

INTERNATIONAL JOURNAL OF BUSINESS AND ECONOMICS INSIGHTS

Volume: 15; *Issue:* 9
Pages: 01–41
Accepted: 02 September 2025
Published: 10 September 2025

AI-Driven Predictive Analytics for Intermittent Demand Forecasting in Wholesale Distribution: A Cost-Based Benchmarking Study

Md Hasanujamman Bari[1]

[1]. *Business Data Analyst, Price & Co., Texas, USA;*
Email: hasanujamman.bari@gmail.com

Doi: 10.63125/kncce104
This work was peer-reviewed under the editorial responsibility of the IJBEI, 2025

Abstract

Wholesale distributors carry tens of thousands of stock-keeping units whose demand is intermittent: most periods record no sales and the remainder record quantities that vary widely. Forecasting such demand decides stock investment and service, yet the literature judges forecasting methods mostly by accuracy statistics that are known to misrank methods for inventory purposes. This systematic review synthesises the literature published before 2020 on classical, artificial-intelligence (AI) and machine-learning (ML) methods for intermittent demand as a literature-based benchmark organised by the type of evaluation used. A reproducible search of the Crossref index with 28 strings returned 5,500 records (2,946 unique); 2,288 titles and 379 abstracts were screened against pre-registered criteria and 232 studies were included and coded with a 40-field form covering context, data, methods, comparative design, accuracy metrics, inventory and cost measures, demand classification and ten reporting items; an independent 10% re-screening estimated consistency. The literature is recent, spread over many venues and concentrated in spare-parts settings; Croston-type estimators remain the reference methods, neural networks are the dominant AI family, and tree ensembles and deep networks appear only at the end of the period. 15% of the studies evaluate forecasts with an inventory or cost measure, and among the studies that compare AI/ML with classical methods, the family reported best differs between accuracy-based and cost-based evaluation in 1 of the 3 cases that report both. Wholesale distribution is explicitly studied in 5 studies. A cost-based benchmarking protocol for a wholesale distributor is derived from the evidence: SKU classification, a classical reference set, rolling-origin evaluation, inventory simulation under a stated policy with holding and shortage costs, and a reporting checklist. The review is limited by single-reviewer, abstract-based screening and by one bibliographic index.

Keywords

Intermittent demand; Croston method; machine learning; neural networks; inventory cost; forecast evaluation; wholesale distribution; systematic review

INTRODUCTION

A wholesale distributor buys in bulk from manufacturers and sells in smaller quantities to retailers, contractors, institutions and other businesses, and holds inventory so that each customer can be served from stock. Its assortment is wide and its demand per item is thin: a distributor of industrial, electrical, automotive or pharmaceutical products may list tens of thousands of stock-keeping units (SKUs), of which a large share sell in only a few periods a year and in quantities that vary from one unit to a pallet (Chen et al., 2007; Rumyantsev & Netessine, 2007; Bacchetti & Saccani, 2012). Demand of this kind, with many zero periods and variable non-zero sizes, is called intermittent, and when the sizes are highly variable, lumpy (Syntetos et al., 2005; Boylan et al., 2008). It is the demand pattern of spare and service parts, of maintenance, repair and operating supplies and of the long tail of any distribution assortment, and it is the pattern on which the standard forecasting and stock-control methods built for smooth demand perform worst (Croston, 1972; Syntetos & Boylan, 2001; Syntetos & Boylan, 2005).

Forecasting intermittent demand has been a research field of its own since Croston showed that exponential smoothing applied to intermittent series produces biased stock levels and proposed to forecast the demand size and the interval between demands separately (Croston, 1972). The field has since produced bias-corrected variants, estimators that update at every period, bootstrapping of the demand distribution, temporal aggregation, count-data models and, from the late 2000s, artificial neural networks and other machine-learning models (Syntetos & Boylan, 2001; Teunter et al., 2011; Willemain et al., 2004; Nikolopoulos et al., 2011; Snyder et al., 2012; Gutierrez et al., 2008; Kourentzes, 2013). The question that a wholesale distributor asks of this literature is practical: which method should drive replenishment, and by how much would a change of method reduce stock or improve service at a given cost. The literature answers it mostly in terms of forecast accuracy, and the accuracy statistics that it uses are known to be unreliable for intermittent series: percentage errors are undefined when the actual is zero, mean absolute error rewards forecasting zero, and the ranking of methods by accuracy need not agree with their ranking by inventory cost or service level (Hyndman & Koehler, 2006; Syntetos & Boylan, 2006; Wallström & Segerstedt, 2010; Kolassa, 2016; Prestwich et al., 2014). Evaluation through an inventory simulation that translates forecasts into orders, stock and shortages under a stated policy and cost structure is the recommended alternative, but it is applied in a minority of studies (Gardner, 1990; Sani & Kingsman, 1997; Syntetos et al., 2010; Tiacci & Saetta, 2009).

The entry of machine learning sharpened the question. Neural networks and tree ensembles can learn demand-size and interval behaviour jointly, use explanatory variables and pool information across SKUs, and the M4 competition showed that well-designed combinations of statistical and machine-learning methods can outperform either family on general time series (Makridakis et al., 2018a; Makridakis et al., 2018b). For intermittent demand the evidence is mixed: some studies report large gains for neural networks over Croston-type methods, others report that the gains vanish under inventory-based evaluation or that simple methods remain competitive (Gutierrez et al., 2008; Mukhopadhyay et al., 2012; Kourentzes, 2013; Lolli et al., 2017; Babai et al., 2014). Narrative reviews of spare-parts forecasting describe the methods and the gap between research and practice but do not quantify how often each family has been evaluated, under which type of evaluation, and with what result (Boylan & Syntetos, 2010; Bacchetti & Saccani, 2012; Hasni et al., 2019a; Van der Auweraer et al., 2019).

This article reports a systematic review of the literature published before 2020 on intermittent demand forecasting with classical, AI and ML methods, designed as a literature-based benchmark: the comparative evidence is organised by the type of evaluation (accuracy-based or inventory-cost-based) and read for the wholesale distribution context. It addresses four research questions. RQ1: what is the volume, growth and distribution of the literature, and which method families does it evaluate? RQ2: how is performance evaluated, with which accuracy metrics, inventory or cost measures, policies and cost parameters? RQ3: what is the comparative evidence on AI/ML against classical methods under accuracy-based and under cost-based evaluation, and which methods are reported best under cost-based evaluation? RQ4: what are the data characteristics and reporting practices, where are the gaps, and which cost-based benchmarking protocol for a wholesale distributor do the findings support? The restriction to literature published before 2020 is deliberate: it fixes the evidence base that preceded the rapid growth of deep-learning forecasting, so that the review describes the foundation on which that

growth builds rather than a moving target.
The contribution is threefold. Methodologically, the review applies a pre-registered protocol with a reproducible search, logged automated and manual screening, a structured extraction form, a reporting-quality score and an independent re-screening of a sample, and it publishes every decision in the supplementary material. Substantively, it quantifies the evaluation practice of the field and shows how often the verdict on AI/ML depends on whether forecasts are judged by accuracy or by inventory cost. Practically, it derives a cost-based benchmarking protocol that a wholesale distributor can run on its own data before changing the method that drives its replenishment.

