

The Future of Demand Forecasting with Generative AI

Integrating Forecasting and Inventory Decisions Using Machine Learning

Types of Forecast Errors and Their Implications

Explainability: A Requirement for Trust in Forecasts

Book Review: *The Art of Uncertainty*

Op-Ed: Overcategorization of Continuous Data

Special Feature:

Revisiting Symmetric MAPE

Special Feature:

UN Sustainable Development Goals





Navigating Uncertainty Calls for Certain Forecasts



Forecast Pro is the leading off-the-shelf software for forecasting professionals

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

Forecast Pro combines:

- ✓ AI-driven automatic forecasting
- ✓ Proven forecasting methods
- ✓ Comprehensive collaboration tools
- ✓ Accuracy tracking and exception reporting
- ✓ Flexible forecast adjustments

—all in one easy-to-use tool!

Proven Results with Forecast Pro:

Aspire Pharma, one of the fastest-growing pharmaceutical companies in the UK, improved forecast accuracy by 40% with Forecast Pro, ensuring more cost-effective and consistent supply continuity.

Learn More & Download a Free Trial:



www.forecastpro.com

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

3 Note from the Editor

ai and machine learning

5 The Future of Demand Forecasting with Generative AI

*Yue Li &
Rachel Pedersen*

14 Integrating Forecasting and Inventory Decisions Using Machine Learning

*Joost F. van der Haar,
Yves R. Sagaert &
Robert N. Boute*

forecasting measures

20 Types of Forecast Errors and Their Implications

Kolja Johannsen

special feature: revisiting symmetric mape

26 Errors on Percentage Errors

Rob J. Hyndman

29 Sparse-Proof sMAPE

Slawek Smyl

31 Know Your Errors!

Stephan Kolassa

forecasting practice

33 Explainability: A Requirement for Trust in Forecasts

Trevor Sidery

39 Commentary: Explanations vs. Explainability

Anne-Flore Elard

41 Commentary: Building Trust through Explainability

Zabiulla Mohammed

special feature:

united nations sustainable development goals

43 The Role of Forecasting in Ending Global Hunger

Lauren Davis

46 Life Below Water

Leo Sadovy

book review

48 *The Art of Uncertainty – How to Navigate Chance, Ignorance, Risk and Luck* by David Spiegelhalter

Ira Sohn

opinion-editorial

51 Overcategorization of Continuous Data

Malte Tichy

*Foresight
Advisory
Board*

Chairman: Jim Hoover, *University of Florida*
Carolyn Allmon, *Forecasting Consultant*
Mark Chockalingam, *Valtitude/Demand Planning LLC*
Cara Curtland, *HP*
Lauren Davis, *NC A&T State University*
Robert Fildes, *Lancaster Centre for Forecasting*
Ram Ganeshan, *College of William and Mary*
Igor Gusakov, *GoodsForecast*
Sevvandi Kandanaarachchi, *CSIRO*
Jonathon Karelse, *NorthFind Management*
Yue Li, *Bain & Company*
Joe McConnell, *McConnell Chase Software*
Polly Mitchell-Guthrie, *Two Halves Consulting*
Dilek Önköl, *Northumbria University*
Steven Pauly, *Slimstock*
Jack Pope, *Investment Economics*
Johann Robette, *Vekia*
Eduardo Romanus, *Ipiranga*
Jerry Shan, *Insightful-Tech Ventures*
Sujit Singh, *Arkiva*
Marina Sologubova, *Estée Lauder*
Eric Stellwagen, *Business Forecast Systems*
Nicolas Vandeput, *SupChains*
Lawrence Vanston, *Technology Futures*
Janina Zittel, *Zuse Institute Berlin*

Editor-in-Chief

Michael Gilliland

Deputy Editor

Stephan Kolassa

Associate Editors

Jeff Baker
Fotios Petropoulos
Evangelos Spiliotis
Aris Syntetos

Column Editors

Simon Clarke
Opinion-Editorial
Shari De Baets
Judgmental Forecasting
Elaine Deschamps
Government & Public Policy
Anne-Flore Elard
Machine Learning & AI
Tao Hong
Energy & Environment
Malvina Marchese
Financial Forecasting
Zabiulla Mohammed
Retail & CPG
Christian Schäfer
Life Sciences
Ira Sohn
Long-Range Forecasting
Simon Spavound
Book Reviews

Foresight Staff

Ying Fry
Marketing & Sponsorship
Liza Woodruff
Design & Production
Ralph Culver
Manuscript Editor
Mary Ellen Bridge
Copy Editor

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.

note from the editor

LAUREN DAVIS ELECTED TO THE IIF BOARD

Congratulations to *Foresight* Advisory Board member Lauren Davis, who was recently elected to the IIF's Board of Directors. Lauren is a Professor in the Department of Industrial and Systems Engineering at North Carolina A&T State University, and her research focuses on stochastic modeling of supply chain systems, particularly with application to hunger relief. See Lauren's article on page 43 of this issue on the role of forecasting to end global hunger.



PREVIEW OF FORESIGHT ISSUE 78

Issue 78 begins with **Yue Li** and **Rachel Pedersen** taking a deep look into the part that generative AI will play in the future of demand forecasting. They note that gen AI is still evolving, but has the potential to enrich existing forecasting frameworks even if not a replacement for forecasting expertise. They also warn that strong AI governance is needed to ensure data quality and reliability.

Continuing in the AI/ML space, **Joost van der Haar**, **Yves Sagaert**, and **Robert Boute** investigate the integration of forecasting and inventory decisions using machine learning. In their study of three Belgian companies in the food industry, they find that better forecasts do not necessarily lead to better inventory decisions. Instead, predicting optimal order quantities directly can result in substantial cost savings for smoother time series.

Forecast errors are inevitable, but not all errors are created equal. So goes the argument by **Kolja Johannsen**, who categorizes four types of forecast errors and provides strategies for responding to them. As he shows, being aware of the drivers behind forecast errors can help improve accuracy as well as make the forecast more useful for decision making.

Forty years ago, the “asymmetry” of mean absolute percentage error was noted by Scott Armstrong. Forecasts that exceeded the actual were penalized more harshly by MAPE than forecasts below the actual, introducing a possible incentive for biasing forecasts to the low side. Various flavors of *sMAPE* – purporting to provide symmetry – were introduced, and **Rob Hyndman** opens our special feature on Revisiting Symmetric MAPE with a recap of that history.

Slawek Smyl continues the discussion with a proposed new metric he calls *Sparse-Proof MAPE* (msMAPE), designed to better handle large-valued as well as sparse (intermittent) time series when forecasts and actuals are non-negative. **Stephan Kolassa** ends the special feature with a commentary on Smyl's msMAPE and a call for using simulation to better understand what any error metric does in a variety of situations.

The “explainability” of a model has become an important element in forecasting. This is especially true with the increased reliance on machine learning models that lack transparency into what variables are driving the forecast. **Trevor Siderly** argues that explainability is a requirement for trust in forecasts and categorizes four types of explainability requirements involving methods, components, drivers, and errors. Since each business user may have a different understanding of explainability, these varied understandings can affect what models the forecaster uses.

In a pair of commentaries, **Anne-Flore Elard** looks at the distinction between explainability and explanations and notes that when models lack a direct mapping with business drivers, this creates a roadblock to their trust and adoption. Then **Zabiulla Mohammed** agrees that explainability is important for building trust, but not at the expense of predictive power or business value.

In *Foresight* issue 74, Bahman Rostami-Tabar and I raised the question of forecasting's role in supporting the United Nations' Sustainable Development Goals. In response, **Lauren Davis** looks at the role of forecasting in ending global hunger, and **Leo Sadovy** addresses its part in life below water.

Frequent book review contributor **Ira Sohn** delivers another, this time examining David Spiegelhalter's *The Art of Uncertainty*. Sohn finds Spiegelhalter to have a singular command of the technicalities of statistics and probability, along with a special talent for communication that exudes confidence and trust. He considers the book an engaging and entertaining read.

Issue 78 concludes with an Op-Ed by **Malte Tichy** on the overcategorization of continuous data. In the most egregious cases, "category hacking" occurs when different category splits are tested until one happens to be statistically significant. Tichy argues that binary decisions don't necessarily require binary categorization of the data and that prematurely classifying continuous quantities is often a lazy shortcut that can impact the quality of the analysis.

IIF COMPETITION PAPERS COMING IN ISSUE 79

The Q4 issue, publishing in October, will feature papers from IIF Competition winner Wayfair and the four other finalists: HP, Ipiranga, Maersk, and OpenGrid Europe.

- **Wayfair** describes their hierarchical forecasting engine that ensembles top-down time series and bottom-up machine learning forecasts to predict monthly demand.
- **HP** shows how it pairs large-scale machine learning forecasts with human insight.
- **Ipiranga** forecasts at multiple levels and different time horizons to manage fuel distribution.
- **Maersk** utilizes a statistical/ML-based automated forecast system to support granular and efficient repositioning of empty containers.
- **OpenGrid Europe** tackles the challenge of forecasting hourly gas flows using a hybrid approach combining ML, time series analysis, and optimization.



"How are the mighty fallen!"

—Mike Gilliland
Dragonfly Farm
Seagrove, N.C.
USA

The Future of Demand Forecasting with Generative AI

YUE LI AND RACHEL PEDERSEN

PREVIEW *Generative AI (gen AI) is an evolving field that shows promising impact on demand forecasting. Li and Pedersen explore how organizations can enhance productivity and decision making with gen AI models – yet warn there is still the need for strong governance to ensure quality and reliability in forecasting uses.*

Generative AI (gen AI) has been transforming business operations and day-to-day life. More and more companies are leveraging gen AI models to enhance productivity and improve decision making. Given the variety of benefits and the rapid evolution in the field, companies are increasingly focusing on the applications of gen AI.

Recent innovations in generative AI have increased the potential for business impact. Large language models (LLMs) are now equipped with more advanced reasoning capabilities, and multimodal models can provide insights from sources such as images and video in addition to text. Meanwhile, agentic AI has made headlines with 2025 called “the year for agentic adoption” (Zhang, 2025). With these developments, companies can apply gen AI to more complex uses. Additionally, organizations can enhance automation by using agents to complete specific tasks and workflows.

The forecasting field has also seen several notable advancements related to generative AI. *Foresight* issue 75 explored the application of LLMs for forecasting and experiments testing LLM forecasting capabilities (Hassani & Silva, 2024; Kolassa, 2024; Bergmeir, 2024). Large companies including Amazon and Google as well as startups such as Nixtla have released forecasting foundation models, which produce numeric forecasts and pre-train across large time series datasets. This model type has been

called “likely the next big thing in time series forecasting” (Bergmeir, 2024, p.33) and offers unique benefits relative to other algorithms.

In this article, we provide a perspective on how generative AI can be used to improve demand forecasting across data sources, algorithms, processes, and technology. For our discussion, we will focus on demand forecasting, which predicts the future customer demand for products and services. This type of forecasting supports resource planning and optimization with applications such as demand planning for production scheduling or inventory management, revenue forecasting for budget allocation, and operational forecasting for workforce planning. The predictions are typically made daily, weekly, or monthly. While our discussion centers on demand forecasting, the conclusions can potentially apply to other types of forecasting as well.

DATA: UNLOCKING NEW DATA SOURCES USING GENERATIVE AI

We have consistently observed in practice that better data beats a better algorithm. Reliable and extensive data sources are a critical foundation for accurate demand forecasts. Generative AI can unlock richer and more reliable data sources for forecasting models. In this section, we outline gen AI’s applications in structuring unstructured data to define valuable forecast drivers and

Key Points

- Generative AI's impact on demand forecasting is promising and still evolving. It has the potential to transform forecasting solutions by enhancing data, expanding algorithm possibilities, increasing interactivity in the forecast process, and redefining forecast technology and software.
- Large language models (LLMs) enhance forecasting data sources by extracting demand signals more accurately and efficiently through improvements in structuring valuable unstructured data sources and generating synthetic data.
- Forecasting foundation models provide advantages in adaptability and efficiency relative to traditional time series models, although the limited explainability of these models might lead to challenges with adoption.
- AI agents and applications using LLMs with retrieval augmented generation (RAG) architecture can provide automation, enhance model explanations, and streamline access to information needed for business decisions.
- Strong AI governance is needed to ensure data quality and reliability across generative AI uses in forecasting.
- Ultimately, gen AI has the potential to enrich existing forecast frameworks but is not a replacement for forecast expertise.

in generating synthetic data to support new product forecasts.

Structuring unstructured data

Companies collect various data sources that hold insights about the market or customers, and some of these data sources might be unstructured in nature – text based and highly variable in format. Demand forecasts can benefit from the information contained in these unstructured data sources, though these data sources must be structured in a column-based format for use in many forecasting algorithms. Examples of valuable unstructured data sources

include news reports containing market trend indicators, sales representatives' emails identifying sales trends or pricing information, and social media posts detailing customer sentiment.

Even though this unstructured data may be useful to forecasting, incorporating this data in forecasting models has been time consuming. Before the rise of LLMs, companies used a technique such as natural language processing (NLP) to structure these unstructured data sources. A traditional NLP pipeline involved extensive data preparation (including steps such as text preprocessing, tokenization, and vectorization) and separate models for each desired insight (such as topic classification or sentiment).

Using prompt engineering with an LLM, this process of structuring data from unstructured data sources is much more efficient. Now, a developer can write a prompt for an LLM to produce the insights from the unstructured data and then parse the LLM responses for the desired structured output. LLMs can be applied for these types of tasks without additional training, which is a key advantage over other methods. The results can be quickly refined and improved through providing a few examples in the prompt (i.e., few-shot learning). Through this technique, companies can unlock insights from unstructured data sources for forecasting in less time.

In addition, companies can access better demand signal from these data sources using LLM prompt engineering. Traditional NLP models do not perceive context as robustly as LLMs and can misclassify sarcasm or incorrectly identify products. In contrast, LLMs benefit from a deepened contextual understanding, allowing for more accurate data labeling. In **Figure 1**, we see an example of a customer review where the LLM prompt engineering technique better identifies sentiment and product

granularity. With improved data labels, forecasters can ultimately achieve better accuracy from the more reliable and more granular signal.

Figure 1. Traditional NLP vs. GenAI for a Sample Social Media Comment

“Wow, this smartphone XXX is just amazing and such an upgrade from the previous YYY version – if you love your calls dropping every two minutes and a battery that dies faster than your enthusiasm. It’s like they designed it to keep my life interesting!”

	Traditional NLP	GenAI
Sentiment	Positive	Negative
Granularity	-	Product XXX

Generating synthetic data

Forecasting new products is difficult due to lack of specific data and relevant history. To collect information before a new product launches, companies can run consumer surveys or panel sessions. However, this process is costly and time consuming, particularly for companies with frequent new product launches. Therefore, this type of customer survey data is not extensively leveraged in new product demand forecasts.

LLMs provide an efficient and cost-effective alternative way for companies to gain insights relevant to forecasting new products. These models can provide synthetic data for new products based on reasoning and defining relevant connections from their extensive training. For example, a company trying to forecast demand for a new product can use AI-generated synthetic personas to simulate survey responses. With the appropriate validation, these simulations can provide a cost-effective way to establish a baseline on potential consumer perspectives. This area is still emerging, and we anticipate significant advances as approaches evolve.

ALGORITHMS: LEVERAGING GENERATIVE AI MODELS TO GENERATE FORECASTS

Forecasting algorithm choices influence the accuracy, flexibility, and scalability of forecasting systems. Two key innovations in generative AI have broadened forecast algorithm frameworks: LLMs and forecasting foundation models. LLMs function as algorithm experimentation assistants, as users can upload data directly into a chat interface and ask for a forecast. Forecasting foundation models produce numeric forecasts directly and can adapt to a range of forecasting tasks based on their extensive training. We explore the different applications for LLMs and forecasting foundation models further in this section.

Large language models

Large language models (like those leveraged in ChatGPT) make forecasting accessible to users with varying levels of expertise. Any stakeholder with interest in forecasting can easily ask an LLM for a forecast through a series of prompts. The previously cited articles in *Foresight* issue 75 explored LLM forecasting.

Since models continue to rapidly evolve, we experimented with more recent LLM versions to identify the latest forecasting capabilities. We conducted experiments using ChatGPT for forecasting tasks, with underlying models GPT-4o (released May 2024) and GPT-o1 (released December 2024).

We observed significant improvements with the recent models. For example, these models now generate Python code, create plots, and provide clearer step-by-step instructions when asked to produce forecasts. In one experiment, we removed the last 12 observations from the well-known dataset of airline passengers, uploaded the data, and asked ChatGPT to forecast the next 12 values. ChatGPT, with the GPT-4o

model, provided both forecast values and Python code for the task. Using the newer GPT-o1 model, ChatGPT responded by providing detailed instructions (including plots and example code) for each step in the forecasting task without generating forecast values. With this recent functionality of producing code, ChatGPT can be used as an experimentation assistant to developers working on forecasting tasks.

However, these improved models still did encounter some of the challenges identified in previous tests of LLMs for forecasting, including hallucinations and result instability (Hassani & Silva, 2024). For example, we found a hallucination in one experiment using GPT-4o: A response claimed to have applied a Box-Cox transformation, but the code encountered an error, and the results did not incorporate the transformation. We also noted challenges with result instability, such as receiving inconsistent forecast results across two GPT-4o sessions using the same prompt. Such issues are known limitations for LLMs broadly and apply to other LLMs beyond those used in these experiments. These observations further highlighted the importance of testing and validating LLM forecasting approaches and output.

Given these observed improvements and limitations, we recommend LLMs can be used by forecasting teams to accelerate experimentation. If developers provide expertise-based validation on the LLM responses, they can leverage the suggested approaches to forecasting tasks. Specifically, we recommend that forecasters mainly use the suggested code instead of the direct forecast results, if possible.

