FORESIGHT

The International Journal of Applied Forecasting

Valedictory from Paul Goodwin
Twenty Years On: How Is Forecasting Faring?

2025:Q2 Issue 77

contents

"Knowledge of truth is always more than theoretical and intellectual. It is the product of activity as well as its cause. Scholarly reflection therefore must grow out of real problems, and not be the mere invention of professional scholars."

JOHN DEWEY, UNIVERSITY OF VERMONT

Len Tashman, *Founding Editor*

FORESIGHT

Foresight is published by the International Institute of Forecasters, with the purpose of advancing the practice of forecasting. We encourage submissions from industry practitioners, software and consulting vendors, and academic researchers. Manuscripts should be written in language accessible to analysts, planners, managers, and students. All manuscripts are peer reviewed and edited for clarity and style.

See the Guidelines for Authors (**forecasters.org/foresight/submit-article/**) for full details on suitable topics, manuscript preparation, and manuscript submission.

Foresight welcomes advertising. However, journal content is solely at the discretion of the editors and will adhere to the highest standards of objectivity. Where an article describes the use of commercially available software or a licensed procedure, the author must disclose any interest in the product. Articles whose principal purpose is to promote a commercial product or service will be rejected.

IIF Membership & Subscriptions
Ying Fry, IIF Business Manager
forecasters@forecasters.org

Foresight Business Office:
8956 Erect Road
Seagrove, NC 27341

note from the editor

IIF EVENTS

On March 3-4, the International Institute of Forecasters hosted parallel events on the University of North Carolina's Charlotte campus:

- **The International Symposium on Energy Analytics** Chaired by Tao Hong, ISEA featured Rafal Weron's IIF Distinguished Lecture Series on electricity price forecasting, and presentations by five previous IIF-SAS research grant winners.
- ***Foresight* Practitioner Conference** Chaired by Matt Schneider, the FPC included pre-conference presentations by the five IIF Forecasting Practice Competition finalists, a panel discussion on Forecast Value Added, and additional presentations by forecasting experts from industry and research.

The **2025 International Symposium on Forecasting** is being held in Beijing, China from June 29 – July 2. Find registration details in the ISF ad on the inside back cover of this issue.

PREVIEW OF *FORESIGHT* ISSUE 77

After nearly 20 years in editorial roles on the *Foresight* staff, **Paul Goodwin** stepped down at the end of 2024. In this issue, Paul leaves us with a farewell address on the current state and future direction of forecasting.

In 2024's VN1 forecasting competition, familiar methods like LightGBM and ensembling performed well, yet surprisingly few participants beat the naïve (no-change) model. Competition organizer **Nicolas Vandeput** shares his key takeaways from the top performers.

Forecasts are not inherently beneficial, but can provide value by improving organizational decision making. To ensure this happens, **James Taylor** argues for a formal, robust, and structured decision model to identify relevant forecasts and their features during decision making. His approach is intended to make forecasters understand how their forecasts impact decision making.

Forecasting performance is typically evaluated by statistical measures of forecast error, ignoring the computational cost of producing the forecast. Yet costs in both computer time and environmental impact can be huge. *Foresight* Associate Editors **Fotios Petropoulos** and **Evangelos Spiliotis** consider the tradeoffs, and show how forecast computation time can be dramatically reduced without significant impact on forecast accuracy.

In the hospitality sector along with some others, the timing of an event's occurrence (e.g., a hotel stay) is distinct from the timing of its initiation (i.e., making the reservation). This complicates the act of forecasting, which must now span multiple time axes. To address this challenge, new *Foresight* contributors **Harrison Katz**, **Erica Savage**, and **Kai Thomas Brusch** describe a two-part forecasting methodology that treats the forecasting process as a *time-shift operator*.

Retrieval-augmented generation (RAG) techniques have enhanced the capabilities of large language models. Building on these advancements, **Ryan Fattini** and **Ryan Young** introduce a novel application of retrieval-augmented forecasting. By integrating natural language processing, this facilitates a conversational approach that enables users to generate and refine forecasts without the need for deep technical knowledge.

It has often been noted that the asymmetry in some performance metrics (including MAPE) might encourage forecasters to "game" the metric by purposely over- or underforecasting. But is this really happening? **Patrick Bower** doesn't think so, and in his opinion-editorial he argues that other factors have much larger influences on forecaster behavior than metric asymmetry.

Our second op-ed is contributed by **Leo Sadovy**, who advocates a systems thinking approach for forecasters seeking to assist in sustainability challenges.

FORESIGHT STAFF UPDATES

After providing several valuable manuscript reviews, Zabiulla Mohammed has been moved from the *Foresight* Advisory Board (FAB) to become *Foresight*'s new Column Editor for Retail & CPG. Zabi is Director of Data Science at Walmart.

Following the FPC, we've added four new members to the FAB:

- Janina Zittel, Head of the Research Campus MODAL EnergyLab at Zuse Institute Berlin. Janina received her doctorate in meteorology with a background in long-term climate projections. She now focuses on optimizing forecasting methods for their impact on decision making.
- Yue Li, Associate Partner at Bain & Company, and based in California. Yue spoke on the impact of Generative AI at the FPC, and specializes in demand forecasting, financial forecasting, and optimization.
- Cara Curtland, Data Science Strategist at HP in Vancouver, Washington. Cara is a senior advisor to executive leadership, and holds degrees in industrial engineering.
- Eduardo Romanus, Data Science Team Leader at Ipiranga in São Paulo, Brazil. Eduardo gives the FAB a presence in South America, bringing industry experience along with a background in electrical engineering (automation and control), data science, and time series forecasting.

POST-FPC EXECUTIVE FORECASTING RETREAT

Following the FPC I was delighted to host several editors and advisors for a forecasting retreat on the farm in Seagrove, North Carolina. The two days included lively discussion on *Foresight*'s future direction and international politics, lunch featuring the local delicacy of chicken fried steak, and trail hiking and tractor training. Thankfully, there was no injury to either tractors or trainees. And even more thankfully, all trainees still have their day jobs in forecasting.

—Mike Gilliland

Dragonfly Farm

Seagrove, NC

USA

Enjoying(?) Chicken Fried Steak

Stephan Kolassa and Jeff Baker in Training

Safety Check for Mark Chockalingam

Twenty Years On: How Is Forecasting Faring?

PAUL GOODWIN

PREVIEW *After nearly 20 years on our editorial staff and contributing over 40 columns, articles, and commentaries, Paul Goodwin is stepping away from* Foresight. *But as he leaves, Paul has generously agreed to provide this farewell retrospective on the current state and future direction of forecasting.*

Eighteen years ago I wrote my first article for *Foresight*'s Hot New Research section (Goodwin, 2007). The International Institute of Forecasters had launched the journal two years earlier with a mission to inform and improve forecasting practice. The need for such a journal was clear: my article reported a study suggesting sales forecasts had become less accurate over the previous two decades (McCarthy et al., 2006). A key reason for this decline, the authors found, was that forecasters of the early 2000s were less familiar with forecasting techniques than their predecessors. So, how has forecasting fared in the subsequent decades? Have accuracy and forecasting practice improved? Are forecasters now employing more appropriate methods?

Of course, accuracy does not directly reflect the quality of forecast practice. Much depends on the *forecastability* of what we are trying to predict. Over time, the world may become less predictable, so accuracy can decline even if forecasting processes improve.

It's difficult to answer the question about accuracy. Even the gloomy conclusions of the McCarthy study were based on just 86 returned questionnaires out of the 480 that had been emailed to companies. There was, therefore, scope for self-selecting bias and distortions that could arise from self-reports of accuracy. The study suggested that average percentage errors (MAPEs) had risen from about 15% to 24%, but this accuracy metric must also be treated cautiously. Product proliferation occurred over the period studied, which may have led to lower sales volumes for individual SKUs and increased volatility – conditions that can amplify the MAPE.

Nevertheless, the evidence we do have suggests there is still much scope for improving accuracy. For example, Steve Morlidge (2013) reported that many forecasts produced by two companies supplying consumer goods were less accurate than naïve forecasts (which simply assume that demand in the next period will be the same as the current period). In a subsequent study, Morlidge (2014) concluded that a large percentage of forecast errors are avoidable. More recently, an analysis of around 147,000 company demand forecasts I conducted with Robert Fildes and Shari De Baets (Fildes et al., 2023) found that forecasters often made judgmental adjustments to computer-based forecasts based on irrelevant information. As a result, the interventions significantly reduced accuracy, especially where the adjustments were upward. Other studies have reached similar conclusions (e.g., Franses and Legerstee, 2010).

Of course, accuracy does not directly reflect the quality of forecast practice. Much depends on the *forecastability* of what we are trying to predict. Over time, the world may become less predictable, so accuracy can decline even if forecasting processes improve. Within the last 20 years, three

major shocks have shaken global supply chains: increasing trade wars and tariffs, COVID-19, and the war in Ukraine. Shocks like these can clearly diminish forecast accuracy, especially where the past is assumed to be a reliable guide to the future. So how do we define good practice, and is there evidence that forecasting standards have improved?

In 2001, Scott Armstrong edited a book called *Principles of Forecasting* (Armstrong, 2001), which laid down what research up to that period had indicated was good practice. The book included a 134-item "forecasting standards" checklist in the form of a questionnaire. The items included:

- select simple methods unless evidence favors complex methods
- avoid biased data sources
- ask experts to justify their forecasts
- provide full disclosure of methods
- test assumptions for validity
- estimate prediction intervals

Mike Gilliland's book *The Business Forecasting Deal* (Gilliland, 2010) took a complementary approach. It identifies bad practices, such as allowing politics to influence forecasts, overusing judgmental interventions, gaming the metrics, and making poor decisions when choosing forecasting software.

Again, obtaining a complete picture of the current quality of forecasting processes across organizations is impossible. It seems likely that forces such as competitive pressures will compel some companies to develop excellent procedures. Many examples of these can be found in *Foresight*. However, recent research papers reveal that some suboptimal practices persist (Karelse, 2021; Fildes & Goodwin, 2021; Goodwin et al., 2023). Judgmental interventions are made all too frequently despite the evidence that their associated biases often damage accuracy and waste time. Many of these adjustments reflect the fact that political interference in forecasting is still pervasive. This often results in forecasts that are *aspirational*, representing what the organization *wants to happen* rather than representing their best guess of what *really will happen*. In addition, formal assessment of uncertainty is rare, so that point forecasts still predominate. Overall, there is a low take-up of significant advances in forecasting techniques, such as new methods for modeling promotions, new hierarchical reconciliation procedures, automatic method-selection procedures, and models based on machine learning. Indeed, spreadsheets still appear to be the most common type of software used in forecasting despite the clearly demonstrated scope for errors that can proliferate over multiple worksheets.

But what of the next 20 years? It's easy to imagine a world where artificial intelligence (AI) has taken over the forecasting function, producing optimal automated forecasts untouched by human hands. After all, AI-based methods performed well in the M5 competition organized by Spyros Makridakis and his colleagues (Makridakis et al., 2022), and the likelihood is that they will get better. However, experts like Stephan Kolassa believe that the practical value of AI is overrated, arguing that simple methods may give satisfactory results at a fraction of the cost of AI implementation (Kolassa, 2022). Moreover, forecasting in organizations is a multifaceted process. It can embrace politics and game playing, wishful thinking, group dynamics, attitudes to risk, the desire for a sense of ownership of forecasts, algorithm aversion, and many other aspects. That's what makes forecasting so interesting to research. While AI's role might expand, these factors are likely to restrict the extent to which it will take over the forecasting function, so the human element will probably persist.

All of this suggests that, 20 years after its launch, *Foresight* will continue to fulfill an essential role in disseminating best practices and the latest findings from forecasting research. It will be fascinating to see how forecasting evolves over the next two decades, and I intend to keep a keen eye on the journal's pages as the ideal place to follow these developments.

REFERENCES

Armstrong, J.S. (ed.). (2001). *Principles of Forecasting: A Handbook for Researchers and Practitioners*. Kluwer Academic.

Fildes, R. & Goodwin, P. (2021). Stability in the inefficient use of forecasting systems: a case study in a supply chain company. *International Journal of Forecasting*, 37(2), 1031-1046.

Fildes, R., Goodwin, P., & De Baets, S. (2023). Judgmental adjustments in demand planning: their motivation and success. *Foresight*, 71, 31-37.

Franses, P.H., & Legerstee, R. (2010). Do experts' adjustments on model-based SKU-level forecasts improve forecast quality? *Journal of Forecasting*, 29, 331-340.

Gilliland, M. (2010). *The Business Forecasting Deal: Exposing Myths, Eliminating Bad Practices, Providing Practical Solutions*. John Wiley & Sons.

Goodwin, P. (2007). Recent studies on forecasting know-how, training and information sharing. *Foresight*, 6, 26-28.

Goodwin, P., Hoover, J., Makridakis, S., Petropoulos, F., & Tashman, L. (2023). Business forecasting methods: Impressive advances, lagging implementation. *Plos One*, 18(12), e0295693.

Karelse, J. (2021). Mitigating unconscious bias in forecasting. *Foresight*, 61, 5-14.

Kolassa, S. (2022). Commentary on the M5 forecasting competition. *International Journal of Forecasting*, 38(4), 1562-1568.

Makridakis, S., Petropoulos, F., & Spiliotis, E. (2022). The M5 competition: Conclusions. *International Journal of Forecasting*, 38(4), 1576-1582.

McCarthy, T.M., Davis, D.F., Golicic, S.L., & Mentzer, J.T. (2006). The evolution of sales forecasting management: A 20-year longitudinal study of forecasting practices. *Journal of Forecasting*, 25(5), 303-324.

Morlidge, S. (2013). How good is a "good" forecast? Forecast errors and their avoidability. *Foresight*, 30, 5-11.

Morlidge, S. (2014). Do forecasting methods reduce avoidable error? Evidence from forecasting competitions. *Foresight*, 32, 34-39.

Paul Goodwin is an Emeritus Professor at the University of Bath (UK), author of numerous books and articles on the use of judgment in forecasting, and was *Foresight*'s longtime Column Editor for Hot New Research.

p.goodwin@bath.ac.uk

Book Signing for *Forewarned*

Corsham 10k Run

With Spyros Makridakis at the M6 Conference

Learnings from the VN1 Forecasting Competition

NICOLAS VANDEPUT

PREVIEW *In 2024's VN1 forecasting competition, participants forecasted the sales of several e-vendors over a 13-week period. Familiar methods like LightGBM and ensembling performed well, yet surprisingly few participants beat the naïve (no-change) model. Now, competition organizer Nicolas Vandeput shares his key takeaways from the top performers.*

I had the pleasure of hosting the VN1 forecasting competition from September to October 2024, with over 250 participants or teams actively participating. The objective of the competition was to forecast the demand (sales) of different e-vendors for the next 13 weeks for a total of over 15,000 item/location combinations. The dataset was provided by Flieber, one of our three sponsors, with the competition spread into two phases:

- **Phase 1:** A warm-up where participants could try out different models, see their scores, and compare themselves with other participants.
- **Phase 2:** The real competition, where participants could only submit a single set of forecasts and couldn't see their scores until the final deadline.

When the competition ended I gathered insights from the top 20 performers, discussing with them their approaches, difficulties, and solutions. These findings are shown in **Table 1**, where submissions are ranked based on their "Score" (= MAE% + |Bias%|). Score is a metric I implement for all my clients, as it offers an excellent trade-off between metric complexity and business value. Here, f represents the forecast and d represents the demand. You can read more about it in my books (Vandeput, 2021; Vandeput, 2023a).

$$Score = \frac{\sum|f-d|}{\sum d} + \frac{|\sum(f-d)|}{\sum d}$$

In the table, the rightmost columns show the time (in minutes) taken for training and inference, and for optimizing parameters using cross-validations or similar techniques. Other than the Score, all data in the table are self-reported by the contestants.

In November I hosted a webinar (viewable at **youtube.com/watch?v=CRGA5mOqSeo**) in which the top five described their methods. In this article I share my learnings from the competition.

LEARNINGS FROM THE VN1 COMPETITION

Disclaimer: Most of these takeaways are based on self-reported information from and my conversations with the top 20 participants. Despite my best efforts, some information could have been lost along the way or misinterpreted.

Beating Simple Models

In the competition setup I provided a function to generate a 12-week moving average for use as a benchmark. Most competitors could easily beat it. While the benchmark achieved a score of 80.5% in Phase 2, the top 20 competitors were all below 53%.