BACKGROUND

Intermittent demand and its forecasting

Intermittent demand is demand that occurs in some periods and not in others; it becomes lumpy when the non-zero sizes are also highly variable. Two statistics describe a series: the average inter-demand interval (ADI), the mean number of periods between non-zero demands, and the squared coefficient of variation of the non-zero demand sizes (CV2). The categorisation scheme that uses cut-off values of 1.32 for ADI and 0.49 for CV2 to separate smooth, erratic, intermittent and lumpy demand was derived from the comparative bias and variance of Croston-type estimators and has become the standard way to describe intermittent-demand datasets and to select methods (Syntetos et al., 2005; Kostenko & Hyndman, 2006; Boylan et al., 2008; Heinecke et al., 2013). Intermittence is a property of the period length as much as of the item: aggregating weekly demand to monthly demand lowers the ADI, which is the basis of the temporal-aggregation methods (Nikolopoulos et al., 2011; Rostami-Tabar et al., 2013; Kourentzes et al., 2014).
Croston's method forecasts demand size and demand interval by separate exponential smoothing and divides the two to obtain a demand-per-period estimate (Croston, 1972). The estimate is biased because the ratio of two smoothed quantities is not the smoothed ratio; the Syntetos-Boylan approximation (SBA) corrects the bias with a multiplicative factor, and the Teunter-Syntetos-Babai (TSB) method smooths the probability of demand rather than the interval, so that the forecast decays during long periods without demand and the risk of obsolescence is recognised (Syntetos & Boylan, 2001; Syntetos & Boylan, 2005; Teunter et al., 2011). Comparative studies on industrial datasets established that SBA and related corrections outperform Croston and simple exponential smoothing on accuracy and on stock-control measures, and that the appropriate benchmark for any new method is this family rather than exponential smoothing alone (Syntetos & Boylan, 2005; Syntetos & Boylan, 2006; Teunter & Duncan, 2009; Boylan & Syntetos, 2007). The statistical models that underlie these estimators were examined and found to be inconsistent with the properties of real intermittent series, which motivated model-based alternatives: count-data models (Poisson, negative binomial, zero-inflated and hurdle models), integer-valued autoregressions and state-space models for non-negative data (Shenstone & Hyndman, 2005; Snyder et al., 2012). Bootstrapping methods resample the empirical demand history, with Markov switching between zero and non-zero states, to produce a distribution of lead-time demand rather than a point forecast; they were introduced for service-parts inventories and have been compared with parametric methods in several studies (Willemain et al., 2004; Syntetos et al., 2015; Hasni et al., 2019a; Hasni et al., 2019c). Reviews of the field describe this development and its application to spare parts, where the installed base and the maintenance policy are additional sources of information (Boylan & Syntetos, 2010; Bacchetti & Saccani, 2012; Van der Auweraer et al., 2019; Pennings et al., 2017).

Artificial intelligence and machine learning for intermittent demand

The AI and ML methods applied to intermittent demand before 2020 fall into a small number of families (Aditya & Ahmed, 2022; Asheke & Maksudul, 2022). Multilayer perceptrons were the first: a network trained on the lagged demand and the elapsed time since the last demand was reported to outperform exponential smoothing (Asheke & Badrul, 2022; Akter & Sumaiya, 2022), Croston and SBA on lumpy demand from an electronics distributor (Imran & Jannatul, 2022; Maksudul & Islam, 2022), and the same group later showed that the advantage depends on the accuracy measure and on the inventory evaluation used (Gutierrez et al., 2008; Mukhopadhyay et al., 2012). Later work examined network inputs (Morshedul & Islam, 2022; Stabak & Shahid, 2022), ensembles and the treatment of zeros, and showed that the gains reported for networks under accuracy measures shrink or reverse under

inventory-based evaluation (Kourentzes, 2013; Lolli et al., 2017). Support vector regression, nearest-neighbour methods (Feroz & Pervaz, 2022; Feroz & Shabbir, 2022), fuzzy and neuro-fuzzy systems, evolutionary optimisation of model parameters and, at the end of the period, random forests, gradient boosting and recurrent networks complete the family (Cortes & Vapnik, 1995; Breiman, 2001; Friedman, 2001; Hochreiter & Schmidhuber, 1997; Salinas et al., 2017). Two conclusions of the general forecasting literature frame the intermittent case. First, machine-learning methods evaluated on large collections of series were, before 2020, no better than and often worse than well-chosen statistical methods, and their advantage appears in combinations and in cross-learning across series (Makridakis et al., 2018b; Makridakis et al., 2018a; Zhang et al., 1998; Zhang, 2003). Second, the relative performance of methods depends on the data characteristics, which is why demand classification precedes method selection and why benchmark results must be read with the dataset in view (Petropoulos et al., 2014; Fildes & Petropoulos, 2015; Kourentzes, 2014).

Evaluating forecasts: accuracy against inventory cost

Forecast accuracy for intermittent demand cannot be measured with the percentage errors used for smooth demand (Islam & Md, 2023; Asheke, 2023), because actual values of zero make the mean absolute percentage error undefined and because the mean absolute error is minimised by a forecast of zero (Hyndman & Koehler, 2006; Syntetos & Boylan, 2005; Wallström & Segerstedt, 2010). The measures proposed in response are scale-free errors relative to a naive benchmark (MASE), scaled errors with a stated denominator, mean error and periods-in-stock as bias measures, and a modified percentage error for intermittent series (Hyndman & Koehler, 2006; Wallström & Segerstedt, 2010; Prestwich et al., 2014; Kim & Kim, 2016). The deeper problem is that any point-forecast accuracy measure judges the forecast against a single number per period while the forecast is used to set an order-up-to level or a reorder point for lead-time demand; the ranking of methods under an accuracy measure need not match their ranking under the stock and service level that result (Gardner, 1990; Syntetos & Boylan, 2006; Syntetos et al., 2010; Kolassa, 2016). The accuracy-implication approach therefore evaluates methods in an inventory simulation (Shahid, 2023; Islam & Ahmed, 2023; Sadia & Mosheur, 2023): forecasts feed a stated policy (periodic review order-up-to, continuous review with a reorder point), lead time and target service are fixed (Adnan, 2024; Islam, 2024; Ashik, 2024), and the outcomes are the achieved service or fill rate (Kynatun & Saba, 2024; Hasinur, 2024; Md Maksudul, 2024), the average stock and the total holding and shortage cost, often presented as trade-off curves (Sani & Kingsman, 1997; Syntetos & Boylan, 2006; Teunter & Duncan, 2009; Tiacci & Saetta, 2009; Babai et al., 2014). The costs of forecast error have been quantified in a warehouse setting and in industrial supply (Shahid, 2024; Islam, 2024; Mainuddin, 2024), and the difference between accuracy-based and cost-based verdicts is the central reason for a cost-based benchmark (Sanders & Graman, 2009; Kerkkänen et al., 2009; Fildes & Kingsman, 2011). The general evaluation literature adds the design requirements: out-of-sample tests with rolling origins, several horizons (Sumaiya, 2024), and error measures that permit generalisation across series (Tashman, 2000; Armstrong & Collopy, 1992; Bergmeir & Benítez, 2012; Davydenko & Fildes, 2013).

The wholesale distribution context

Wholesale trade is a large and distinct sector: the United States Census Bureau classifies it as sector 42 of the North American Industry Classification System and surveys it annually, and econometric studies of American firms show that wholesalers' inventory turnover and its relation to margins and lead times differ from those of retailers and manufacturers (Office of Management and Budget, 2017; U.S. Census Bureau, 2019; Chen et al., 2005; Chen et al., 2007; Rumyantsev & Netessine, 2007). Three features of the sector matter for intermittent demand forecasting. The assortment is wide and demand per SKU is thin, so that a large share of items is intermittent by any categorisation. The wholesaler sits between manufacturers with long lead times and customers who expect immediate availability, so that service-level commitments and holding costs are both high and the trade-off between them is the operational objective. And the wholesaler's orders are the demand signal that its suppliers see, so that its forecasting method shapes the bullwhip effect upstream (Lee et al., 1997; Chen et al., 2000; Cachon et al., 2007). Collaborative arrangements such as vendor-managed inventory and collaborative planning, forecasting and replenishment were developed partly to replace the distributor's noisy orders with shared demand information, and their benefits depend on the forecasting method applied to the shared

data (Cachon & Fisher, 2000; Disney & Towill, 2003; Aviv, 2001; Sari, 2008). The spare-parts literature, which supplies most of the intermittent-demand evidence, concerns distributors in all but name: aftermarket and service-parts organisations stock thousands of slow-moving items for many customers under service commitments (Bacchetti & Saccani, 2012; Boylan & Syntetos, 2010; Turrini & Meissner, 2019). Reading the intermittent-demand benchmark for wholesale distribution therefore means attending to the scale of the assortment, to the cost structure of the evaluation and to the policy that the forecasts feed; Table 1 summarises the method families that the benchmark compares, with their defining references.