Forecasting specific foundation models

Foundation models are a significant development in the field of forecasting, as they leverage large-scale, time-series-specific knowledge bases for prediction. Like large language models, they are

built on transformer-based model architecture and benefit from pretraining across data spanning various industries and time horizons. With their extensive pretraining, these models can work in a zero-shot context (making predictions without requiring retraining) and provide adaptability across uses. Some recent releases include TimeGPT from Nixtla (Garza et al., 2024), TimesFM from Google Research (Sen & Zhou, 2024), and Chronos from Amazon (Ansari et al., 2024).

Recent publications have detailed the model performance on benchmark data, though there is interest in further exploring model performance across business applications. For example, blog posts by Amazon Web Services (Biso et al., 2025) and Google Research (Sen & Zhou, 2024) detailed the performance advantages of Chronos and TimesFM, respectively, compared to traditional forecast algorithms using benchmark datasets. However, this performance might not reflect the behavior in business applications due to potential data leakage. Specifically, these foundation models might have seen this benchmark data in training, which could inflate performance results (Bergmeir, 2024).

To provide an additional perspective on the performance of these foundation models across various contexts, we designed an experiment to compare model performance without data leakage concerns. We produced forecasts using sanitized, private data that would not have been available in forecasting foundation model training. We compared the performance of some forecasting foundation models and traditional models. The forecasting foundation models are the latest models for timegpt, timesfm2, and chronos-bolt-base as of March 2025. The traditional models used are relatively out of the box with default parameters – some include a simple overwrite on seasonality due to known knowledge about the data

Figure 2. Model Performance by OWA and WMAPE Across Uses

Target type	Dataset	Frequency	Horizon	Top 3 models by OWA	Top 3 models by WMAPE
Units/Volume	Retail (710 series)	Daily	14	chronos, timesfm, prophet	chronos, prophet, auto_arima
	Petrochemical (197 series)	Monthly	3	timesfm, chronos, auto_ets	timesfm, chronos, timegpt
	CP (336 series)	Monthly	2	chronos, timesfm, theta	auto_arima, timesfm, auto_ets
	CP (669 series)	Weekly	6	timegpt, timesfm, chronos	timesfm, timegpt, chronos
\$ Sales	Fast food (461 series)	Daily	7	exponential_smoothing, auto_ets, timesfm	exponential_smoothing, auto_ets, timesfm
	Construction (3 series)	Monthly	3	auto_theta, theta, timegpt	timegpt, auto_theta, theta

frequency. For the comparison we used two metrics: OWA (overall weighted average), the benchmark metric used in the M4 competition, and WMAPE (weighted mean absolute percentage error), a metric that weights the absolute percentage error based on the actual value and assigns greater importance to higher-value series. The summarized results from these experiments are presented in **Figure 2**.

In our experiment, we observed that forecasting foundation models (highlighted in red) performed quite well overall across industry use cases. In general, we noted comparable performance with traditional models, with forecasting foundation models appearing in the top three models for both metrics across all selected datasets. When examining the accuracy values, the forecast foundation models show accuracy very similar to traditional models. Based on our evaluation in this experiment, there is no superior forecasting foundation model across problem contexts.

In addition to the promising performance of forecasting foundation models, these models also offer advantages in adaptability, efficiency, and speed at scale. These models are highly adaptable and generalizable due to their large training base. With this base, these models can provide reasonably good accuracy out of the box without requiring additional steps like demand

pattern assessment, seasonality tests, or parameter tuning. These models also show efficiency from a data perspective and do not have limitations due to degree-of-freedom constraints when used on small sample sizes like some traditional models. In addition, forecasting foundation models have advantages in efficiency over traditional forecasting models as they provide predictions in a zero-shot context, performing only inference in their run time. Further, forecasting foundation models benefit from computational power (running on CPU or GPU): users can submit multiple series at once to the models without implementing parallelization in code. For example, when running a test on the retail dataset with 710 series, we saw that traditional models ran in sequence took on average eight times longer than the average time for foundational models. At scale, the compute approach of the foundation models provides efficiency advantages over model training performed in sequence.

In the future, forecasting foundation models could potentially boost forecast performance through the incorporation of multimodal data sources. This extension in functionality can be compared to the evolution of LLMs, which have moved beyond text inputs to also consider data sources like images, audio, and video through multimodal models. Now, a multimodal model can

perform tasks such as interpreting a chart or summarizing events in a video. Forecasting foundation models have similar potential for multimodal extensions, directly incorporating a variety of data sources to improve forecast accuracy. For example, if a company is forecasting when machines might need servicing, a multimodal forecasting foundation model could assess images of the machines in addition to time series data to make better predictions on when servicing might be needed. This expanded capability to assess various data sources can enhance the accuracy of these forecasting foundation models.

Despite their advantages, forecasting foundation models lack explainability and transparency, which can make it difficult for stakeholders to trust the results. An important consideration for organizations when selecting forecast algorithms is the trade-off between accuracy and explainability. For some organizations, the improvements in accuracy from these forecasting foundation models might not outweigh the limitations in explainability. These models offer little transparency into how forecasts are generated – similar to other deep learning forecasting models – providing stakeholders with limited visibility into the reasons for forecast outcomes. Additionally, stakeholders might challenge the model inherently leveraging data from other industries (in the model pretraining), questioning the relevance of these sources and if this extra information might cause a *garbage in, garbage out* effect.

Given these considerations, forecasting foundation models are currently best positioned as enhancements to, rather than replacements of, traditional forecasting techniques. Forecasting foundation models could be particularly advantageous for companies that need to quickly scale forecasting capabilities but do not have stringent requirements for explainability. For companies

interested in using forecasting foundation models, some practical approaches for implementing these models while easing adoption challenges include:

1. Using a forecasting foundation model as a baseline forecast in a stacked model structure, so that it works together with a more explainable model to generate a final forecast;
2. Providing a forecast from a forecasting foundation model as a quick reference forecast for stakeholder decision making;
3. Incorporating foundation model options in a forecast pipeline (i.e., an automated machine learning [AutoML] model selection approach).

PROCESS AND TECHNOLOGY: ENABLING SOLUTIONS AND PROVIDING DECISION SUPPORT WITH GENERATIVE AI

Effective forecasting relies not only upon relevant data and effective forecast algorithms, but also integrated processes and technology. Generative AI can transform process and technology across the forecasting solution life cycle, affecting experimentation, development, automation, and operationalization. This transformation is enabled by gen AI techniques including retrieval augmented generation (RAG), coding assistants, and agentic AI. In this section, we explore each of these innovations and the corresponding influence on process and technology.

Enhancing forecast interpretability

One common challenge that business users face when applying forecasts is a limited understanding of how forecasts are generated and the key factors. Given the lack of transparency, organizations might not access all the benefits of forecast-informed decision making.

Businesses can take steps to overcome this challenge by leveraging generative AI to make the forecast more accessible to business users. This application of

generative AI can enable users (like financial planners) to better understand forecasts and better assess the business impact of forecast scenarios.

Specifically, businesses can enhance forecast understanding by making explanations more accessible using gen AI applications within RAG architecture. In RAG architecture, an LLM is equipped with a series of reference documents and can be instructed by prompt engineering (such as chain-of-thought prompting) to reference these documents when answering questions. To apply this principle to forecasting, organizations can build applications that give LLMs all documentation from the forecast model (produced by the forecasting and data science teams), equipping an LLM to provide explanations within the scope of those reference documents. For example, an organization can develop a chatbot powered by an LLM where a financial planner can ask questions about a forecast. This type of application can help the planner better understand predictions through a self-service interface. In this case, if the financial planner questioned why a forecasting model predicted a drop in revenue, the LLM could answer, referencing the model structure and summarizing the effect of key drivers in the model from feature importance artifacts.

Organizations can also improve decision making through enhanced scenario analysis using gen AI tools. As stakeholders review a forecast, they are often interested in how key factors and model assumptions affected the results. For example, a financial planner might be interested in understanding the impact of operational decisions or market indicators on the revenue forecast outlook. Using a generative AI-powered application, a stakeholder could use natural language queries to adjust key factors in the model and view forecast scenarios. To generate predictions for

these scenarios, the LLM would make an API call to run the forecast model given with the scenario-specific assumptions. This natural language approach offers more flexibility and a broader range of scenarios relative to other scenario analysis tools. Ultimately, this LLM-enabled scenario analysis methodology supports more informed strategic and operational choices through more dynamic scenario generation.

Streamlining forecast development using coding copilots

Forecasting teams can save time on code development using AI-powered coding technologies. These coding assistants, such as Cursor (cursor.com/en), ClaudeCode (claudecode.org), and GitHub Copilot (github.com/features/copilot), integrate directly with IDEs (integrated development environments) to provide real-time code suggestions and autocompletions. These tools can provide developers with syntax completion for tasks such as data preparation and modeling, documentation (such as docstrings), and automated unit tests. Overall, these copilots benefit forecasting teams by accelerating code development, freeing up human resource time to focus on other forecast-enhancement tasks.

Automating forecasting processes and streamlining decision making

Agentic AI offers unique capabilities compared to other types of generative AI, with agents capable of functioning autonomously, working toward goals, and making decisions. Although the implementations are still evolving, we believe scalable agentic AI solutions will emerge in 2025. There are ample opportunities to apply agents across the forecasting domain, and agents have the potential to revolutionize forecasting technology by facilitating automation and supporting decision making.

Agentic AI can facilitate efficient workflow orchestration. As agents can make decisions autonomously, agents can

leverage forecasting tools to create an autonomous forecasting pipeline. For example, a series of agents can be used to forecast product demand. In this setting, one agent would automatically collect the data, one agent would preprocess the data, one agent would select a model, and one agent would help interpret results. To ensure reliability in agent-driven forecasting processes, teams can establish a series of checks on the output (either human-driven manual checks or agentic-based automatic checks). In this way, agents can span the forecast workflow: providing a forecast, offering an explanation, and answering follow-up questions as needed.

Agentic AI can also streamline decision making from forecasts, providing suggestions for routine tasks. There is potential for agents to automatically optimize supply chains, adjust production, and place orders without human intervention. For example, an agent can directly provide guidance on how many products to manufacture based on supply chain considerations. This guidance would be based on reasoning across sources, which could follow this type of logic: *“For the upcoming month, forecast demand will be X based on our model, the organization already has Y units of inventory according to inventory data, and the organization targets a particular service level of Z based on supply chain guidelines. Given these considerations and capacity constraints, the Product Manufacturing agent recommendation is that we produce S units in factory A and T units in factory B.”* Then, the agent can implement this recommendation and place the order with user approval, increasing automation in downstream forecast-related processes. In these ways, agentic AI can transform both consumption of the forecast and the related decision-making processes. Ultimately, these innovations will transform current forecasting approaches and processes through

changing the way people interact with forecasting software and models.

AI GOVERNANCE

Throughout all the applications mentioned above, governance is critical to successful implementation of generative AI in forecasting. To guard against unintended outcomes, teams must prioritize testing gen AI outputs throughout development. For example, when using generative AI to enhance data, developers should review the output to validate the sample data aligns with expected results. To ensure reliable outputs from AI systems, organizations must prioritize guardrails and human-in-the-loop design. Examples of human-in-the-loop design might include a requirement for users to approve AI-suggested decisions or giving users the opportunity to provide feedback on responses from chat-based applications. Ultimately, AI governance supports the alignment of generative AI-enabled forecasting solutions with organizational standards and goals.

CONCLUSIONS

Generative AI is a powerful yet evolving tool within demand-forecasting applications. Forecasting solutions can incorporate gen AI capabilities to increase efficiency, streamline decision making, and improve results. Prompt engineering with LLMs can be used to unlock richer data sources for forecasting tasks. Recent innovations in LLMs can accelerate experimentation, and forecasting foundation models provide promising levels of accuracy with increased adaptability and speed benefits. AI agents and LLMs with RAG architecture can power solutions to support decision making and automation within the forecasting processes.

Despite these advancements, generative AI implementation alone is not enough. Strong AI governance is essential to

ensure reliability, accuracy, and alignment with standards. Generative AI is a powerful accelerator, though its impact ultimately depends on thoughtful implementation. Human expertise and judgment in forecasting will remain critical as gen AI is increasingly implemented in forecasting processes.

REFERENCES

Ansari, A.F. et al. (2024). Chronos: Learning the language of time series. arxiv.org/pdf/2403.07815. *Transactions on Machine Learning Research*, Volume 2024.

Bergmeir, C. (2024). Commentary: Can LLMs provide good forecasts? *Foresight*, 75, 18-20.

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

Biso, N., Chan, A., & Masood, M. (2025, March 5) Time series forecasting with LLM-based foundation models and scalable AIops on AWS, AWS Machine Learning Blog. aws.amazon.com/blogs/machine-learning/time-series-forecasting-with-llm-based-foundation-models-and-scalable-aiops-on-aws/

Garza, A., Challu, C., & Mergenthaler-Canseco, M. (2024). TimeGPT-1. arxiv.org/pdf/2310.03589.

Hassani, H. & Silva, E.S. (2024). Large language models as benchmarks in forecasting practice. *Foresight*, 71, 55-61.

Kolassa, S. (2024). LLMs, data leakage, bullshit, and botshit. *Foresight*, 75, 11-16.

OpenAI. (2025). ChatGPT. chatgpt.com/

Sen, R. & Zhou, R. (2024, February 2). A decoder-only foundation model for time-series forecasting. Google Research blog. research.google/blog/a-decoder-only-foundation-model-for-time-series-forecasting/.

Wang, J., Hoecker, A., & Whitten, C. (2025). What Is Agentic AI? Bain & Company. bain.com/insights/what-is-agentic-ai/.

Zhang, J. (2025, March 27). Why Agentic AI Is the Next Frontier of Generative AI. *Forbes*. forbes.com/councils/forbestechcouncil/2025/03/27/why-agentic-ai-is-the-next-frontier-of-generative-ai/?utm_source=chatgpt.com.



Yue Li specializes in demand forecasting, revenue management, and optimization. She leads Bain's Demand Forecasting Center of Excellence team, where she has advised numerous Fortune 500 clients across industries in diagnosing, enhancing, developing, and deploying forecasting solutions. Yue earned a

PhD in operations research, and holds five+ forecasting-related patents. Prior to Bain she worked in SAS Forecasting R&D, conducting research and developing enterprise solutions such as Merchandise Intelligence and Visual Forecasting. Yue is a member of the *Foresight* Advisory Board.

Yue.Li@bain.com



Rachel Pedersen is an Expert Senior Manager in Data Science with Bain & Company's AI, Insights, and Solutions team, specializing in predictive modeling and building generative AI applications. Rachel applies her expertise to enhance client capabilities in areas such as demand forecasting, marketing analytics, and revenue management. As a leader in

Bain's Demand Forecasting Center of Excellence, she collaborates with clients across various industries to incorporate best practices and improve forecast performance. Rachel holds a MS in analytics and a BS in statistics.

Rachel.Pedersen@bain.com

Integrating Forecasting and Inventory Decisions Using Machine Learning

JOOST F. VAN DER HAAR, YVES R. SAGAERT, AND ROBERT N. BOUTE

PREVIEW *Can inventory ordering decisions be improved by integrating forecasting and inventory decisions using machine learning? That is the question addressed in this study of three large Belgian companies in the food industry. Van der Haar, Sagaert, and Boute investigate the performance of methods that predict optimal order quantities directly, instead of first forecasting and then calculating optimal inventory quantities. Their results show that using an integrated approach can lead to substantial cost savings for smoother time series, yet the opposite holds when applying it to erratic and lumpy time series.*

A forecast is only as good as the decision it informs. Good demand forecasts allow organizations to maximize their customer service levels while minimizing inventory-related costs. Defining what makes a good forecast, however, is tricky. Demand forecasting traditionally focuses on predicting demand with maximum accuracy and uses these predictions to inform inventory ordering decisions. Recent research suggests that it may be better to instead directly predict optimal inventory ordering decisions, which maximize service levels and minimize costs. **Figure 1** visualizes the difference between these two approaches.

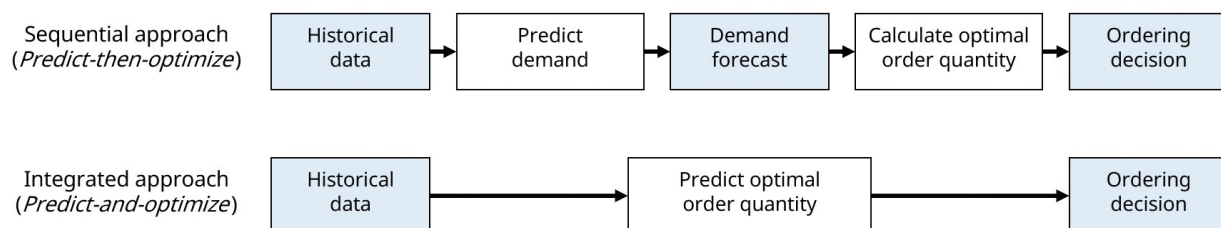
To understand when an integrated forecasting and inventory optimization approach works best, we first need to understand where sequential approaches fail. Sequential approaches, also referred to as predict-then-optimize approaches, typically involve creating a forecast and assuming a probability distribution on

the forecast errors. For example, we can assume that the forecast errors are independent and identically normally distributed. If the assumed distribution fits well, predict-then-optimize approaches can derive replenishment policies that perform well. If not, it may be better to use an integrated predict-and-optimize approach that does not require assumptions on forecast errors.

To see where such assumptions can break down, consider the following examples:

- *Independence:* A series of warmer days in a row can lead to consecutive under- or overestimation of demand, leading to correlation between forecast errors.
- *Identically distributed:* Variance of errors tends to be higher at demand peaks, for example when discounts are offered on products. As a result, errors are often not identically distributed.
- *Choice of distribution:* In practice, forecast errors rarely follow a nicely defined probability distribution such as the normal distribution.