Perhaps the main surprise of the competition was that only a few competitors could beat the 50.7% score of the naïve model (which forecasts future sales as unchanged from the last observed sales). Naïve forecasts were much more accurate than moving averages for Phase 2 because of a quirk in seasonality due to the competition timing. During Phase 1, participants had access to sales and price data from July 6, 2020 to October 2, 2023, and had to forecast sales from October 9, 2023 to January 1, 2024. During Phase 2, they had sales through January 1, 2024

Table 1. Ranking of Top Performers and Benchmarks

Score	Name	Model	Outliers	Segmentation	Language	Libraries Used	Lines of Code	Hardware	Training & Inference (minutes)	Parameter Optimization (minutes)
46.4%	Philip Stubbs & Jakub Figura	LGBM, SARIMA, Anchored Multiplicative Seasonal Indexing	Few outliers fixed manually/post processing	Yes for AMSI	Python	Pandas, Polars, Numpy, MLForecast, LightGBM, Statsmodels	690	16GB RAM, i5	262	-
46.6%	Justin Furlotte	LGBM Recursive, LGBM Direct, Seasonal Theta			Python	Pandas, Scipy, Statsmodels, Numba, Nixtla, LightGBM, Numpy	297	32GB RAM, M1 Max	10	120
47.6%	Arsa Nikzad	LGBM Recursive, MFLES			Python	Pandas, Numba, Nixtla, LightGBM, Numpy	307	384GB RAM, n2-standard-96	10	3,000
47.7%	Antoine Schwartz	Temporal Fusion Transformer (TFT)	Linear interpolation for obvious shortages or lockdowns periods		Python	Nixtla, Optuna	600	32GB RAM, 8 CPUs, 1 GPU (16GB mem)	3	600
48.1%	An Hoang	LGBM, TimesFM		Basic segmentation based on the number of zeros and seasonal strength	Python	Polars, Nixtla, LightGBM, JAX	2,184	16GB RAM, M3	5	120
49.6%	Quan Pham & Tung Bach	LGBM, DynamicTheta, MFLES			Python	Pandas, Nixtla, LightGBM	200	16GB RAM	1	60
49.9%	Teodor Georgiev	LGBM Recursive			Python	Pandas, LightGBM	290	16GB RAM	7	25
49.9%	Davide Burba	LGBM			Python	Pandas, MLflow, Pydantic, LGBM	800	16GB RAM, M3	360	6,000
50.5%	Colin Catlin	AutoTS	Yes	Ensemble by "profile" (7 buckets) and timesteps	Python	AutoTS	500	32GB RAM, 7840HS	2	10,800
50.6%	Nitin Menon	LGBM			Python	Pandas, LightGBM, Numpy	200	32GB RAM	5	10
50.7%	Wagner Assis	LGBM		Segment as feature (seasonality): [4, 12, 24, 52]	Python	Pandas, LightGBM, Numpy, Optuna	1,300	16GB RAM, i7	8	45
50.7%	Evandro Cardozo da Silva	Naive			Python	Nixtla, Pandas		8GB RAM, i7	1	9
50.7%	**BENCHMARK**	**Naïve (no-change)**								
51.0%	Daniel Emery	Convolutional Neural Net			Python	Tensorflow	300	32GB RAM	1	15
51.1%	Moad Charhbili	MSTL, MFLES			Python	AutoMFLES, MSTL, StatsForecast, Nixtla	125	32GB RAM		180
51.2%	Tyler Blume & Jose Morales	LGBM Recursive, MFLES, DynamicTheta		For LGBM: replaced forecasts with 0's if recent history is all 0	Python	Statsmodels, Nixtla, MLforecast	300		1	20
51.3%	Emmy									
51.8%	Diego Fernandez	LGBM, Naive 0	Outlier intervals	3 groups - interval and standard deviation based and "bad series" group	Python	Pandas, Nixtla, LightGBM, Numpy	500	16GB RAM, M3	5	15
51.9%	José Morales	LGBM Recursive	Used robust scaler	Series with last two values constant vs rest	Python	Pandas, Polars, LightGBM, MLForecast	300	16GB RAM, i7	0	1
52.2%	Neo Anderson	LGBM			R	LightGBM, Tidymodels	300	64GB RAM	26	-
52.7%	Jack Rodenberg	LGBM, MSTL, Theta		Forecasted 1200 Sku combos as zeros (discontinued items)	Python	Pandas, Polars, Nixtla, LightGBM		18GB RAM	10	20
80.5%	**BENCHMARK**	**12-week Moving Average**								

Key Points

- The VN1 forecasting competition ran from September through October 2024, with over 250 individuals or teams participating. Contestants forecasted sales for 15,000 e-vendor/item combinations across 13 weeks.
- As has been observed in other recent forecasting competitions such as the M5, LightGBM was a strong performer, taking most of the top rankings.
- Also consistent with past competitions, ensembling proved a worthwhile approach. And one simple method (naïve forecast) took most participants by surprise by finishing 12th.

Insights from Top Competitors

Based on my interviews with top competitors, their main skill was being able to evaluate, fine-tune, and select models easily. Most of them tried out different approaches, features, and parameters. None got simply lucky at trying out a model that was successful by default. The key here is to iterate quickly using a robust testing framework. A model that delivers good results in your setup is likely to deliver good results in the future.

Here are some additional insights from my interviews:

Tools. All the top participants but one used the programming language Python (and its multiple libraries) to analyze the data and create their forecasts. No one reported using Excel, VBA, Matlab, or SQL. Most participants used Pandas (a library within Python) to manipulate data (with a few going for Polars, a faster but lesser-known alternative), and half reported using Nixtla's models or util functions. My advice for anyone who wants to do forecasting at scale is to learn Python and skip Excel, VBA, Matlab, and SQL.

Based on my interviews with top competitors, their main skill was being able to evaluate, fine-tune, and select models easily.

and had to forecast sales from January 8 to April 1, 2024. Since moving averages included year-end sales, this resulted in overforecasts for Phase 2. If the results had been evaluated a few weeks later, it's unlikely that a naïve forecast would have performed so well compared to the moving average.

One competitor understood this and, after trying out multiple models, achieved 12th position by sticking to the naïve model. Top competitors tried out multiple models before choosing their final solution, and the value of naïve forecasts wasn't obvious to everyone. Among the 250+ participants, only one concluded that naïve was delivering better accuracy than (most) other methods. It took skill and knowledge to understand that a solution as simple as naïve delivered good results.

Note that it's unusual for naïve forecasts to beat moving averages. That's why I advise against using naïve forecasts as benchmarks, as they are often too easy to beat.

Outlier detection. Only two participants reported flagging outliers. I personally don't use any statistical method to flag outliers, and I don't advise my clients to do so. I previously published an article (Vandeput, 2023b) and hosted a webinar (***youtube.com/watch?v=VQDXNhAXSEc***) to explain why I don't detect outliers and what I do instead.

Shortage and zeroes. Three participants reported flagging shortages or end-of-life products. However, the way they inferred these situations (e.g., from zeroes in the data) was not specified. Unfortunately, the competition didn't provide inventory data to automatically flag shortages, so we couldn't demonstrate the importance of using inventory data. I plan to include this in my next competition (VN2). Even when my clients don't provide their historical inventory data, I still spend time

flagging shortages and end-of-life: simple methods usually provide tangible added value.

Segmentation. A few participants segmented products. Most often, the segmentation was based on seasonal patterns.

Code complexity. Half of the participants delivered a solution in less than 300 lines of code. This highlights that if you know what you're doing, you can deliver high value with minimal code complexity.

Complexity is also dependent on the libraries used and if the participants included the code required to optimize their models. Putting models in production in a live environment will also require more code to enhance robustness and cope with most edge cases. Moreover, within the competition the participants received structured data and didn't have to deal with promotions and shortages.

Business knowledge. No participant reported using specific business insights or manually reviewing forecasts.

Models

VN1 – like most data science competitions – is a social competition: people share ideas and notebooks and communicate. So there is an organic aspect to the models that end up being used in the competition. It's likely that if someone shared a notebook achieving a reasonable score on day one, many competitors would have used it. Competitors have limited time, so if they find a working technique and a ready-to-use notebook, they will use it. I classified the models used (or not used) from D to A.

D Models (Not used by anyone in the top 20):

- Facebook Prophet (and its neural version). Facebook Prophet has been debunked over the last few years as a poor model for forecasting supply chain demand. Nevertheless, you can often see people advocating for it on the internet (or using it as a benchmark). I wouldn't advise any of my clients to use it.
- XGBoost and CatBoost. When it comes to boosted trees, all the participants preferred the LGBM implementation. In my experience, XGBoost delivers a similar performance as LGBM but is usually (but not always) slower. On the other hand, CatBoost is less reliable in my limited experience.
- No one in the top 20 used exponential smoothing models (aka Holt-Winters) in their "usual" implementation. I personally like these models: they are easy to understand, implement at scale, and can deliver relatively good results if you tune them adequately.
- Neural Networks. No one used simple (feed-forward multi-layers) neural networks. Unfortunately, these are often

Half of the participants delivered a solution in less than 300 lines of code. This highlights that if you know what you're doing, you can deliver high value with minimal code complexity.

Running time. Nearly all solutions could deliver forecasts within 10 minutes – machine learning is definitely fast – and only two teams reported much longer running time. The winning team required nearly 4 1/2 hours due to their use of ARIMA (a famously slow model that I will address below). The second "slow" model was because the participant squeezed extra accuracy by ensembling his model 30 times (that is, rerunning his model 30 times and averaging the results).

Parameter optimization. Most machine learning models need hyperparameter tuning. Nevertheless, some competitors stuck to using their models' library default, whereas others took up to 200 hours of cross-validation time to optimize them. It is a surprise that using default values didn't impact the accuracy of some models much. Feature engineering seems to be more important than parameter optimization.

relatively slow to run while requiring extensive hyperparameter tuning. At SupChains, we haven't used neural networks to forecast demand since 2019.

C Models (Barely used, and usually only used as part of an ensemble):

- ARIMA was only used by a single team (the winning one), resulting in an extensive running time of 262 minutes, whereas most other solutions ran in less than 10 minutes. I don't use ARIMA and don't advise my clients to use it (actually, I coached multiple companies out of ARIMA). It is extremely slow, and beyond this result in VN1, I have not seen ARIMA delivering value compared to regular exponential smoothing on any supply chain dataset I have seen. Moreover, ARIMA struggles with 0 values (they are everywhere in supply chains) and will have a difficult time understanding shortages and promotions. Note that the winning team used ARIMA in an ensemble of models where ARIMA only accounted for 30% of the overall ensemble.
- A single participant used an automated time series machine learning framework (AutoTS). It didn't perform especially well compared to other solutions despite requiring an optimization time of 180 hours and more lines of code than most other solutions. I wouldn't advise my clients to use ML automated framework: it's too slow, and often doesn't result in accurate forecasts.
- A single competitor used a Convolutional Neural Net.

B Models (Demonstrated added value by multiple competitors):

- Two competitors used Transformer models to forecast demand, a new type of model that (to the best of my knowledge) was used for the first time in this forecasting competition. It's likely that we'll see more and more transformers in the next competitions.
- Four top competitors reported success using a new model, MFLES (introduced in 2024 based on gradient-boosted time series decomposition). MFLES got traction as the author (Tyler Blume) was one of the participants who successfully used it in VN1 and shared some of his notebooks to the community.
- Four competitors used the Theta model, which dates back to early 2000. Theta became famous after achieving the winning position at the M3 competition. For VN1, it seems that the DynamicTheta implementation from Nixtla got a lot of traction and achieved great results with little computation time. As far as I can tell, Theta wasn't used (successfully) during the M5 or Intermarché competition. Thus it is unclear if Theta is making a solid comeback or just happened to work well on this specific dataset.

A Models (Best of the best):

- Light Gradient Boosted Machine (LGBM)! Most top competitors used it, reporting good accuracy and fast execution. LGBM has already been used by many participants in previous forecasting competitions (M5 and Intermarché), and it is also our favorite model at SupChains. I would advise that it should be the backbone of your forecasting efforts as well.
- Lastly, most participants relied on ensembling. Rather than sticking to a single model, they combined the forecasts of different models. They also combined different instances of a single model. Since most ML models are inherently stochastic, you can rerun the same underlying model multiple times and take the average. Many participants used this technique with LGBM. Ensembling models is nearly guaranteed to deliver better results; it's as close as you can get to a free lunch.

CONCLUSION AND NEXT STEPS

The success of most of the top participants in the VN1 forecasting competition can be attributed to these key factors:

- **Structured Framework to Evaluate Models:** Establishing a clear and systematic framework to assess model

performance was crucial for selecting the best forecasting methods. If you can't properly assess the quality of your models, you can't make a great forecasting tool – it's as simple as that.

- **Fast Experimentation:** The ability to iterate quickly allowed competitors to try out more models, select better features, and fine-tune their models more effectively.
- **Model Exploration:** Top performers tested various models and techniques before selecting their final approach.
- **Feature Engineering:** Creating, testing, and selecting meaningful features proved to be a major differentiator in achieving superior forecast accuracy.
- **LGBM:** LightGBM emerged as the dominant model, delivering great accuracy and speed while being easy to use.
- **Ensembling**: Combining different solutions.

Running a forecasting competition isn't as easy as I suspected – Phase 1 was launched during my honeymoon! – but we had a lot of fun, we learned much, and I'm looking forward to the next one. The plan for VN2 is to include data about shortages and find a supply chain with promotions. Stockouts and promotions have a massive impact on demand forecasting – that's the first data I go after in all my projects – and good models should be able to cope with them. Please reach out to me if you think you have the right supply chain data for VN2.

Beyond the inclusion of promotions and shortages, let's review some of the aspects of VN1 and how they would change (or not) in VN2:

- **Data Scope:** VN1 was big enough to be realistic but didn't penalize participants for lack of computation power.
- **Metrics:** While Score isn't a perfect gauge of forecasting quality, it's the best compromise I can think of when it comes to evaluating forecasts – and no one complained about it! We could imagine using the value-weighted Score (Score = €MAE% + |€Bias%|) in VN2 depending on the dataset.
- **Two Phases:** I think we found the right balance with two phases: All participants appreciated Phase 1 emulation and notebook sharing (there was a prize for the most community-endorsed public notebook), while Phase 2 was the real competition (with a single submission). Unfortunately, evaluating forecasts on a single phase is always prone to luck. We could potentially go for a Phase 3 (similar to Phase 2), but this would substantially increase workload for participants.
- **Teams:** Many participants enjoyed (and learned a great deal from) competing as teams.

Acknowledgments

Thanks to Philip Stubbs, Ruben van de Geer, Jacopo De Stefani, Thierry Azalbert, and Carmen González Camba for their contributions to this article.

REFERENCES

Vandeput, N. (2021). *Data Science for Supply Chain Forecasting*. de Gruyter.

Vandeput, N. (2023a). *Demand Forecasting Best Practices*. Simon and Schuster.

Vandeput, N. (2023b). Outlier Detection and Correction. Medium (July 5). ***nicolas-vandeput.medium.com/outlier-detection-and-correction-694f9f474c2d***

Nicolas Vandeput helps supply chain leaders achieve demand and supply planning excellence. He founded his consultancy company, SupChains, in 2016, and in 2018 founded SKU Science, an online platform for supply chain forecasting. Passionate about education, Nicolas is both an avid learner and a teacher. Since 2020, he has been teaching demand forecasting and inventory optimization to master students in CentraleSupelec, Paris, and guest teaching in various universities worldwide. He has published three books: *Data Science for Supply Chain Forecasting* in 2018 (second edition in 2021), *Inventory Optimization: Models and Simulations* in 2020, and *Demand Forecasting Best Practices* in 2023.

nicolas.vandeput@supchains.com

Decision Modeling to Increase Forecast Usability

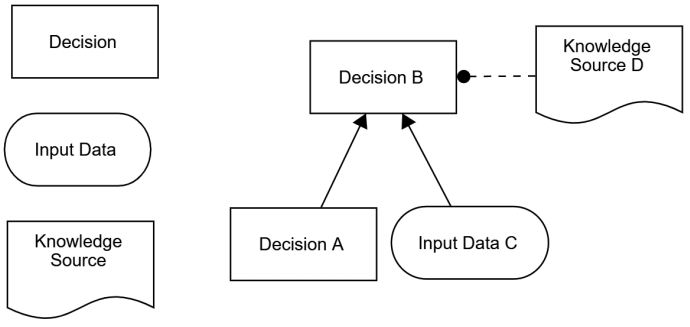
JAMES TAYLOR

PREVIEW *To ensure forecasts add value to organizational decision making, James Taylor argues for a formal, robust, and structured decision model to identify relevant forecasts and their features during decision making. He proposes a Decision Model and Notation (DMN) approach to align and focus forecast development, and to ensure forecasters understand how their forecasts impact decision making.*

Forecasts are powerful inputs for changing the behavior of an organization. Good forecasts can lead to reduction in costs and waste, streamline operations, and increase sales. Yet the act of creating the forecast is only the first step – the organization must change its behavior in response to the forecast. After all, if nothing changes because of the forecast, what good was it? Specifically, organizations must make *decisions* in the light of the forecasts. Decisions such as restocking, pricing, discounts or marketing offers, as well as many others could be made differently depending on the specifics of the forecasts. Robette (2023) highlights that forecast accuracy improvements can sometimes lead to degradation of decisions being made, and urges organizations to focus on usefulness of the forecast rather than forecast accuracy.

To increase visibility of a forecast's relation to a decision and to ensure forecast usefulness, we need a robust, formal model of the decisions that are affected by the forecast. A decision model can show exactly how the forecast impacts the decision, what considerations besides the forecast are relevant, and how all the pieces fit together. It can show what information is required and what policies or constraints affect decision making.

Figure 1. Visual Building Blocks for a Decision



DECISION MODEL AND NOTATION

The Decision Model and Notation (DMN) standard is published by the Object Management Group, an organization responsible for standards. DMN provides a robust framework for building decision models that address the challenges of developing and deploying forecasts.

Core Elements of DMN

DMN has two facets: a visual modeling notation and a representation of decision logic. The core of the visual modeling notation is a set of three shapes and two lines, as shown in **Figure 1**.