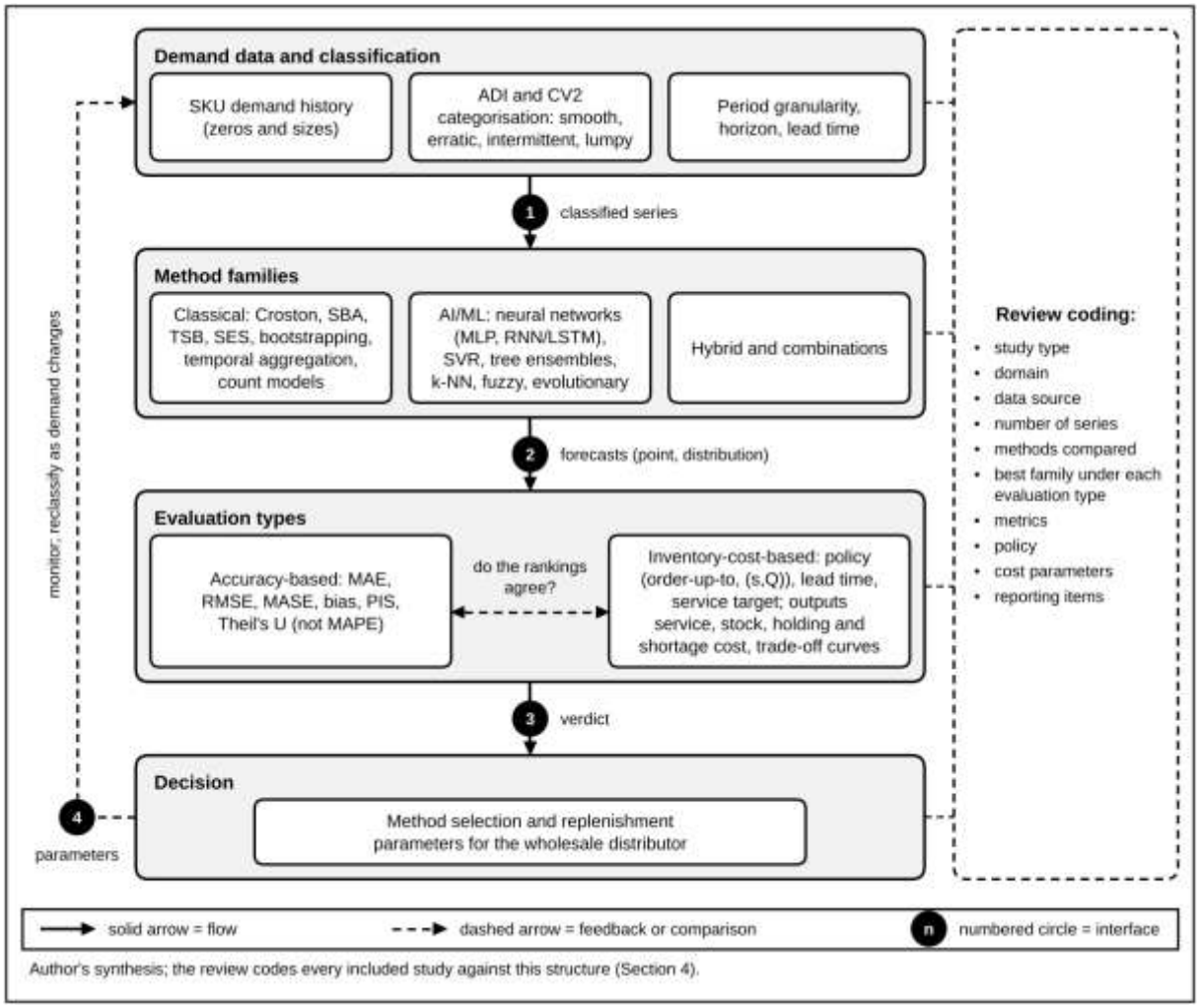
Table 1: Method families for intermittent demand forecasting, with their defining references in the pre-2020 literature

Family	Methods	Principle	Defining references (pre-2020)
Classical: size-interval estimators	Croston; Syntetos-Boylan approximation (SBA); Syntetos-Boylan-Johnston variant; Teunter-Syntetos-Babai (TSB)	Demand size and demand interval (or demand probability) smoothed separately; bias correction; probability updated every period to recognise obsolescence	(Croston, 1972; Syntetos & Boylan, 2001; Syntetos & Boylan, 2005; Teunter et al., 2011; Boylan & Syntetos, 2007)
Classical: smoothing and ARIMA	Simple exponential smoothing, moving averages, ARIMA, state-space exponential smoothing	Standard smooth-demand methods used as benchmarks; biased on intermittent series	(Gardner, 2006; Hyndman et al., 2002; Box et al., 2015)
Classical: count and state-space models	Poisson, negative binomial, zero-inflated and hurdle models; integer-valued autoregressions; non-negative state-space models	A probability model for counts with zeros; lead-time demand distribution follows from the model	(Shenstone & Hyndman, 2005; Snyder et al., 2012; Kolassa, 2016)
Classical: bootstrapping	Willemain-type Markov bootstrap; empirical and parametric bootstraps	Resampling of the demand history to obtain the distribution of lead-time demand	(Willemain et al., 2004; Syntetos et al., 2015; Hasni et al., 2019a; Hasni et al., 2019c)
Classical: temporal and hierarchical aggregation	ADIDA, MAPA, multiple temporal aggregation; cross-sectional hierarchies	Aggregate the series to reduce intermittence, forecast, disaggregate; combine levels	(Nikolopoulos et al., 2011; Rostami-Tabar et al., 2013; Kourentzes et al., 2014; Hyndman et al., 2011)
AI/ML: neural networks	Multilayer perceptrons; recurrent and long short-term memory networks; deep probabilistic forecasters	Non-linear mapping from lagged demand, elapsed time and covariates to demand; cross-learning over series	(Gutierrez et al., 2008; Mukhopadhyay et al., 2012; Kourentzes, 2013; Lolli et al., 2017; Hochreiter & Schmidhuber, 1997; Salinas et al., 2017)
AI/ML: kernel, tree and instance methods	Support vector regression; random forests; gradient boosting; nearest neighbours	Supervised learners on feature vectors built from the series and item attributes	(Cortes & Vapnik, 1995; Breiman, 2001; Friedman, 2001; Makridakis et al., 2018b)
AI/ML: fuzzy and evolutionary	Fuzzy and neuro-fuzzy systems; genetic and swarm optimisation of model parameters	Rule-based or population-based approximation and parameter search	(Zhang et al., 1998; Min, 2010)

Family	Methods	Principle	Defining references (pre-2020)
Hybrid and combinations	Classical plus neural components; forecast combinations; selection by demand class	Combine families or select among them by data characteristics	(Zhang, 2003; Petropoulos & Kourentzes, 2015; Petropoulos et al., 2014; Fildes & Petropoulos, 2015)
Evaluation methods	Scale-free and bias measures (MASE, PIS, modified percentage errors); inventory simulation; trade-off curves	Judge forecasts by error measures suited to zeros, or by the stock and service they produce under a policy	(Hyndman & Koehler, 2006; Wallström & Segerstedt, 2010; Prestwich et al., 2014; Kim & Kim, 2016; Syntetos & Boylan, 2006; Syntetos et al., 2010; Tiacci & Saetta, 2009)

Figure 1 states the conceptual framework that organises the review. Demand data are classified; candidate methods from the classical, AI/ML and hybrid families produce forecasts; the forecasts are evaluated either by accuracy statistics or by an inventory simulation that reports service and cost; and the evaluation type decides the verdict. The review codes every included study against this structure and asks, for each comparison of AI/ML with classical methods, which evaluation type produced the verdict.