Figure 1. Sequential vs. Integrated Approach to Forecasting and Inventory Control



Some of these assumptions can be avoided when using predict-then-optimize methods. For example, the assumption that the forecast error follows a specific probability distribution can be avoided by using the empirical error distribution. Quantile regression can be used to avoid both this assumption and the assumption that forecast errors are identically distributed. Eliminating the independence assumption, however, is more challenging.

In this study, we show when predict-and-optimize methods work well, and when predict-then-optimize methods prevail. To this end, we use historical sales datasets from three large Belgian food companies to compare the performance of several predict-then-optimize models to that of a predict-and-optimize machine learning method. We proposed the latter method in van der Haar et al. (2024), which builds on related works by Ban and Rudin (2019) and Huber et al. (2019), among others.

Our results show that the choice of method can have an enormous impact on the bottom line. The predict-and-optimize method can lead to cost savings of more than 60% in the best case, but it can also lead to cost increases of at most 43% in the worst case. We find that a predict-and-optimize method works best for smoother time series, whereas predict-then-optimize methods perform better for more erratic and lumpy time series.

APPROACH

Whenever you train a forecasting model using supervised learning, you identify a set of forecasting model parameter values such that the model's predictions minimize some error metric over the training data. For example, most models are trained to minimize the mean squared error (MSE) of the model predictions. However, when we optimize our predict-and-optimize models on the cost-based loss metric described in van der Haar et al. (2024), the loss function measures all current and future costs that follow directly from the inventory ordering decision. For example, if the model overestimates demand, we do not measure

Key Points

- This study compares the performance of different methods for forecasting and inventory control on data from three large Belgian food companies.
- Better forecasts do not necessarily lead to better inventory decisions. Correctly identifying the distribution of forecast errors is at least as important as minimizing the errors when the goal is to make inventory decisions.
- It may be better to directly forecast optimal order quantities ("predict *and* optimize") instead of first forecasting demand and then determining order quantities ("predict *then* optimize").
- Predict-and-optimize methods can lead to substantial cost savings for smoother time series, whereas predict-then-optimize methods achieve the best results for erratic and lumpy time series.

by how much demand is overestimated, but instead measure the amount of costs incurred as an effect of this overestimation (e.g., holding costs in terms of cost of capital, and perishing costs in terms of raw material costs).

To provide some intuition on what this means, we compare the cost-based loss function to the MSE and the pinball loss in **Figure 2**. The MSE penalizes large errors much more than small errors, and has the nice property that it leads to the maximum likelihood model if errors are normally distributed. When a forecaster wants to obtain quantile forecasts, they can use the pinball loss, which asymmetrically penalizes errors. For example, to obtain a 90% quantile forecast, this function penalizes underestimation nine times as much as overestimation. The cost-based loss is similar to the pinball loss, but penalizes based on prediction outcomes instead of prediction accuracy. For example, if losing sales is nine times as costly as having leftover inventory, this function penalizes ordering too little nine times as much as ordering too much.

Figure 2. Comparison of the Mean Squared Error, Pinball Loss and Cost-Based Loss

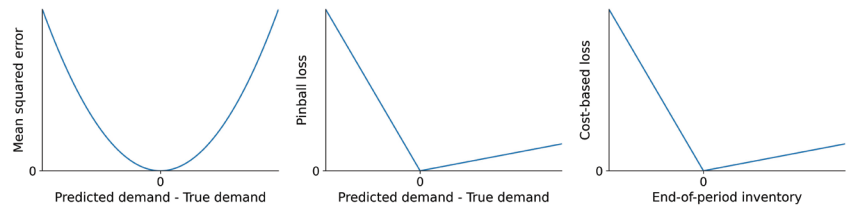


Figure 3. Example of Cost-Based Loss for Perishable Goods

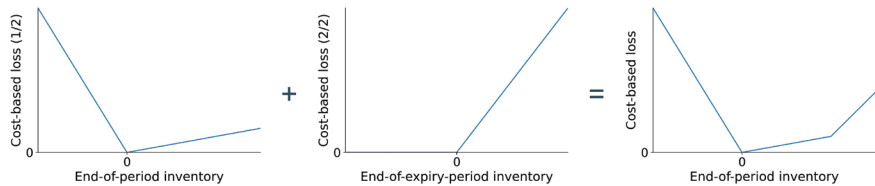


Figure 2 shows what the pinball loss and cost-based loss may look like when aiming for a 90% service level.

Cost-based loss, however, is much more flexible than pinball loss. The cost structure visualized in Figure 2 makes sense when dealing with nonperishable goods, where we trade off a risk between having too much or too little in the period for which we’re ordering. For perishable goods, we might also want to look at how much of what we order is left at its expiry date, such that we can account for expiry-related costs. Cost-based loss allows us to do exactly this, as visualized in **Figure 3**. This figure shows that the cost-based loss for nonperishable goods (left panel) can be augmented with a term that penalizes ordering so much inventory that it leads to products expiring (middle panel). The resulting function (right panel) consists of three parts: The leftmost part gives a linear penalty for each unmet demand. The middle part penalizes leftover

inventory that does not lead to perishing, e.g., to account for refrigeration costs. The right part penalizes leftover inventory that is bound to perish.

CASE STUDIES

We evaluate the performance of this predict-and-optimize approach on proprietary data from three large Belgian companies in the food industry. These were a ready-made meal company, a catering company, and a supermarket. Company C1 has 10 years of weekly sales records and data on ingredients and materials contained in their prod-

ucts, company C2 has 3.5 years of weekly sales data, and company C3 has 4 years of daily sales data on products grouped into four distinct product categories (a, b, c, and d). Product categories a and b involve nonperishable (frozen) products, whereas categories c and d involve perishable products. **Table 1** provides a brief overview of the different datasets in terms of average demand interval (ADI), squared coefficient of variation (CV2), number of products, and number of data rows.

To compare the performance of the predict-and-optimize against the sequential predict-then-optimize methods, we simulate how the inventory systems of the different companies would have performed under each of the methods. For nonperishable goods, we track holding costs for leftover inventory at the end of each period, and penalty cost for each sale that was lost because no inventory was available to meet demand. For perishable goods, we also track how much inventory was lost due to expiries. To obtain a realistic ratio between these costs, we set the holding cost at 1, the perishing cost at 8 and vary over the lost sales cost “ p ” to simulate the effect of different service levels. Products in category C3c expire after two days in stock, whereas products in category C3d expire after four days.

We compare the performance of eight different methods on these datasets:

Table 1. Summary of Data Characteristics for the Different Companies and Product Groups

Company/Category	C1	C2	C3a	C3b	C3c	C3d
ADI	1.00	1.00	1.25	1.00	1.08	1.00
CV2	0.45	0.14	1.48	0.60	0.87	0.57
Products	17	2	3	3	3	3
Data rows	5,596	372	4,383	4,383	4,383	4,383

- **ASL:** The predict-and-optimize method discussed in the previous section. We use ASL-LS (Approximate Supervised Learning for Lost Sales) to refer to the approach for nonperishable goods and ASL-PG (ASL for Perishable Goods) to refer to the one for perishable goods. We implement both in LightGBM, but note that our method is model-agnostic and can also be used for other model types such as XGBoost and neural networks.
- **QR:** A predict-and-optimize variation of quantile regression (QR) where a LightGBM model is trained to predict the optimal demand quantile using the pinball loss. The order size is then given by the prediction minus the current stock. Pinball loss coefficients are given by the penalty costs for underpredictions, and by the holding/perishing costs for overpredictions.
- **LGBM:** A predict-then-optimize method where a LightGBM (LGBM) model is used to create a forecast, after which safety stock is added based on the empirical distribution of the residuals (LGBM-E) or under an assumed normal distribution (LGBM-N).
- **ETS:** A predict-then-optimize method where an Error-Trend-Seasonality (ETS) model (Hyndman & Khandakar, 2008) is used to create a forecast, after which safety stock is added based on the empirical distribution of the residuals (ETS-E) or under an assumed normal distribution (ETS-N).
- **LR:** A predict-then-optimize method where a local linear regression model (i.e., a linear regression model specific to a single time series) is used to create a forecast, after which safety stock is

added based on the empirical distribution of the residuals (LR-E) or under an assumed normal distribution (LR-N).

For predict-then-optimize methods, both the empirical and the normal distributions are fitted on the forecast residuals from the preceding 52 weeks of data.

For each method except for LR, we obtain cost estimates using cross-validation with prequential train/validation/test splits. We use two years of validation data and two years of test data for C1. For C2, we use eight months of validation data and eight months of test data. Finally, we use one year of validation data and one year of test data for C3. Hyperparameter tuning is performed using Optuna. The test procedure is the same for LR, but no validation runs are performed as LR does not have hyperparameters to tune. Models are retrained at the start of every month for each dataset and each method. An illustration of the prequential train/validation/test procedure for company C1 is given in **Figure 4**.

For ease of interpretation, we present our results in terms of standardized costs in **Figures 5, 6, and 7**. Figure 5 provides the results for the nonperishable goods in datasets C1 and C2, Figure 6 for the nonperishable goods in C3a and C3b, and Figure 7 for the perishable goods in C3c and C3d. We obtain these standardized costs by dividing the total costs for a given method by the total costs under ASL for any given combination of dataset and cost of lost sales. Consequently, any standardized cost below 1.00 indicates that a method outperforms ASL, while any standardized cost above 1.00 indicates that a method is outperformed by ASL.

Figure 4. Setup of Our Train/Validation/Test Procedure for Company C1

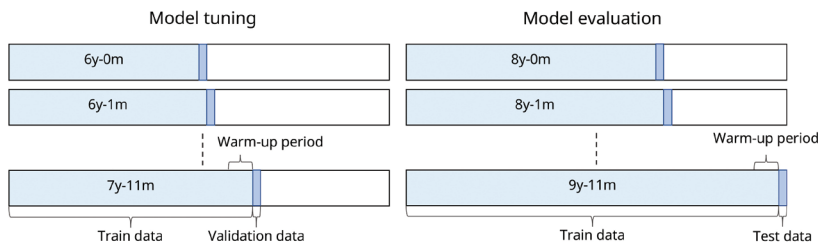


Figure 5. Standardized Costs on Nonperishable Ready-Made Meal and Catering Products

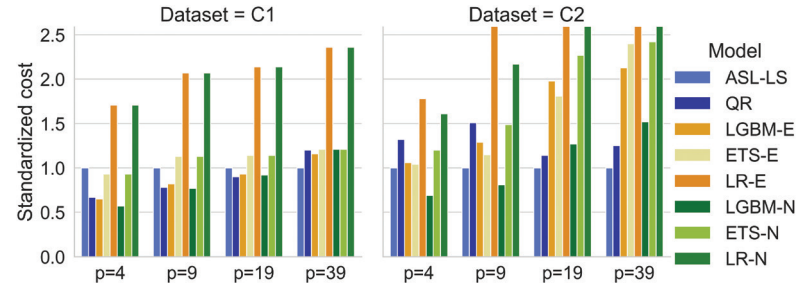


Figure 6. Standardized Costs on Nonperishable Retail Products

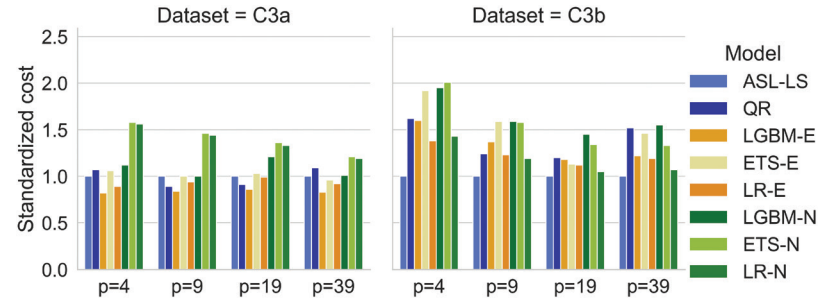
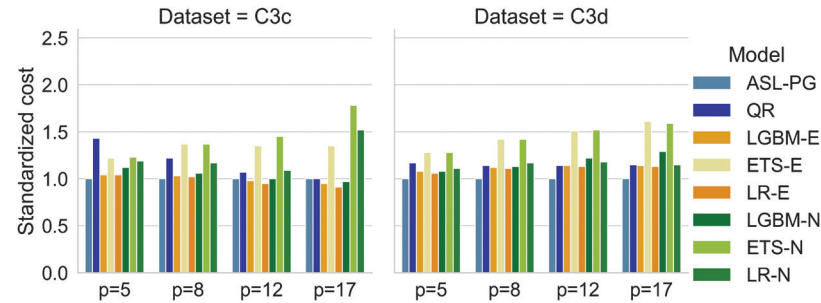


Figure 7. Standardized Costs on Perishable Retail Products



DISCUSSION

The results show that there is no single method that always performs best. We find that the choice for a predict-and-optimize method can lead to either considerable cost savings of more than 60% compared to any benchmark, or considerable cost increases of at most 43%. Cost differences are largest for nonperishable goods (Figures 5 and 6) and appear to be moderated by the penalty p for unmet demand. For companies C1 and C2, the integrated predict-and-optimize method is the best-performing method for the higher service levels (where p is higher). For datasets C3a and C3c, a predict-then-optimize consistently performs on par or best, whereas for datasets C3b and C3d, a

predict-and-optimize model consistently performs on par or best.

To understand where performance differences may come from, it is easiest to consider the datasets and their characteristics pairwise. First, both C1 and C2 involve weekly nonintermittent sales data for different nonperishable food offerings with relatively low variation in sales, as measured by the CV2. Second, C3a and C3b contain daily sales data for different nonperishable product categories sold by the same supermarket. Third, C3c and C3d contain the same type of data for different perishable product categories. Note that for each pair, performance of the predict-and-optimize approach is best for the dataset with the lower CV2.

More generally, ASL performs best for 75% of cases where $CV^2 < 0.7$ (datasets C1, C2, C3b and C3d) and is outperformed for 75% of cases where $CV^2 \geq 0.7$ (datasets C3a and C3c). For the case with smooth demand where $CV^2 < 0.7$, ASL leads to average cost savings of 18% compared to the best-performing benchmark (QR). In contrast, for the case with more erratic or lumpy demand where $CV^2 \geq 0.7$, it would lead to an average cost increase of 8% compared to the best-performing benchmark (LGBM-E).

CONCLUSION

The quality of a forecast depends on more than its accuracy. The ability to accurately describe the distribution of its errors is oftentimes at least as important, as the decisions a forecast informs can be just as dependent on the error distribution as on the forecast accuracy. When the assumed distribution of the forecast errors (e.g., the normal distribution) differs from the true distribution, methods that integrate forecasting and inventory optimization tend to perform well, as they do not involve forecast errors.

The key idea of integrated methods, also known as predict-and-optimize methods, is to directly forecast the optimal order decision instead of first creating a demand forecast and then using it to determine the order decision. We investigate the performance of one such integrated method and find that it can substantially outperform methods that treat forecasting and inventory optimization as two distinct tasks for smooth time series. In contrast, the latter tend to outperform the investigated predict-and-optimize method for more erratic and lumpy time series. We thus conclude that a predict-and-optimize method can aid companies dealing with smoother demand patterns in simultaneously reducing costs and maximizing product availability.

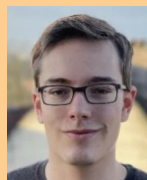
REFERENCES

Ban, G.-Y. & Rudin, C. (2019). The big data newsvendor: Practical insights from machine learning. *Operations Research*, 67(1), 90–108.

Huber, J., Müller, S., Fleischmann, M., & Stuckenschmidt, H. (2019). A data-driven newsvendor problem: From data to decision. *European Journal of Operational Research*, 278(3), 904–915.

Hyndman, R.J. & Khandakar, Y. (2008). Automatic time series forecasting: The forecast package for R. *Journal of Statistical Software*, 27, 1–22.

Van der Haar, J.F., Wellens, A.P., Boute, R.N., & Basten, R.J.I. (2024). Supervised learning for integrated forecasting and inventory control. *European Journal of Operational Research*, 319(2), 573–586.



Joost F. van der Haar is a PhD Researcher in machine learning for operations management at the Faculty of Economics and Business at KU Leuven. His research focuses on the interface between forecasting and decision optimization, with application areas ranging from inventory control to maintenance optimization. The objective of his research is to align forecasting and decision optimization for these areas using techniques such as cost-sensitive learning and deep reinforcement learning.

joost.vanderhaar@kuleuven.be



Yves R. Sagaert is head of the research group of Predictive AI and Digital Shift at the VIVES University of Applied Sciences in Belgium. He is currently a researcher at KU Leuven and an Adjunct Professor at the IÉSEG School of Management in Lille (France). His research focuses on incorporating market intelligence in demand forecasting through leading indicators and the effects on supply chain management, and especially on the presence of limited historical business data. His expertise lies in supply chain management, leading indicators, business forecasting, inventory, variable selection, and shrinkage.

yves.sagaert@vives.be



Robert N. Boute is a Professor of Operations Management at the Faculty of Economics and Business at KU Leuven and Vlerick Business School, and a principal investigator at Flanders Make. His research focuses on inventory control and supply chain management. Recent works include digital operations, smart logistics, and predictive analytics for service maintenance applications. He was named one of the 40 Best Professors Under 40 by Poets & Quants in 2016, and was laureate of the 2017 Franz Edelman Award for Achievement in Operations Research and Management Sciences.

robert.boute@kuleuven.be

Types of Forecast Errors and Their Implications

KOLJA JOHANNSEN

PREVIEW *Forecast errors are inevitable, but not all errors are created equal. How you react to forecast errors can be as important as choosing the right forecasting model to begin with. Kolja Johannsen moves beyond the typical discussion of forecast error metrics by categorizing four types of forecast errors, explaining why differentiating them matters, and providing strategies on how to respond to them.*

THE GOAL OF FORECASTING

The goal of forecasting generally is to help make better decisions, whether it's where and how much to invest, how much inventory to hold, or whether to bring along an umbrella. In practice, forecasting involves balancing two often competing objectives:

1. **Accuracy:** How closely predictions match actual outcomes.
2. **Explainability:** The ability to understand the factors behind forecast deviations.