- **Decisions**, shown as rectangles, capture the questions that must be answered to make a decision, and can be decomposed into sub-decisions (and sub-sub-decisions, etc.).
- **Input Data**, shown as flattened ovals, represent information provided to the decision-making context and with which the decision is to be made. Each Input Data has an information structure and typically represents an entity in a data model.
- **Knowledge Sources**, shown as document shapes, represent the policies, analyses, regulations, or best practices that guide or constrain the decisions in the model.

There are also annotations and group shapes, additional ways to link things together, and some technical objects to support the development of executable models. However, the core of the model is as described. These shapes can be linked together using two kinds of connectors:

- **Information Requirements** are solid arrows showing the inputs required by a decision. The input data or decision at the blunt end of the arrow is required by the decision at the sharp end. This defines dependency – it states that the decision cannot be made without those input data or sub-decision outcomes.
- **Authority Requirements** are dashed lines with round ends. They show which decisions are constrained or guided by each knowledge source, so you can see the impact of a policy or regulation on the model.

This simple notation is robust. Even very complex decisions can be modeled by decomposing into smaller decisions. The dependency network created by the information requirements shows exactly how information is used to make the decisions represented in the model.

A Decision Table

For a decision to be made by a human, the model described above is often enough. With documentation of each node, a human decision maker has a clear picture of how they should proceed. For decisions to be automated, however, an additional layer of decision logic needs to be specified.

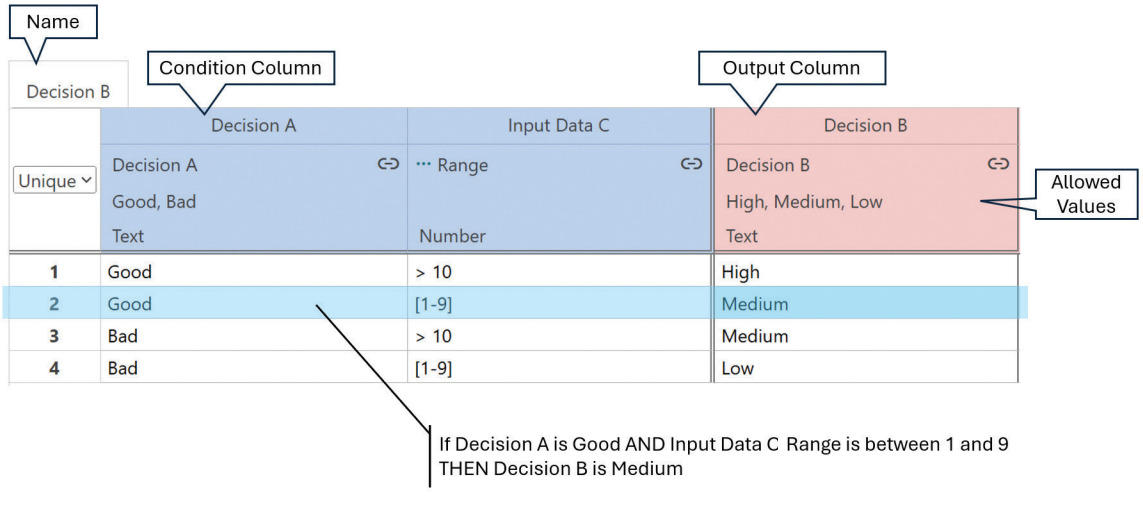
Key Points

- Forecasts are not inherently beneficial to an organization. Rather, a forecast provides value by improving organizational decision making.
- Forecasters need to understand the decisions that might be impacted by their forecasts, how those decisions are made today, and how those decisions are judged as "good" or "bad."
- Understanding and designing the decision *first* changes the assumptions about what forecasts will help in a wide variety of ways. This is useful when embedding forecasts (and other predictions) into automated operational decision processes.
- To maximize the value they provide to their organizations, forecasters should conduct decision modeling of a forecast's potential impact before they build their forecasts. An emerging standard framework for this is the Decision Model and Notation (DMN).

The most common way to do this is to develop decision tables for each decision in the model.

Figure 2 shows a table with the logic and rules for the decision in Figure 1. The columns are based on the information requirements in the diagram and each requirement (one to Decision A and one

Figure 2. Table with Decision Logic



to Input Data C) results in at least one column. Input Data C is a single attribute – a range. The rows show how to combine the values from each information requirement to make the decision. Every column is ANDed together on a row and each row is generally an alternative way to make the decision. The highlighted row, for instance, states that if the result of Decision A is Good *and* the range is between 1 and 9 then Decision B's output is Medium. Constraints on allowed values can be specified (as illustrated by "High, Medium, Low" under Decision B).

The DMN standard allows for various ways to manage these tables and supports other ways to represent logic, sufficient to define a wide range of decision types.

The Value-Add of DMN for Forecasting

DMN decision models add value in two ways to those developing and deploying forecasts: they help ensure that an existing forecast is being used correctly to influence decisions, and they can ensure the right forecast is developed in the first place.

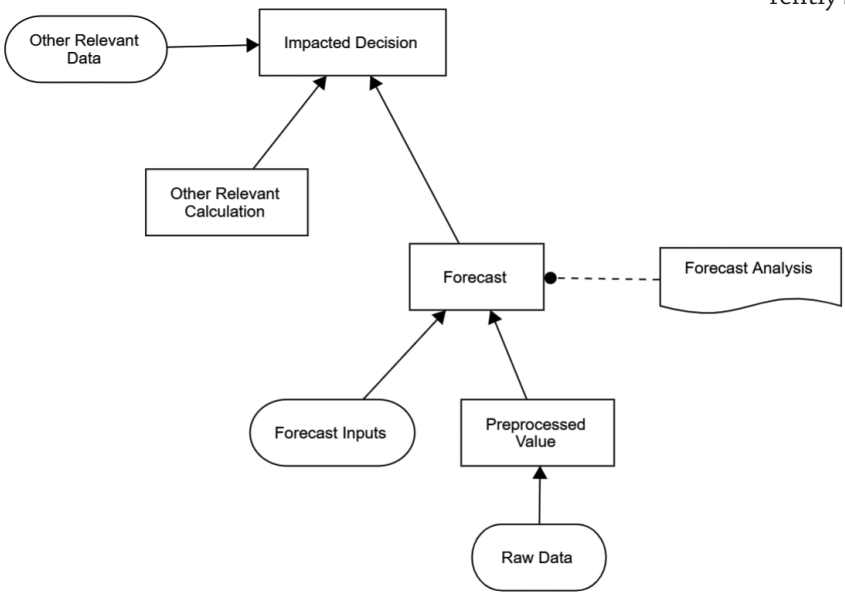
Once a team has developed a forecast, it can create decision models that show how the forecast is expected to be used in decision making. One decision in the model represents the decision as to what comprises the forecast itself. The information needed to calculate the forecast might be input data or sub-decisions. Raw data used by the forecast will be input data. If data is preprocessed before being input into the forecast, sub-decisions will capture those calculations. The analysis work done to create the forecast will be shown as one or more knowledge sources. The overall visual model informs the organization how the forecast works, as shown in **Figure 3**.

This model is then integrated into one or more models representing the decisions to be influenced by the forecast. In these additional models, one of the sub-decisions is the forecast itself, and all other relevant data and calculations involved are modeled as data input or sub-decisions. The role of the forecast is clearly shown, the rationale for different parts of the decision is captured as knowledge sources, and all the data required for the overall decision is apparent. If the decision needs to be automated, in whole or in part, the specific logic for the decisions involved can be specified.

Experience with the notation has shown another benefit. Building such a model *before* completing the development of a forecast provides valuable insight into exactly what kind of forecast will be most useful. It is often most effective to first build a model of how the decision is currently being made – without the proposed forecast. This may be an existing, well-understood decision-making approach, or it may require multiple iterations to capture a standard way to make decisions that have not been standardized previously.

Once a decision model is agreed to, a discussion can take place about which decisions could be changed and improved using forecasts, and what kind of forecasts would enable such a change. Decision points such as the accuracy that would be required by the use case, the time horizon, and the granularity can all be described based on the decision

Figure 3. A Decision Model Including a Forecast



model. The decision model precisely frames the need for the forecast.

DMN APPLIED TO A REPLENISHMENT USE CASE

Using DMN, the following simplified example shows how a replenishment decision is redefined to include a forecast. Three steps are used: current state modeling, opportunity identification, and future state modeling.

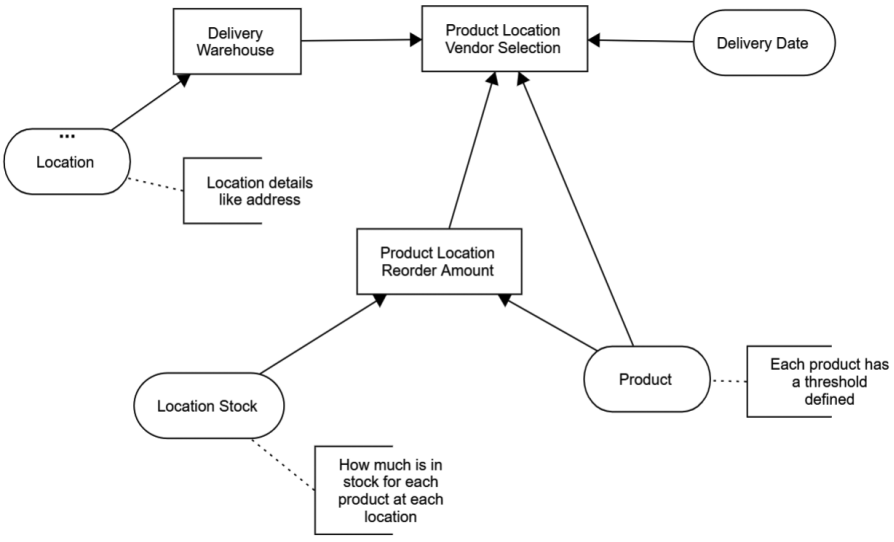
Current State Modeling

The most effective way to build a decision model with DMN is to begin with top-down analysis. Subject matter experts (SMEs) can describe how decisions are made today. Then, the DMN model can capture how these decisions are made. The model provides a visual blueprint and resolves inconsistencies and differences of opinion. SMEs generally find it easy to break down their decision making into its component parts. This approach rapidly outlines any differences between experts, as well as commonly reused sub-decisions.

Imagine a simple replenishment scenario. Today, a check is performed three working days before the end of the month at each location. If a product's stock level has dropped below a defined threshold, a replenishment order is placed. The order size is calculated, and a vendor is selected based on order size, required delivery date, and the location of a relevant warehouse. This decision is made for each product and each store every month. However, the company has found that it gets a significant number of stockouts or overstocking events, so it wants to reconsider the decision process to include a forecast.

The initial decision model might look like **Figure 4**. The reorder amount is calculated based on product threshold and location stock data. This reorder amount, as well as the date and delivery warehouse location, are used to pick a vendor.

Figure 4. Current State Decision Model Showing the Initial Replenishment Decision



Opportunity Identification

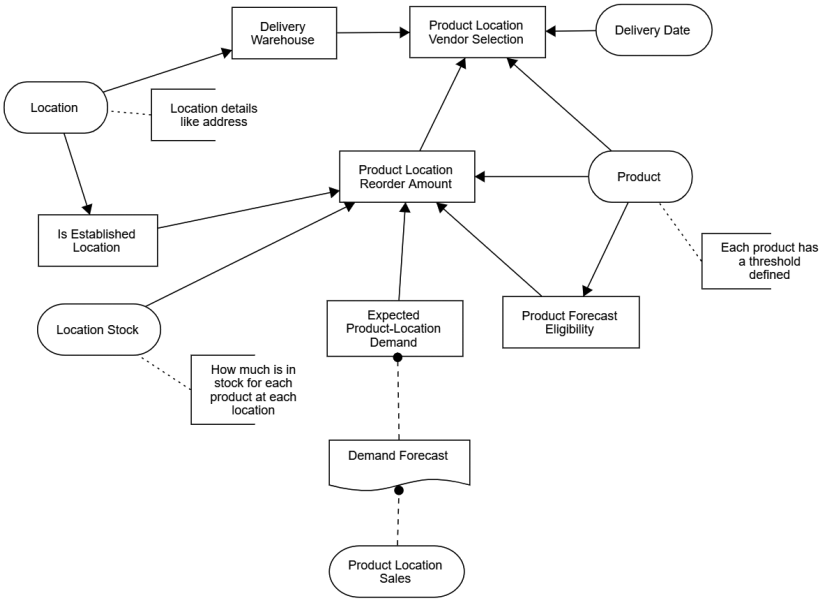
With the model defined, we can have a discussion of how the decision could be improved with forecasts. Specific decisions can be identified that could be improved by having a discussion like "If only we could forecast X we could make decision Y more accurately." The requirements for the forecast in terms of accuracy, timeliness, time horizon, etc. largely follow from how and when that decision is made.

In this example, the obvious way to improve the decision making would be to forecast demand and use that to set the size of the orders. This would require a forecast for sales of a given product in a given location. The forecast would need to be for the following month, and would have access to sales data for most of the current month so that it is accurate when the decision is made.

Future State Modeling

Assuming a forecast can be built, the decision model is adjusted to show how it will be used. The forecast(s) and supporting details (input data or sub-decisions to calculate interim values) are added to the model, and changes are made to show how the decisioning would change. Such a model is shown in **Figure 5**.

Figure 5. Future State Decision Model Showing How the Forecast Is Used



In our example, the forecasting team finds it can build forecasts of sufficient accuracy for most but not all products. It also finds that forecasts can't be usefully developed for locations until they have operated for 12 months. The forecast models will be continually updated, and the decision model needs to reflect these limitations. A new forecast decision is added with its inputs. Decisions to check the age of a location and forecast eligibility for a product are added. The reorder sizing decision is revised to use the forecast if the product and location are acceptable to the team and otherwise remain the same.

In the future, the logic for the product eligibility decision can be updated if the forecasting approach improves. Similarly, the requirement for 12 months of operations can be changed if that ceases to be necessary for useful development of the forecast.

DMN can be used beyond clarifying the impact of forecasts on decision making and transparently documenting forecasting and decision attributes. The new model can also be used to assess the impact of the forecast on historical decisions to see when the new approach would have resulted in a different reorder level. If the forecasts are effective, these differences should to a large degree correspond with the stock outages or overages that led to the initial discussion. Assessing the impact can be done manually or using simulation if the decision has been automated. Using DMN to both frame and monitor forecasting in this way helps an organization continually improve the usability and effectiveness of its forecasts.

James Taylor is the Founder and an Executive Partner at Blue Polaris (formerly Decision Management Solutions). He is a leading expert in how to use decision modeling, business rules, machine learning, and artificial intelligence to deliver business impact. For the last 20 years, James has provided strategic consulting to companies of all sizes and across all sectors to improve decision making and effectively adopt advanced technology. He is the author of several popular books, including *Digital Decisioning: Using Decision Management to Deliver Business Impact from AI*, and, with Jan Purchase, *Real-World Decision Modeling with DMN*. He is a member of the DMN Revision Task Force.

james@bluepolaris.com

REFERENCES

Robette, J. (2024). Forecast desirability: Is better the enemy of good? *Foresight*, 74, 24-29.

Decision Model and Notation standard available from the Object Management Group **omg.org/dmn/**

The Trade-Offs between Forecasting Performance and Computational Cost

FOTIOS PETROPOULOS AND EVANGELOS SPILIOTIS

PREVIEW *Forecasting performance is typically evaluated by statistical measures of forecast error, ignoring the computational cost of producing the forecast. Yet these costs, in both computer time and environmental impact, can be huge. Building upon previous research, Petropoulos and Spiliotis show how forecast computation time can be dramatically reduced without significant impact on forecast accuracy.*

Forecasting is a vital component of modern business operations, as every decision rests on an explicit or implicit forecast. As the size of organizations increases and company operations become more complex and intertwined, forecasting becomes even more necessary. Proper measurement of forecasting benefits and costs is critical.

Research by Yardley and Petropoulos (2021) suggests that the evaluation of forecasts should shift from simply measuring forecast error via traditional error measures (such as the MSE or MAPE) to more broadly determining a forecast's *utility* to the organization. In supply chain management, for instance, utility can be expressed in terms of reduced inventory, reduced backlog, increased customer service level, or in direct monetary terms. In financial forecasting, utility can be directly measured on the realized profits standardized by the investment risk.

In this article, we expand on prior discussions around "beyond forecast-error measures" to consider inclusion of the cost of producing the forecasts. Widely recognized are costs associated with the collection and maintenance of raw data, personnel costs for processing and analyzing the data (data analysts), model-development costs of new forecasting models (data scientists), costs of integrating new models into the production systems (data engineers), and costs of assessing and adjusting system forecasts using judgment (demand planners). However, frequently overlooked is another critical element: the computational cost (time required) for producing the forecasts.

Computation cost is of particular concern for organizations dealing with large amounts of data. Retailers such as Walmart and Tesco offer 50,000 to 200,000 different stock keeping units (SKUs) in their brick-and-mortar stores while operating 5,000 to 10,000 stores. Demand forecasts (for replenishment, inventory, and ordering) are required for every combination of SKU and store. The size of these challenges explodes for the online marketplaces. Seaman (2018) mentions that Walmart deals with about two billion combinations of SKUs and postcodes. Even if only a small fraction of these needs to be regularly forecast, the computation costs can be staggering.

This problem is of still greater concern with the growing use of machine learning (ML) solutions within forecasting systems. Powerful as they are, such solutions also impose additional computation costs. Producers and users of forecasts must recognize that there may be trade-offs between the incremental benefits of improved forecast performance and the corresponding increases in computation costs.