Figure 1: Conceptual framework of the review: demand classification, method families, evaluation types and the decision they support

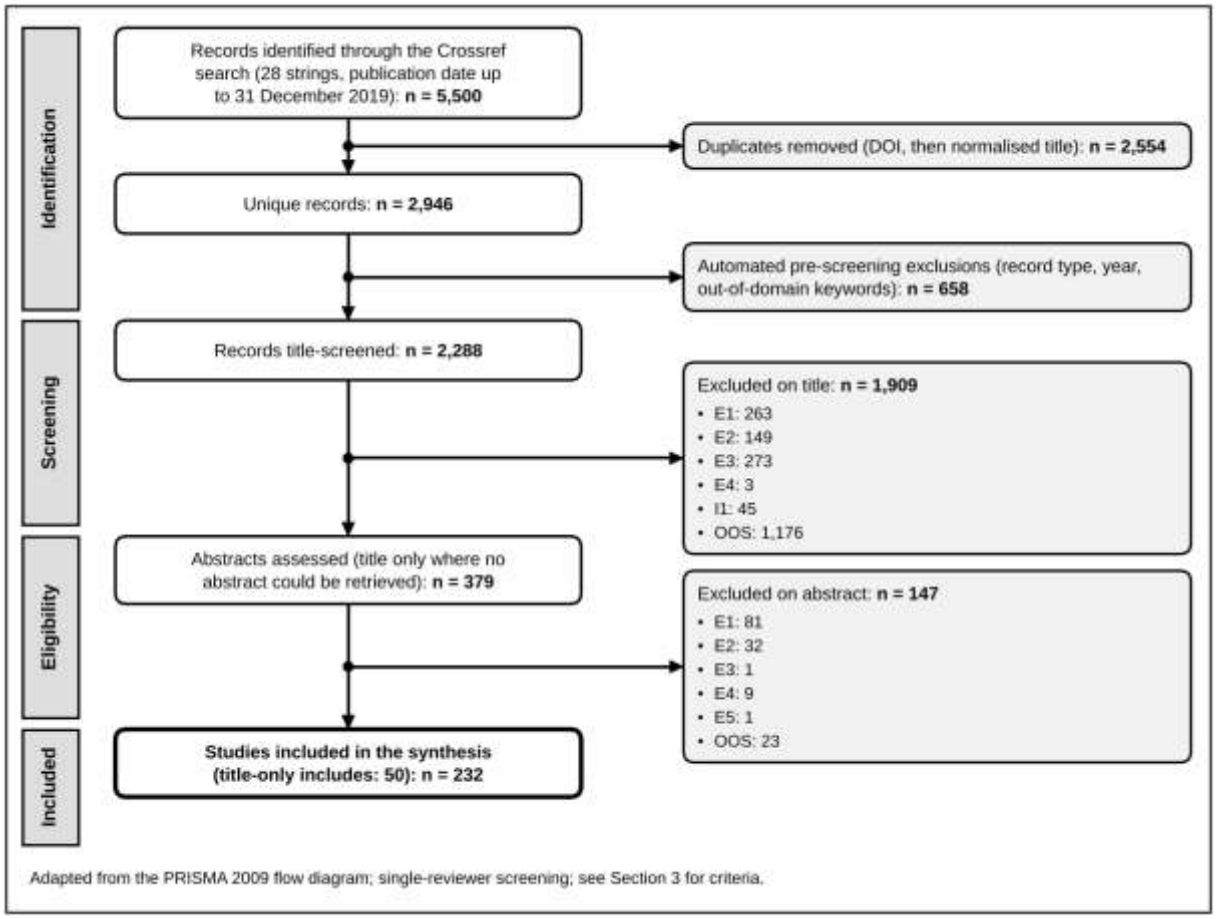


METHODOLOGY

Protocol and reporting

The review followed a protocol written before the search (file protocol.md in the supplementary material) and is reported with the PRISMA 2009 items that apply to a review of a bibliographic index by a single reviewer (Moher et al., 2009; Liberati et al., 2009). The protocol fixed the research questions, the information source, the search strings, the eligibility criteria, the screening stages, the extraction form, a ten-item reporting-quality score and the synthesis plan. Conventions for systematic reviews in management and operations research, where the evidence consists of methodological and applied studies without a common outcome measure, were followed for the structure of the synthesis (Tranfield et al., 2003; Durach et al., 2017; Seuring & Müller, 2008; Webster & Watson, 2002; Petticrew & Roberts, 2006). Figure 2 gives the flow of records.

Figure 2: PRISMA-style flow of records through identification, automated pre-screening, title screening and abstract screening



Information source and search

The information source was the Crossref works index (https://www.crossref.org), queried through its public application programming interface on 31 October 2025. Crossref holds the metadata that publishers deposit when they register digital object identifiers (DOIs) for journal articles, proceedings papers and book chapters, and it can be queried reproducibly by uniform resource locator without a subscription. It was used because the OpenAlex index, the source of the author's earlier reviews, had suspended its list and search endpoints at the time of the search; OpenAlex (https://openalex.org) remained available for single-record lookups and was used to retrieve abstracts by DOI. Crossref's bibliographic query searches the title, author, container and year fields and ranks records by relevance; it does not search abstracts, which is a limitation stated in Section 6.4. Twenty-eight search strings (Table 2) were run with the filter restricting the publication date to 31 December 2019 or earlier and no language or document-type filter at search time; for each string the first 200 records by relevance were

retrieved and the total number of matching records reported by the index was recorded. The strings covered the vocabulary of intermittent demand (intermittent, lumpy, sporadic, erratic, slow-moving, spare parts), of the classical and AI/ML method families (Croston, bootstrapping, temporal aggregation, neural network, machine learning, deep learning, support vector, random forest), of inventory-based evaluation (inventory cost, service level, stock control) and of the wholesale distribution context (wholesale, wholesaler, distributor, distribution centre, SKU). The index throttled concurrent requests: 16 of the 112 result pages could not be retrieved in the first pass, 14 of them were retrieved in a clean-up pass on the same day, and two pages (the third page of the string on Croston's method and the second page of the string on spare-parts forecasting with neural networks) could not be retrieved at all and are recorded in Table 2; because every page beyond the first is a lower-relevance page and most strings overlap, the loss is unlikely to have removed a study found by no other string, but it cannot be excluded. Reference lists of included reviews were not searched.

