0.3733, Ambient achieves an sMAE of 0.5843, Non Foods yields an sMAE of 0.7516. Does it mean

that the model is performing very differently in these product groups? Not at all: the prediction-bucketed representation shows that these large differences do not witness a very different forecasting performance but are mostly due to Fresh consisting of more fast-movers (circles lean towards the right-hand side) and Non Foods of more slow-movers (circles lean towards the left-hand side) than Ambient. For a given prediction bucket below 100, Fresh still achieves slightly

The prediction-bucketed representation of Figure 1 and 2 can be done for any KPI. Depending on the KPI properties, achieved KPI values of fast- or slow-movers are then rated more gently or more strictly. For example, when one plots the relative bias ($[\text{mean prediction} / \text{mean observation}] - 1$) per bucket, one can rate the calibration of the forecast in the sense that a mean prediction of a certain value yields a mean observation of approximately the same value.

Clearly, precision scaling is the most important factor that determines forecast accuracy when measured with traditional KPIs, and the Poisson distribution is a useful benchmark to guide the expectation of what is possible.

smaller, thus better values of sMAE than the other product groups.

REFERENCE FORECAST QUALITY

How close to the Poisson case does one need to be for a “Good,” “OK,” or “Fair” rating? The colored curves in Figure 1 and 2 are drawn using a simple parametrization (using a negative-binomial model); they define “constant forecast quality” across velocities. The parameters of these curves were fitted empirically. To gauge our expectation of what an “unacceptable” forecast quality means, we use an artificially “bad” forecast that produces the same value for every product-location-day, i.e. which does not differentiate at all but always produces the overall mean sales. This purposely coarse-grained forecast is rated with the quality “unacceptable” – it is the most noisy-yet-calibrated forecast that one can think of. The different intermediate ratings were given based on experience from many different projects. The Sainsbury’s forecast matches the “excellent” line across more than four orders of magnitude. One also clearly sees that something is off for the fast-moving Ambient (nonperishable) products, which has triggered a targeted investigation. Clearly, precision scaling is the most important factor that determines forecast accuracy when measured with traditional KPIs, and the Poisson distribution is a useful benchmark to guide the expectation of what is possible.

REAGGREGATE FOR GLOBAL FORECAST RATING

Figure 1 is a powerful tool to understand whether a set of forecasts behaves as we expect it across velocities, for any thinkable KPI. It is, however, quite hard to grasp when you’re not used to it. To summarize the result into a single number, we re-aggregate the achieved quality for the different buckets into a single number, the *forecast quality rating*. The overall rating procedure is intellectual property of Blue Yonder and protected by filed patents, with details in Tichy and colleagues (2022). The rating amounts to 100% when the forecast is perfectly Poissonian across all velocities: it is then essentially perfect, all circles lie exactly on the blue line. The rating yields catastrophic 0% when the forecast is unacceptable for all velocities: that is, when all circles lie on the grey line or above. The rating thus summarizes the overall closeness of the set of circles to the perfect line into a single number – weighting each circle by its size (the number of measurements that it contains). Monitoring forecasting performance can thereby be automated to a great degree because the unavoidable ups-and-downs of traditional accuracy KPIs due to precision scaling are taken into account.

On the other hand, we can also use the reference lines to produce values of sMAE that can then be rated “Good,” “OK,” or “Fair” for that particular set

of product-location-days, and put the achieved KPI value into that context. An achieved 55% sMAE immediately seems less problematic when one knows that a perfect Poisson forecast would achieve 53%, and yet it becomes a clear call to action when the perfect Poisson forecast would achieve 2%.

CONCLUSION

Forecasting performance goals need to be corroborated by data in a bottom-up fashion, not by gut feeling and top-down management.

Precision scaling is the most prominent effect on achieved forecasting KPIs that under no circumstances can be ignored. Data-agnostic, top-down accuracy targets like “we need to achieve 10% error” are doomed to fail. Without accounting for the specific selling rates, arbitrary targets

Data-agnostic, top-down accuracy targets like “we need to achieve 10% error” are doomed to fail. Without accounting for the specific selling rates, arbitrary targets are just that: arbitrary. ...No forecast quality conclusion at all can be drawn from sMAE values alone.

are just that: arbitrary. In other words: No forecast quality conclusion at all can be drawn from sMAE values alone.