It is easy to over-index on forecasting accuracy or how close the forecast is to the actual values. However, "in itself, a more accurate forecast delivers no intrinsic value [as] its worth comes by facilitating better decisions and actions that do generate monetary and other organizational benefits" (Robette, 2023, p. 12).

When solely relying on forecast accuracy as a target, you not only forgo additional insights forecasts can provide, but also risk lower forecasting accuracy in the future. Moreover, when used well, forecasts are not mere predictions but tools for detecting, understanding, and explaining business dynamics.

THE NATURE OF FORECAST ERRORS

Forecast errors are the difference between actual and predicted values. They matter in two ways:

1. **In-sample errors:** Used to set the parameters of the forecasting model

and thereby shape predicted future values.

2. **Out-of-sample errors:** Measure how well the forecast performs vs. actuals.

A skilled forecaster manages the in-sample errors so that the model captures signals and filters out noise, while using out-of-sample errors to identify business changes and communicate reasons for deviations. "The nature of the forecast error can make a big difference" (Morlidge, 2023, p. 33) because it is not solely its size that can influence the quality of decisions, but what created the error in the first place.

TYPES OF FORECAST ERRORS

When treating all forecast errors as equal, we disregard that their source and properties can have vastly different implications for business decisions.

Hendry (2000) differentiates between seven types of forecast errors based on their statistical origins. Although this framework offers valuable insights for theoretical analyses, its direct application in business practice is less straightforward.

In this article, I introduce a taxonomy of forecast errors that is aimed at addressing how error types affect an organization's decision making. This framework is motivated by my experience forecasting daily business metrics at a quarterly cadence with regular reviews and a special focus on quarterly forecast accuracy. While the daily forecast is used to determine

the health of the business, the quarterly accuracy is crucial for planning and goal setting.

Key questions I commonly answer in this context are: How are we performing vs. the forecast? What explains the difference between the forecast and actuals? What does this mean for the rest of the quarter?

As a result, a crucial feature of forecast deviations is whether they primarily impact the daily accuracy (“transient” forecast errors), or if they affect the quarterly accuracy as well (“persistent” forecast errors).

Within these two categories, I differentiate between four distinct error types: noise, consciously omitted features, deviations in predicted inputs, and misspecifications. This classification is illustrated in **Figure 1**.

Although there is a natural parallel between the dichotomy of persistent and transient errors and the differentiation between bias and variation (see e.g., Morlidge, 2023, p. 35), not all persistent errors indicate a bias in the forecasting model, as explained below.

TRANSIENT DEVIATIONS

Transient deviations are forecast errors that are short-lived and should not affect the forecast’s performance over longer horizons. These are errors that may impact the forecast accuracy for a few days or a week, but do not affect the quarterly (average) forecast accuracy.

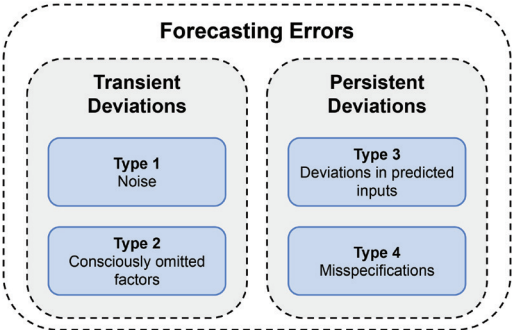
Type 1: Noise

- *Description:* Random, unpredictable fluctuations inherent in any data.
- *Example:* Day-to-day variations in sales due to individual consumer behaviors.
- *Implications:* Such fluctuations are normal and not indicative of systemic issues. The level of noise in a time series limits the level of forecast accuracy that can be achieved.
- *Action:* To avoid overreactions, educate stakeholders that noise is inevitable and unavoidable. Where possible, seek

Key Points

- Forecast errors are not merely technical issues; they have significant implications for decision making.
- Errors can be classified into four error types based on their persistence and origins.
- Understanding the source of errors allows forecasters to refine models, improve communication, and align stakeholder expectations.
- Practical guidelines for distinguishing error types include trend and performance tracking, collaboration with business partners, and understanding business drivers.
- Effective error analysis ensures that future forecasts build on strengths rather than repeating past errors.

Figure 1. Classification of Forecasting Error Types



to reduce the noise in the data, which can sometimes be achieved by changes in policies and practices that drive the data.

This first type of forecast error is what most people think of as the forecast error: the randomness that is inherent to the data generating process (DGP). This inherent noise represents the unavoidable day-to-day fluctuations between forecasts and actuals. Over longer periods of time these fluctuations should largely cancel out.

Type 2: Consciously omitted features

- *Description:* Factors intentionally excluded from the model to avoid overfitting or instability.

- *Example:* Including all possible regional, national, or religious holidays would introduce excessive parameters and risk overfitting; however, omitting them will lead to temporary deviations. There is a trade-off between reduced complexity and accuracy.
- *Implications:* These deviations carry information about future patterns and, while not explicitly modeled, may need to be considered when evaluating business performance.
- *Action:* Document impactful factors and communicate expected deviations to stakeholders.

This second type of forecast error receives less attention but is a common challenge when dealing with forecasts. No matter how sophisticated a model is, there are factors that are not fully built into the model, even if it's known that they have an impact. In contrast to the Type 1 “noise,” these deviations are predictable.

One solution is to document which holidays have historically had an impact and to communicate proactively to stakeholders that a temporary deviation of x% is expected over this period. Because this deviation is temporary, it does not indicate a risk to forecast accuracy on other dates. Another option is to include a second step in the forecasting process to adjust predictions based on these factors.

Importantly, consciously omitted features leading to Type 2 errors only have a temporary impact. Features that explain larger shifts in the level or trend and thereby affect forecast deviations over longer horizons are not transient, but persistent.

PERSISTENT DEVIATIONS

In contrast to the errors above, there are also forecast deviations that have a lasting impact on our expectations and require different responses.

Type 3: Deviations in predicted inputs

- *Description:* Errors stemming from inaccuracies in external variables used in the forecast or other variables within a multivariate time series model.
- *Example:* Lower-than-expected marketing expenditure leading sales to underperform vs. forecast.
- *Implications:* When inputs, like marketing budgets or predicted weather conditions, are subject to forecast errors themselves, downstream forecasts inherit these inaccuracies.
- *Action:* Reassess input forecasts and communicate forecast deviations proactively.

Using forecasts as explanatory variables (inputs) introduces a new source of error: inaccurate input predictions. The resulting errors are distinct from errors caused by the forecasting model itself. While they can result in increasing deviations over time, these errors do not indicate a misspecification of the forecasting model as this still reflects the DGP. Type 3 errors are generally not the result of a bias in the forecasting model.

Type 3 errors are helpful from an explainability perspective. We can quantify how much of the forecast miss was due to the forecast error in each input and how much remains unexplained by them. This explainability supports the goal stated by Robette (2023, p. 19) that “the forecast and the decision process [should be] as well aligned as possible to create value for an organization.”

The concept of Type 3 errors is less directly applicable to univariate time series models that lack predicted exogenous regressors. For example, in any univariate autoregression, any lasting deviation can ultimately be interpreted as a misspecification of the model.

Type 4: Misspecifications

- *Description:* Persistent errors due to model assumptions such as missing factors or incorrect parameters.
- *Example:* Using a linear trend to forecast growth when the growth is exponential.
- *Implications:* These errors undermine both accuracy and explainability, leading to poor decision making. They often signal the need for fundamental adjustments to the model or its assumptions.

- *Action*: Revisit and refine assumptions, incorporating deeper understanding of underlying dynamics.

These are the most consequential forecast errors because the model does not correctly capture underlying dynamics hurting both explainability and accuracy. Type 4 errors are not short-lived and often become worse over time. This includes structural breaks in deterministic trends which Hendry (2002, p. 25) identified as the “primary cause of forecast failure.”

Because these errors are often caused by fundamental changes to the business – like supply disruptions or shifts in consumer behavior – forecasts can help identify these disruptions and alert stakeholders proactively.

ILLUSTRATIVE EXAMPLE

The stylized example in **Figure 2** shows the forecast errors of a revenue prediction made on 12/31. It illustrates how the different types of errors contribute to the overall observable forecast error and how their identification helps reveal underlying dynamics.

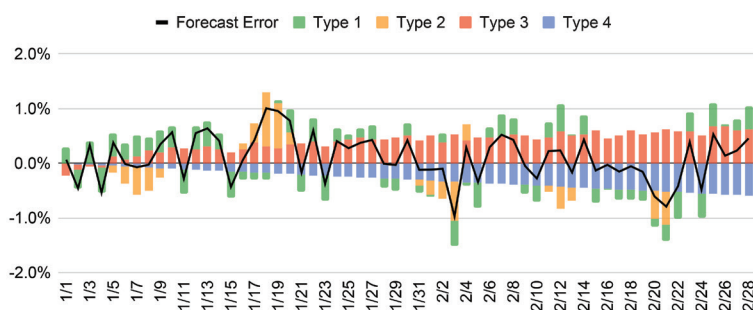
The *Type 1 errors* (noise) are relatively minor and have an average close to zero, meaning they do not impact the quarterly average accuracy.

The *Type 2 errors* (consciously omitted factors) cluster around omitted holidays. While they account for the largest portion of the forecast errors on 1/7 and 1/18, these deviations are short-lived.

The *Type 3 errors* (deviations in predicted inputs) result from higher-than-expected marketing expenditure. These errors grow over time as actual spending growth exceeds what was assumed in the forecast. Whether this bias persists or reverts depends on whether elevated spending levels continue. In this case, the source of the forecast error is known and reflects updated business decisions rather than a flaw in the forecasting model.

The *Type 4 errors* (misspecifications), by contrast, are increasingly negative and require further investigation. If they are driven by incorrect trend assumptions,

Figure 2. Breakdown of Forecast Deviations



the errors will likely continue to grow, not only this quarter but also in next quarter’s forecast. Simply rerunning the model with updated data is unlikely to resolve this completely.

Focusing solely on the total forecast error can create a misleading impression that the model is unbiased and performing well, with average errors near zero. However, as Hendry (2000, p. 15) cautions, “careful analysis as to why [a forecast is accurate] is strongly recommended.” In reality, Type 3 and Type 4 errors can occur simultaneously, masking the underlying dynamics that drive the deviations.

By breaking the overall error into its components, it becomes clear that unless we adjust the model to address structural misspecifications, future forecasts will continue to be biased.

PRACTICAL GUIDELINES FOR IDENTIFYING AND MANAGING ERROR TYPES

As shown in the example, being aware of the drivers behind forecast errors can help improve forecast accuracy as well as making the forecast more useful for decision making.

While it is rarely possible to perfectly determine how much each error type contributes to the overall forecast errors, a detailed error analysis can provide valuable insights for stakeholders and forecasters. The classification introduced above provides a conceptual framework for this analysis and the following guidelines.

1. Understand Business Drivers

- Context is crucial for evaluating if a forecast deviation provides a signal or is due to noise. Knowing where the business is expanding helps, for example, to distinguish between an expected forecast error due to shifts in the impact of regional holidays (Type 2) and other errors. This understanding helps not only to explain forecast errors, but also to manage in-sample forecast errors by improving modeling assumptions.

2. Track Trends in Forecast Errors

- Keeping track of how forecast errors evolve (in-sample and out-of-sample) helps to identify when errors drift over time, indicating persistent errors: Type 3 or 4.
- Keeping a track record of what explains past deviations helps identify omitted drivers such as holidays (Type 2) and informs whether additional features need to be included in the model going forward.

3. Evaluate Predicted Inputs

- If the forecast relies on predicted exogenous variables, keeping track of the performance of these predictions is a key consideration. This allows us to differentiate between Type 3 and Type 4 errors and helps evaluate if the accuracy of predicted inputs changes over time.
- Exogenous variables that cannot be accurately predicted should be identified and omitted from the model.

4. Analyze Past Forecast Performance

- In practice, forecasts are rarely one-off but rather happen regularly with iterations on methodology. Tracking how outdated forecasts perform in the long run helps evaluate whether fundamental assumptions of these forecasts were reasonable, e.g., assumptions around trendlines (linear, logarithmic, exponential, etc.). This can help reveal patterns that otherwise would be missed on shorter horizons.

- Backtesting the current version of the model can be a helpful tool if the current version can be applied to more limited historical data. This also allows us to quantify what share of persistent deviations were Type 3 and Type 4.

5. Review Omitted Factors

- Type 2 errors can be quantified by analyzing in-sample forecast errors. With hundreds of holidays and events worldwide, modeling all of them explicitly would lead to overfitting. However, by analyzing their impact on in-sample deviations, you can derive an expected impact.
- While exogenous variables that cannot be predicted accurately may need to be removed from the model, they can still be a valuable source of information when analyzing past forecast errors. This can help quantify Type 4 errors that stem from omitted variables.

6. Avoid Overreacting to Transitory Deviations

- It is tempting to overinterpret the noise inherent in forecast errors and to give in to human biases and tendencies to see patterns where there are none. Similarly, it can be tempting to give up on interpreting forecast deviations altogether. The key is to accept that small day-to-day deviations are normal (Type 1 and 2 errors) and do not necessarily reveal information about long-term accuracy.

7. Educate Stakeholders

- Stakeholder confidence is key to maximizing the benefits forecasts have for decision making, and educating stakeholders is crucial for building confidence. This includes providing a general understanding of the forecasting approach, and also regularly sharing updates on how the forecast is performing and why deviations exist. Quantifying the impact of Type 2 and 3 errors can be a valuable tool to build confidence in the forecast.

CONCLUSION

Forecasting errors are not merely a technical concern but a resource for refining forecasting models and decision-making processes. By categorizing errors and understanding their causes, forecasters can improve both the quality of predictions and their communication to stakeholders. Future forecasts should build on these insights, ensuring progress rather than replicating past errors.


REFERENCES

- Hendry, D. F. (2000). A general forecast-error taxonomy. Mimeo, Nuffield College, Oxford. [bc.edu/RePEC/es2000/0608](https://www.bec.edu/RePEC/es2000/0608)
- Hendry, D.F. (2002). Forecast failure, expectations formation and the Lucas critique. *Annales d'Économie et de Statistique*, 67/68, 21–40.
- Morlidge, S. (2023). Measuring the cost of forecast error. *Foresight*, 68, 31–35.
- Robette, J. (2023). Does improved forecast accuracy translate to business value? *Foresight*, 68, 12–19.




Kolja Johannsen is a Leading Data Scientist at Duolingo Inc., specializing in forecasting and business intelligence. He previously worked as an Economic Consultant at Cornerstone Research and received a PhD in finance from the University of Warwick (UK).

kolja@johannsen-web.com




From Supply Chain Chaos to S&OP Excellence


Redefine efficiency with AI-driven predictive insights, robust replenishment planning, and real-time analytics—all on a single, agile cloud platform. Deploy a smarter, more resilient supply chain in weeks with PlanVida.





Contact us today for your free three-month trial!


Inventory Planner


 Inventory Tracking


 Replenishment Tracker


 S&OP / IBP


 Executive View


 Purchase Order Details


 S&OP Chart


 Warehouse Utilization

 Container Tracking

 Alerts

 Purchase Order Tracker


 Inventory Decomposition


 Inventory Performance


S&OP / IBP

	Category	S&OP Category	Warehouse	Supply SKU	Country						
Period: Year	2025-05	2025-06	2025-07								
Week Start	4/21/2025	4/28/2025	5/5/2025	5/12/2025	5/19/2025	5/26/2025	6/2/2025	6/9/2025	6/16/2025	6/23/2025	6/30/2025
Total											
Opening Inventory	3,649,777	3,903,488	4,011,769	4,098,054	4,153,643	4,126,756	4,210,239	4,020,766	3,868,748	3,705,422	3,60
Forecast	258,033	257,993	258,993	262,150	257,874	257,748	260,864	260,937	255,629	256,022	26
Order Receipts	508,332	364,616	343,624	316,154	229,014	341,066	68,708	104,902	89,664	147,902	5
Inbound Containers	96	66	48	51	26	27	13	18	16	25	
Confirm Inbound Containers	96	66	48	51	26	27	13	18	16	25	
Projected Order Release	67,220	12,396	54,506	46,830	63,738	81,182	126,114	113,870	126,974	375,966	22
Projected Inbound Containers	0	0	0	0	0	0	0	0	0	0	
Total Order Receipt	508,332	364,616	343,624	316,154	229,014	341,066	68,708	104,902	89,664	147,902	5
Total Inbound Containers	96	66	48	51	26	27	13	18	16	25	
Closing Inventory	3,903,488	4,011,769	4,098,054	4,153,643	4,126,756	4,210,239	4,020,766	3,868,748	3,705,422	3,600,723	3,40
Weeks on Hand	15.1	15.5	15.9	16.1	16.0	16.3	15.6	15.0	14.4	14.0	
Pallets	21,826	22,560	22,248	22,032	20,859	20,059	18,832	17,737	16,705	15,960	1
Square Feet	291,005	300,801	296,638	293,758	278,123	267,447	251,095	236,488	222,736	212,796	19
Next 12W Average	258,560	258,455	258,563	258,267	258,115	257,962	257,840	257,732	257,606	257,438	25
Avg Minimum WOS	2	2	2	2	2	2	2	2	2	2	
Cuts	-3,412	-1,659	-1,654	-1,585	-1,973	-164	-2,683	-4,017	-2,640	-3,421	-

USA | UK | South East Asia

 www.planvida.ai

 +1(781)995-0685

 valitude@valuechainplanning.com

SPECIAL FEATURE:

REVISITING SYMMETRIC MAPE

PREVIEW For 40 years the “asymmetry” of MAPE has been discussed, debated, and at times seemingly resolved with definition of the Symmetric MAPE (sMAPE) metric. However, some versions of sMAPE turned out to be not quite as symmetric as intended, and negative forecasts or actuals posed unforeseen challenges. In 2014 Rob Hyndman explored these issues in his Hyndsight blog (an updated adaptation of which is published here). And now, M4 winner Slawek Smyl proposes a new variation he calls Sparse-proof MAPE (msMAPE), suitable for both large-valued and sparse (intermittent) time series when forecasts and actuals are non-negative. Stephan Kolassa closes the feature with a commentary on Smyl’s msMAPE, advocating for the use of simulation to better understand what an error metric does in a variety of situations.