Table 2: Search strings run against the Crossref bibliographic fields (publication date up to 31 December 2019), with the total matches reported by the index, records retrieved, unique records contributed and included studies retrieved by each string

String	Terms (bibliographic fields, relevance-ranked)	Matches	Retrieved	Unique	Included	Note
Q01	intermittent demand forecasting	199,230	200	171	55	
Q02	intermittent demand Croston	128,692	150	149	56	1 page(s) not retrieved (index throttling); 1 re-fetched
Q03	intermittent demand neural network	972,173	200	191	16	
Q04	intermittent demand machine learning	1,178,680	200	196	26	1 page(s) re-fetched in the clean-up pass
Q05	lumpy demand forecasting	166,533	200	166	20	
Q06	spare parts demand forecasting	257,138	200	195	113	1 page(s) re-fetched in the clean-up pass
Q07	spare parts forecasting neural network	1,005,810	150	142	56	1 page(s) not retrieved (index throttling)
Q08	intermittent demand inventory cost	434,447	200	189	11	2 page(s) re-fetched in the clean-up pass
Q09	intermittent demand service level stock control	2,767,076	200	193	19	1 page(s) re-fetched in the clean-up pass
Q10	intermittent demand bootstrapping	130,908	200	180	61	1 page(s) re-fetched in the clean-up pass
Q11	slow-moving items demand forecasting	350,088	200	171	20	
Q12	sporadic demand forecasting	179,937	200	163	8	
Q13	intermittent demand classification categorisation	384,300	200	193	57	
Q14	intermittent demand temporal aggregation	362,037	200	190	35	

String	Terms (bibliographic fields, relevance-ranked)	Matches	Retrieved	Unique	Included	Note
Q15	wholesale demand forecasting	173,663	200	161	5	1 page(s) re-fetched in the clean-up pass
Q16	wholesaler distributor demand forecasting SKU	169,170	200	166	7	1 page(s) re-fetched in the clean-up pass
Q17	distribution center demand forecasting SKU	1,081,482	200	168	7	
Q18	demand forecasting machine learning SKU retail	1,241,558	200	193	8	
Q19	forecast accuracy inventory performance intermittent	1,380,077	200	167	4	
Q20	forecasting inventory cost evaluation intermittent	1,580,991	200	187	32	
Q21	intermittent demand deep learning LSTM	1,213,663	200	198	22	1 page(s) re-fetched in the clean-up pass
Q22	intermittent demand support vector regression	671,147	200	197	4	
Q23	intermittent demand random forest gradient boosting	606,092	200	197	11	
Q24	spare parts inventory forecasting cost service	898,447	200	197	54	1 page(s) re-fetched in the clean-up pass
Q25	erratic demand forecasting	166,601	200	162	7	1 page(s) re-fetched in the clean-up pass
Q26	count data demand forecasting zero-inflated inventory	1,720,026	200	191	7	
Q27	intermittent time series forecasting	2,349,229	200	178	1	1 page(s) re-fetched in the clean-up pass
Q28	demand classification spare parts forecasting methods	1,635,988	200	187	110	1 page(s) re-fetched in the clean-up pass

Eligibility criteria

Table 3 lists the inclusion and exclusion criteria. A record was eligible if it was a journal article, proceedings paper or book chapter published between 1990 and 2019 in English, concerned the forecasting of intermittent, lumpy, sporadic, slow-moving or erratic demand (or was a methodological contribution addressed to such demand), and reported, evaluated or compared at least one forecasting approach, evaluated forecasts of such demand with inventory or cost measures, or proposed an evaluation method for such forecasts. Studies of smooth or aggregate demand without an intermittent component, inventory models that assume a known demand distribution and evaluate no forecasting method, uses of the vocabulary outside demand forecasting (medicine, energy supply, hydrology, computing), non-journal record types and forecasting at the aggregate level only were excluded. Reviews and surveys of intermittent demand forecasting and methodological papers on metrics and classification were included and coded by study type.

Table 3: Eligibility criteria and the screening stage at which each was applied

Criterion	Definition	Stage applied
I1 Record type, period, language	Journal article, proceedings paper or book chapter; Crossref issued date on or before 31 December 2019 and year 1990 or later; English	Automated (type, year), title, abstract
I2 Subject	Forecasting of intermittent, lumpy, sporadic, slow-moving or erratic demand (spare and service parts, MRO items, retail or wholesale SKUs with intermittent sales, generic intermittent series), or a methodological contribution addressed to such demand (metrics, classification, evaluation design, inventory implications)	Title, abstract
I3 Method or evaluation	Reports, evaluates or compares at least one forecasting approach (classical or AI/ML), or evaluates forecasts of such demand with inventory or cost measures, or proposes an evaluation method for such forecasts	Abstract
E1 Not intermittent	Smooth or aggregate demand only (energy load, tourism, traffic, product-family totals) with no intermittent component in method or data	Automated (keywords), title, abstract
E2 No forecasting	Inventory models that assume a known demand distribution and evaluate no forecasting method	Title, abstract
E3 Other meaning	The words intermittent, sporadic, lumpy or demand used outside demand forecasting (medicine, energy supply, hydrology, computing)	Automated (keywords), title, abstract
E4 Record type or language	Dissertations, whole books, editorials, errata, patents, posted preprints, conference abstracts without a paper, non-English records	Automated (type), title, abstract
E5 Aggregate only	Forecasting of aggregate or family-level demand with no SKU-level intermittent component and no intermittent-demand method	Abstract
OOS Out of scope	Unrelated fields retrieved by the generic terms of the strings (demand classification in other senses, SKU studies without forecasting, wholesale trade economics)	Title, abstract

Screening

Records were de-duplicated by DOI and then by normalised title, and passed through three stages (Figure 2). Stage 0 was automated: records of types other than journal article, proceedings paper or book chapter, records dated 2020 or later by the index, and records whose titles matched keyword rules for the medical, energy, water, transport and computing uses of the vocabulary were removed, and every automated exclusion was logged with its rule (merge3x.py in the supplementary material). Stage 1 was a title screen by the author; every title was read and each excluded title received a reason code (E1 to E5 or out of scope). Stage 2 was an abstract screen with data extraction, performed by screening agents working from a written brief (extract_brief.md) that fixed the retrieval routes for abstracts (OpenAlex by DOI, then the Crossref record, Semantic Scholar and, as a last resort, the publisher's landing page for the DOI, one attempt each), the order in which the criteria were applied, the category strings of the extraction form and the rule that nothing may be recorded that is not in the fetched text; the author resolved the records that the agents marked as uncertain. When no abstract could be retrieved and the title indicated eligibility, the record was included on the title alone and its extraction fields were coded as not reported; the share of such records is reported in Section 4.1 and treated as a limitation. Full texts were not consulted. An independent 10% sample of the abstract-stage records was re-screened and re-extracted by a second agent that had not seen the first decisions, and agreement is reported in Section 4.1.

Data extraction and appraisal

A structured form (Table 4) was completed for every included record from its abstract. The form records the bibliographic data; the study type; the application domain, data source, dataset name, number of series, period granularity, horizon, evaluation design and whether a wholesale or distribution context is mentioned; nineteen method indicators grouped into the classical and AI/ML families of Figure 1; the comparative design (whether AI/ML is compared with classical methods, the family reported best under accuracy-based and under inventory-cost-based evaluation, whether the two rankings disagree, the best single method named and the reported magnitude of its advantage); eleven accuracy-metric indicators; seven inventory or cost-measure indicators; whether an inventory simulation was run, under which policy, and whether cost parameters are stated; the demand classification used; and whether the level of intermittence is reported. Fields not addressed in the abstract were coded as not reported, which makes every count a lower bound of the practice in the full texts. Ten reporting-quality items (Table 5) were scored 1 when the record explicitly reports the item and 0 otherwise; the sum gives a reporting score between 0 and 10. The score describes what the abstract reports and is not a judgement of the study's quality.