The scaling-aware forecast rating I have described creates confidence and a good degree of trust between project teams and customers about how well a model is doing. It builds a bridge from statisticians who comfortably navigate probability distributions to project managers who need to communicate go/no-go decisions to management. At a glance, catastrophic, mediocre, good, and excellent models (in the sense of how close they come to the Poissonian ideal) can be identified. In a second step, one can investigate the groups of forecasts that do not behave in a Poissonian way, search for reasons for that behavior, and possibly implement improvements.

Just like engineers know in advance how fast a car they build will be, forecasters want to know how good a forecast can become beforehand. But the forecasting

field is not there yet. The question of *forecastability* in the context of precision scaling then becomes: To which extent can the gap to the Poisson ideal be closed? This will probably never be answered universally, and whether the remaining uncertainty in a given setting is true unavoidable noise or a curable effect of other yet-to-be-incorporated features depends on the specific situation under scrutiny.

We can improve our quantitative understanding of how good a forecast can become by focusing on those areas (product groups, locations, situations, etc.) that could harbor improvement potential. Yet we must keep in mind that forecasts are only the means to an end: the optimization of processes along the supply chain. When investing time and effort to improve forecasts, therefore, we must focus on those areas that also have an actual impact on the business.

REFERENCES

- Kolassa, S. (2008). Can We Obtain Valid Benchmarks from Published Surveys of Forecast Accuracy? *Foresight*, Issue 11, 6-14.
- Tichy, M. (2023). The Technological Limits to Forecasting, *Foresight*, Issue 70, 64-65.
- Tichy, M. (2024). The Forecaster’s Evaluation Dilemma, *Foresight*, Issue 72, 10-14.
- Tichy, M. et al. (2022). Scaling-aware Rating of Count Forecasts. <https://arxiv.org/abs/2211.16313>
- Wick, F. et al. (2021). Demand Forecasting of Individual Probability Density Functions with Machine Learning, *Operations Research Forum*, 2, 1-39.



Malte Tichy has a research background in quantum physics, with a PhD from University of Freiburg. He is a Fellow and Director of Data Science 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.

mc.tichy@gmail.com

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.

Commentary: A Good Correction for Forecastability

STEFAN DE KOK

Forecastability is a key factor when considering acceptable levels of forecast error – and an important contributor to forecastability is the overall volume of demand. Malte Tichy refers to demand volume’s impact on forecastability as *precision scaling*, and his approach is both interesting and novel. There is much to like, some to caution, and a little to criticize. This commentary will cover a few different angles.

Let’s start with application to traditional deterministic forecasts, or point forecasts, and traditional error metrics that are commonly applied to them such as MAE and WMAPE (which Tichy refers to as sMAE). For these types of metrics, the scaling effect is very much an issue. Tichy’s method is a very clever correction mechanism, taking metrics that cannot be fairly compared across different product segments and making them more fair.

But there are a few cautions here, as *more fair* is not yet *perfectly fair*. First, demand volume (or velocity) is not the only contributor to forecastability. Over extended periods of time, three items can have the same velocity, but one may be intermittent, another lumpy, and the third erratic. If the forecast method produces smooth predictions, such as exponential smoothing does, they may have the same mean forecast, but their forecastability will differ significantly. If the forecast produced is frequency based, like Croston’s method, then due to the mechanism of grouping values into buckets of individual days, the lumpy item becomes indistinguishable from a fast mover on the rare days the lumpy item has demand. (For example, a lumpy item with a 100-unit forecast on an occasional day looks the same on those days as a continuous fast mover with 100 units on every day.) Again, their forecastability is very different. Getting the 100 units exactly on the right day for the lumpy item is much harder than predicting the consistent 100 units of the fast mover.