Containing computation costs leads not only to direct monetary savings but also has environmental benefits including

Key Points

- Evaluation of forecasting performance has traditionally focused on forecast error, without considering the forecast's utility to the organization or the cost of producing it.
- Producing forecasts entails both personnel costs (such as data analysis, model development, and reviewing/adjusting system forecasts) and computational costs.
- A "fast and frugal" approach considers suboptimal models, suboptimal parameters, and less frequent updating of models and parameters. This approach is found to yield comparable forecast accuracy while reducing computational cost.
- Evaluation of the forecast should include assessment of the computational costs of producing it. The financial and environmental savings can be huge for organizations, such as large retailers, that require millions of forecasts per period.

decreased CO2 emissions. Simplification of forecast generation – applying Occam's razor – may be warranted.

THREE COMPONENTS OF FORECAST-COMPUTATION COSTS

We can distinguish three components of computation cost:

- Cost of moving from a single model to a pool of models. (Do we need a large pool of candidate models?)
- Cost of identifying optimal parameters for forecasting models. (Does it matter if our selected parameters are slightly suboptimal?)
- Cost of updating forecasting models. (How often do we need to refit/reparameterize models or change from one model to another?)

In all three cases, it may be possible to reduce computation costs significantly with little deterioration in forecasting performance.

Suboptimal Models

A dominant view in forecasting literature is that we need to identify the optimal model form for the data in hand. If we're dealing with univariate time series forecasting, we could consider the 30 available exponential smoothing models and check which one works best for each of our time series data. These 30 models comprehensively cover different types of trend, seasonality, and error form. But do we really need to apply all 30 to each time series, or can we limit the test to a smaller set of alternative models?

Petropoulos et al. (2024) proposed a reduced set that consists of just the eight smoothing models shown in **Table 1**. They then applied these to each of the 50,000 monthly real time series from the M1, M3, and M4 Makridakis forecasting competitions. Their key finding was that this reduced set of models decreased

Table 1. The Eight Exponential Smoothing Models of the Reduced Set

Model	Description
ETS(ANN)	Additive errors, no trend, no seasonality
ETS(MNN)	Multiplicative errors, no trend, no seasonality
ETS(AAdN)	Additive errors, additive damped trend, no seasonality
ETS(MAdN)	Multiplicative errors, additive damped trend, no seasonality
ETS(ANA)	Additive errors, no trend, additive seasonality
ETS(MNM)	Multiplicative errors, no trend, multiplicative seasonality
ETS(AAdA)	Additive errors, additive damped trend, additive seasonality
ETS(MAdM)	Multiplicative errors, additive damped trend, multiplicative seasonality

computation time in the program R by 70%. Surprisingly perhaps, this impressive reduction in computational cost was accompanied by a 10% increase in accuracy based on the mean absolute scaled error (MASE) measure.

Given the popularity of exponential smoothing models in practice, we would advise demand planners to carefully consider a well-defined subset of the available models. These provide a cost-effective way to capture all the possible data patterns and simultaneously maintain accuracy while reducing computation time.

Similar results were seen when the reduced set was tested on the M5 dataset. There, accuracy of the reduced set was on par with the full 30 methods as well as with the well-regarded LightGBM machine learning ensemble. Computation time was reduced by at least 30%.

Given the popularity of exponential smoothing models in practice, we would advise demand planners to carefully consider a well-defined subset of the available models, such as the one presented in Table 1. These provide a cost-effective way to capture all the possible data patterns and simultaneously maintain accuracy while reducing computation time.

The subset test was also applied on the ARIMA family of models. Here, Petropoulos et al. created a reduced set by limiting the maximum order of the autoregressive and moving average terms. They found that for a large set of monthly data there was not much to gain in terms of point forecast accuracy beyond a maximum order of three. Additionally, for the estimation of the uncertainty parameters in the ARIMA models, a maximum order of just two gave the best performance while being 100 times faster than a maximum terms order of five.

Suboptimal Model Parameters

Theory suggests that we should seek an optimal set of parameters for our forecasting models. But Nikolopoulos and Petropoulos (2018) argued that, to the contrary, an optimal set of parameters is not necessarily the ultimate goal because it is a goal impossible to achieve. We can never know the true data generating process for our data; we can only fit models to the data we have already observed. This means we can identify parameters that are optimal for the past data, but this provides no guarantee that this set of parameters will also be optimal for the future data. Thus, it seems reasonable that modeling the data while taking shortcuts in optimizing the parameters could be beneficial, if doing so (a) reduces computation cost and (b) does not significantly sacrifice forecasting performance.

Nikolopoulos and Petropoulos tested this argument on over 300 monthly nontrended and nonseasonal time series from the M3 competition. For these series, they specified a simple exponential smoothing model that has one parameter: the

Grid-search optimization, also known as parameter sweep or exhaustive search, is a technique that identifies optimal parameters by exploring all possible values within a defined range, using a fixed increment. In our case, we aim to optimize a single parameter, alpha, which takes values in the range [0, 1]. Therefore, the search begins at 0, ends at 1, and proceeds with a step size of m, resulting in n=1/m incremental steps. The value of alpha corresponding to the lowest forecast error (typically according to the MSE) is selected as the optimal solution.

Trial-and-error optimization also identifies optimal parameters by repeatedly testing different values and observing the forecast error produced. Its difference lies in the fact that the testing is iterative, relying on experimentation that progressively narrows the search space for a predetermined number of steps rather than a predefined set of values. In our case, the algorithm starts by testing the values 1/3 and 2/3. The current optimal solution, s, would then serve as a focal point for defining the next two values to be tested

$$s - 1/(3 * 2^{(k-1)}) \text{ and } s + 1/(3 * 2^{(k-1)})$$

where k=2. For every subsequent step (k=3, 4, ...), a new focal point is defined and two new values are tested for optimality.

Note that none of these techniques guarantees finding the global optimal value.

smoothing constant alpha. To optimize alpha, they implemented two algorithms: grid search and trial-and-error.

In grid search, their results showed that as few as two optimization steps lead to forecasting performance that is, in statistical terms, as good as any other (higher) number of steps. Five optimization steps were enough to achieve performance that is top-ranked on average. This five-step increment procedure corresponds to selecting the value of alpha as either 0, 0.2, 0.4, 0.6, 0.8, or 1. Any further granularity in terms of identifying the optimal alpha (such as 0.24 instead of 0.2) has virtually no impact on the out-of-sample forecasting performance (as measured by the symmetric MAPE). Similar findings stand for the trial-and-error algorithm, providing the same performance (from a statistical point of view) for all optimization steps tested.

Thus, without implementing the full optimization, forecast accuracy was not sacrificed. But the computation time needed to produce forecasts decreased dramatically – on the order of 90% for the grid-search and 50% for trial-and-error algorithm. The reduced computation time would translate to direct monetary savings in terms of the usage of cloud computing services.

While the results of this research are restricted to a simple exponential smoothing model and a single parameter, we would expect that the insight of the low (or, virtually, no) difference between optimal and suboptimal solutions holds for more sophisticated models. We would argue that, in each case, the users need to decide the appropriate amount of sacrifice in forecasting performance if optimization is not performed to the fullest.

Infrequent Updating

Machine learning advocates argue that global ML models are computationally more efficient (and thus less costly) than conventional forecasting models. This is because once the global model is estimated, it immediately produces forecasts for several time series and across several time periods. Most of the computation cost associated with global ML models is one-off, whereas traditional univariate time series approaches need to be re-estimated (arguably) every period.

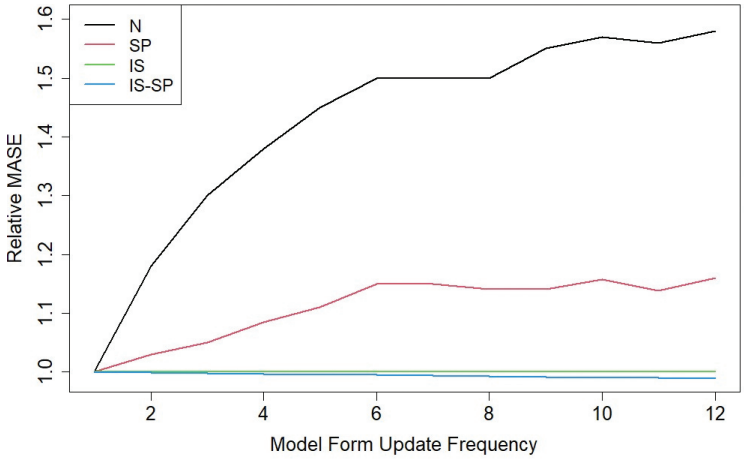
Is this the case, though? How often do we really need to update our univariate models? And, when we do, do we need to update only the model parameters, or should we also reconsider the best choice of model (such as adding/removing trend or seasonality)?

In Spiliotis and Petropoulos (2024), we examined the effects on forecasting performance and computational cost of the frequency that a new model form (such as a model with trend but no seasonality, or a level-only model) is selected. We allowed this frequency to vary so that models could be estimated as frequently as every single period (every time a new data point becomes available) or as infrequently as once a year. When this frequency equals 1, then a new model (and its parameters) is re-estimated in every single period. However, as this frequency decreases (i.e., model updating is done less frequently; as in every two or every three periods), the model form will be kept fixed for several periods while its parameters may still be updated every single period. We considered four updating scenarios for the model parameters:

- (N) No updating (neither initial states nor smoothing parameters).
- (SP) Only the smoothing parameters are updated.
- (IS) Only the initial states of the model are updated.
- (IS-SP) The initial states and the smoothing parameters of the model are updated.

For each of these four scenarios, we explored the effect on computation time and forecasting performance for how frequently updates are performed. When the updating frequency is equal to 1, then all scenarios are identical as a new model form (and its parameters) is specified every single period.

Figure 1. Relative Forecast Accuracy (MASE) of the Four Model-Updating Scenarios



Our analysis of nearly 50,000 real time series showed that for three of the four scenarios (all except N), updating less frequently does not have an adverse effect on forecasting performance. Interestingly, simply updating the initial states (keeping model form and smoothing parameters fixed) results in significant computational savings and these are proportional to the inverse of the updating frequency. This was to be expected, as most of the cost for univariate forecasting modeling is related to the number of the models that we opt to estimate every period. If the cost is reduced to simply re-estimating one model rather than refitting all possible models, then the computational burden becomes trivial.

Figure 1 shows the relative forecast accuracy (MASE) of the four updating scenarios (N, SP, IS, and IS-SP). It was created from the M4 monthly data using the ETS method under 12 different model form update frequencies (1-12 months).

Importantly, we found that the optimal update frequency for the model selection is between four to eight months for a majority of these time series. Since a global ML model would possibly also need to be re-estimated at least once a year, its computational cost would be comparable to that of univariate models.

MONETARY SAVINGS AND THE ENVIRONMENT

In our consideration of three ways to save on computation costs, each intervention is applied in isolation from the others. But one could well consider a forecasting support system that simultaneously (i) employs a reduced/suboptimal set of models, (ii) suboptimally estimates model parameters, and (iii) less frequently updates the models. Applied jointly, these three interventions could multiply the savings in computation costs without deterioration in forecasting performance; however, this is to be confirmed empirically through future research. Also, while our discussions focused on conventional univariate forecasting models (such as exponential smoothing and ARIMA), we expect that the same principles apply to other forecasting methods.

Many modern organizations rely on cloud computing services to complete data science-related tasks, including the production of forecasts. For them, the computational gains described earlier can be translated into direct and significant monetary savings. In all three studies mentioned above, the authors perform back-of-the-envelope calculations and estimate the savings that could be achieved by large retailers like Walmart or Tesco. Perhaps even more important,

the benefits of reduced computational costs extend directly to the environment. Petropoulos et al. (2024) mention that the adoption of the reduced set of exponential smoothing models by large retailers operating online marketplaces and producing daily forecasts could result in a yearly reduction of carbon footprint that equates to the CO2 absorbed by 3.2 million trees, while also saving about $1,000,000 in computational resources. The associated monetary and sustainability benefits would significantly increase if the other two interventions were to be applied at the same time.

Based on this evidence, it's time to stop evaluating forecasts purely on statistical error measures. Instead, we should extend our forecast assessment to computational costs and other costs associated with producing the forecasts.

Fotios Petropoulos is Professor of Management Science at the University of Bath. He is Editor of the *International Journal of Forecasting* and Associate Editor of *Foresight*. His research focuses on time series forecasting, judgmental forecasting, model selection, and integrated business forecasting processes.

f.petropoulos@bath.ac.uk

Evangelos Spiliotis is Assistant Professor at the School of Electrical and Computer Engineering, National Technical University of Athens. He is the co-organizer of the M4, M5, and M6 forecasting competitions and an Associate Editor of both the *International Journal of Forecasting* and *Foresight*.

spiliotis@fsu.gr

REFERENCES

Nikolopoulos, K. & Petropoulos, F. (2018). Forecasting for big data: Does suboptimality matter? *Computers & Operations Research*, 98, 322–329.

Petropoulos, F., Grushka-Cockayne, Y., Siemsen, E., & Spiliotis, E. (2024). Wielding Occam's razor: Fast and frugal retail forecasting. *Journal of the Operational Research Society*, 1–20.

Seaman, B. (2018). Considerations of a retail forecasting practitioner. *International Journal of Forecasting*, 34(4), 822–829.

Spiliotis, E. & Petropoulos, F. (2024). On the update frequency of univariate forecasting models. European *Journal of Operational Research*, 314(1), 111–121.

Yardley, L. & Petropoulos, F. (2021). Beyond error measures to the utility and cost of the forecasts. *Foresight*, 63, 36–45.

Wayfair wins IIF Forecasting in Practice Competition

On March 2, 2025, finalists for the IIF's *Impact of Forecasting in Practice Award* gave their presentations before an adjudication panel at the *Foresight* Practitioner Conference. The IIF is pleased to announce Wayfair as winner of the $10,000 prize and congratulates the other finalists from Ipiranga, Maersk, OpenGrid Europe, and HP. In a forthcoming special feature, *Foresight* will publish papers from the finalists describing their approach and its impact.

Award competition chair Chris Fry (left) announcing the winner

Two-Part Forecasting for Time-Shifted Metrics

HARRISON KATZ, ERICA SAVAGE, AND KAI THOMAS BRUSCH

PREVIEW *Many commercial sectors (such as hospitality) face the challenge of forecasting metrics that span multiple time axes – where the timing of an event's occurrence is distinct from the timing of its recording or initiation. In this paper, Katz, Savage, and Brusch present a novel two-part forecasting methodology that addresses this challenge by treating the forecasting process as a* time-shift operator. *The approach combines univariate time series forecasting to predict total bookings on booking dates with the Bayesian Dirichlet Auto-Regressive Moving Average (B-DARMA) model. The aim is to forecast the allocation of future bookings across different trip dates based on the time between booking and trip (lead time). This approach provides a sensible solution for forecasting demand across different time axes, offering interpretable results, flexibility, and the potential for improved accuracy. The efficacy of the two-part methodology is illustrated through an analysis of Airbnb booking data.*

TIME-SHIFTED METRICS

Accurate demand forecasting is essential across industries for optimizing operations, managing resources, and strategic planning. Yet many sectors face the challenge of *time-shifted metrics*, where the event date (e.g., booking, order, or trade date) does not match the date the service is consumed or settled (trip date, delivery date, settlement date). Such temporal separation often causes inconsistencies and reduced accuracy when traditional, single-axis forecasting methods are applied.

Figure 1 illustrates the *time-shifted* nature of the data, where a single booking date might correspond to a trip starting immediately or up to weeks later. In hospitality settings, the total nights (or stays) initially forecast from the booking perspective can change by the time the trip date arrives, often due to cancellations or modifications. This time-shifted nature can also appear in supply chain (purchase vs. delivery), retail (sale vs. shipment), healthcare (appointment vs. consultation), and finance (trade vs. settlement). Traditional forecasting approaches that

Figure 1. Heatmap of Simulated Booking Counts for Each (booking date, trip date) Pair

operate within a single temporal framework struggle to track how the metric transitions from one origin to another.

While hierarchical time series methods indeed aim to reconcile forecasts at different organizational or product levels (Hyndman et al., 2011), they do not directly handle the "two-axis" structure arising from time-shifted metrics. Here, the primary challenge is that one axis (e.g., booking date) must be aligned with a distinct time axis (e.g., trip date), rather than just different levels or categories on a single timeline. Similarly, temporal aggregation approaches address scaling forecasts up or down in time granularity (Silvestrini & Veredas, 2008), but they do not typically involve distributing one metric across *another* time dimension. Hence, while our framework is conceptually adjacent to these literatures, time-shifted forecasting remains a unique problem requiring distinct methods.

In contrast, compositional data analysis provides a way to model proportions that sum to a whole (Aitchison, 1986; Zheng & Chen, 2017), but its application to lead-time distributions in a multi-axis setting has been limited. Hybrid strategies that combine univariate and compositional techniques (Armstrong, 2001) can help bridge these gaps.

Our paper introduces a two-part forecasting methodology that treats the process as a time-shift operator. We first project total demand on the booking axis, then translate those forecasts to the trip axis using a compositional time series model. For Airbnb, these predictions inform a wide range of decisions – such as dynamic pricing, host recruitment, staffing programs, and marketing campaigns – so that supply can be aligned with expected guest stays. Even a seemingly small decrease in forecast error can translate into substantial cost savings or revenue optimizations at scale. After describing this approach in more detail, we present our analysis of Airbnb data and conclude with broader insights on how it can apply to other industries.