Table 4: Data extraction form completed for each included record (fields and categories)

Field group	Fields and categories
Bibliographic	Record identifier, DOI, year, source, study type (original or empirical study; methodological or metrics; review or survey; case or industry report; simulation-only study)
Context and data	Application domain (spare or service parts; automotive; aerospace or defence; electronics or high-tech; retail; wholesale or distribution; pharmaceutical or healthcare; industrial MRO or manufacturing; utilities or transport; generic benchmark data; other; not reported); data source (company data, public dataset, synthetic or simulated, mixed, not reported) and dataset name; number of series (fewer than 50; 50 to 499; 500 to 4,999; 5,000 or more; not reported); period granularity (daily, weekly, monthly, other); forecast horizon stated; evaluation design (rolling origin, fixed origin, not reported); wholesale or distribution context mentioned (yes, no)
Methods evaluated	Nineteen indicators: Croston; SBA or SBJ; TSB; simple exponential smoothing or moving average; ARIMA; count or integer-valued models; bootstrapping; temporal aggregation; hierarchical aggregation; Bayesian; multilayer perceptron; recurrent or deep network; support vector machine or regression; tree ensemble; nearest neighbours or regression tree; fuzzy or neuro-fuzzy; evolutionary or swarm optimisation; hybrid; other
Comparative design	AI/ML compared with classical methods (yes, no); family reported best under accuracy metrics and under inventory or cost measures (AI/ML; classical; hybrid; mixed or no clear winner; not applicable); accuracy and cost rankings disagree (yes, no, not applicable); best single method named; reported magnitude of the advantage
Evaluation	Accuracy metrics (MAE or MAD; MSE or RMSE; MAPE or sMAPE; MASE; RMSSE or scaled errors; ME or bias; PIS or CFE; Theil's U; relative or percent-better measures; other; none); inventory or cost measures (service level or fill rate; stock or holding cost; backorders or stock-outs; total cost; trade-off curves; achieved versus target service; none); inventory simulation performed; policy (order-up-to; (s,S); (s,Q); base-stock; other; none); cost parameters stated (holding rate, shortage cost, lead time)
Data characteristics	Demand classification used (ADI and CV2; ABC or XYZ; other; none); intermittence level reported (share of zeros, ADI or CV2 values)
Outcome and appraisal	Main reported outcome (one line); limitations stated; ten reporting-quality items (Table 5); basis of the record (abstract, title only, full text); abstract source (OpenAlex, Crossref, Semantic Scholar, none)

Table 5: Ten-item reporting-quality score, with the research question that each item serves

Item	Scored 1 when the record reports	Research question
1 Data source described	The origin of the demand data (company, sector, public dataset or simulation) with enough detail to characterise it	RQ1, RQ4
2 Number of series stated	The number of SKUs or series evaluated	RQ1, RQ4
3 Evaluation design stated	Rolling-origin or fixed-origin out-of-sample evaluation with the size of the training and test windows	RQ2
4 Forecast horizon stated	The horizon or lead time over which forecasts are evaluated	RQ2
5 Intermittent-appropriate accuracy metric	At least one scale-free, absolute, bias or relative measure (MASE, MAE, RMSE, ME, PIS, Theil's U, percent better); MAPE or sMAPE alone does not score	RQ2
6 Inventory or cost evaluation performed	Forecasts translated into stock, service or cost through an inventory simulation or an accuracy-implication metric	RQ2, RQ3
7 Cost parameters stated	Holding rate, shortage or backorder cost, lead time and service target stated	RQ3
8 Croston-type benchmark included	Croston, SBA, SBJ or TSB among the compared methods	RQ3
9 Code or data availability	Code, data or supplementary material stated to be available	RQ4
10 Limitations stated	Limitations of the study or of the approach acknowledged	RQ4

Synthesis

The synthesis is descriptive and takes the form of a literature-based benchmark. For RQ1, records were counted by year, source, study type, domain, data source and scale, the yield of each search string was computed, and the method families were counted overall, by period (1990 to 2009 against 2010 to 2019) and in co-occurrence within studies. For RQ2, the accuracy metrics, inventory and cost measures, policies and cost parameters were counted overall and by study type. For RQ3, the comparisons of AI/ML with classical methods were tabulated as vote counts of the family reported best under accuracy-based and under cost-based evaluation, the agreement between the two verdicts was computed for the studies that report both, and the methods reported best under cost-based evaluation were listed study by study with the dataset and the magnitude reported; evidence gap maps cross the method families with the evaluation types and the domains with the evaluation types. For RQ4, the data characteristics and the reporting items were counted and the score distribution summarised. No meta-analysis of effect sizes was attempted: the studies differ in data, horizon, metrics and policies, and most abstracts report no comparable effect size. The benchmarking protocol of Section 5 was derived by stating, for each step of a cost-based benchmark, what the coded evidence supports, what it leaves open, and what the inventory-forecasting literature prescribes.

RESULTS

Search and screening outcome

The 28 search strings returned 5,500 records in total, of which 2,946 were unique after de-duplication by DOI and normalised title (Figure 2). Automated pre-screening removed 658 records: 367 because the record type was not a journal article, proceedings paper or book chapter (whole books, book series entries, reference-work entries, datasets, reports and posted content), and 312 because the title matched one of the out-of-domain keyword rules listed in the protocol (medical, energy, hydrological and computing uses of the word intermittent and its neighbours). The remaining 2,288 titles were screened by the reviewer; 1,909 were excluded, the most frequent reasons being unrelated fields (1,176 records), the use of intermittent, lumpy, sporadic or demand outside demand forecasting (273), non-intermittent demand (263), inventory models without a forecasting element (149) and publication before 1990 (45). 379 records proceeded to abstract screening and extraction, 147 were excluded at that stage (E1 81, E2 32, OOS 23, E4 9, E5 1, E3 1), and 232 studies form the included corpus. For 174 of the included studies

the decision and the extraction rest on an abstract retrieved from OpenAlex, Crossref, Semantic Scholar or the publisher landing page, for 8 on an open-access full text, and for 50 on the title alone because no abstract could be retrieved from any source; the title-only group consists largely of book chapters whose publishers deposit no abstract, and its extraction fields are mostly "not reported", which lowers every coverage statistic reported below.

An independent re-screening and re-extraction of 38 records (10% of the abstract stage) agreed with the primary decision in 37 records (97%) and, for the records included by both, agreed on 319 of 350 extracted fields (91%); disagreements concentrated in the replenishment policy, the application domain, the accuracy verdict and the demand-classification fields, which abstracts often describe only implicitly, and the primary extraction was retained in every case.

Volume and distribution of the literature (RQ1)

Table 6 and Figure 3 describe the corpus. The 232 included studies were published between 1990 and 2019, with 56 (24%) appearing before 2010 and 176 (76%) from 2010 onwards; the single most productive year was 2019. The growth parallels that of the whole screened set (Figure 3(a)), so it reflects the expansion of the literature indexed by Crossref as much as any change in research interest. By study type, the largest group was original or empirical study (159 studies, 69%), followed by methodological or metrics (25) and simulation-only study (19). Reviews and surveys (11 studies) were retained because they document the methods and evaluation practices of their time and because they are the sources through which practitioners encounter the field; they are counted with the other studies in the descriptive tables and contribute no comparative verdicts.