Another caution is the presentation of the Poisson distribution as being “Perfect.” There are a few underlying assumptions here that do not hold in practice. Notably, Poisson applies only if every customer purchases exactly one unit of an item; it’s typical, though, for consumers to purchase multiples of items. A reason for choosing Poisson is that the true distribution is impossible to know and with more realistic theoretical distributions, intractable to work with. So, yes: in lieu of having a better alternative, assuming Poisson is a great choice – but it should not be presented as being “Perfect.” Rather, the reference Poisson curve in Tichy’s Figures 1 and 2 should be considered an “Estimated Best.”

Likewise, the other reference curves – “Excellent,” “Good,” “OK,” “Fair,” “Insufficient,” and “Unacceptable” – are even more arbitrary. It seems self-contradictory in an article highlighting the precision differences between items within a single company to then use benchmarks from across many different companies. This is reminiscent of earlier days, when commercial best-fit algorithms were based on empirically found weighting factors across other companies. Tichy’s assumption of a negative binomial distribution for his reference curves with some arbitrary fitted parameters across whole segments of items across many different companies brings back bad memories of those bygone days. I would do away with all these unnecessary curves, and instead use a simple but defensible range from “Naïve” to “Estimated Best,” where “Naïve” could be any of the standard options, such as a random walk.

As distribution assumptions go, Poisson is not a terrible one, but we should not need to assume a distribution in a metric calculation at all. This is my only real objection. When generating a forecast, model assumptions are natural, and you try to find the most convenient assumptions to generate reliable forecasts across the target domain. But metrics that measure

forecasts should not embed any assumptions. They should give unbiased verdicts. Their purpose is to judge forecasts for the impact of the model assumptions that have been made. In cases where the metric made a worse assumption than the forecasting model did, it will penalize the forecast for a flaw in the metric itself rather than a flaw in the forecast. While a given error metric may not have any intrinsic assumptions, by scaling it using a value that does have embedded assumptions, the resulting scaled metric also has them embedded.

Lastly, Tichy mentions that the NMRPS metric generalizes the WMAPE to probabilistic forecasts. As a foremost expert on probabilistic forecasting, I consider this an overreach. It is untrue. NMRPS is an equivalent to WMAPE for forecasts that assume a Poisson distribution applies, rather than that a Gaussian (or normal) distribution applies. True probabilistic forecasts make no limiting assumptions about the distribution at all. The same is true about probabilistic metrics. The NMRPS metric is still deterministic: it compares one forecast point to one actual point at a time, where the forecast point is considered to be the mean of the assumed Poisson distribution.

A probabilistic metric such as Total Percentile Error (de Kok, 2017) compares forecast distributions to actual distributions, not points to points. As a result, the forecastability issue diminishes and precision scaling is unnecessary. Forecasted distributions simply have longer tails for slower moving items and a higher weight at zero for more intermittent items. Those distributions are not any harder to predict than smooth fast demand. Compare that to predicting the probability of an earthquake happening to predicting the timing of one happening. The first is forecastable, the second is not. Different metrics are needed to measure the first than to measure the second. For NMRPS, Tichy demonstrates that precision scaling is still called for, a direct consequence of it still being deterministic.

Perfect is the enemy of good. While my commentary may sound critical, I applaud

Tichy for this innovation. In the age-old dilemma of making fair comparisons between traditional forecast error metrics over different product segments, I believe this to be the first good approach I have encountered. In the perfect world, all companies would replace their deterministic forecasts and deterministic metrics with probabilistic equivalents, and these issues would disappear. But history has shown that people prefer the devil they know over the one they don't, and perfect is not going to happen overnight.

Tichy provides a good solution for all those wed to their existing metrics and KPIs. The grouping of items per day solves a problem other approaches could not address: how to deal with non-stationarity, such as trend, seasonality, and promotional impact. The switch from a Gaussian to a Poissonian assumption removes a division-by-zero issue some other forecastability calculations suffer from. Combined, this approach should work well for industries with both high intermittency and high overall sales volumes and that forecast in daily granularity – such as retail. While this approach cannot be directly applied to other industries, it can certainly serve as a template. Rather than trying to find a general-purpose approach for all industries, Tichy demonstrates that finding a built-to-purpose approach is viable. All that is required is to swap some of the retail-oriented assumptions for ones that hold in the industry of choice.