Errors on Percentage Errors

ROB J. HYNDMAN

The following is an adaptation of Rob Hyndman’s Hyndsight blog post from April 16, 2014, available at robjhyndman.com/hyndsight/smape/. Notation has been updated to bring consistency across all three articles in this special feature.

The MAPE (mean absolute percentage error) is a popular measure for forecast accuracy and is defined as

$$\text{MAPE} = 100 * \text{mean}[|A - F| / |A|]$$

where A denotes the actual (observed value) and F denotes its forecast. The mean is taken over all points in the time frame under consideration.

Armstrong (1985) was the first (to my knowledge) to point out the asymmetry of the MAPE, saying that “it has a bias

favoring estimates that are below the actual values” (p. 348). A few years later, Armstrong and Collopy (1992) argued that the MAPE “puts a heavier penalty on forecasts that exceed the actual than those that are less than the actual.” Makridakis (1993) took up the argument, saying that “equal errors above the actual value result in a greater APE than those below the actual value.” He provided an example where A = 150 and F = 100, so that the relative error is 50/150 = 0.33, in contrast to the situation where A = 100 and F = 150, when the relative error would be 50/100 = 0.50.

Thus, the MAPE puts a heavier penalty on errors when A < F than when F < A.

To avoid the asymmetry of the MAPE, Armstrong (1985) proposed the “adjusted MAPE,” which he defined as

$$\text{adjusted MAPE} = 100 * \text{mean}[2 * |A - F| / (A + F)]$$

There is considerable debate on whether forecast error should be defined as $\epsilon = A - F$ or as $\epsilon = F - A$ (Green & Tashman, 2008). The former is preferred by statisticians, while the latter is more common among practitioners. Since the absolute error term $|A - F|$ is equivalent to $|F - A|$, such discussion is not germane to this article.

By that definition, the adjusted MAPE can be negative (if $A + F < 0$), or infinite (if $A + F = 0$), although Armstrong claims that it has a range of (0,200). Presumably he never imagined that data and forecasts can take negative values. Strangely, there is no reference to this measure in Armstrong and Collopy (1992).

Makridakis (1993) proposed almost the same measure, calling it the “symmetric MAPE” (sMAPE), but without crediting Armstrong (1985), defining it

$$\text{sMAPE} = 100 * \text{mean}[2 * |A - F| / |A + F|]$$

However, in the M3 competition paper by Makridakis and Hibon (2000), sMAPE is defined equivalently to Armstrong’s adjusted MAPE (i.e., without the absolute values in the denominator), again without reference to Armstrong (1985). Makridakis and Hibon claim (incorrectly) that this version of sMAPE has a range of (−200,200).

Flores (1986) proposed a modified version of Armstrong’s measure, defined as exactly half of the adjusted MAPE defined above. He claimed (again incorrectly) that it had an upper bound of 100.

Of course, the true range of the adjusted MAPE is $(-\infty, \infty)$ as is easily seen by considering the two cases $A + F = \varepsilon$ and $A + F = -\varepsilon$, where $\varepsilon > 0$, and letting $\varepsilon \rightarrow 0$. Similarly, the true range of the sMAPE defined by Makridakis (1993) is $(0, \infty)$. I’m not sure that these errors have previously been documented, although they have surely been noticed.

Goodwin and Lawton (1999) point out that on a percentage scale, the MAPE is symmetric and the sMAPE is asymmetric. For example, if $A = 100$, then $F = 110$ gives a 10% error, as does $F = 90$. Either would contribute the same increment to MAPE, but a different increment to sMAPE.

Koehler (2001), in a commentary on the M3 competition, made the same point, but without reference to Goodwin and Lawton.

Whether symmetry matters or not, and whether we want to work on a percentage or absolute scale, depends entirely

on the problem, so these discussions over (a)symmetry don’t seem particularly useful to me.

Chen and Yang (2004), in an unpublished working paper, defined the sMAPE as

$$\text{sMAPE} = \text{mean}[2 * |A - F| / (|A| + |F|)]$$

They still called it a measure of “percentage error” even though they dropped the multiplier 100. At least they got the range correct, stating that this measure has a maximum value of two when either A or F is zero, but is undefined when both are zero. The range of this version of sMAPE is (0,2). Perhaps this is the definition that Makridakis and Armstrong intended all along.

As will be clear by now, the literature on this topic is littered with errors.

If all data and forecasts are non-negative, then the same values are obtained from all three definitions of sMAPE. But more generally, the last definition above from Chen and Yang (but with a multiplier of 100) is clearly the most sensible, if the sMAPE is to be used at all. In the M3 competition, all data were positive, but some forecasts were negative, so the differences are important. However, I can’t match the published results for any definition of sMAPE, so I’m not sure how the calculations were actually done.

Personally, I would much prefer that either the original MAPE be used (when it makes sense), or the mean absolute scaled error (MASE) be used instead (Hyndman & Koehler [2006]; en.wikipedia.org/wiki/Mean_absolute_scaled_error). There seems little point using the sMAPE except that it makes it easy to compare the performance of a new forecasting algorithm against the published M3 results. But even there it is not necessary, as the forecasts submitted to the M3 competition are all available in the Mcomp package for R (cran.r-project.org/web/packages/Mcomp/), so a comparison can easily be made using whatever measure you prefer.

Thanks to Andrey Kostenko for alerting me to the different definitions of sMAPE in the literature.

REFERENCES

Armstrong, J.S. (1985). *Long-range forecasting: From crystal ball to computer*, Wiley.

Armstrong, J.S. & Collopy, F. (1992). Error measures for generalizing about forecasting methods: Empirical comparisons. *International Journal of Forecasting*, 8(1), 69-80.

Chen, Z. & Yang, Y. (2004). Assessing forecast accuracy measures. [researchgate.net/publication/228774888_Assessing_forecast_accuracy_measures](https://www.researchgate.net/publication/228774888_Assessing_forecast_accuracy_measures).

Flores, B.E. (1986). A pragmatic view of accuracy measurement in forecasting. *Omega*, 14(2), 93-98.

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

Green, K. & Tashman, L. (2008). Should we define forecast error as $e = F - A$ or $e = A - F$? *Foresight*, 8, 38-40.

Hyndman, R.J. & Koehler, A.B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4), 679-688.

Koehler, A.B. (2001). The asymmetry of the sAPE measure and other comments on the M3-Competition. *International Journal of Forecasting*, 17, 537-584.

Makridakis, S. (1993). Accuracy measures: Theoretical and practical concerns. *International Journal of Forecasting*, 9(4), 527-529.

Makridakis, S. & Hibon, M. (2000). The M3-Competition: Results, conclusions and implications. *International Journal of Forecasting*, 16(4), 451-476.



Rob J. Hyndman is a Professor of Statistics in the Department of Econometrics and Business Statistics at Monash University. He is an elected Fellow of the Australian Academy of Science, the Academy of Social Sciences in Australia, and the International Institute of Forecasters. Rob is the author of over 200 research papers and five books in statistical science and was Editor-in-Chief of the *International Journal of Forecasting* from 2005-2018.

rob.hyndman@monash.edu



Forecasting Impact

explores the fascinating ways forecasting is shaking up the world. Wanna talk business, economy, healthcare or education? The International Institute of Forecasters' monthly podcast hosts prominent academics and practitioners to discuss how the science and practice of forecasting is making a real impact.

New foresights await:

- Expand your forecasting field knowledge
- Learn from the best forecasting minds
- Hear real-life forecasting stories
- Add must-have books to your forecasting reading list

<https://forecasters.org/publications/forecasting-impact-podcast/>

Sparse-Proof sMAPE

SLAWEK SMYL

Symmetric MAPE (sMAPE) was introduced in Flores (1986). It is a popular point forecast metric with some good features, including being range limited. This note proposes a modification to the sMAPE metric, suitable for both large-valued and sparse time series.

While Hyndman (2025) describes several variations of sMAPE, the most familiar version was provided by Makridakis (1993). This definition of sMAPE for a single point is

$$\left[\frac{|A-F|}{(|A|+|F|)} \right] * 200$$

when $A \neq 0$ or $F \neq 0$, otherwise 0

where A is an actual (observed) value and F is the forecasted value. The metric is then averaged over all data points in the forecasted horizon.

The sMAPE is typically used for non-negative series. Hyndman pointed out problems with various alternative sMAPE definitions when applied to data with negative values. To avoid these problems, we will proceed with this assumption of non-negative values, so the formula for a single point becomes

$$\left[\frac{|A-F|}{(A+F)} \right] * 200$$

when $A > 0$ or $F > 0$, otherwise 0.

This works well for larger-valued time series, but not for sparse (“intermittent”) series where nonzero values appear sporadically and the rest of the values are zero. Let’s say that $A = 0$; then unless the forecast $F = 0$ and we invoke the special rule that in such a case the metric becomes zero, no matter how close F is to zero, the error will be 200% (Boylan & Syntetos, 2006).

PROPOSED msMAPE

To address the sMAPE issue when the actual value $A = 0$, I am proposing a new variation of the metric:

$$\text{msMAPE} = \left[\frac{|A-F|}{(\max(A,m)+F)} \right] * 200$$

where m is a small value, larger than zero. When forecasting counts, m will typically be set to one. But, as discussed below, the value used for m is situation-dependent.

EXAMPLES

With m set to one, if actual $A = 0$ and we quite correctly forecast some small number, say $F = 0.01$, the error is approximately 2%. While this is twice the MAD, it is close to zero, as we would intuitively expect. For a larger forecast, say $F = 0.5m$, error grows to 67%, and for $F = m$ it is 100%. This is again quite intuitive.

For $A \geq m$, msMAPE becomes the standard sMAPE.

Note that for $A < m$, the size of the error gets reduced compared to standard sMAPE. For example, if $m = 1$, $A = 0.5$, and $F = 1$, then for that point $\text{msMAPE} = 50\%$ and $\text{sMAPE} = 67\%$. This is useful when averaging over the many data points in a series when large values are more important, as in revenue forecasting at a retailer.

In the retail situation, 100% error on a low-revenue item is less important than 20% error on a high-revenue item. Simple averaging of sMAPEs is not helpful in this case, as less consequential errors on the many low-revenue items are weighted the same as the more consequential errors on the few high-revenue items. This leaves a distorted reflection of business reality.

It is possible to view the m as a threshold, causing reduction of error for series with smaller actuals. Thus, for $m = 100$,

- $A = 10$ and $F = 5 \rightarrow \text{msMAPE} = [5/(100+5)] * 200 \cong 10\%$ (while $\text{sMAPE} = 67\%$)
- $A = 200$ and $F = 240 \rightarrow \text{msMAPE} = [40/(200+240)] * 200 \cong 18\%$ (while $\text{sMAPE} = 18\%$)

Here, we do not need to apply any additional weighting scheme.

REFERENCES

- Boylan, J.E. & Syntetos, A.A. (2006). Accuracy and accuracy-implication metrics for intermittent demand. *Foresight*, 4, 39-42.
- Flores, B.E. (1986). A pragmatic view of accuracy measurement in forecasting. *Omega*, 14(2), 93-98.
- Hyndman, R.J. (2025). Errors on percentage errors. *Foresight*, 78, 26-28.
- Makridakis, S. (1993). Accuracy measures: Theoretical and practical concerns. *International Journal of Forecasting*, 9(4), 527-529.



Slawek Smyl works for Walmart as a Distinguished Data Scientist. He holds a MSc degree in physics from Jagiellonian University, Poland, and the MEng degree in information technology from RMIT University, Australia. He has won time series competitions, including the M4 Forecasting Competition in 2018. His main professional interest is in using neural networks for forecasting.

slawek.smyl@walmart.com

67% of SMBs

**say tariffs are
disrupting
their supply
chain plans.**

NETSTOCK



Is your forecast up to the task?

With Netstock Predictor IBP, forecast with precision – no matter the disruption.

Why Netstock Predictor IBP?

- ✓ Forecast by customer, channel, or region
- ✓ Simulate scenarios like tariff-driven cost spikes
- ✓ Balance supply, demand, and capacity
- ✓ Keep teams aligned with one plan

Scan the QR
to explore
Netstock IBP



Visit
www.netstock.com

Better forecasting. Better decisions.

Know Your Errors!

STEPHAN KOLASSA

Slawek Smyl has proposed an interesting variant on the symmetric Mean Absolute Percentage Error (sMAPE), designed to avoid an issue with the sMAPE that was first identified by Boylan and Syntetos (2006): If the actual is zero, then *any* nonzero forecast will give rise to an sMAPE of 200%, no matter how bad the forecast is. Of course, if we actually sell zero widgets, we would prefer a forecast of one widget over a forecast of a thousand – and the fact that the sMAPE does not differentiate between the two forecasts is indeed a major issue.

Smyl's proposal addresses this issue by essentially changing all zero actuals to some small number m in the denominator of the evaluation. The result is the metric msMAPE that is now sensitive to the forecast even when the actual is zero. It is still bounded between 0 and 200% like the original sMAPE – so one could actually take one half of either one to have an error metric between 0 and 100%, which people may be more comfortable with. Of course, while this bounding may be a feature to some, it could be a bug to others who believe that truly bad forecasts should lead to errors that are not bounded by some value.

However, a few points about the msMAPE should be kept in mind. For one, it loses the symmetry that was one of the main selling points of the sMAPE, i.e., that exchanging each actual with its associated forecast preserves the sMAPE. Note that this is a different kind of symmetry than treating over- and underforecasts equally, as pointed out by Goodwin and Lawton (1999). More precisely, the msMAPE is still symmetric in this sense whenever the actual is larger than the parameter m – but not if it is smaller.

Next, suppose we forecast zero. Would we ever do so? Sometimes products move very slowly, so a zero forecast might indeed make sense...and surprisingly often it is “optimal” in the sense of minimizing an error metric (see below)! If the actual is

also zero, the msMAPE is a perfect 0%, so let's assume the actual is larger than zero, more precisely larger than our parameter m . It turns out that we then always get an msMAPE of 200%, regardless of the actual. But of course a zero forecast can be bad to quite different degrees, depending on whether the actual is one or one thousand. Slow movers can suddenly turn into fast movers because of promotions or seasonality, so a high actual with a zero forecast can certainly happen. Importantly, the sMAPE shares this phenomenon of also being 200% if the forecast is zero, regardless of the size of the (nonzero) actual. So while the msMAPE at least fixes half the issues of the sMAPE with zeros (the case of a zero actual and a nonzero forecast), it still exhibits the other half (zero forecast and nonzero actual).

Another thing we need to keep in mind is that to calculate the msMAPE we must fix the parameter m . If we choose this parameter wisely, then we can indeed compare the msMAPE between different series, as Slawek points out. But what is a “good” value of m ? Ideally, we would pick an m that helps the msMAPE reward good forecasts. But now we are back to figuring out what a “good” forecast is. As always, we first need to answer this question to pick (and, in the case of the msMAPE, parameterize) an appropriate error metric, a point I made in Kolassa (2020).

Over the years, I have developed a habit: Whenever I see a new forecast accuracy metric I want to understand, I play around with simulated data to see which forecast the metric would prefer for these data, and actually did this for the sMAPE at stats.stackexchange.com/q/145490/1352. So let's do this here, with a special view to checking how important the choice of the parameter m is. For instance, **Figure 1** shows simulated actuals that follow a Poisson distribution with a parameter of 0.5, so their long-term average or expectation is 0.5, and they are quite intermittent. This latter point is important in view

Figure 1. Simulated Actuals

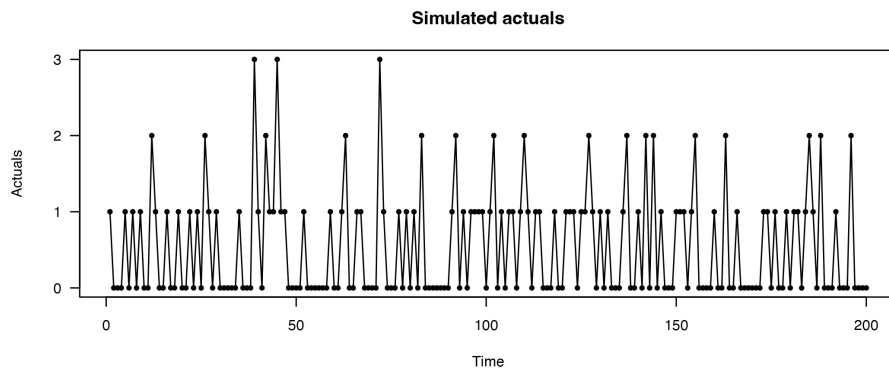
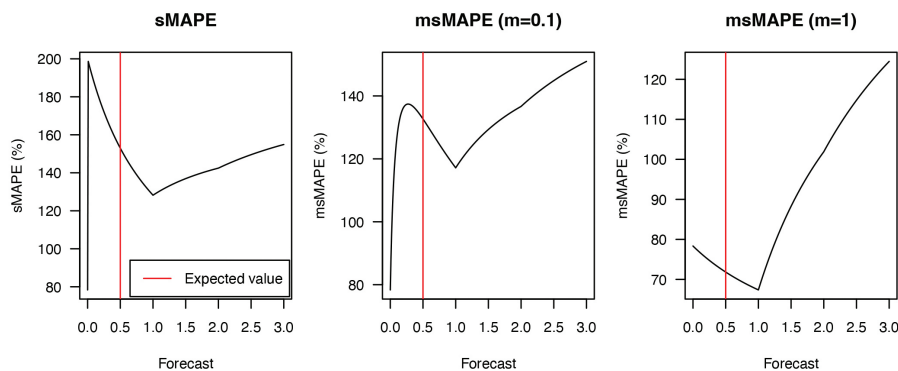


Figure 2. sMAPEs and msMAPEs for Various Forecasts



of the issues the sMAPE has with zero actuals. I simulated 10,000 such actuals, with Figure 1 showing the first 200.