Table 6: Characteristics of the included studies: period, study type, application domain and data source

Characteristic	Studies	Share
Studies included	232	100%
Published 1990 to 2009	56	24%
Published 2010 to 2019	176	76%
Study type: original or empirical study	159	69%
Study type: methodological or metrics	25	11%
Study type: simulation-only study	19	8%
Study type: case or industry report	18	8%
Study type: review or survey	11	5%
Application domain: spare or service parts	76	33%
Application domain: not reported	57	25%
Application domain: aerospace or defence	38	16%
Application domain: automotive	17	7%
Application domain: utilities or transport	13	6%
Application domain: industrial MRO or manufacturing	11	5%
Application domain: electronics or high-tech	9	4%
Application domain: other	3	1%
Application domain: retail	3	1%
Application domain: pharmaceutical or healthcare	2	1%
Application domain: wholesale or distribution	2	1%
Application domain: generic benchmark data	1	0%
Data source: not reported	116	50%
Data source: company data	90	39%
Data source: synthetic or simulated	16	7%
Data source: mixed	8	3%
Data source: public dataset	2	1%

Characteristic	Studies	Share
Wholesale or distribution context mentioned	5	2%

The application domains in Table 6 confirm the origin of the field in spare and service parts: spare or service parts was the most frequent domain (76 studies, 33%), followed by aerospace or defence (38) and automotive (17). Studies that used generic benchmark data (the M-competition series, the Royal Air Force and automotive spare-parts collections that circulate in the literature, or the public retail datasets of the later period) numbered 1. A wholesale or distribution context (a wholesaler, a distributor or a distribution centre named as the setting) was mentioned in 5 of the 232 studies (2%), and the retail domain contributed 3 studies. This is the first substantive finding for the practitioner audience of this paper: the empirical evidence on intermittent demand forecasting comes overwhelmingly from maintenance and after-sales environments, in which the cost structure (expensive parts, high shortage penalties, long lead times) differs from that of a distributor whose intermittent SKUs are often inexpensive, numerous and sold alongside fast movers. Company data were the dominant data source (90 studies, 39%), with 2 studies using a public dataset, 16 relying on synthetic or simulated series only and 8 combining sources; the data source was not stated in the abstract of 116 studies.

Figure 3: Bibliometric profile of the included studies: (a) records per publication year against the whole screened set, (b) study types, (c) application domains

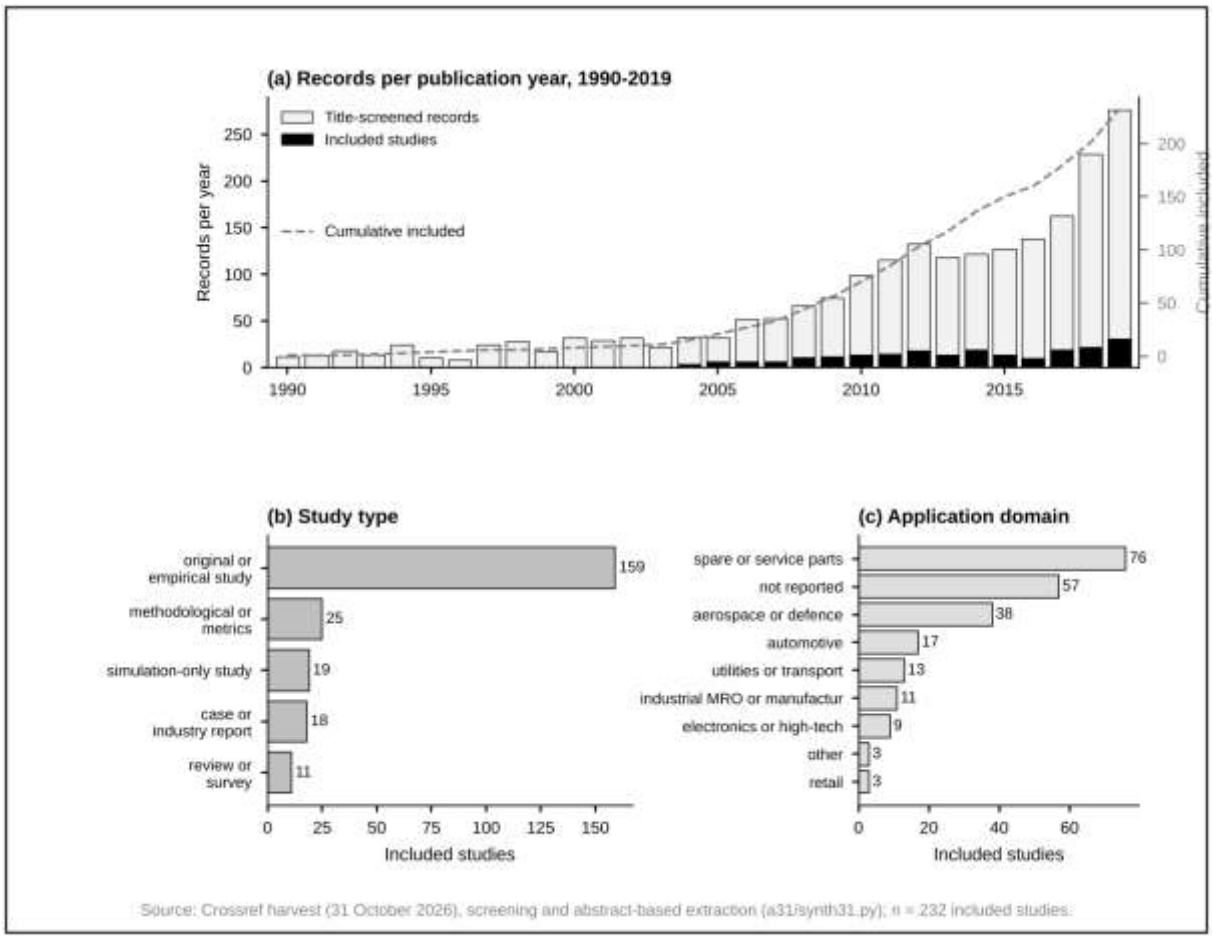


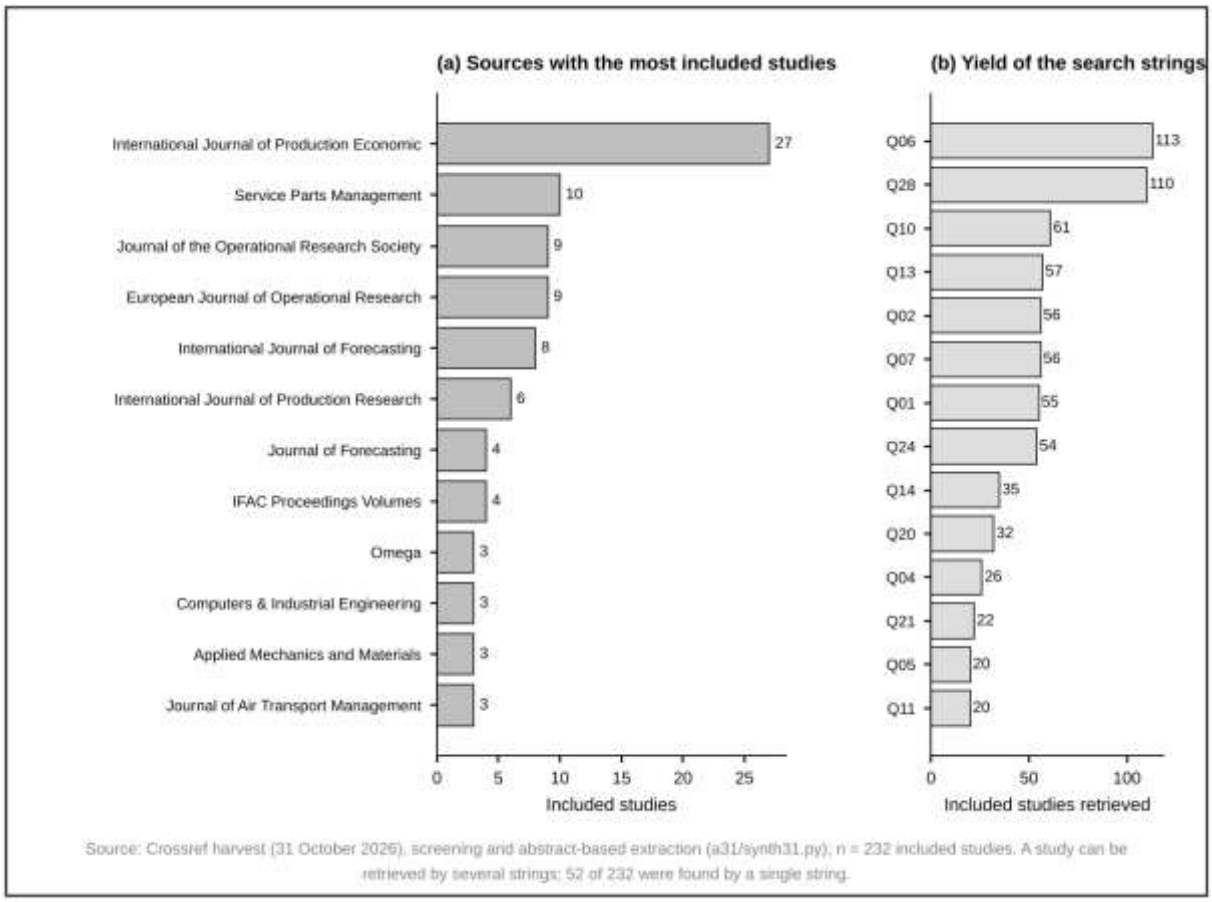
Table 7 and Figure 4 report the publication venues and the scale of the evidence. The corpus is spread over 136 distinct sources; the twelve most frequent venues account for 89 studies (38%), led by International Journal of Production Economics (27 studies), Service Parts Management (10) and Journal of the Operational Research Society (9). The concentration in the operational research and forecasting journals reflects the field's methodological orientation, while the long tail of conference proceedings and regional journals, many of them engineering outlets, carries most of the AI and machine-learning