Key Points

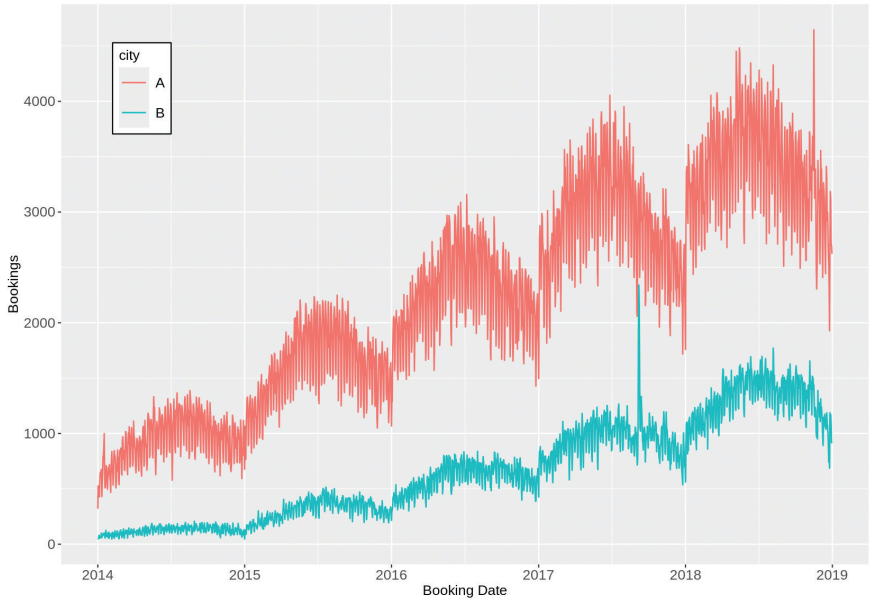
- This paper introduces a two-part methodology that combines univariate time series forecasting with the B-DARMA model to address the challenge of forecasting time-shifted metrics in industries where timing of events and their recording differ.
- The B-DARMA model is designed for compositional time series data, modeling lead time distributions by capturing temporal dependencies and compositional constraints inherent in such data.
- By decoupling the forecasting process into two components, the methodology allows for independent adjustments and incorporation of external variables enhancing adaptability to changing conditions without overhauling the entire system.
- In testing on Airbnb booking data, this approach has delivered forecasts that are interpretable and more accurate than a bottoms-up benchmark. It can be adapted to other sectors facing similar challenges with time-shifted metrics, including supply chain management, retail, manufacturing, healthcare, and finance.

METHODOLOGY

Full mathematical derivations of the Bayesian Dirichlet Auto-Regressive Moving Average (B-DARMA) model, including the additive log-ratio transformations, can be found in our online supplement. Katz et al. (2024) provides additional technical details on the methodology and our analysis.

The two-part model structure begins with a forecast of total bookings made on a booking date regardless of trip date. For this, a univariate time series model (e.g., ARIMA, Prophet, or exponential smoothing) is applied to historical daily bookings. Given its robustness to trend changes and seasonality, we used Prophet (Taylor &

Figure 2. Daily Airbnb Bookings by Booking Date Over Training Period (2014-2018)



Letham, 2019) to obtain our forecast of the total bookings for each future date.

The second part of the model uses B-DARMA for lead-time allocations. It begins by creating a vector representing the proportions of bookings made on each day that fall into each lead-time bucket (e.g., 0-1 month, 1-2 months, etc.). Then we utilize B-DARMA, which provides a structure for the means in the transformed space.

Having computed the forecasts of total bookings on each date, along with the proportions of bookings each day falling into each lead-time bucket, we multiply them. This combines the two parts of the model structure.

DATA

Our example employs two anonymized Airbnb datasets:

City A: A large metropolitan market with strong seasonal variability.

City B: A midsized leisure destination with more moderate seasonality.

Each dataset spans six full years (January 2014 to December 2019, before the COVID lockdowns) at a daily granularity. Each contains the number of bookings made on day t, the trip date (or month) of each booking, and lead time in months. We create *monthly lead-time buckets* from 0 to 12, forming 13 possible lead times. The daily bookings are shown in **Figure 2** while the lead-time proportions are shown in **Figure 3**.

In City B, we observe a notable spike on the booking-date axis around September 1, 2017. In City A, there is a lesser spike around November 2018. These surges are not due to data errors; instead, they likely stem from external triggers that prompted many reservations within a short window. Typical examples include extreme weather advisories leading to last-minute changes, sporting events where the final location is confirmed late in a playoff series, or major concerts/music festivals that announce dates and release tickets at once, causing a rapid influx of bookings when fans learn the time and venue. Such real-world events can produce abrupt jumps in the booking-date series, even if the trips themselves occur on future dates.

Our training period was five full years (2014 through 2018) with test period January 1 through December 31 of 2019.

We fit all models on the *training window*, then compare forecast accuracy over the *test window* (one full calendar year).

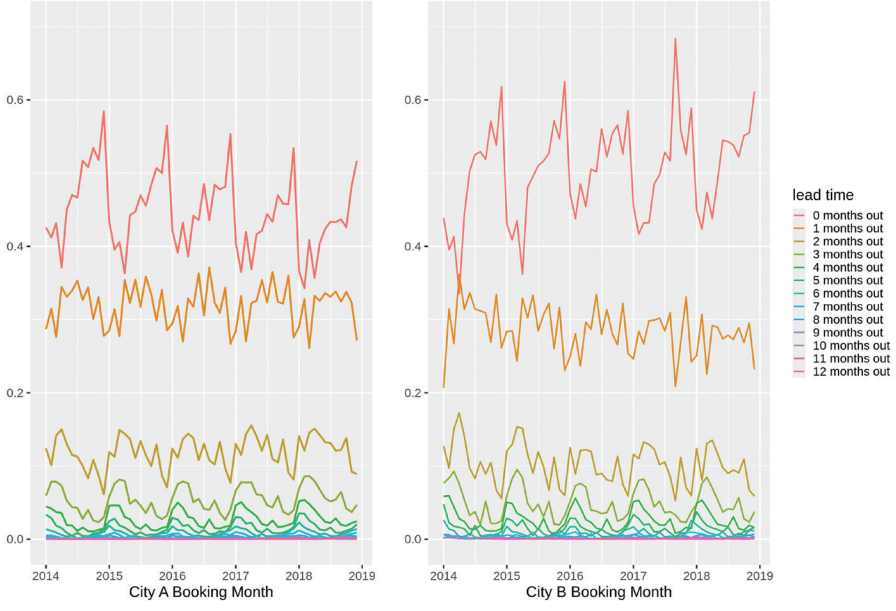
TWO-PART METHOD IMPLEMENTATION

In our approach, we first forecast total daily bookings on the booking-date axis for each city by applying Prophet (Taylor & Letham, 2018) to the five years of training data (2014-2018). For City A, the model incorporates weekly and annual seasonal components, as well as holiday factors reflective of a large metropolitan market. City B, by contrast, is a midsized leisure destination and therefore employs slightly different holiday and event indicators to account for its unique patterns. From these daily forecasts, we produce monthly total bookings for the one-year test window.

Having obtained the expected number of bookings for each day, we then address how those bookings spread across various lead times. Specifically, we record the proportion of bookings – ranging from zero to 12 months in advance – on each booking month. To model these allocations, we use a B-DARMA(1,0) approach in which this month's lead-time distribution depends on recent patterns. We also incorporate monthly seasonality via Fourier terms and include a linear trend to capture shifting booking behaviors throughout the calendar. This yields a monthly sequence of lead-time proportions for 2019, providing insight into how the share of last-minute versus long-term bookings evolves over time.

With total daily bookings and monthly lead-time proportions in hand, we first aggregate our daily booking-date forecasts into monthly totals. We then multiply each month's total bookings by the corresponding lead-time proportions and shift these results forward by the appropriate monthly offset (0 to 12). By summing across all booking months that align with a given trip month, we obtain the final forecast of how many stays (or similar events) will occur in that month. This step effectively translates the original booking-date perspective into the trip-date perspective, revealing when actual consumption or usage is expected to take place.

Figure 3. Proportion of Monthly Bookings by Lead Time Over Training Period (2014-2018)



BENCHMARK: BOTTOM-UP PROPHET APPROACH

To highlight the potential benefits of modeling lead times compositionally, we compare our method with a simpler bottom-up Prophet approach. In this benchmark, we create a separate univariate Prophet forecast for each monthly lead-time bucket. For instance, one model predicts the number of bookings on day t that are for a trip starting within the same month, another model does so for bookings with a trip date that falls in the following month, and so on up to 12 months out from the booking month. Summing these daily forecasts across all buckets gives an estimate of total daily bookings, which we can compare to the original Prophet forecast in our Part 1. Converting each bucket's forecast to a proportion of the monthly total also provides a compositional perspective we can pit against the B-DARMA outputs.

Additional Monthly-Prophet Benchmark. In the supplementary materials, we also provide results from a benchmark that applies Prophet at the monthly level (by lead-time bucket), allowing us to compare performance under a coarser temporal aggregation. Full details of the model setup and comparison metrics for this monthly Prophet approach can be found in Supplementary Material S3.

METRICS AND RESULTS

We assess both the booking-date and trip-date forecasts using several criteria. MAPE captures how far off the forecast is in relative percentage terms, while MAE shows absolute differences between forecasted and actual totals. We also include a normalized L^1 distance for compositional vectors, sometimes called Manhattan distance, to measure how closely the forecasted proportions align with actual lead-time distributions for each booking month. In situations where the specific breakdown of bookings across lead times is critical – such as staffing hotels or planning supply chains – this compositional accuracy can be as important as the raw total demand forecast.

BOOKING-DATE AXIS

Table 1 shows monthly aggregated forecast errors on the *booking-date axis* for 2019. It compares our two-part method to the bottom-up Prophet approach over the test window, in terms of both MAE and MAPE. It also shows mean normalized L^1 distance for lead-time distributions to the trip-date axis. Lower values in each metric indicate better performance.

Additionally, we tested a Prophet-based benchmark at the monthly level (rather than daily) by lead-time bucket; the supplementary materials provide full details. Results were largely consistent with the daily bottom-up forecasts once aggregated to a monthly scale.

For both City A and City B, the two-part approach (Part 1 total + lead-time from B-DARMA) generally tracks overall daily bookings well, with average MAPE around 4.8% for City A and 3.1% for City B. The bottom-up Prophet sum occasionally lags behind changes in overall level demand, having marginally higher MAPE (5.1% for City A and 3.2% for City B).

LEAD-TIME DISTRIBUTION

The far-right column in Table 1 compares the mean normalized L^1 distance for the lead-time distributions in Cities A and B.

Table 1. Performance Metrics Over the Test Window (2019)

City	Method	Booking Date MAE	Booking Date MAPE	Lead-Time Mean Normalized L^1
A	Two Part	5083	4.8%	0.0229
A	Bottom-Up	5336	5.07%	0.0389
B	Two Part	1406	3.07%	0.0300
B	Bottom-Up	1455	3.15%	0.0499

Figure 4. Monthly Lead-Time Proportions in 2019 for Cities A (top) and B (bottom)

Figure 5. Normalized L^1 Distance by Booking Month in 2019

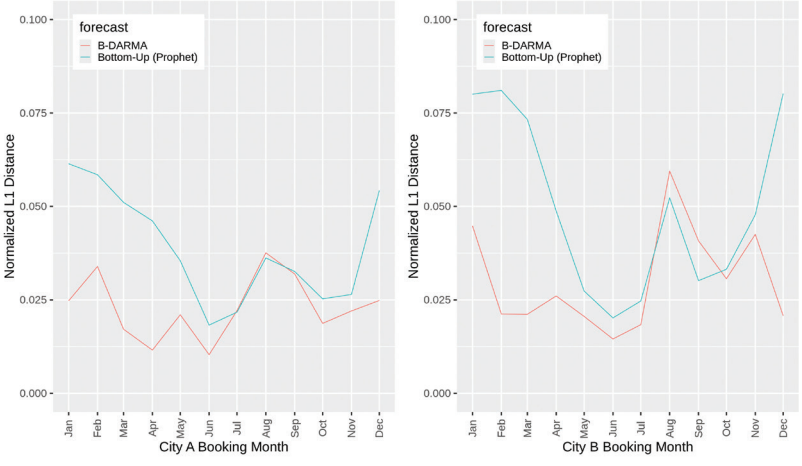


Figure 4 shows the forecasts for the lead-times and the normalized L^1 for each booking month in the 2019 test window.

Each small subplot corresponds to one of 13 lead-time buckets, where "0" indicates bookings within the same month, "1" is next-month bookings, "2" is two-months-ahead, and so forth. Actual proportions appear in red, with forecasts from B-DARMA in green and bottom-up Prophet in blue, illustrating how each method captures the evolving share of bookings across different lead times.

Figure 5 shows the Normalized L^1 distance by booking month in the 2019 test window for each forecast method (B-DARMA vs. bottom-up Prophet). The left panel shows results for City A, and the right panel for City B. Lower values indicate closer agreement between forecasted and actual lead-time distributions.

In both markets, the two-part (B-DARMA) approach outperforms the bottom-up Prophet model. For City A, B-DARMA achieves a mean normalized L^1 distance of 0. 0229 (vs. 0.0389)). City B, while exhibiting higher overall volatility, shows a similar pattern: 0.030 (B-DARMA) vs. 0.0499 (bottom-up) for the L^1 distance. These results indicate that the two-part

method's compositional framework captures cross-bucket correlations more effectively, leading to more accurate lead-time allocations than independent univariate forecasts.

DISCUSSION

Accurate and consistent forecasting across multiple time axes remains a challenge in many industries. By treating the forecasting process as a time-shift operator, our two-part methodology not only provides coherent forecasts but also improves interpretability and flexibility. In our analysis of City A and City B, separating the forecasts into total bookings (Part 1) and compositional lead-time distributions (Part 2) led to lower error rates on both the booking-date and trip-date axes compared to a bottom-up Prophet approach. This is because B-DARMA captures the cross-bucket correlations that univariate bucket-by-bucket models often overlook.

Accurate and consistent forecasting across multiple time axes remains a challenge in many industries. By treating the forecasting process as a time-shift operator, our two-part methodology not only provides coherent forecasts but also improves interpretability and flexibility.

A major advantage of this two-part method lies in its modularity and adaptability. Adjusting total forecasts in response to macroeconomic or event-driven shocks, for instance, can be done without refitting the lead-time model, allowing rapid scenario analyses when unexpected changes occur. This modular structure also extends naturally to short-horizon forecasting, where some future bookings are already known. If, for example, today is January 21 and occupancy is forecast for January 30–31, a portion of those stays may already be booked. In such cases, these existing reservations serve as a baseline or "backfill" on the trip date, while the univariate booking-date forecast and B-DARMA compositional vectors project any *additional* bookings that might still materialize. Because lead-time allocations remain anchored to the booking date, it is straightforward to align incremental forecasts with the appropriate trip dates, preserving both the scenario-testing capability for total demand and the advantage of having prior knowledge about near-term reservations.

Moreover, the B-DARMA model can incorporate exogenous covariates not just for the booking date (e.g., day-of-week, macro factors) but also for the trip date. One could, for instance, add a Super Bowl or Easter indicator to the relevant trip-date bucket, thereby shifting proportions if that event drives higher demand at certain lead times. By adding holiday/event covariates in the compositional (alr) space, the model can directly link a "trip date" feature to the observed lead-time allocations, ensuring that special events feed back into both total demand and how that demand is distributed over the booking horizon. This flexibility helps unify the perspective of forecasting "for" a particular day (trip date) with the perspective of forecasting "on" a particular day (booking date) under one cohesive framework.

However, this methodology is not without limitations. Splitting the forecasting process into two parts may overlook interactions between total demand and lead-time behavior that a unified model could potentially capture. Additionally, B-DARMA's compositional foundation generally assumes that proportions remain strictly positive, so extremely sparse or zero-valued lead-time buckets can pose modeling challenges. If lead times are of no particular interest or remain largely static, the added complexity of a compositional model may not justify its use, and a simpler bottom-up or univariate strategy could suffice. Another potential

enhancement is to incorporate probabilistic intervals, given that B-DARMA's Bayesian framework naturally supports credible intervals for lead-time proportions.

Despite these caveats, our results suggest that for scenarios where dynamic lead times meaningfully influence resource allocation or demand planning, a two-part compositional framework delivers valuable improvements in accuracy and interpretability. Extending the approach to hierarchical structures (e.g., city vs. region) or further temporal aggregations offers promising avenues for future research.

Acknowledgements

The authors thank Sean Wilson, Liz Medina, Jenny Cheng, Jess Needleman, and Peter Coles for helpful discussions, Ellie Mertz for championing the research, and Lauren Mackevich and Lori Callo for their indispensable operational support.

Data and Code Availability

The primary dataset used in our study "Two-Part Forecasting for Time-Shifted Metrics" is not publicly available due to confidentiality constraints. However, the supplementary material containing the full model specifications, including the Stan code for the Bayesian Dirichlet Auto-Regressive Moving Average (B-DARMA) model, is available for public access. This supplementary material can be found at our GitHub repository: ***github.com/harrisonekatz/consistent_forecasting_bdarma_paper.***

REFERENCES

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. Chapman & Hall.