Now, “reasonable” forecasts for this series might be somewhere between 0 and 3. So I next calculated the sMAPE and the msMAPE for such forecasts (in steps of 0.1) over my 10,000 simulated actuals.

Figure 2 shows the results and indicates the expected value of the actuals by a red vertical line.

We see a couple of things:

1. If we want to minimize the sMAPE, a zero forecast is “optimal,” in the sense

of giving us the smallest expected sMAPE. Of course, a flat zero forecast may not be optimal for subsequent decisions, so the sMAPE incentivizing us to output this might be an issue.

2. The choice of m for the msMAPE matters. If $m=0.1$, we are incentivized to output a zero forecast to minimize the msMAPE, but if $m=1$, we prefer a forecast of one.

3. None of these error metrics reward us for forecasts of the true expectation of 0.5. This is not really surprising, as I discussed in Kolassa (2020).

Fun fact: If our series becomes even more intermittent, with a Poisson parameter below about 0.42, the msMAPE seems to be minimized by a zero forecast for

all values of m ! This is not shown in these plots, but I would absolutely encourage you to play around with simulated data like this, which can be done in any environment, from R or Python to Microsoft Excel.

Bottom line: There is no perfect error metric. All measures have advantages and disadvantages. One should always tailor the error metric to the situation at hand, and to what kind of forecast we want to elicit. The practicing forecaster should try to understand what their error measure does, which can be especially surprising for intermittent series. Finally, simulation is one way of getting this understanding, and can easily be done even in MS Excel.

REFERENCES

- Boylan, J.E. & Syntetos, A.A. (2006). Accuracy and accuracy-implication metrics for intermittent demand. *Foresight*, 4, 39-42.
- Goodwin, P. & Lawton, R. (1999). On the asymmetry of the symmetric MAPE. *International Journal of Forecasting*, 15(4), 405-408.
- Kolassa, S. (2020). Why the “best” point forecast depends on the error or accuracy measure. *International Journal of Forecasting*, 36(1), 208-211.



Stephan Kolassa is Deputy Editor of *Foresight*, a Data Science Expert at SAP Switzerland AG, and an Honorary Researcher at the Centre for Marketing Analytics and Forecasting, Lancaster University (UK) Management School. In 2023 he was named a Fellow of the International Institute of Forecasters.

stephan.kolassa@sap.com

Explainability: A Requirement for Trust in Forecasts

TREVOR SIDERY

PREVIEW *Building trust in your forecasts is a fundamental part of a forecaster's job, and a key element of trust is to have "explainable" forecasts. Yet each business user may have a different understanding of what is meant by explainability. In this article, Trevor Siderly categorizes explainability based on types of user expectations, purposes, and trade-offs. He argues that considering the types of user expectations allows for a better up-front discussion on the requirements of a forecasting product, thus making sure that model development will be fit for purpose. Foresight editors Anne-Flore Elard and Zabiulla Mohammed follow up with commentaries.*

As forecasters, we naturally want to produce the best forecasts we can, but "best" can mean different things depending on the context. When making forecasts for business there are often other requirements that need to be considered.

One common requirement for our business colleagues – the users of the forecast – is the need to understand how the forecast was constructed. Knowing the data that was utilized (e.g., sales history, pricing, promotional activity) and how these factors impacted the forecast give users the confidence to rely on the forecast to make decisions. They will often need to defend the forecast to additional colleagues and therefore need results that are explainable. Yet the definition of "explainable" forecasts is very specific to the use case, and sometimes the individual.

must address these issues with stakeholders before starting a project.

WHAT DOES EXPLAINABILITY MEAN?

Each forecasting problem may have different explainability requirements. In some cases, there will be specific policies that guide us, but for others the business may just know that it needs to trust the forecast. To discuss the range of expectations, I have categorized the types of explainability by the following statements:

- Methods – I want to understand the methods used!
- Components – I want to understand how the forecast was constructed!
- Drivers – I need to understand how business events impact the forecast!

These requirements from business colleagues need to be taken seriously. Otherwise, we may find that our forecasts are being ignored.

For some, the proven performance of the forecast is enough for confident usage, while for others every aspect of the forecast is heavily scrutinized.

These requirements from business colleagues need to be taken seriously. Otherwise, we may find that our forecasts are being ignored. Any effort made to find the most accurate method, or set up production systems, could be wasted. We

- Errors – I need to understand what caused the forecast to deviate from reality!

Each of these versions of explainability will put constraints on the forecast methods we can choose, with those further down the list tending to add more restrictions. Typically, when data scientists discuss model explainability they are referring to the ease of showing the impact

Key Points

- *Explainability* is a requirement for building trust with the business users of a forecast.
- Different users may have different understandings of explainability, and this impacts how forecasters can model data.
- Developing a common understanding of explainability will ensure delivery of trustworthy and usable forecasts.

of data features on the resultant forecast. This aligns closely with the “Drivers” category above. When engaging with decision makers, I’ve found this isn’t always what they mean or need. Conversely, there are cases of being asked to explain the method when the business really needs to interrogate the drivers. By understanding the true needs of the business, we can make sure that we don’t unnecessarily constrain the models we use or have a model that is intractable to the questions being asked.

METHODS

Our business colleagues will often ask to understand the approach being taken so that they can get an impression of a method’s suitability. They need to trust that the forecast will behave in an expected way and won’t become unstable. To some extent, this is both the easiest definition of explainability to address as well as the least helpful. While high-level training sessions on the forecasting method can give a sense of comfort, full acceptance of the forecast will inevitably come down to seeing its live performance and observing the results. In cases where the business will have to assess the method formally – such as against written policies or even government regulations – we would still need to involve an expert. A high-level understanding of the methodology is not sufficient.

It may seem that there are very few constraints that will need to be imposed by

having to explain the methods used. But in practice, it’s clear that business colleagues need some intuitive understanding of the model to have trust in the results. While models like LSTMs, XGBoost, and other machine learning (ML) methods have proven to be good choices for modeling complex patterns and improving forecast accuracy, it may still be prudent to choose methods that don’t feel like a black box.

COMPONENTS

Sometimes the business will want a deeper understanding of how the forecast was constructed, to make an assessment of whether each of these parts makes sense. It is common for business users to already have an appreciation of trends and seasonal patterns, and thus to check if our models are behaving as expected. The review can be quite subjective, but necessary for providing confidence.

Many forecasting methods are comprised of components that are combined in either an additive or multiplicative way to make the final forecast (e.g. linear regression, ARIMAX, or Holt-Winters). Each component can be presented to the business for review alongside the full forecast. As an example, consider a method that outputs trend, seasonal, and events components. By showing how the trend both fits historical data and projects forward, the business user can consider if this agrees with their experience of recent operations and assumptions on future trends. Note that we didn’t need to say how the trend was found, just that it can be showcased. Similarly, if an event in question has an impact value attached to it, the business can be more informed if they need to make judgmental adjustments to the size of the event.

The requirement for outputting components will restrict our method choice to models that can be broken down into parts. Moreover, these components should represent business events or concepts. Talking about trends and seasonality are concepts that will be familiar to business colleagues, but something like autoregressive terms may not be. It is possible to include these more abstract

components, and when their impact on the forecast is small the business can overlook them. But if the impact is large, our colleagues are likely to want an explanation.

DRIVERS

The goal for driver-based forecasting is to clearly show how the forecast is constructed and what observable impacts have gone into the model. The business will already have a good idea of the things that affect operations and will expect to see them as components. To meet these expectations, we must show a breakdown of how the model was constructed from these business drivers. For example, if we need to make a forecast of next month's sales in a grocery store, we probably need to include the impact of weather, holidays, events, and inflation, among other factors. The impact of each needs to be shown separately and should sum up to the full forecast. In this respect it is quite similar to the "Components" form of explainability, but with much more focus on observable events. Depending on the use case, we may find that components like "trend" are not sufficiently explainable and need to relate to other observables (e.g. inflation, number of customers, etc.).

By surfacing information on what kinds of data (and their impacts) have already been included in the model, user adjustments to the forecast can be made accordingly – for example, if the model didn't include events, then adjustments for an upcoming promotion would have to be made manually. As well as ensuring trust by surfacing all this information, we are providing ways for the business to analyze our results. This version of explainability is often required when we are replacing an existing process that was itself often a simplified (or speculative) form of a driver-based forecast.

The "Drivers" form of explainability can be quite restrictive in the types of methods that can be employed. It very naturally lends itself to generalized linear models (including standard linear regression) but not much else. Alongside this, there has been a lot of discussion around

how to extract similar information from more complex ML models. With tree-based models we can extract feature importance, and for many other models we can do sensitivity analysis to understand the local correlation between a data feature and our forecast variable. These are all useful tools for the practitioner but may not be sufficient for the strictest version of explainability. However, each use case is different, and these alternative approaches may still give the end user enough confidence that the important business drivers are present in the model.

ERRORS

Why did the forecast performance not meet expectations (as measured in MAPE or other familiar metric)? This is not a question on the composition of the forecast, but about what caused any gap between the forecast and actuals. Error analysis is a useful tool for forecasters to diagnose when their method left something important out of the model. (See Johannsen on page 20 of this issue for a useful taxonomy of forecast errors.) For business users, reporting such errors pinpoints when forecast deviations cannot be explained. This lets them consider what in the business might have changed, or if there was an operational problem that can be addressed.

This retrospective analysis is a difficult problem and can end up as a project on its own. While we have access to actualized data for both inputs and outputs, we probably don't know the true impacts of any particular event. For example, if we look at ice-cream sales and see an increase in sales, we would hypothesize that this is typically from weather, promotions, and holiday impacts. We can model each impact separately, but there is no source of truth for whether our models are correct, as we only know the combined impact. We can supplement the above modeling with validating and experimentation (e.g., with A/B testing) to get a better handle on the true impact of specific events. But this will only work where we have access to counterfactuals, which is often not the case.

Having modeled our historical impacts, we must relate them to the forecast. For each driver in the retrospective analysis, we need to find the impact of making wrong assumptions when we built the forecast. The combination of all the effects of these assumptions helps us explain the total forecast error. However, we will never explain everything, as we have neither perfect models nor data – and there is always some element of randomness or noise.

We may also find we have an accuracy-explainability trade-off. Taking the ice-cream example again, if the forecast we are analyzing has a three-month horizon, it is likely that we get more accurate forecasts by combining weather, holiday, and promotions impacts into a “seasonal” component, as we don’t have weather or promotional data out that far. Conversely, if we are to fully explain the forecast error, we need to include all the impacts to provide the required level of explainability.

In terms of which models to use, nonlinearity can be a real problem. If impacts become interconnected or don’t have a consistently positive or negative correlation with the forecasted variable, then they will not seem trustworthy. Taking the weather example, it will look strange if the impact on ice-cream sales usually increases with temperature but has a couple of dips, or the uplift due to higher temperatures is adversely affected by a promotion.

IMPLEMENTATION IN PRACTICE

Having understood the possible options, we must understand our business colleagues’ requirements. I’ve found that the following types of questions can be helpful as a starting point:

- How do you interact with the current forecast?
- How do you discuss current forecasting errors with your stakeholders?
- What information do you need in order to trust the forecast?
- Does trusting the forecast help you make better decisions?

By trying to understand how decisions are made in practice, we can align our explainability options with business needs. There will be many ways in which these requirements can be implemented as well as pitfalls we may encounter, so the following are some suggestions from my experience.

METHODS

The core difficulty with helping business colleagues understand any methodology is that our processes will likely outlast a given person being in a particular role. Business stakeholders who set up a project will have provided their own expertise and have a personal investment in the solution. Handovers to new colleagues, though, can end up in “explainability drift” where new requirements are added to prove trustworthiness. By focusing on key performance metrics, clear documentation, and training, we can maintain a smooth handover and keep the forecast providing valuable decision support.

COMPONENTS

When designing a solution or system we must consider how to surface the components to the end user. Is it more appropriate to output a simple table, or is some form of user interface needed to interrogate the model? Is the user of the forecast familiar with data analytics, or do they need some form of standardized report? As with the “Methods” category of explainability, we should consider the long-term needs of the business, and not just the current interactions.

DRIVERS

When we start a new project, we always get our initial list of important business drivers from our stakeholders. While we need to test the validity of these drivers, and investigate what may have been missed, the business expertise should always be the starting point. More than that, our stakeholders will have an expectation that we can show what impact each of these drivers is having – be they internal events like promotions, or external forces like weather or national holidays.

Unfortunately, not all forecasts can be explicitly modeled by drivers, and we may still need to have a component that represents things we cannot model. While it is easy to declare that the seasonality of ice-cream sales is driven by weather, it may be impractical to create such a model that delivers the required accuracy. We may need to negotiate with the business as to what drivers are key to their trust in the model and what can be absorbed in an underlying model. For example, our colleagues may be more concerned with understanding deviations from “normal” behavior than explicitly modeling everything. Going back to the ice-cream case, understanding unexpected weather effects may be more important than modeling the whole seasonal weather pattern. We could design our model accordingly and have a seasonal component and a business driver of “unusual weather” impacts.

ERRORS

As with the driver-based model, we need to understand the expected impacts by first asking the business. Unlike the driver-based forecast, however, we will have to make sure that we are analyzing every impact, not just those that improve the forecast. As described before, we may have more ability to interrogate each event that happened in the past than was practical when making the forecast. For some variables we may be able to use control groups for events that were not global to understand their impact. Other variables may have the ability to be modeled when looking backwards, but were not viable inputs for the forecast due to poor future information. Some examples of this are unique one-off events, and those events that we have no idea when they will happen (such as a major storm that impacts commerce).

For drivers included in the forecast model, it is helpful to understand deviations of impact from our expectations. For those drivers that we didn’t explicitly include in the forecast, we will have to calculate the amount that missing information affected our forecast. This is not

a straightforward task – we will be trying to calculate how much of an event is “hidden” within the forecast model, and what is “extra.” Taking the ice-cream example, you might model the expected sales for typical weather for a particular time of year and assume that was what the forecast assumed. If you then observe unexpectedly hot weather and model its impact, you can assume the difference contributed towards the forecast error. This is only an approximation but should give an indication to the business of why the forecast didn’t match the actualized data. Another approach would be to try and put into the forecast model all drivers you need to analyze, accepting that you may have problems with overfitting and inaccuracies. Depending on the business use case and the impact on model accuracy, this may be an acceptable trade-off.

CONCLUSION

Unless our forecasts are trusted by decision makers, the accuracy or elegance of our solution can be irrelevant. Spavound and Kourentzes (2022) emphasized that trustworthiness in forecasting practice is essential. They argued for wider exposure of data science students to the practical realities of forecasting – including skills in communication with stakeholders and understanding the demands of businesses. Ultimately, a large part of building trust is to understand the full set of requirements for any solution, so that we can have up-front conversations about the trade-offs.

In recognition of this, there is often talk in the forecasting community about the importance of explainability. As discussed in this article, we should not think of explainability as a single problem, but one that is molded by the context being considered.

Extra requirements will always have costs attached, be these more complex solutions or a trade-off with accuracy. While the latter may make any of us forecasters wince, it may still be the right thing to do if it results in improved outcomes. Similarly, just because you would prefer to have the freedom to explore more techniques, you

may still improve the accuracy of the current business process with a constrained set of models. Rather than being fearful of constraining our pool of model candidates, Petropoulos et al. (2024) showed


that a reduced set of models can still perform well – while having additional benefits in cost savings. This aligns nicely with the most important priority of all: adding tangible value to the business’s decision-making process.



Trevor Sidery is a Principal Data Scientist at Tesco, leading the data science team’s forecasting capability. Before joining Tesco, he was a university researcher in the area of parameter estimation using Bayesian statistics.
Trevor.Sidery@tesco.com

REFERENCES

- Johannsen, K. (2025). Types of forecast errors and their implications. *Foresight*, 78, 20-25.
- 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.
- Spavound, S. & Kourentzes, N. (2022). Making forecasts more trustworthy. *Foresight*, 66, 21-25.




NETSTOCK

Bring your forecast to life with the BI Cube

Netstock Predictor IBP’s BI Cube lets you harness data instantly in your BI tool of choice.

No additional setup.
No steep learning curve.
Just deeper insights and faster forecasting decisions.




With the BI Cube, you can:

- ✓ Visualize data in custom dashboards
- ✓ Blend with ERP, CRM, and e-commerce data
- ✓ Share insights across teams

Scan the QR code to explore Netstock Predictor IBP

Visit www.netstock.com



Commentary: Explanations vs. Explainability

ANNE-FLORE ELARD

Trevor Sidery argues that developing a common understanding of explainability is key for delivering usable forecast. While working with hundreds of companies on the topic of demand planning for supply chains, I find that a common understanding of explainability would be better identified relative to the use of AI techniques in the consensus process.

During the consensus process, demand planners create a consensus forecast considering multiple inputs from the sales team, the marketing team, and others. These inputs are critical because the sales team knows about upcoming deals, and the marketing team knows about new products being launched. Differences between their forecasts and those of the demand planners need to be discussed and aligned, and demand planners must explain and justify the variances between their forecasts and the forecasts of those teams. All parties to the consensus process should understand what causes the differences between their forecasts. If a demand planner cannot explain their numbers, the process will be difficult to handle.

Besides the consensus process, the level of understanding of the forecast can also vary between roles. Planners and data scientists have the training to drill down into details, while executives might ask about the errors at a high level (“How much accuracy is good?”), and managers might ask about the forecasting methods used. Thus, there can be various kinds of requests to explain the forecast and validate that it’s trustworthy. The requests will be based on the various roles and backgrounds of the persons asking.