REFERENCES

de Kok, S. (2017). The Quest for a Better Forecast Error Metric: Measuring More than the Average Error, *Foresight*, Issue 46, 36-45.



Stefan de Kok, CEO and co-founder of Wahu-pa, has a long career in supply chain planning, with broad experience across many roles and industries. He has been a trailblazer in educating the industry on the need for probabilistic methods to deal with uncertainty. He is passionate about applying probabilistic algorithms to demand planning, inventory optimization, manufacturing, and integrated business planning solutions. Stefan earned an MSc in applied mathematics from Delft University of Technology, is a regular speaker and blogger, and is currently writing his first textbook on probabilistic planning and forecasting.

stefan.dekok@wahupa.com

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.

Commentary: A Major Milestone for Forecast Accuracy Summarization

TIM JANUSCHOWSKI

I first came across Malte Tichy and his work at ISF 2023 in Charlottesville, where he gave a wonderful talk on the content of this paper. It was an eye-opening moment for me – and a bit of a shock. How to summarize accuracy metrics beyond taking averages and histograms is an incredibly important topic, to which I had only given just enough thought.

Bucketing data by velocity was a standard practice I'd learned back in my early forecasting days at Amazon. Also, having a lower bound on forecast accuracy provided by a baseline is something I'd long incorporated and passed on to others. But having an upper bound on forecast accuracy was something new. Clearly, Tichy had thought much more deeply about this than I had, and his contribution is incredibly useful in forecasting practice. When his ISF talk ended, I emailed my teams at Zalando the slides, videos, and papers, and we went to work on implementing this immediately.

Tichy's work is impactful because it provides a visual inspection tool that answers and preempts many questions at the same time. For example: Where are your problems in forecasting? What does your data distribution look like? What is the width of the band between bad and

excellent forecasts? And what is the best possible forecast for the segment that you're interested in? Tichy's approach vastly improves upon the common practice of setting accuracy objectives based on what management thinks the goal should be.

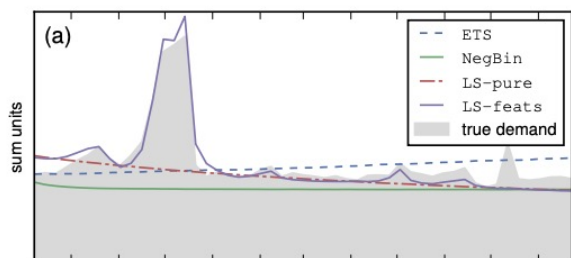
Despite its immediate utility, this is not yet the silver bullet of forecast evaluation – and Tichy doesn't claim it to be. Additional aspects of forecast evaluation are important and need some variations of this work or other tools.

For example, bucketization by *time* is something I have found to be useful. Not all time periods are created equal, e.g., times when you have a promotion, the time before and after a promotion, and the all-important Q4 peaks in many retail sectors. Slightly varying Tichy's work to bucket by time would provide some new insights, but unfortunately, the visualization would not be as beautiful. Bucketization by time doesn't have the same natural ordering as bucketing by velocity – where faster selling items are usually easier to forecast. But the same tool still has some value.

There is another more subtle question not addressed by Tichy's approach: How to ensure whether the forecast “has captured the effect of interest” especially in the face of effects that are only visible on an aggregation level and not on the SKU level. For example, slow movers also sell more during Q4 than in regular times. If you sum up all slow movers, this effect will be obvious. But for an individual SKU, it's not so obvious. Here, I like to plot aggregated demand versus aggregated forecast to get a *qualitative* understanding which I believe must complement quantitative evaluations.

For example, **Figure 1** shows true demand for Amazon retail aggregated for

Figure 1. True Demand for Amazon Retail Aggregated Versus Different Forecasts



(Source: Seeger and colleagues, 2016, and appears with permission of the authors)

an entire marketplace (shaded grey) versus different forecasts (colored lines). Ranking of the accuracy metrics is not obvious for the different forecasts.