Armstrong, J.S. (2001). Combining Forecasts. In: Armstrong, J.S. (eds) *Principles of Forecasting*. International Series in Operations Research & Management Science, vol 30. Springer, Boston, MA.

Hyndman, R.J., Ahmed, R.A., Athanasopoulos, G., & Shang, H.L. (2011). Optimal combination forecasts for hierarchical time series. *Computational Statistics & Data Analysis*, 55(9), 2579–2589.

Katz, H., Brusch, K.T., & Weiss, R.E. (2024). A Bayesian Dirichlet auto-regressive moving average model for forecasting lead times. *International Journal of Forecasting*, 40(4), 1556–1567.

Silvestrini, A. & Veredas, D. (2008). Temporal aggregation of univariate and multivariate time series models: A survey. *Journal of Economic Surveys*, 22(3), 458–497.

Taylor, S.J. & Letham, B. (2018). Forecasting at scale. *The American Statistician*, 72(1), 37–45.

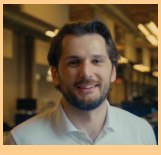
Zheng, T. & Chen, R. (2017). Dirichlet ARMA models for compositional time series. *Journal of Multivariate Analysis*, 158, 31–46.

Harrison Katz holds a PhD in statistics from UCLA. Currently, he serves as a Tech Lead Data Scientist on Airbnb's forecasting team. Prior to Airbnb, Harrison held research positions at the Federal Reserve Board of Governors in risk analysis, specializing in credit default swap markets. His applied research interests lie in financial markets and economics, while his statistics research focuses on Bayesian forecasting and compositional time series.

harrison.katz@airbnb.com

Erica Savage is an Associate Principal of Forecasting at Airbnb, where she contributes to the development of advanced statistical models to support strategic business decisions. Her research interests lie at the intersection of data science and finance, forecasting within the hospitality industry, lead time analysis, and revenue management.

erica.savage@airbnb.com



Kai Thomas Brusch is a Senior Manager at Airbnb. He leads teams responsible for forecasting, competitive strategy, and executive reporting. Kai holds a master's degree in computer science from the University of Hamburg and has over 13 years of experience in data science and finance at Airbnb. His research interests are in marketplaces, forecasting, and Bayesian methods.

kai.b@airbnb.com

Retrieval-Augmented Forecasting: Bridging Human Insight and Machine Precision

RYAN FATTINI AND RYAN YOUNG

PREVIEW *The rapid evolution of retrieval-augmented generation (RAG) systems has profoundly enhanced the capabilities of large language models through a technique known as* grounding. *This technique enriches foundation models with verifiable sources of contextual information they were not originally trained on, thus reducing model guesswork (hallucinations) and improving the accuracy of model inferences. Building on these advancements, Ryan Fattini and Ryan Young introduce a novel application of retrieval-augmented forecasting. Their system leverages natural language processing (NLP) to address common practitioner challenges, such as ease of use, data scarcity, and the complexity of modeling interactive effects. Additionally, it transforms conventional ensemble techniques into a multiplayer, reward-based system where user-generated (player) scenarios are evaluated and weighted according to past accuracy and bias.*

The man-versus-machine dilemma is nothing new to forecasters. It has existed for decades within the teams responsible for making judgmental forecasts under uncertainty (Tversky and Kahneman, 1974). Within an organization's operational, planning, and strategic teams, there is a mixture of advanced statistical models run by data scientists, and manual overrides driven by expert intuition. This is a domain where human and machine forecasting regularly intersect, with the two camps locked in perpetual debate – weaponizing varying degrees of statistical rigor, nuanced qualitative arguments, politics, rank, and ego (Vandeput, 2021). In many cases, this debate results in suboptimal conclusions.

Recent advancements in large language models (LLMs) underscore a pivotal shift towards conversational artificial intelligence (AI), and forecasting is no different. Natural language processing (NLP) transforms every machine interaction and transforms our ability to predict the future. By integrating LLMs and NLP technologies, this article describes a system where conversational AI becomes a cornerstone of modern forecasting practice.

THEORETICAL UNDERPINNINGS

The application of LLMs to forecasting has focused on the self-attention mechanism of transformers and has produced encouraging but mixed results (Bergmeir, 2024). The primary objectives had been improving accuracy over the benchmarks, reducing training time, and *zero-shot forecasting* (which is the ability to generalize new time series or domains not included in its training set). Our proposed RASOR (Retrieval Augmented Semiotic Recursion Framework) similarly focuses on improving forecasting accuracy. However, its target priorities are elements at the heart of the practitioner's dilemma:

- How to optimize the hybridization of humans and machines?
- How to make it easy for any subject matter expert (SME) to produce sophisticated, yet accurate forecasting scenarios?
- How to reduce forecast bias?
- How to measure human intent when new variables are considered provoking judgment overrides?

To illustrate the RASOR approach, we analyze the revenue of a retail organization in the first quarter of 2022. This period was marked by unique macroeconomic twists and turns, and therefore serves as a practical example of how our system can capture the variability of real-world data. Specifically, we will show how some key macroeconomic indicators can be incorporated into the forecasting process without adding additional features, more historical data, or retraining the entire model.

A CONCEPTUAL RETAIL EXAMPLE

In the first quarter of 2022, quarter-over-quarter U.S. GDP dropped by 2%, inflation increased by 2.2%, the unemployment rate dropped 8.8%, and median home values increased by 2.3%. There was also an 8% decrease in the organization's revenue. A typical way to model the effect of these indicators would be to use a linear or nonlinear function with feature coefficients and relevant interaction terms that describe their influence on the dependent variable (organization's revenue). Conventional approaches estimate these coefficients and terms by training a model on historical data collected across several disparate data sources.

Now consider that same window of time, 2022 Q1, but from the perspective of a human analyst. How would an economic analyst describe the influence given the context of the same macroeconomic indicators? The summary could go like this:

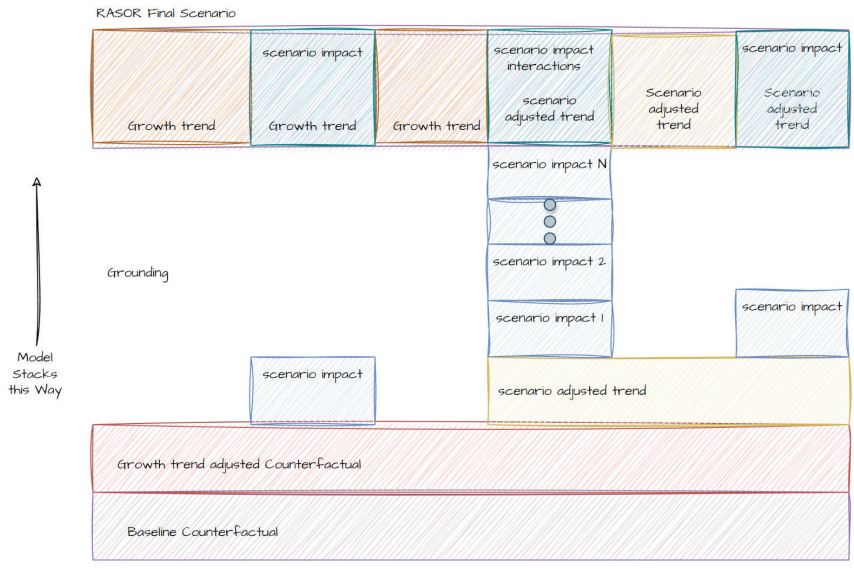
> *GDP dropped by 2% due to supply chain disruptions and ongoing pandemic effects. However, the job market remained strong, with unemployment dropping 8.8%, indicating robust job growth. Inflation hit a 40-year high increasing 2.2%, propelled by heightened consumer demand and rising energy costs. Meanwhile, the housing market cooled, with higher mortgage rates slowing home sales and moderating price increases.*

The human analyst's description covers the same macroeconomic indicators. However, there is added nuance, with the interactions between different indicators explained as "propelled by heightened consumer demand and rising energy costs" or "due to supply chain disruptions and ongoing pandemic effects." Classical models cannot easily include these interactions. The nuanced analysis presents the challenges of encoding these data points numerically and capturing the forecaster's intuition. It is often a difficult task to mine the disparate data sources and engineer features to represent these interactions. Deep learning methods could capture the interactions for us, but

Key Points

- Using advancements in RAG techniques within LLMs, a forecasting system becomes more adaptable to data changes, including real-time changes. The system can incrementally train a forecasting model based on new information, reducing the latency typically associated with incorporating recent events into forecasts.
- The system introduces a novel approach to traditional numerical data processing by integrating natural language processing (NLP) into forecasting. This enhances the system's ability to process and interpret complex, multivariate data by using the more nuanced understanding and interpretative flexibility offered by natural language.
- Utilizing natural language inputs, the system introduces a conversational approach to forecasting, enabling users to generate and refine forecasts without the need for deep technical knowledge. This democratizes advanced planning analytics, making it accessible to a broader range of decision makers in various industries.
- Profiles summarize the bias and insights behind user judgments. The user's past accuracy and bias inform ensemble weights rewarding stronger planners. This multiplayer ensemble improves accuracy by tapping into the collective intelligence of humans and machines.

Figure 1. Baseline Counterfactual



at the price of low interpretability, and the disparate data mining problem remains. We begin to lose sight of a primary objective: *ease of use* for planners. The recursivity of natural language is well equipped to capture and convey these nuanced interactions in a format that is accessible to a much wider audience. This observation is the fundamental principle behind the RASOR system. We capture and apply nuanced language using RASOR, by starting with a new "what if" scenario.

THE COUNTERFACTUAL: CREATING A "WHAT IF" SCENARIO

All forecasts produced by the RASOR system begin with a single *counterfactual scenario*: a fictional (but possible) approximation of business as usual. In our current retail example, we might want to understand the impact of an upcoming financing event.

A "business as usual" counterfactual scenario would answer questions like, "What would our sales have looked like if we did not run the financing event?" By analyzing a "what if" scenario against what happened, we can better understand the effect of our decisions or external factors (Neuberg, 2003).

To gauge the impact of past events within the retail context described, it is essential to explore potential alternatives – what might have occurred under different external circumstances. The human analyst provided a nuanced macroeconomic summary for 2022 Q1, but to use this information in a meaningful way, we first need to understand the divergence of outcomes. To this end, we apply a counterfactual forecast to construct a "what if" scenario. In our example, this scenario includes "What if all relevant levers affecting retail sales distributed around their means?" or, equivalently, "What if all our basic retail assumptions held?" This what-if counterfactual scenario forms the backbone of the system.

The counterfactual layer operates as a synthetic control capturing seasonality patterns. This multilevel seasonal approach is a necessary and sufficient condition of the structural framework at the model's foundation. **Figure 1** illustrates a single baseline counterfactual that captures multilevel seasonality at the system's core. This includes reproducing daily, weekly, and monthly seasonality trends within a tight range of possible parameters and removing impact from historical treatments and events (e.g., promotions and natural disasters). The growth trend adjusted layer represents the trend applied to the baseline counterfactual for each dynamic user scenario. This allows the system to infer causal relationships. With a control stripped of treatments, it is reasonable to assume the relationship between the treatment and control will consistently hold (Brodersen, 2015).

In practice, there are several methods for creating counterfactual scenarios. A practitioner might prefer Frequentist approaches with confidence intervals, Bayesian methods with prior beliefs, time series decomposition-like techniques, or a synthetically augmented, ready-to-use dataset. Any of these methods can

be effective, as the system's general case requires at minimum a single synthetic control to produce the baseline.

Continuing with the 2022 Q1 window of time, layering the revenue values (actuals) into the diagram will introduce periods of divergence from the what-if scenario. Considering the 2022 Q1 macroeconomic summary provided by the human analyst, we possess a reliable explanation of how the macroeconomic indicators likely contributed to the divergence of the actual sales from the counterfactual. A classical model with a robust historical dataset covering all the indicators and modeled interactions will explain a useful amount of the variance (Kolassa, 2024). But what if we don't have such a model or enough data? What if we don't understand interactions or have the resources to model them at all?

Our system addresses these problems in two main steps. This two-step process generates functions that, when retrieved, transform the growth-trend-adjusted counterfactual layer into the scenario-adjusted layer (Figure 1). First, we identify the indicators impacting the time window of interest and label this period with the event descriptions (e.g., Hurricane Ian). Then we take both the observed and counterfactual values as dependent and independent variables to train a model and assign the label. The labeled functions allow the planner to easily reproduce linear and nonlinear relationships. This is accomplished by stitching the linear parts back together using natural language, which will be discussed below in the retrieval section. This also allows the user to create entirely new scenarios by rearranging the order of the piecewise functions or mixing in functions from different time periods. Segmenting the linear components can be accomplished through linear spline interpolation, which ensures continuity and captures the piecewise linear trends of the time window (De Boor, 1978). Or it can be approximated by "eyeball analyses" of the historical behavior.

Figure 2 shows how the analyst's macroeconomic economic summary of 2022 Q1 conditions is used as the label for the 2022 Q1 model. Label A is our analyst's Q1 economic summary example, and labels B, C, and D would be similar summaries for different time periods in Q2.

Figure 2. Example of Counterfactual vs. Actual for 2022 Q1-Q2

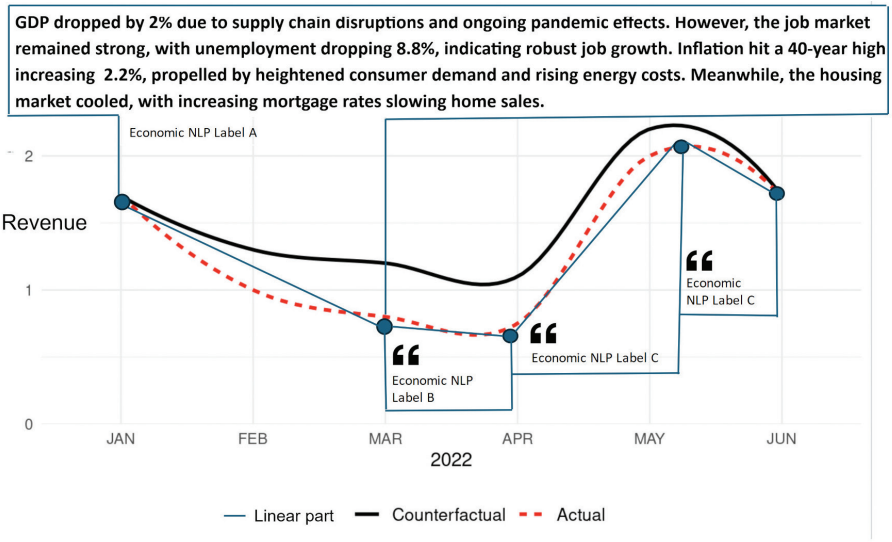


Figure 3. Flow of Impact Captures

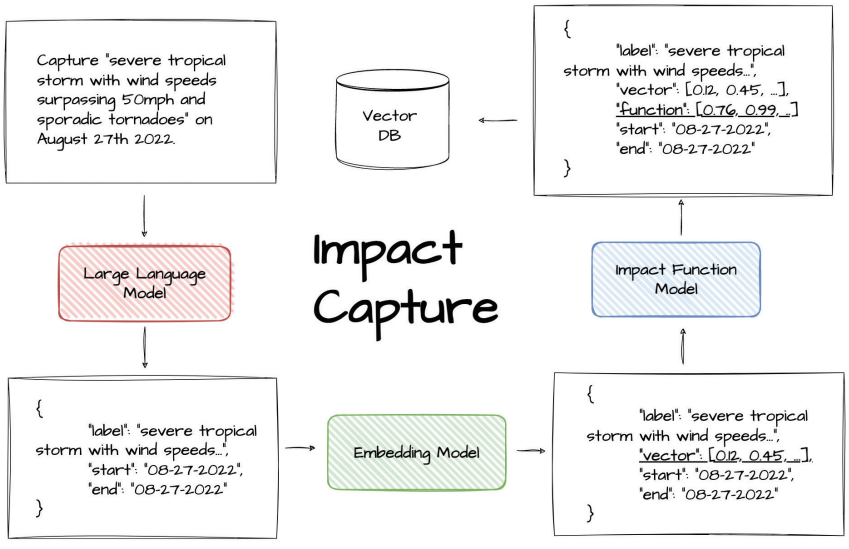
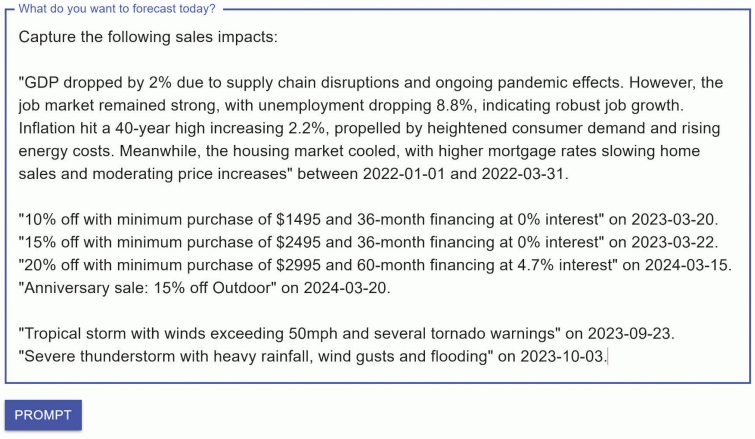


Figure 4. Example User Prompt Used to Capture Several Impacts and Trends



The function values are then stored in the system database along with the date stamps, period description in plain text, and vector representation (seen below in **Figure 5**).