EXPLANATIONS VS. EXPLAINABILITY

The concept of explainability, however, is different than the explanations that varied audiences require to be convinced about a forecast (which methods, the

error, etc.). Explanations, as required and useful as they are, are not equivalent to the core concept of explainability.

Explainability has become a more critical and more formal concept with the introduction of artificial intelligence (AI) and machine learning (ML) techniques applied to forecasting. When machine learning techniques are applied to time series forecasting, the ML model gets retrained every time a new forecast is generated, because the statistical properties of the data distribution change as new actuals come in. In the case of a weekly forecast, the next week’s model is retrained; the model is then not the same. Imagine a tree-based ML model that is retrained weekly to incorporate the latest data. Every week a new tree is built. The planners should not compare the previous week to this week, unlike what they were used to doing with statistical forecasting. It is not the same tree. It becomes quite hard to explain why a forecast is up or down from last week when there has been a change in the underlying model.

Also, machine learning models rely on methods like Shapley values or Local Interpretable Model-agnostic Explanations (LIME) for feature-related explanations. The aggregate contribution of a feature (which could be approximated to a “driver” of the forecast) is not the direct sum of the contribution of that feature for the lower-level forecast. For instance, take 10 products that make up a brand, with a forecast for each product. Each of these forecasts has a promotion feature with a contribution calculated through Shapley values. These 10 promotion feature contributions do not naturally aggregate to the contribution of the promotion feature at the brand level. This requires normalization to aggregate forecasts across different products of the same category.

Furthermore, many users have been accustomed to driver-based forecasting with clearly identified business-driven building blocks. Not only can these blocks be summed up, but they can also be used to explain what pushed a forecast up or down from week to week. The learning curve can hence be steep when transitioning from a statistical forecast to machine learning techniques, as this involves moving from business drivers to ML features. And while neural networks could be used in industries where trend changes are common and critical to incorporate, they are even more challenging to explain.

All these challenges have led to the crystallization of the concept of explainability – which has become a key requirement

in most organizational forecasting today when using AI. AI applied to forecasting techniques can create the dreaded black box, where no one knows what goes on inside. In response to this, researchers and practitioners have been striving to make AI forecasts explainable. These efforts have led to new approaches referred to above, such as SHAP (SHapley Additive exPlanations) values – a method for explaining the output of a machine learning model by quantifying the contribution of each feature to the model's prediction – and LIME. There is also a new research area called “Explainable AI” or “XAI.”

As a practitioner, I have seen explainability become a critical concern. With the rise of the AI/ML techniques, planners who have been used to “building blocks” and “drivers” now try to map business understanding to “features” (the technical components of AI/ML).

Practitioners now face a paradox. Organizations hunger to utilize best-in-class ML and deep learning techniques, and to benefit from the adaptability of such models. But these models lack a direct mapping with explainable business drivers, creating a roadblock to their trust and adoption. That is where the core definition of explainability in practice resides: as a requirement for building trust with end users.



Anne-Flore Elard is a Practitioner in Data Science and AI/ML, with an MBA from MIT Sloan. Her thesis focused on collective intelligence frameworks, after which she developed applications leveraging text similarity measurement. In her role at Kinaxis, Anne-Flore has created AI service offerings for multiple companies. Since 2018, she has deepened her expertise in forecasting by utilizing machine learning techniques and is currently leading AI Product Management at Kinaxis. Anne-Flore is *Foresight* Column Editor for Machine Learning and AI.

anneflore.elard@gmail.com

Commentary: Building Trust through Explainability

ZABIULLA MOHAMMED

Forecasting plays a crucial role in helping businesses make informed decisions about inventory, staffing, supply chain management, and daily operations. Trevor Sidery's paper highlights key aspects of forecasting, and this commentary expands on those ideas – agreeing with some points, challenging others, and offering additional insights based on my hands-on experience in retail forecasting.

WHY EXPLAINABILITY MATTERS

Explainability is essential for forecasts to be trusted and used effectively. Even the most advanced forecasting models are meaningless if decision makers don't have confidence in them. Leaders across different departments – such as executives, supply chain managers, and store operators – need to know how forecasts

to overstate the limitations this creates. Some argue that driver-based forecasts fit best with traditional approaches like generalized linear models (GLMs), but modern machine learning techniques offer even more flexibility. Methods like SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations) make it possible to extract insights from complex models, including tree-based algorithms and neural networks.

The idea that businesses will reject models simply because they are harder to interpret isn't entirely accurate. Instead, the key is making sure models provide enough transparency to satisfy business needs while still delivering strong predictive performance.

The idea that businesses will reject models simply because they are harder to interpret isn't entirely accurate. Instead, the key is making sure models provide enough transparency to satisfy business needs while still delivering strong predictive performance.

are built so they can rely on them for strategic and operational decisions.

Sidery rightly points out that different people have different expectations when it comes to explainability. Some may want to understand the methods used, while others care about the factors driving the forecast or the reasons behind forecast errors. For example, merchandising teams may focus on how demand growth aligns with business goals, while supply chain teams want to know how forecasts affect day-to-day operations. Sidery's classification of explainability into methods, components, drivers, and errors helps structure these discussions effectively.

RETHINKING CONSTRAINTS ON MODEL SELECTION

While aligning model selection with explainability is important, it's possible

EMBRACING BLACK-BOX MODELS IN RETAIL FORECASTING

Sidery suggests that organizations should avoid black-box models, but that perspective doesn't fully reflect how businesses are evolving. Machine learning models like XGBoost and LightGBM are widely accepted in retail forecasting because they offer strong predictive power. Many businesses are willing to adopt these models as long as they consistently deliver reliable results. The focus shouldn't be on avoiding black box models altogether, but rather on enhancing them with tools that explain their outputs – such as “feature importance” analysis or SHAP values. Trust is increasingly built on real-world performance rather than just how easily a model can be explained.

LEARNING FROM FORECAST ERRORS

Sidery brings up an important point about analyzing forecast errors, but this could be explored even further. Understanding why a forecast was off isn't just about explaining past mistakes – it's about making future predictions more accurate. Businesses that continuously refine their models based on errors can improve forecasting over time.

Additionally, modern forecasting tools now allow for real-time adjustments. Some retailers use adaptive learning techniques where models self-correct based on recent performance. For instance, an AI-driven grocery demand forecasting system could adjust its predictions if unexpected weather conditions, supply shortages, or promotional events impact sales.

BALANCING ACCURACY VS. EXPLAINABILITY

Sidery briefly mentions the trade-off between accuracy and explainability, but this is a major challenge in forecasting. Advanced models, particularly deep learning techniques, can offer improved accuracy in certain scenarios, though they often pose challenges in terms of interpretability. In such scenarios, simpler models like exponential smoothing

may be easier to explain but deliver less accuracy.

Retailers need to find the right balance based on their needs. For short-term forecasting – such as predicting sales for fresh food – transparency may be more important than absolute precision. Small buffers can be maintained to account for any uncertainties. However, for long-term strategic forecasts where accuracy has a significant financial impact, businesses may be more open to using black-box models if they have a tested and proven record of success.

BRINGING BUSINESS EXPERTISE INTO THE EQUATION

One aspect missing from the discussion is the role of business expertise in shaping forecasting models. While explainability frameworks are useful, data scientists must also integrate domain knowledge into their forecasts. The best models aren't just driven by historical data – they also account for real-world factors like promotions, supply chain disruptions, and competitor behavior.

For example, a retail forecast that ignores major industry trends or shifts in consumer behavior won't be useful in practice. Business stakeholders can provide valuable insights that help refine forecasting models and make them more actionable.

FINAL THOUGHTS

Sidery's article provides a strong foundation for understanding explainability in retail forecasting, but it could benefit from a more practical perspective on model flexibility, continuous improvement, and business integration. The key takeaway is that explainability is important, but it shouldn't come at the expense of predictive power or business value. By balancing technical sophistication with business usability, organizations can develop forecasting solutions that are not only effective but trusted and used by decision makers.



Zabiulla Mohammed is a Director of Data Science at Walmart, where he leads a team in implementing forecasting, machine learning, and experimentation projects. With over 15 years of experience, he has utilized data science techniques to drive business growth in retail, eCommerce, supply chain and transportation, product development, opera-

tions, business strategy, marketing, and customer analytics. Zabi has also worked for such companies as Sam's Club, Honeywell, and IBM. He holds a master's degree in management information systems and a bachelor's degree in computer science and engineering. Zabi is *Foresight* Column Editor for Retail and CPG.

mohammedzabiulla@gmail.com

SPECIAL FEATURE:

UNITED NATIONS SUSTAINABLE DEVELOPMENT GOALS

PREVIEW *In 2015 the United Nations published its 2030 Agenda for Sustainable Development. This document established 17 goals for societal advancement, representing a vision then shared by all member states. In Foresight issue 74, Bahman Rostami-Tabar and I solicited commentaries on how forecasting may inform SDG-related decisions. We sought to increase awareness and enhance the impact of forecasting on these critical goals that have far-reaching consequences for society and the environment. In the following, Lauren Davis and Leo Sadovy discuss forecasting's role in two of the goals, #2 Zero Hunger and #14 Life Below Water.*

The Role of Forecasting in Ending Global Hunger

LAUREN DAVIS

Global hunger is a persistent problem, with approximately 2.33 billion people worldwide unable to consistently access nutritious food to live an active and healthy life (United Nations, 2024). The term *food insecurity* is commonly used to describe this condition. Factors contributing to the growing food-insecure population are multifaceted and encompass challenges in two broad areas when examined through the lens of supply chain management: (1) food availability, which captures the supply-side obstacles associated with insufficient food production to meet the need; and (2) food accessibility, which highlights the demand-side issues reflecting situations where people who need the food cannot access it.

Food availability is impacted by climate, agricultural practices, and supply chain disruptions (e.g., political conflict, disasters, pandemics) that can reduce quantities or temporarily stop the production and movement of food. Food accessibility challenges limit one's ability to obtain food and are influenced by factors related to socioeconomic status (e.g., income levels), geographic placement, and physical infrastructure (e.g., transportation pathways and cold storage). Physical infrastructure and geographic barriers isolate

people from food markets while making it difficult to properly transport and store nutritious food for people to access. Lack of available and accessible food can cause high food prices, limiting the purchasing power of vulnerable populations and in turn contributing to malnutrition and poor health.

The 2024 UN Sustainable Development Goals report calls for the development of resilient, sustainable, and equitable food systems to achieve the global hunger, nutrition, and sustainability outcomes associated with SDG2. The forecasting community can play a key role in helping to achieve these objectives, as many of the barriers to success are intertwined with the design and management of supply chain systems.

Forecasting has a rich history in supply chain management, particularly in demand prediction in the for-profit sector where forecasts help to inform decisions on the production, storage, and distribution of supply to meet demand. Recently, forecasting techniques are being used to tackle problems in humanitarian areas with respect to acquisition and distribution of relief items (Altay & Narayanan, 2022). However, the literature in this area is still quite sparse. Expanding upon

the research in the humanitarian relief domain can help move society closer to achieving zero hunger. As key supply chain stakeholders adopt more sophisticated methods of data collection and integration, forecasters can leverage this data to develop predictions that influence changes in food policy. A brief discussion of how the forecast community might consider their potential impact is presented below.

Forecasting food availability on the farm, at retail markets, and in nonprofit food-assistance organizations can help drive sustainable and equitable food systems. For example, in the context of food aid, forecasting can be used to predict the quantity of food available for rescue. This informs decisions on the timing of food rescue activities, resulting in reduced food waste. Additionally, advanced information about potential supply shortages can influence the types of interventions selected to mitigate the impacts of these events. Such interventions may include finding alternative sources of food supply, soliciting or acquiring additional financial support to address the root causes of the supply shortage, or providing a means to enhance purchasing power (in food-aid supply chains) when donated food sources are not available. Forecasts about food availability can also inform interventions that leverage local resources, particularly during times of global supply disruptions, thus empowering small-scale producers and farmers to be used more effectively. These types of interventions help to promote supply chain resilience by incorporating redundancy into the network. Predicting food availability by quantity, food type, and location can assist with better matching of available food with food need in an equitable manner. This is extremely important for the most perishable food items. Lastly, food supply is a highly uncertain commodity affected by climate, agricultural practices, and human behavior. Therefore, integrating some of these external factors into forecasting models is also a fruitful area of investigation.

Forecasting food need is also an important problem to consider, as it enables effective matching of supply with demand. Predicting demand for food is not trivial, particularly when considering geographic and demographic factors that influence food preferences. Ignoring dietary and cultural preferences could result in edible food being discarded as the right food is not reaching the right location at the right time. Forecasters can draw from the rich literature on demand prediction in for-profit settings to understand how to develop models of global food demand. Specifically, how can we create better tools to estimate true need, at scale, which is population- and location-specific? Furthermore, how can we connect the forecast models to decision support systems and make them interpretable for decision makers? Answering these two questions can spur adoption of developed forecast models by practitioners.

In addition to food availability, forecasting resource availability can help limit the impact of supply shortages. Human resources (volunteer and paid labor) are essential for smooth production and distribution of food. COVID-19 highlighted the impact of human resources on food availability as worker absences limited food production, causing stockouts of essential items in downstream retail markets. Additionally, driver shortages contributed to transportation delays between key points of distribution within the supply chain. Resource shortages impact both food availability and accessibility. Insufficient human and material capacity constrains the food system, making it less resilient to sudden shocks.

Increasing the resilience of food systems can also be furthered by predicting areas of vulnerability linked to climatological or human-caused events. Linking these location predictions with forecasts of potentially displaced populations can help provide a better understanding of the spatial demand for food.

This topic list serves as a starting point for the conversation on the role of forecasting

in achieving the UN SDGs. Problems related to forecasting food supply, demand, and resource availability are complex and require integrating data from multiple sources. In the presence of limited or unavailable data, methods for developing forecasts in these data-sparse environments are also needed. Achieving SDG2 requires using forecasts to help increase the available food production, match the produced food with food demand, and distributing the food to the right locations at the right time utilizing human and material resources effectively. These activities must be conducted to minimize food waste, ensuring fairness when distributing to food-insecure populations. This also means leveraging the existing infrastructure (small-scale producers) and large-scale global food producers so that, when disruptions occur, there are redundancies in place to ensure access is not disrupted for extended periods of time.

To connect forecasting research with practice, forecasting models need to be integrated with decision models to enable interpretable scenario analysis. Researchers should work collaboratively with key stakeholders to ensure predictions align with the business-use case. This will ensure that, as a research community,

we are not predicting something that is interesting academically while not being useful to real-world situations.

REFERENCES

Altay, N. & Narayanan, A. (2022). Forecasting in humanitarian operations: Literature review and research needs. *International Journal of Forecasting*, 38(3), 1234-1244.

United Nations (2024). "The Sustainable Development Goals Report, 2024" retrieved from unstats.un.org/sdgs/report/2024/The-Sustainable-Development-Goals-Report-2024.pdf

Lauren Davis received her BS in computational mathematics from Rochester Institute of Technology, her MS in industrial and management engineering from Rensselaer Polytechnic Institute, and her PhD in industrial and systems engineering from North Carolina State University. Prior to joining the faculty at North Carolina A&T, she spent 12 years as a senior software engineer at IBM supporting SAP implementations in the U.S., Mexico, and China. Her research focuses on stochastic



modeling of supply chain systems in for-profit and nonprofit settings. She currently serves as a board member of the Second Harvest Food Bank of Northwest North Carolina. Lauren is a member of the *Foresight* Advisory Board and was recently elected to the IIF Board of Directors.

lbDavis@ncat.edu

Life Below Water

LEO SADOVY

“No water, no life. No blue, no green.” — Dr. Sylvia Earle

While it would be difficult to single out any one of the UN’s Sustainable Development Goals (SDGs) as being the most important, “Life Below Water” would make a good contender for the top spot. Seventy percent of the planet is ocean – why we called it “Earth” rather than “Water” was certainly a missed opportunity. True, thanks to woody trees, 80% of the planet’s biomass is on land. On the other hand, there is 10 times more carbon stored in the oceans than on land and the atmosphere combined, including 35% of all human-emitted carbon dioxide.

At the highest overall ranking, the oceans are involved in some of the most dramatic aspects of climate change. Melting glaciers and warming temperatures are causing sea levels to rise. Sea-level rise is inundating coastal communities and estuaries, and causing salt water intrusion into coastal aquifers. Carbon dioxide absorption is causing acidification – global warming’s evil twin – which in turn negatively impacts coral reefs and all calciferous shell-building creatures such as mollusks, crustaceans, and plankton. Rising sea surface temperatures affect the production and strength of cyclones and the monsoonal rain patterns. And perhaps most consequential of all, sea surface warming combined with changes to temperature and salinity from glacial meltwaters runs the risk of disrupting the global conveyor belt known as the Atlantic Meridional Overturning Circulation, shutting down the Gulf Stream with a potentially disastrous effect on global weather patterns.

Even excluding global warming, life below water is already threatened in numerous other ways by human activity. The oceans’ fisheries are depleted beyond recognition

compared to a century ago, in terms of population sizes and the near elimination of the larger individuals of each species. Marine mammals that have not yet been hunted to near extinction face threats from shipping and oil exploration. Agricultural nutrient runoff causes rampant algal blooms, which turn deadly when these blooms deplete the oxygen (hypoxia/anoxia) other marine species depend on for respiration. Dredging, bottom trawling, and seafloor mining destroy habitats and disrupt local ecosystems. Macro-plastics from fishing and human waste kill marine life by entanglement or ingestion. Microplastics work their way back up the food chain and onto our grocery shelves. Coastal development destroys critical wetlands – the most productive of all ecosystems, marine or otherwise.