For quantitative evaluations, Tichy’s work on summarization of metrics is state-of-the-art. Another recent, and potentially underappreciated, advance in forecast accuracy metrics is decomposition of the continuous ranked probability score (CRPS) into three highly interpretable parts (Bracher and colleagues, 2021). **Figure 2** shows how CRPS can be decomposed into penalties for overforecasting, for underforecasting, and for width of the probabilistic forecast interval. These represent the main sources of forecast error.

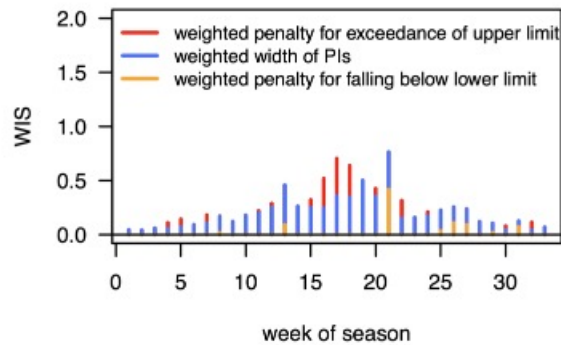
I’m grateful that Tichy has written up his ideas in different versions – both rigorous and technical for the research audience, and in a more accessible version for practitioners in *Foresight*. I hope this becomes standard best practice for everyone working in forecasting.

What Malte Tichy describes should change your approach to forecast evaluation. If it doesn’t, you’re either very smart (because you’ve had similar or better ideas already), or you’re missing out on a transformational practice – which wouldn’t be very smart.

REFERENCES

- Bracher, J., Ray, E.L., Gneiting, T., & Reich, N.G. (2021). Evaluating Epidemic Forecasts in an Interval Format, *PLoS Computational Biology*, 17(2), e1008618.
- Seeger, M.W., Salinas, D., & Flunkert, V. (2016). Bayesian Intermittent Demand Forecasting for Large Inventories, *Advances in Neural Information Processing Systems*, 29.

Figure 2. CRPS Decomposition: Height of the Bar Is the Weighted Interval Score (WIS) Which Approximates CRPS. (Source: Bracher and colleagues, 2021)



Tim Januschowski is a Director of Engineering and the Site Lead for Berlin at Data-bricks. Opinions expressed here are his own and do not necessarily reflect the views of the author’s employer, company, institution, or other associated parties. Before joining Data-bricks, Tim ran the pricing platform for Zalando SE and spent many years at Amazon and AWS, where he led the time series organization for AWS AI. He is passionate about teaching forecasting to students and professionals alike – for example, via maven.com, where he runs “Modern Forecasting in Practice” with his longtime collaborator Jan Gasthaus.

tim.januschowski@gmail.com



IIF Forecasting Practitioner Section (FPS) Bridging Forecast Professionals and Researchers



Attend panels, webinars & workshops to see practitioners & researchers join forces to improve forecasting.

Establish New Relationships • Drive New Business Partnerships • Go to FPS & IIF Events

To join the FPS, email Section Chair Elaine Deschamps at elained@forecasters.org.

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.

Commentary: The Scaling Trap in Retail Forecasting

ZABIULLA MOHAMMED

Malte Tichy addresses several critical aspects related to forecast accuracy, goal setting, and scaling traps. My commentary delves into some of the key points raised, shedding light on their significance and implications.

as Normalized Mean Ranked Probability Score and MAE do not adequately address this issue without further investigation. For instance, in the context of inventory forecasting, fast-moving items in high-growth categories may benefit from a reasonable degree of over-forecasting. The

In the context of inventory forecasting, fast-moving items in high-growth categories may benefit from a reasonable degree of over-forecasting. The projected high growth can compensate for the costs associated with maintaining excess inventory.