CAPTURING NONPERIODIC EVENTS: TRAINING ON SMALL PERIODS OF TIME

Capturing event impact represented with a single data point (promotions, storms, etc.) follows the same two-step process as the longer time windows such as Q1 2022, with some differences (**Figure 3**). For instances involving only one data point, we proceed with some assumptions and generate synthetic data by treating the observed data point as the mean of a distribution (e.g., normal distribution) while defining variability based on a desired variance. However, this approach generalizes beyond specific distributions. What matters is the ability to generate a function that maps the counterfactual distribution to the observed distribution, regardless of their underlying types. These functions are labeled with semantic descriptions, and unlike Gaussian Mixture Models where the weights represent probability densities, RASOR assigns weights based on the semantic similarity to the user's intent. This shift enables RASOR to generalize over meaning rather than purely statistical densities. Conceptually, as more functions are labeled and weighted by their relevance to intent, the system converges in distribution and meaning, aligning its forecasts more closely with the user's goals.

The user's impact description and times are extracted by an LLM. Then the label is vectorized by an embedding model. A model is trained on the time period to capture the impact and stored in the database. This flow shows a single label, but multiple labels can be added to the same prompt as seen in **Figure 4**.

We made additional assumptions for our general retail case. The first assumption is that the dependent variable does not have a baseline value when the independent variable is zero. This assumption applies in contexts where no operations occur, such as a closed showroom or offline website, eliminating any potential impacts. The second assumption is that a degree of impact proportionality is expected,

Figure 5. System Database Entries

Label	Function	Start_date	End_date	Vector
GDP dropped by 2% due to supply chain disruptions and ongoing pandemic effects. However, the job market remained strong, with unemployment dropping 8.8%, indicating robust job growth. Inflation hit a 40-year high increasing 2.2%, propelled by heightened consumer demand and rising energy costs. Meanwhile, the housing market cooled, with higher mortgage rates slowing home sales and moderating price increases.	$[B_1, B_2, \ldots]$	2022-01-01	2022-03-31	$[v_1, v_2, \ldots]$
10% off with minimum purchase of $1495 and 36-month financing at 0% interest	$[B_1, B_2, \ldots]$	2023-03-20	2023-03-20	$[v_1, v_2, \ldots]$
15% off with minimum purchase of $2495 and 36-month financing at 0% interest	$[B_1, B_2, \ldots]$	2023-03-22	2023-03-22	$[v_1, v_2, \ldots]$
20% off with minimum purchase of $2995 and 60-month financing at 4.7% interest	$[B_1, B_2, \ldots]$	2024-03-15	2024-03-15	$[v_1, v_2, \ldots]$
Anniversary sale: 15% off Outdoor	$[B_1, B_2, \ldots]$	2024-03-20	2024-03-20	$[v_1, v_2, \ldots]$
Tropical storm with winds exceeding 50mph and several tornado warnings	$[B_1, B_2, \ldots]$	2023-09-23	2023-09-23	$[v_1, v_2, \ldots]$
Severe thunderstorm with heavy rainfall, wind gusts and flooding	$[B_1, B_2, \ldots]$	2023-10-03	2023-10-03	$[v_1, v_2, \ldots]$

aligning with the majority of business dynamics. This framework for capturing nonperiodic events provides adaptability and responsiveness.

For example, if a new promotion (e.g. 10% off with minimum purchase of $1,495 and 36-month financing at 0% interest) is launched, the event is captured at the close of business. We can then apply this promotion immediately to any subsequent forecast. The model generalizes over time as more occurrences of the event and event variations are added to the system, learning on the fly. With the system's ability to interpret causal relationships, when variations of the promo (e.g. 15% off with minimum purchase of $2,495 and 36-month financing at 0% interest) are captured, planners can further analyze causation in addition to improving forecast accuracy and dimensionality reduction. The functions and their labels are then stored in the system database. Figure 5 illustrates the system database entries showing the period label, the event functions with start and end dates, and vector representations.

RASOR: RETRIEVAL-AUGMENTED TIME SERIES SYSTEM

How these functions are added to the forecast is similar to the LLM RAG pattern. The objective of most RAG implementations is to improve inference accuracy and reduce hallucinations by adding extra context relevant to the original prompt that the LLM has not been trained on (Fan et al., 2024). This happens by matching the user's input against vector representations of the context using a distance metric (e.g., cosine or Jaccard similarity). The RASOR system differs from the RAG pattern in that rather than add context to the LLM, RASOR adds context relevant to the original time series that the time series has not been trained on. Thereby the RASOR system effectively "grounds" the forecast. **Figure 6** shows the flow from multiplayers to application, LLM, vector database, time series model, ensemble weights informed by the leaderboard, and final scenario.

Figure 6. High-Level Diagram of the Retrieval-Augmented Forecasting System

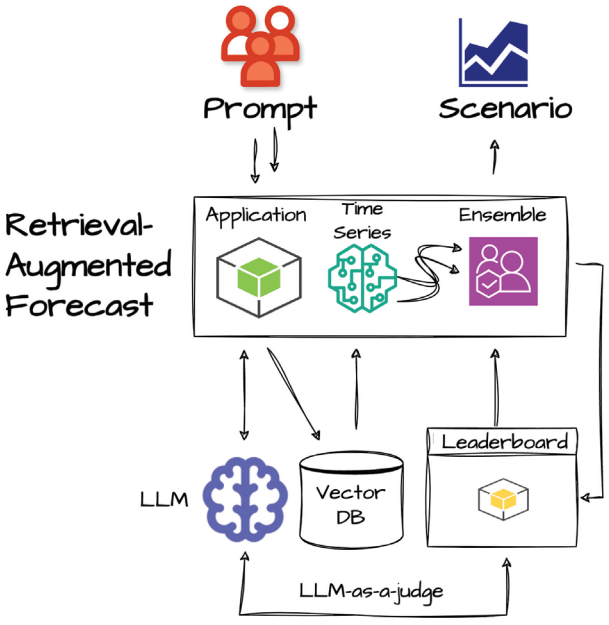
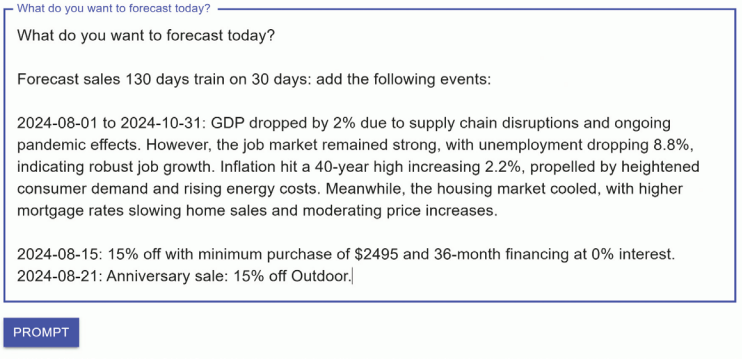


Figure 7. Example of User Prompt Creating a New Scenario

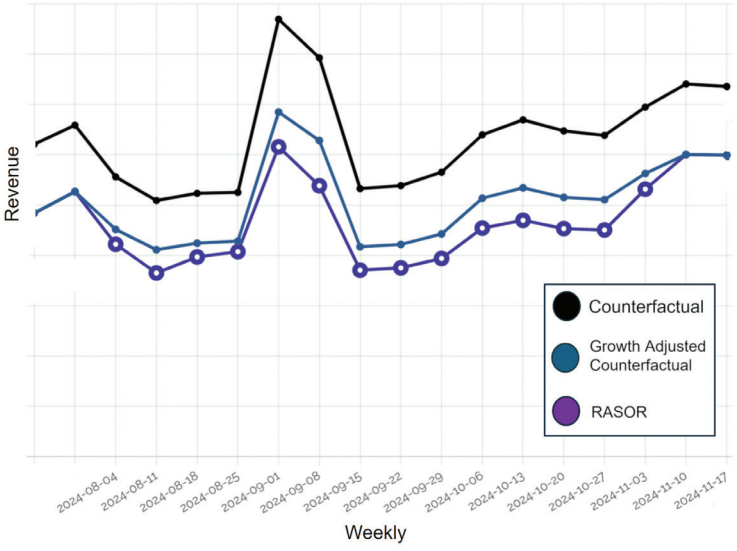


This retrieval mechanism reduces cost and time incurred from model retraining, improving adaptability and accuracy. Understanding the semantic meaning of language is critical (Yadkori et al., 2024) for retrieval systems like RAG and RASOR. A successful "grounded" time series requires aligning user intent with appropriate function impacts, beyond lexical similarity. The retrieval system must discern meanings accurately, even when words or phrases sound similar or differ significantly in context. For example, RASOR uses embedding language models (Li and Yang, 2018) to deeply understand and process the user's intent. This capability allows for precise adjustments based

Figure 8. Impact Event Calculations Using Functional Programming

$$\left(\cdots f_{\text{nhlfinals}}\left(g_{\text{vipevent}}\left(h_{\text{thunderstorm}}\left(\sum_{i=1}^{N} w_i \cdot f_i(x)\right)\right)\right)\cdots\right)$$

Figure 9. Example of Retrieval-Augmented Forecast



on the semantic content of user inputs, improving accuracy and relevance.

PRACTICAL APPLICATION OF RASOR

Having outlined the theoretical underpinnings of RASOR, we now demonstrate its practical application. To begin forecasting, a planner at any level of technical skill simply creates a prompt with any captured events, assumptions, or economic projections they might consider. **Figure 7** provides a user prompt example. This prompt creates a new scenario by specifying a training horizon, forecast horizon, economic adjusted period, and two single-day promotional events within the adjusted period.

Once submitted, the prompt is sent to an LLM for text to JSON transformation and is returned to the application. (JSON is a text-based format for storing and exchanging data that's both human-readable and machine-parsable.) The application parses the user-defined forecasting horizon and training horizon (Figure 7) from the JSON. The observed data in the training horizon is used to compute a linear trend that is applied to the baseline counterfactual, transforming the baseline counterfactual to a growth-trend-adjusted counterfactual (Figure 1). The growth-trend-adjusted counterfactual now includes multilevel seasonality, trend, and noise components.

The system will then "ground" this forecast in the following three steps. The first step is creating an impact score for each retrieved function. The impact score starts with the semantic meaning score output from the embedding model (a score of 1 would be an exact match). A survival function is then applied, giving more weight to recent events and less to older ones. Two optional parameters can also be applied. Option one is a lexical penalty that can be used to tighten the variability of the model to ensure closer context alignment. Option two is an off-period penalty applied when the forecast period differs from the origin (e.g., when applying an impact captured on a Tuesday to a Friday, etc.).

The second step converts these final impact scores into function weights when multiple impact functions are retrieved. For example, if multiple variations of stormy weather are returned, weights will be applied based on the strength of their respective impact scores.

The third step is applied when multiple impact types are assigned to the same period; for example, a promotion, a weather event, and a sporting event assigned to the same day. A functional programming method is used to produce the final effect. Rooted in lambda calculus, functional programming shows how basic operations can be combined into more complex expressions to produce complex interactions, as shown in **Figure 8**.

The final scenario is produced when all calculations are complete. **Figure 9** provides sample output showing the growth-adjusted time series against the counterfactual, along with the retrieval-augmented scenario.

Figure 10. Scenario Planning in Multiplayer Mode

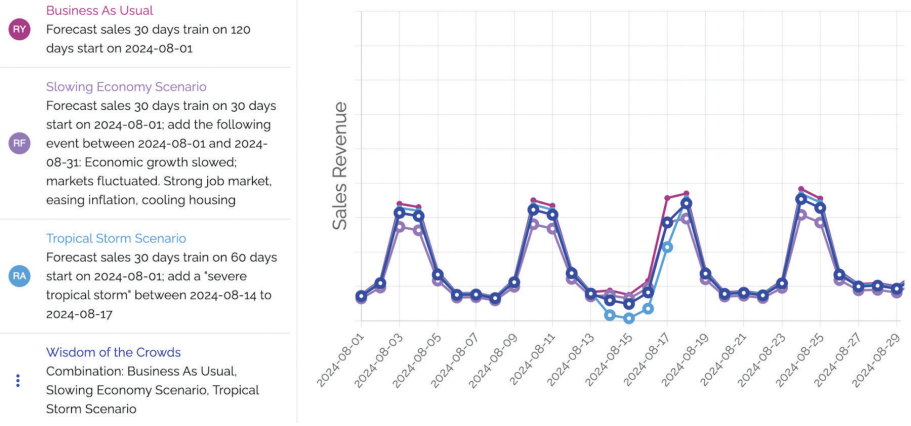


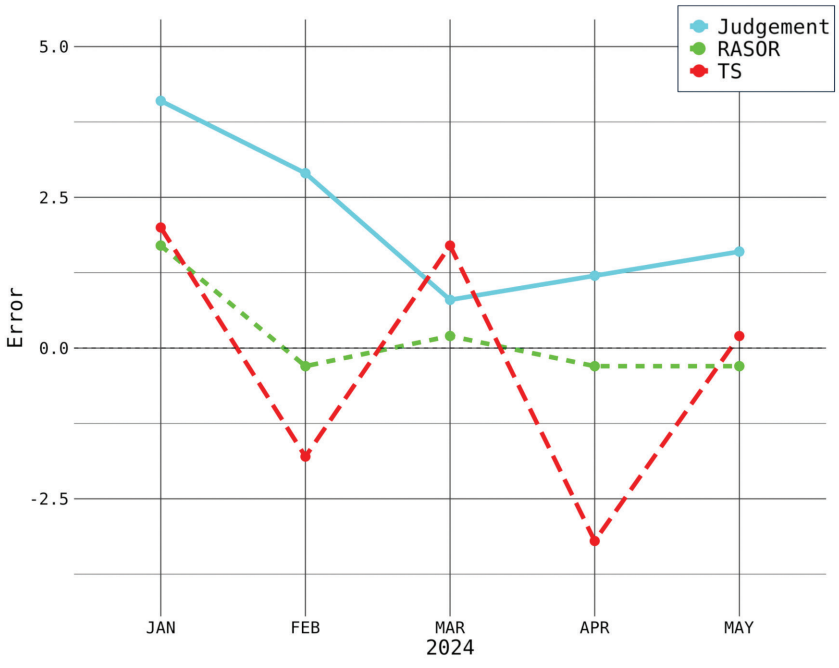
Figure 11. RASOR Planner Leaderboard

User	Summary	Bias	Accuracy
RY	This forecaster showed a moderate level of accuracy but had a significant positive bias in their predictions. They made a modest number of forecasts compared to others in the group. Their forecasts frequently included specific promotional events like discounts and financing offers, suggesting an emphasis on sales promotions that may have contributed to the overestimation in their bias.	60.1	80.57%
RA	This forecaster achieved a high level of accuracy but exhibited a noticeable negative bias. They made the most forecasts among the group, indicating a strong commitment to forecasting efforts. Their forecasts often accounted for negative factors such as hurricanes and reduced consumer spending, reflecting a cautious approach that may explain the negative bias.	-25.3	90.18%
RF	This forecaster attained the highest accuracy with minimal positive bias. However, they provided only a single forecast, which makes it difficult to fully assess their forecasting capabilities relative to others. Their forecast incorporated extensive macroeconomic variables and promotional events, demonstrating depth of analysis but lacking in frequency.	5.7	95.34%
RY+RF+RA	This ensemble forecaster achieved very high accuracy but displayed a positive bias, indicating a consistent overestimation of sales figures. By combining the strengths of the individual forecasts, it effectively incorporated factors like promotional events, economic fluctuations, and unforeseen disruptions. However, while the ensemble delivers superior performance, the limited number of forecasts from some individual models suggests potential for further refinement in scenario coverage and scalability.	3.2	96.21%

MULTIPLAYER FORECASTING: GAMIFYING THE CROWD

Evidence shows that crowds outperform individuals in accuracy, as seen in expert aggregation (Petropoulos et al., 2022, p. 738) and prediction markets (Wolfram, 2024). The "wisdom of the crowds" concept is something machine learning engineers have applied for years, where the "crowd" is a family of machine learning models outputting complementary errors. These errors are then stacked or ensembled together to improve model accuracy and generalization (Ganaie et al., 2022). As the RASOR system lowers the barrier to enter, the inclusion of multiple planners into the ensemble will diversify and complement insight, as illustrated in **Figure 10**. Various leaders and SMEs can apply their assumptions and desires transparently, with their intent and bias tracked through the LLM-as-a-judge evaluation system (Zheng et al., 2024). This includes human-in-loop oversight of the LLM judgment. Here, users RF, RY, and

Figure 12. Management System Results from the First Five Months of 2024



RA merge scenarios completing a "wisdom of the crowds" ensemble.

Figure 11 provides an example of user evaluation. The Planner Leaderboard contains user profiles recording insight, bias, and accuracy.

Our POC was launched on January 1, 2024, as part of the forecasting management system. The results reflecting the model operating "in the wild" demonstrated a balanced output that should further balance out over time as the crowd grows. **Figure 12** shows the baseline time series forecast errors compared to those of pure judgmental forecasts and RASOR. The time series is an ARIMAX-like approach using ridge regression for exogenous variables and a transfer function for temporal dependencies. Judgments were applied weekly along with baseline time series forecasts and RASOR updates. Results were aggregated monthly with mean percentage error (MPE) used to highlight bias. Here we see that the judgmental forecasts are considerably more biased than RASOR, with the errors of the latter being also less volatile over time compared to the statistical benchmark.