If you’re wondering if there were some aspect of sustainable development of the oceans where you could apply your analytic and forecasting skills, you can take your pick from a hundred different problems or opportunities. The ocean needs you. More specifically, here’s a summary of a range of forecasting applications in the marine domain:

- Fish stock assessments and seafood demand/consumption
- Marine pollution tracking, such as oil spills, plastic waste, and chemical contaminants
- Estimates of the spread and concentration of harmful algal blooms or hypoxic zones (dead zones) caused by nutrient pollution
- Biodiversity conservation, including changes and threats to habitats and population dynamics

- Coastal erosion and flooding
- Invasive species spread and control
- Valuation of ecosystem services, such as nutrient cycling, seafood, and storm-surge protection
- Predicting pH levels and population trends for affected species
- Evaluating the effectiveness of various mitigation strategies
- Predicting frequency, concentrations, and dispersal of point and nonpoint pollution sources and flows
- Predicting the various inputs to the Representative Concentration Pathway (RCP) scenarios that drive most climate-change models
- Predictive models to provide early warnings of impending thresholds, tipping points, and ecological collapse
- Simulation of various management scenarios for addressing any of the above problem areas

To be effective in this endeavor, you are going to need to attach and embed yourself in a larger team or project that will be operating within a comprehensive systems model of the issue being studied. I discuss this in my recent op-ed, “Systems Thinking to Address Sustainability” (Sadovy, 2025). Your primary role as a forecasting expert will involve predictions for the various inputs and components of the overarching systems model, such as sources, sinks, flows, and delays. An accurate and useful outcome depends both on the quality of that overall model – the responsibility of the principal investigator – and the quality of all the inputs to the model, which is where you come in.

While your work might be in the assistance of scenario generation, it could also find application in assessing economic impact, guiding mitigation strategies, or setting the stage for policy debate and discussions. Or you might be involved with early warning systems, identifying vulnerable ecosystems, or planning for extreme events.

By now it might also have occurred to you that a knowledge of geospatial analytics,

remote sensing, and satellite data will likely come in handy. In addition to whatever domain knowledge you acquire (i.e. fisheries, plastics, nutrients, etc.), the data under analysis will often have both temporal and spatial components.

My own studies in this discipline have ranged across multiple areas: climate refugees displaced by rising sea levels, evaluating the sustainability of fish meal as a protein source for cattle and hogs, the interaction of the Haber-Bosch-driven nitrogen cycle’s joint impact on atmospheric CO₂ and marine nutrient runoff, and a vulnerability assessment of Jakarta as the sea rises and the land subsides. While these were merely secondary research efforts, they all depended on primary sources where forecasts of global warming, dredging impacts, alternative protein sources, wetland eutrophication, and population relocation costs were key assumptions and drivers of the resulting assessment and recommendations. In addition to the list of the larger institutions addressing sustainability in my op-ed, if you want to get plugged into marine-specific organizations, I can suggest the Ocean Conservancy, the Coral Reef Alliance, the Pew Charitable Trusts – Ocean Conservation, or NOAA-Fisheries. There are dozens more like them once you begin your search. And the sooner you start, the better: the ocean needs you.

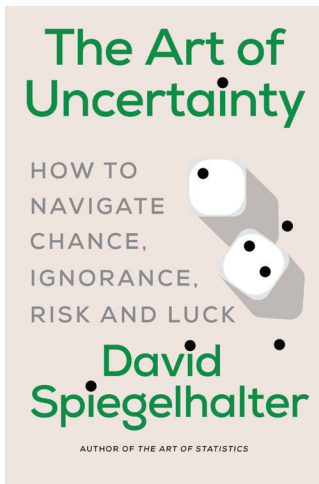
REFERENCES

Sadovy, L. (2025). Systems thinking to address sustainability. *Foresight*, 77, 46-47.



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 adaptation, climate refugees, and water management / coastal and marine concerns.

leosadovy@gmail.com



The Art of Uncertainty – How to Navigate Chance, Ignorance, Risk and Luck by David Spiegelhalter

REVIEWED BY IRA SOHN

For those still deciding on their summer beach read, David Spiegelhalter's *The Art of Uncertainty* will not disappoint. This book is devoted to data collection, data analysis, statistics, and probability – all raw materials for forecasters, and all normally indicative of a dull and difficult tome. But Spiegelhalter's newest offering is none of that. *The Art of Uncertainty* is engaging, entertaining, and, after getting past some of the more complicated mathematical material, an easy read.

The book is firstly a primer on statistics and probability. It includes a brief history of the field and is appended with a glossary of the subject's technical terms and major tools. Short biographical sketches of the central and often colorful personalities who developed the core principles and “products” of the discipline are embedded in the chapters. They include, among many others in an eclectic list, Reverend Thomas Bayes, members of the Bernoulli family, Alan Turing, Richard Feynman, and Daniel Ellsberg of the infamous 1971 Pentagon Papers report.

Secondly, the book is a memoir and chronicle, comprising serious and entertaining work-related encounters with data analysis and probability theory. Key elements of risk and chance in insurance and lotteries are discussed, as are everyday coincidences and luck – things seen by some as acts of the supernatural. Throughout the book, Spiegelhalter reflects on timeless existential questions that involve life and the future, employing statistics as the focal point.

Almost every chapter discusses errors and lesser gaffes – ranging from alarming to amusing – that the author and his colleagues have encountered over the years. These mistakes often begin with data collection and analysis. They continue through questionable or flawed assumptions in statistical testing and modeling and culminate in misinterpretation of results. Faulty forecasts and/or policy errors often ensue from these errors of judgment.

As one example of such an error, in the United Kingdom in June 2021 it was reported that “the majority of people dying from COVID-19 had been fully vaccinated” (p. 187). This statistic could imply that the vaccines were grievously ineffective or even harmful, and predictions about the makeup of people dying from COVID-19 using this data would be wrong. Analyzing the data more carefully revealed that, at the time, the earliest recipients of the vaccine were the elderly and clinically most vulnerable. It is important to recall that the vaccine is not 100% effective at preventing death from COVID-19. Thus, “if enough people get vaccinated, the ‘breakthrough’ deaths will outnumber the deaths in the unvaccinated group” – even though the latter group is at higher risk.

ABOUT THE AUTHOR

Sir David Spiegelhalter is an emeritus professor of statistics at the University of Cambridge (UK) and is considered one of Britain's most eminent statisticians. His earlier book, *The Art of Statistics* (2019), achieved bestseller status – surprisingly,

given the leaden nature of its subject matter. Due to his exceptional ability to communicate complicated statistical information to the general public, Spiegelhalter was made a non-executive Director of the United Kingdom Statistical Authority during the COVID-19 pandemic.

Throughout the pandemic, he often appeared at briefings to interpret the latest statistical data to the British public on the course of the virus and its consequences. As a result of his clear and convincing communication and analysis he acquired “national treasure status” in the words of one commentator (Seagull, 2024). Since this newest book was written following the worst of the pandemic, Spiegelhalter was well aware of the need to collect, present, and analyze data carefully, lest erroneous conclusions result in tragic policy decisions.

A worthwhile introduction to the wisdom and enthusiasm that Spiegelhalter imparts can be found in a 2020 conversation celebrating his receipt of the Royal Society’s Michael Faraday Prize ([youtube.com/watch?v=JW9pIVfanjo](https://www.youtube.com/watch?v=JW9pIVfanjo)). This award is bestowed annually on an outstanding scientist who also has an exceptional ability to communicate scientific ideas to the general public in “a clear and engaging way.”

WIELDING TOOLS OF THE TRADE

The Art of Uncertainty begins with a few philosophical assertions about the omnipresence of uncertainty in daily life: “our very existence depends on a fragile chain of unforeseeable events”; “we all have to live with uncertainty”; “we may prefer to ignore uncertainty, but it would be better to acknowledge it” (p. 15). Throughout the book the author provides examples of the many “derivatives” of uncertainty that we encounter day to day at work and at home, such as risk, chance, future predictions, surprises, coincidences, and luck.

He emphasizes how the tools of his trade – statistics and probability – can be deployed to assess vaccine safety. In hindsight we know there was much ignorance

and uncertainty surrounding the early months of the pandemic. The spread, vulnerability, severity, and mortality from the illness, along with the initial responses by public health officials, were all subject to unknowns. Consequently, statisticians were compelled to refine their probability assessments about the disease as new data were assembled and analyzed. Bayes’ theorem was enlisted to provide updated judgments (with revised probabilities) regarding the transmission of the virus and the likely success of alternative treatment mechanisms.

For non-statisticians, Prof. Spiegelhalter provides a very down-to-earth explanation of Bayes’ theorem. He notes that “it can be considered as a basis for learning from experience” (p. 188), and that “it underlies what happens when humans react to new information” (p. 189). In short, employing Bayes’ theorem allows statisticians to refine their probability judgments about an event as new evidence is introduced, with the expectation that improved estimates will result. Readers interested in refreshing their understanding of the theorem and its applications to problems in criminal cases, political polling, artificial intelligence, medical research, and self-driving cars can listen to a recent podcast with Spiegelhalter (podcasts.apple.com/us/podcast/the-history-of-revolutionary-ideas-the-bayesian/id1682047968?i=1000699344442).

The incorporation of new data to revise algorithms is the essence of machine learning, a branch of artificial intelligence (AI). Updating “prior probabilities” using Bayesian analysis with newly acquired data unlocks unknown or previously hidden patterns of correlation and/or causality. These techniques play a major role in improving the prospects for more accurate forecasts. Considering the proliferation of AI in almost all facets of modern life, it would not be an exaggeration to declare that it is only a matter of time until we are all Bayesians (Joiner, 2025).

Spiegelhalter has a singular command of the technicalities of his discipline, along with a special talent for communication that exudes confidence and trust. He is a

passionate advocate of enlisting common sense and practicing humility, especially when the data, statistical tests, probability assessments, and models yield nonsensical results. He is comfortable with saying, in effect, “Based on the information I have, and the assumptions employed, I cannot explain the results obtained. Therefore, we must look elsewhere.” Common sense needs to be pressed into action when explaining a “tail” probability such as “perfect storm” events, which are often confused with “black swan” events: “*Perfect storms* are an extreme version of a familiar event in the far tails of the distribution... while *black swans* are qualitatively different types of events that had not even been thought of” (p. 356).

Spiegelhalter’s traits are on full display in the amusing example of the “case of the double-yolked egg box” (p. 125). Someone bought a box of six eggs and found all were double-yolked. An inquisitive statistician would immediately ask, “What is the probability of this happening?” According to industry experts, only one in 1,000 eggs are double-yolked. Therefore, the probability of buying a box of six double-yolked eggs is practically zero. Based on the number of six-egg boxes sold per year in the UK, this event would be expected to occur once every 500,000 years!

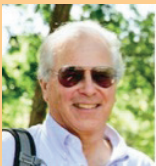
But since this event has happened, the claimed probability might appear to be wrong. Except the math is correct! Spiegelhalter then suspected the problem was with the assumption that the eggs in the box are independent of each other. He went to his neighborhood shop and purchased another box of eggs, clearly labeled “double-yolked eggs.” All were indeed double-yolked. His conclusion: the original box of eggs containing double-yolked eggs was mistakenly packaged in an ordinary egg box, instead of a box labeled “double-yolked.” QED!

The Art of Uncertainty is chock-full of serious and amusing examples of other commonplace errors such as conflating correlation with causality. Chapters 10 and 11 are replete with recent applications of statistics and probability in developing forecasts in the legal and judicial system, sports (in the United States, the NFL’s Next Gen Stats comes to mind), weather and climate predictions, political polling and campaign funding, detecting bank and credit card fraud and, not least, in the military, all of which are rapidly becoming AI-intensive.

For a productive, instructive, and enjoyable vacation experience, *The Art of Uncertainty* and the beach await *Foresight* readers this summer.

REFERENCES

- Joiner, S. (2025). At work, a quiet AI revolution is under way. *Financial Times*, February 11.
- Seagull, B. (2024). As luck would have it. *Financial Times*, October 15.
- Spiegelhalter, D. (2019). *The Art of Statistics: Learning from Data*, Pelican.
- Spiegelhalter, D. (2025). *The Art of Uncertainty: How to Navigate Chance, Ignorance, Risk and Luck*, W.W. Norton.



Ira Sohn is Emeritus Professor of Economics and Finance, Montclair State University, and *Foresight* Column Editor for Long-Range Forecasting.
imsfinc@gmail.com

Overcategorization of Continuous Data

MALTE TICHY

Forecastable/unforecastable. High-margin/low-margin. Fast-seller/slow-seller. We love categories, especially binary ones that represent two clearly distinct, dichotomous cases. Discrete categories simplify our thinking and ease our decisions.

Accept vs. reject: we put our signature below a job offer, or not. This candidate or that one: we make exactly one cross on the ballot. In or out: a product is listed, or it is not. When we make binary decisions, dichotomous categories are a necessity.

When analyzing data in general and forecasts in particular, introducing the right categories and splitting data along these categories can reveal important structures and relations. It can make sense to carve out distinct groups from continuous data – such as when a histogram identifies several well-separated peaks. This could indicate an underlying structure that justifies drawing a boundary.

But not everything that resembles a well-defined category is one. How can we unambiguously distinguish a village from a town, or a town from a city? These questions should not be dismissed as philosophical distractions, because the practical impact of the misuse of too coarse-grained categories is severe. Parsing a continuous selling rate distribution may not be justified by the data, but is often done anyway, and in an arbitrary way. Such categorization can lead to sub-optimal decisions, thereby jeopardizing business value.

I have often witnessed exaggerated, unnecessary, and harmful categorizations of forecasting datasets, especially of selling rates. For example, a large assortment of diverse products is binarized into “slow-moving” and “fast-moving” items. Two totally different models are then applied.

But what makes a product a slow-mover? Often I see arbitrary but “convenient” definitions, such as that the product sells less than five times a day on average. But why is five important, rather than four or six or some other value? Insisting on having only two categories of slow- and fast-movers makes us treat similar cases very differently, and quite distinct cases similarly. The logic that applies to the super-slow-mover that sells once every few weeks is the same as for the one whose velocity is 4.9 per day (which almost promotes it to the “fast-moving” group). On the other hand, this product selling 4.9 times a day and the one with 5.1 daily sales are treated differently. It is unlikely that the essence of what makes one model more suitable for one or the other group can be captured by a single variable being below or above a certain threshold.

I’m not saying “slow” and “fast” are never reasonable classifications within a dataset. But the distributions of selling rates, forecast accuracies, and margins are often continuous, without well-separated peaks. There is often no clear and nonarbitrary boundary between “slow” and “fast,” “low” and “high,” “weak” and “strong.” A binary picture of black-and-white pixels lacks depth: it is only a shallow representation of what a greyscale – yet alone a full-color picture – could achieve. Similarly, enforcing binary categories hides the richness of the spectrum. Interesting nonlinear relationships between the selling rate and some other quantity (e.g., forecast accuracy) inevitably get lost when such binary categorization is applied.

As a rule of thumb, the results of a data analysis should not depend on the definition of the categories that are used to perform it. The reliability of statistical analyses can be jeopardized by “p-value

hacking” – where many different hypotheses are tested until one happens to be statistically significant. This risks interpreting mere chance as true effect. In close analogy, I have witnessed “category-hacking.” Suppose you want to prove that slow-movers differ from fast-movers in forecast quality. Even if no such effect is truly present in the data, scanning the range from 0.1 sale/day to 10 sales/day as the slow-mover/fast-mover boundary could likely yield the desired result of showing a significant difference. Yet other values for the boundary may show no difference between the groups. One should cultivate at least a moderate degree of skepticism against statements asserting fundamental differences between one group and the other, when that grouping is based on a continuous variable.

For some decisions, it is unavoidable to eventually categorize the products into two distinct classes. However, I would always challenge the supposed binary

nature of any business decision. “Select all the items that will be marked down by 50% of their current price!” Why apply such coarse-grained treatment in the first place? Why not apply 10%, 20%, 30% markdowns to reflect the individual situation of each item? Your business will run much better when you do not tie yourself to artificial and false binary divisions. Treating a continuous spectrum of cases with only two strategies is clearly suboptimal.

Some choices are, of course, truly binary. Still, it is not necessary to binarize the quantities that inform that decision! For example, when you consider which products to list next season, there is little value in first applying binary labels (slow-seller/fast-seller, high-margin/low-margin) and then making decisions. The raw continuous quantities can better support a well-informed decision that considers all facets of the data.

Prematurely classifying continuous quantities into two or few categories is too often a lazy shortcut for more convenient data analysis or business policy. Careless categorizations can impact the quality of the analysis, by missing the fine-grained and differentiated aspects that could have been exploited. Setting up classifications must be done with care, taking advantage of established data mining and statistical techniques, and should be challenged continuously. Appropriate classifications can add business value by providing meaningful differentiations for analysis and decision making. Arbitrary classifications likely will not.



Malte Tichy has a research background in theoretical quantum physics, with a PhD from the University of Freiburg. He learned the nuts and bolts of applied data science and forecasting within various hands-on and leadership roles at the supply chain software company Blue Yonder. As a Senior Key Expert in Data Analytics & AI, he works on forecasts for wind-turbine component reliability and maintenance expenditures at Siemens Gamesa Renewable Energy.

mc.tichy@gmail.com

So you want to write for *Foresight*?

Foresight seeks submissions from forecasting professionals in all areas of business and public service, and from academic researchers.

Find manuscript preparation and submission details at
[**forecasters.org/foresight/submit-article/**](https://forecasters.org/foresight/submit-article/)

Advanced Forecasting, Integrated S&OP, SCM Optimization



Drive supply chain excellence with Valtitude!

- **Strategic Advisory:** A proven leader in demand forecasting, S&OP, and supply chain optimization, enhancing forecast accuracy and service levels.
- **Tailored Solutions:** Designed to address your specific pain points.
- **Data-Driven Insights:** Use real-time data and analytics for proactive decision-making.
- **End-to-End Support:** From strategy to execution, we ensure seamless integration and continuous improvement.

Contact us today for
a complimentary
Quick Diagnostic!

PLANVIDA
Planning Meets Analytics

SAP IBP

📍 USA | UK | South East Asia
☎ +1(781)995-0685

✉ valtitude@valuechainplanning.com
🌐 www.valuechainplanning.com



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