THE CHALLENGE OF SCALING IN RETAIL FORECASTING

I agree with Tichy on the significance of scaling issues in forecasting. Retail forecasting faces the scaling trap – particularly at a granular level – where accurate prediction becomes increasingly challenging as the number of items grows. It is evident in practice that high-velocity items tend to be more forecastable, while low-volume items pose a challenge due to limited data availability and patterns that are less predictable. Consequently, the accuracy burden falls heavily on forecasting fast movers, as they contribute significantly to a retailer's profit and loss. Even slight fluctuations in volume, despite high forecast accuracy, can quickly accumulate losses. On the other hand, although slow movers may have lower expectations of accurate forecasting, an improved forecast compared to the current version can still bring value. Problematically, however, aggregating accuracies including numerous long-tail items and reporting simple key performance indicators (KPIs) such as MAE, MAPE, SMAPE, etc., fails to illuminate the overall quality of the forecast.

OVER-FORECASTING VERSUS UNDER-FORECASTING IMPLICATIONS

It is essential to determine the directionality of the forecast error: whether the error is biased towards over- or under-forecasting. Metrics in the article such

projected high growth can compensate for the costs associated with maintaining excess inventory. Conversely, a high degree of under-forecasting can aggravate the out-of-stock rate, resulting in missed sales opportunities and increased losses. For slow-moving items though, excess forecast inventory may lead to elevated storage and transportation costs, and eventually clearance or liquidation, thereby reducing the retailer's profit margin. Therefore, it is imperative to analyze the implications of over-forecasting and under-forecasting errors separately, considering the specific characteristics of the business needs.

TANGIBLE IMPACT ON BUSINESS

Tichy rightly points out that when improving forecasts, we must focus on those areas that also have an actual impact on the business. Forecasting teams must establish a connection between forecast accuracy and tangible business outcomes. While using complex metrics is necessary to ensure the reliability and quality of forecasts, it is equally important to translate these metrics into tangible opportunities that can benefit the business. These opportunities may include increased sales, reduced costs, higher margins, enhanced productivity, or improved customer service and satisfaction. By doing so, forecasters and data scientists not only gain trust but also facilitate the acceptance of forecasts within the business.

UNREALISTIC ACCURACY GOALS

I completely support Tichy's viewpoint that simple KPIs should not be solely relied upon for determining forecast accuracy goals. Neglecting the diverse implications of forecasting needs, simple KPIs can lead to unfavorable outcomes and fail to highlight the strengths of forecasts. For example, MAPE is a simple KPI that is easily understood. In retail, it is often a small fraction of items that drives the majority of sales (consistent with the Pareto Principle). If we rely solely on the average MAPE across all item forecasts, we are favoring the forecasts of 80% of items that generate lesser sales. This can result in the selection of a forecast that may not perform as well for the 20% of items that are critical to the business. Such a scenario can lead to poor decision making and have a detrimental impact on overall business performance. To avoid falling into this trap, we need to employ more sophisticated techniques, illustrated by Tichy in Figures 1 and 2. These techniques allow us to gain deeper insights into the performance of individual items and make more informed decisions based on their specific impact on sales.

CONCLUSION

Scaling issues in forecasting pose a significant challenge, especially with a large

number of items faced by retailers. High-velocity items are particularly burdened by the need for high accuracy due to potential losses from slight fluctuations. Low-volume items can still benefit from improved forecasting even with lower accuracy expectations. Aggregating long-tail items and using simple KPIs fails to capture the quality of the forecast, and directionality of forecasting errors is crucial. Setting unrealistic accuracy goals based solely on simple KPIs is not advisable, but organizations can bridge the expectation mismatch by fostering trust and collaboration between business and data science teams.



Zabiulla Mohammed is a Director of Data Science at Walmart, where he leads a highly skilled team in implementing machine learning and experimentation projects. With over 14 years of experience, Zabiulla has successfully utilized cutting-edge data science techniques to drive business growth in various sectors such as retail, e-commerce, product development, operations, business strategy, marketing, and customer analytics. Throughout his career, he has also worked for esteemed companies like Sam's Club, Honeywell, and IBM. Zabiulla holds a master's degree in management information systems and a bachelor's degree in computer science and engineering.

mohammedzabiulla@gmail.com

\$95
50% discount
for IIF members

Keep topical forecasting information at your fingertips with *Foresight* Guidebooks.

These carefully curated collections of articles feature the best information to date as presented by today's experts in the field.

Try *Foresight* Guidebooks, available at: forecasters.org/foresight/guidebooks/

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.