CONCLUSION AND NEXT STEPS

There are several potential enhancements that can be made to this POC. For the model, this involves testing different statistical approaches and challenging assumptions. The current architecture makes use of Meta's large Llama 3 70b model (Meta AI, 2024) to extract user forecast specifications. This is almost certainly overkill and could likely be accomplished using a smaller, fine-tuned model variant.

There are also many opportunities for automation. As LLMs continue to get smarter, it is quite likely that an LLM researcher will be able to identify and label nonperiodic events with minimal human oversight. This concept can be extended further to include the entire forecasting problem, where an LLM agent cross-references recent news, weather, trends, etc. with the existing events database to build its own forecasting scenarios. This also positions the system for full conversational speech-to-text scenario planning between human and machine.

Finally, while raw semantic similarity is a good approximation in terms of matching relevant events, it sometimes fails to distinguish between good and bad variations of a similar event impact, e.g., interest rate hikes and interest rate cuts. These failures get worse as event description length increases. It is suspected that LLMs might be better equipped to differentiate between relevance and impact, although they might need some fine-tuning. Fine-tuned embedding models might also suffice.

We have plans to run different stages of this model on benchmarks like M5 (Makridakis et al., 2022) to ease comparison with other state-of-the-art (SOTA) models in the community. We are also exploring further research in the context of information theory. Leveraging the nuances of language to capture complex

interactions and dense information in a computationally efficient manner may contribute insights into computational linguistics, AI, and semantic compression.

REFERENCES

Meta AI. (2024). Llama 3. ***ai.meta.com/blog/meta-llama-3/***

Bergmeir, C. (2024). LLMs and foundational models: Not (yet) as good as hoped. *Foresight*, 73, 33-38.

Brodersen, K.H. et al. (2015). Inferring causal impact using Bayesian structural time-series models. *Annals of Applied Statistics*, 9(1), 247-274.

De Boor, C. (1978). A practical guide to splines. (Vol. 27, p. 325). Springer.

Fan, W. et al. (2024). A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 6491-6501.

Ganaie, M.A. et al. (2022). Ensemble deep learning: A review. *Engineering Applications of Artificial Intelligence*, 115, 105151.

Jiang, W. et al. (2023, May). An empirical study of pre-trained model reuse in the hugging face deep learning model registry. In *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*, 2463-2475. IEEE.

Kahneman, D. & Tversky, A. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157), 1124–1131.

Kolassa, S. (2024). Interactions in forecasting. *Foresight*, 73, 63-64.

Li, Y. & Yang, T. (2018). Word embedding for understanding natural language: A survey. In Srinivasan, S. (ed) *Guide to Big Data Applications*. Studies in Big Data, vol 26. Springer.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2022). M5 accuracy competition: Results, findings, and conclusions. *International Journal of Forecasting*, 38(4), 1346-1364.

Neuberg, L.G. (2003). Causality: Models, reasoning, and inference, by Judea Pearl, Cambridge University Press, 2000. *Econometric Theory*, 19(4), 675-685.

Petropoulos, F. et al. (2022). Forecasting: Theory and practice. *International Journal of Forecasting*, 38(3), 705-871.

Vandeput, N. (2021). Forecast value added. Medium. ***nicolas-vandeput.medium.com/forecast-value-added-ebc163d7ccd***

Wolfram, T. (2024). The accuracy of prediction markets. *Foresight*, 73, 48-53.

Yadkori, Y.A., Kuzborskij, I., György, A., & Szepesvári, C. (2024). To believe or not to believe your LLM. ***arxiv.org/pdf/2406.02543***

Zheng, L. et al. (2024). Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36, 46595-46623.

Ryan Fattini is the VP of Data Analytics & AI at CITY Furniture, specializing in team building and hybridizing the core disciplines of business intelligence, engineering, data science, and AI. He led the initiative at CITY Furniture to develop their real-time streaming architecture and data fabric, as well as to close the gap between the rigors of academic data science and practical enterprise machine learning models. Ryan serves on the Miami Dade College advisory board for Data & Analytics.

ryanfattini@gmail.com

Ryan Young is a staff Data Science Engineer at CITY Furniture, with extensive experience in developing full-stack applications that drive impactful results. His expertise spans machine learning, data analysis, and software development, enabling him to deliver innovative solutions tailored to diverse business needs.

rayoung1093@gmail.com

The Mythical Influence of Metric Asymmetry

PATRICK BOWER

Did you ever hear something that you know isn't true, but is so embedded in popular culture that it's hard to convince folks otherwise? The television character Ted Lasso told us that goldfish have a three-second memory span, while popular U.S. history has it that George Washington had wooden teeth. Neither statement is true, but they are widely *believed* to be true and thus often accepted as fact without question.

In our profession, a myth I often see and hear suggests that the selection of a forecast accuracy measure promotes or rewards over- (or under-) forecasting. This statement is almost always paired with words that suggest stakeholders are trying to game the system to improve forecast accuracy. This is a profoundly silly statement, if only because – ironically – overforecasting has the potential to negatively impact other metrics like inventory carry and obsolescence, while underforecasting can negatively impact customer service fill levels. It is objectively a falsehood that stakeholders try to game the system based on asymmetry in the forecast accuracy measure. Yet it is believed to be true, and no one challenges the belief.

Worse, this gaming-the-system narrative is often used in debates about which forecast accuracy metric is the best. In doing so, the falsehood is cascaded into some of the most dogmatic rhetoric in our field.

Accuracy measures do not promote under- or overforecasting. In fact, if you asked the average commercial (sales and marketing) team member, I suspect few would have a clue about the calculation of forecast accuracy measures. And even fewer would understand the distinction between different measures. From my perspective, conflating the nuance of measures with nefarious gamesmanship is a problem.

About 20 years ago, I led a consulting team performing a demand-curve analysis across multiple businesses. In every business observed – regardless of the accuracy measure employed – demand was overforecast. This was not a surprise.

When we did a closer inquiry of the reasons behind these analytical results, it became clear that demand was overforecast because of the aspirational bias of the commercial teams – typically sales and marketing. These people simply believed in their market research, the marketing and sales plans, the media and creative plans, and their ability to execute. To be fair to the commercial teams, organizations want these teams to reach forward and be aspirational. But they also want them to demonstrate prudence, realism, and measured approaches when forecasting. The second half of that message is often missed.

Interestingly, as demand planners we beg for inputs from commercial teams – and in nearly every demand-curve analysis we found that these inputs were treated additively, instead of incrementally. Planners layered these inputs on top of each other without determining the interaction between the base demand and other inputs. Planning organizations' management of external factors and inputs were often part of the problem.

In the analysis, we did find occasional examples of gaming the forecast. Some forecasts were held to a higher-than-reasonable plan despite negative results because the commercial teams wanted to delay telling leadership bad news. Or they might have felt they didn't have enough information to lower the forecast based on a few months of subpar results. Or maybe they just wanted to hold onto their ability to spend their budget on interesting projects, as lowering the forecast normally means this sort of brand-enhancing

spending is cut. In each case, the forecast was overstated, without any direct or tangential connection to the forecast accuracy measure being used.

Similarly, I have observed a few cases of underforecasting by those who were overly conservative or angling to over-deliver – commercial team members with hero complexes trying to pull volume bunnies out of a hat.

I most often saw these instances of underforecasting habits at a lower level of aggregation – a key account or a product family – yet here again, none of these behaviors had anything to do with the forecast accuracy measure. They were motivated by other reasons.

I hope that once and for all we can abandon this line of thinking, this forecast accuracy measurement fallacy. By assuming that adjustments to the forecast are compelled by metrics instead of other influences, planners misidentify the problem and delay or forestall potential improvement.

Patrick Bower is Senior Director of North American Supply Chain at Actylis, a manufacturer and distributor of specialty chemicals for life sciences. He previously held leadership roles at Combe, Cadbury, Kraft Foods, Unisys, and Snapple, among others. Patrick is a frequent writer and speaker on supply chain subjects, has been recognized four times by *Supply and Demand Chain Executive* magazine as a "Pro to Know," and was named by *Consumer Goods Technology* magazine as one of their 2014 Visionaries. In 2012 he received the inaugural Excellence in Business Forecasting and Planning Award from the Institute of Business Forecasting, and in 2023 received their Lifetime Achievement Award. Patrick also writes the S&OP feature for ASCM's "SCM Now" blog.

plbowerone@yahoo.com

Systems Thinking to Address Sustainability

LEO SADOVY

The greatest threat to our planet is the belief that someone else will save it. — Robert Swan

So you want to tackle sustainability problems like poverty, energy, or climate change? Good, because the planet needs you. But before you jump in you need to understand how the scientists, researchers, practitioners and policy gurus in the field work, which comes down to a simple two-word phrase: *systems thinking*.

Sustainability is a complex field, and the professionals involved deal with it in exactly that manner, building models of complex adaptive systems (CAS) to simulate economic, social, physical, and biological environments. What makes these models "complex" is that they are nonlinear, involve feedback loops, and exhibit emergent behavior. What makes them "adaptive" is that they are self-organizing, evolving, and consist of individual networked agents that learn and adapt via interactions and feedback from each other and from the environment. Their boundaries are typically open, exchanging energy, information, or matter with their environment, and can be fluid or permeable, influencing the system's development.

These CAS models in turn are built on a systems framework, and it is critical that you understand the concepts and terminology of systems thinking if you are going to make a meaningful contribution in your chosen domain. Sustainability professionals think and build their models in terms of sources and sinks, stocks and flows, buffers and levels, positive and negative feedback loops, delays, and stock versus flow-limited resources. Their nonlinear nature can often result in chaotic behavior such as period doubling and fold-bifurcation tipping points.

As an experienced and capable forecaster, where do you fit in? It should be clear from the above that it will likely not be as a sustainability systems model builder – going in, none of us has either the model building nor domain expertise required of such a role. Your first step would be to identify which domain and which sustainability goal stokes your interest and passion, and dig into both the popular and peer-reviewed literature. Perhaps you are already embedded in the CPG industry – the next logical step would be to bone up on sustainable production, consumption, supply chains, and the circular economy. Or maybe you're involved in your local community, in which case a deeper understanding of sustainable transportation or water supply would best support your cause.

Still, you're going to need to get a grasp on the principles of systems thinking, and here the gold standard would be Donella Meadows' book, *Thinking in Systems: A Primer*. Then watch a few of the many YouTube videos of her lectures and talks.

Next up is the matter of linking up with the researchers and teams doing this work. If you want to aim high, start with the UN Sustainable Development Solutions Network (SDSN), the International Institute for Environment and Development (IIED), the Intergovernmental Panel on Climate Change (IPCC), or the World Resources Institute (WRI). Then there will be the dozens if not hundreds of organizations associated with your chosen domain, readily located by searching LinkedIn organizations by subject matter. Lastly, universities around the world are doing sustainability-related research by the metric ton – academic opportunities abound.

Once attached and embedded into a team or project, your primary value is going to lie in providing forecasts, but not for the overall model objective (i.e., global temperature increase, etc.). Instead, for the various model inputs and components – the sources, sinks, flows and delays. Examples are seafood consumption, residential solar panel installations, plastic packaging waste, and employment opportunities by educational discipline. Not the model itself, but everything that drives the model.

There are three obvious areas, however, where you are likely to have a more immediate and significant impact, derived from your analytical wheelhouse. The first is simply your data preparation, transformation, and analysis expertise – dealing with outliers and missing values, ranges and cardinality, seasonality effects and non-stationarity – not to mention basic data sourcing and acquisition.

Second will be your choice of the appropriate prediction technique. While you may be most familiar with time-series forecasting, most sustainability studies are focused on outcomes in a more distant future where other analytic approaches might be better suited to the variability around such a long-term trendline.

Third, there is the expertise you naturally bring to quantifying risk and uncertainty. Professionals outside of analytics do not have an intuitive awareness of how widely variable the data actually is. They have no appreciation of how large those prediction intervals really are (especially for time series data), and what that means for the range of possible outcomes and prediction interpretation. An understanding of confidence intervals aids in scenario planning by considering a range of possible outcomes rather than just a single estimate, allowing for flexible strategies that can adapt to varying future conditions. The discussion and the magnitude of the hard decisions to be made will be very different depending on whether some critical threshold lies outside two standard deviations, or inside of one standard deviation.

After that, it should be a simple matter of asking, "How can I help?" Researchers and practitioners will be more than eager to share and explain their project with you – manage the conversation towards an objective of understanding what their analytic and forecasting needs are from a systems modeling perspective. You should then be able to settle on a mutually agreed approach that has the best chance of your work having an impact on the overall outcome.

When earlier I stated "the planet needs you," that is not entirely accurate. Planet Earth will continue on, with or without us, regardless of what we do to the environment. It's humanity that needs you. It is we humans, and the economic, social, and natural ecosystems that support us, who require a sustainable approach to our activities. Getting there is not a foregone conclusion, but it will not happen without a dedicated effort.

Leo Sadovy holds an MBA in finance and a master's in analytics, and makes his living as a director of analytics in the media and marketing industry. With a commitment to leave this planet a better and sustainable place for his children and all the world's children, Leo is currently pursuing another master's in sustainability. His focus is on climate adaption, climate refugees, and water management / coastal and marine concerns.

leosadovy@gmail.com

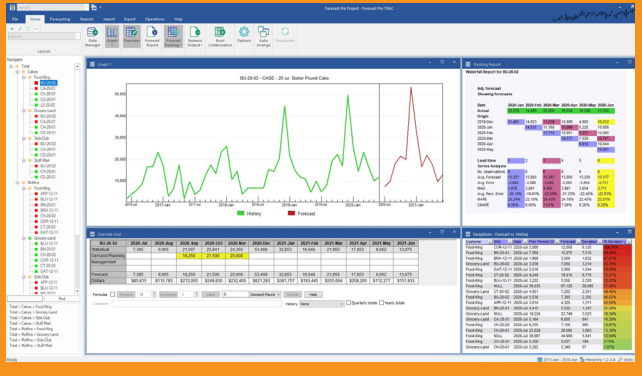
Forecast Pro TRAC

The Power of AI, Machine Learning and Statistical Forecasting - all at your fingertips

Forecast Pro TRAC is the top-rated demand forecasting solution that is powerful yet easy to use.

This comprehensive forecasting system includes:

- ✔ AI-driven automatic forecasting
- ✔ Proven forecasting methods
- ✔ Flexible forecast adjustments
- ✔ Multiple conversions and hierarchies
- ✔ Accuracy tracking and exception reporting
- ✔ Forecast Value Add Reporting
- ✔ Team forecasting and Excel collaboration



"Forecast Pro TRAC has freed me from number crunching and given me more time to analyze my business."

Michael Pan
Wakefield Canada, Inc.

Trusted by more than 35,000 users worldwide, Forecast Pro improves your forecasts, supports your S&OP process and easily integrates with your existing software systems.

forecast pro®

www.forecastpro.com

Do You & Your Crew Want More Info?

- **Download a free trial**
- **Request a live demo using your own data**

The 45th International Symposium on Forecasting

Beijing, China, June 29-July 2, 2025

ISF 2025

Keynote Speakers

Valentina Corradi, Professor of Econometrics, University of Surrey

Azul Garza, CTO and Co-Founder, Nixtla

Yan Liu, Professor and the Director of the Machine Learning Center, University of Southern California

Anastasios Panagiotelis, Associate Professor and Deputy Head of Business Analytics, University of Sydney Business School

Pierre Pinson, Professor, Imperial College London and Chief Scientist, Halfspace

Shouyang Wang, Director of AMSS Center for Forecasting Science, Chinese Academy of Sciences

Practitioner Speakers

Sven Crone, Assistant Professor, Lancaster University; CEO & Founder, iQast

Chris Fry, Director of Data Science, Google Cloud

Mohsen Hamoudia, CEO and Founder, PREDICONSULT

Boyi Hou, Huawei Cloud BU Algorithm Innovation Lab

Max Mergenthaler, CEO and Co-Founder, Nixtla

Yijie Peng, Guanghua School of Management, Peking University

Shiyu Wang, Senior Staff Researcher and Director of the Supply Chain Algorithm Team, ByteDance Inc.

Qingsong Wen, Head of AI & Chief Scientist, Squirrel Ai Learning

Zhou Ye, Principal Researcher and Managing Director of the Supply Chain Algorithm Team, ByteDance Inc.

Contacts

Yongmiao Hong, Chinese Academy of Sciences, *General Chair*

Jue Wang, Chinese Academy of Sciences, *Program Chair*

Yanfei Kang, Beihang University, *Program Chair*

Ying Fry, *IIF Business Manager*
isf@forecasters.org

Venue Beijing Friendship Hotel

About the ISF

The ISF is the premier forecasting conference, attracting the world's leading researchers and practitioners. Through a combination of speaker presentations, academic sessions, workshops, and social programs, the ISF provides many excellent opportunities for networking, learning, and fun.

For more information: ***https://isf.forecasters.org/***

FORESIGHT
The International Journal of Applied Forecasting
Business Office: 8956 Erect Road
Seagrove, NC 27341

ISSUE 77

LAST ISSUE ALERT?

If it says **Last Issue Alert** above your name, take a couple of minutes to renew your membership now and keep *Foresight* coming your way.

Renew or start your IIF membership:
forecasters.org/membership/join/

Email our Business Manager at
forecasters@forecasters.org