applications of the 2010s. The scale of the datasets is modest by the standards of a distributor's assortment: 16 studies evaluated fewer than 50 series, 5 between 50 and 499, 10 between 500 and 4,999 and 8 evaluated 5,000 or more, with the number of series not stated in the abstract of 193 studies. Monthly series were the most common granularity (12 studies), followed by weekly (3) and daily (2) data. The search strings differed sharply in yield (Figure 4(b)): the intermittent demand forecasting, Croston and spare-parts strings retrieved most of the included studies, while the strings aimed at machine learning, deep learning and wholesale distribution retrieved few, and 52 of the 232 included studies (22%) were found by a single string only, which indicates that a narrower search would have missed a material share of the evidence.

Table 7: Publication venues, number of series and period granularity of the included studies

Source or characteristic	Studies	Share
International Journal of Production Economics	27	12%
Service Parts Management	10	4%
Journal of the Operational Research Society	9	4%
European Journal of Operational Research	9	4%
International Journal of Forecasting	8	3%
International Journal of Production Research	6	3%
Journal of Forecasting	4	2%
IFAC Proceedings Volumes	4	2%
Omega	3	1%
Computers & Industrial Engineering	3	1%
Applied Mechanics and Materials	3	1%
Journal of Air Transport Management	3	1%
Other sources (124)	143	62%
Number of series: fewer than 50	16	7%
Number of series: 50 to 499	5	2%
Number of series: 500 to 4,999	10	4%
Number of series: 5,000 or more	8	3%
Number of series: not reported	193	83%
Period granularity: not reported	212	91%
Period granularity: monthly	12	5%
Period granularity: weekly	3	1%
Period granularity: other	3	1%
Period granularity: daily	2	1%

Figure 4: Sources and search yield: (a) venues with the most included studies, (b) included studies retrieved by each search string



Method families evaluated (RQ2)

Table 8 and Figure 5 report the forecasting methods evaluated in the included studies, coded in nineteen categories and summarised in two periods. Any classical method (Croston and its variants, exponential smoothing or moving averages, ARIMA, count models, bootstrapping, temporal aggregation, hierarchical aggregation or Bayesian methods) appeared in 105 studies (45%), and any AI or machine-learning method (neural networks of any kind, support vector machines or regression, tree ensembles, nearest neighbours, fuzzy systems, evolutionary optimisation or hybrids) in 58 studies (25%). Croston's method was evaluated more often than feed-forward neural networks (50 against 29 studies), and the Croston family as a whole (Croston, SBA or SBJ, TSB: 50, 24 and 10 studies respectively) remained the reference set throughout the period, as the protocol of Section 5 assumes. Simple exponential smoothing or moving averages were evaluated in 48 studies, often as the naive comparator that Croston's method was designed to replace (Croston, 1972; Johnston & Boylan, 1996). Bootstrapping methods appeared in 12 studies and temporal aggregation (ADIDA, MAPA and their relatives) in 8, both concentrated in the later period (9 and 8 of their studies from 2010 onwards), consistent with the chronology of their defining references (Willemain et al., 2004; Nikolopoulos et al., 2011; Kourentzes et al., 2014; Hasni et al., 2019a). Count or integer-valued models were evaluated in 15 studies and Bayesian methods in 11.

Table 8: Forecasting methods evaluated in the included studies, by family and period

Method	Family	Studies	Share	Before 2010	From 2010
Croston	Classical	50	22%	15	35
SBA or SBJ	Classical	24	10%	4	20
TSB	Classical	10	4%	0	10
Simple exponential smoothing or moving average	Classical	48	21%	15	33
ARIMA or ARMA	Classical	9	4%	1	8
Count or integer-valued models	Classical	15	6%	7	8
Bootstrapping	Classical	12	5%	3	9
Temporal aggregation (ADIDA, MAPA)	Classical	8	3%	0	8
Hierarchical or cross-sectional aggregation	Classical	6	3%	1	5
Bayesian methods	Classical	11	5%	3	8
Neural network (MLP, feed-forward)	AI/ML	29	12%	5	24
Recurrent or deep network	AI/ML	4	2%	1	3
Support vector machine or regression	AI/ML	21	9%	2	19
Tree ensemble (random forest, boosting)	AI/ML	6	3%	0	6
Nearest neighbours or regression tree	AI/ML	7	3%	0	7
Fuzzy or neuro-fuzzy	AI/ML	8	3%	1	7
Evolutionary or swarm optimisation	AI/ML	4	2%	0	4
Hybrid (classical plus AI)	Hybrid	24	10%	2	22
Other	Other	138	59%	29	109
Any classical method		105	45%		
Any AI/ML or hybrid method		58	25%		
AI/ML compared with classical in the same study		23	10%		

Within the AI and machine-learning families, feed-forward neural networks (multilayer perceptrons and their variants) were the most frequent (29 studies), followed by support vector machines or regression (21), fuzzy or neuro-fuzzy systems (8), tree ensembles such as random forests and gradient boosting (6), recurrent or deep networks (4), nearest-neighbour or regression-tree methods (7) and evolutionary or swarm optimisers used to tune a forecasting model (4). Hybrids that combine a classical structure with a learned component, in the manner of the Croston-and-neural-network schemes of the 2000s (Gutierrez et al., 2008; Hua et al., 2007; Mukhopadhyay et al., 2012), were coded in 24 studies. The period split in Figure 5 is instructive: 5 of the 29 neural-network studies were published before 2010, so the neural-network literature on intermittent demand predates the deep-learning wave, while recurrent and deep architectures (3 of 4 studies from 2010) and tree ensembles (6 of 6) are almost entirely a phenomenon of the 2010s and are therefore under-evaluated relative to their current use in practice (Makridakis et al., 2018b; Salinas et al., 2017).

Figure 5: Method families evaluated in the included studies, by period

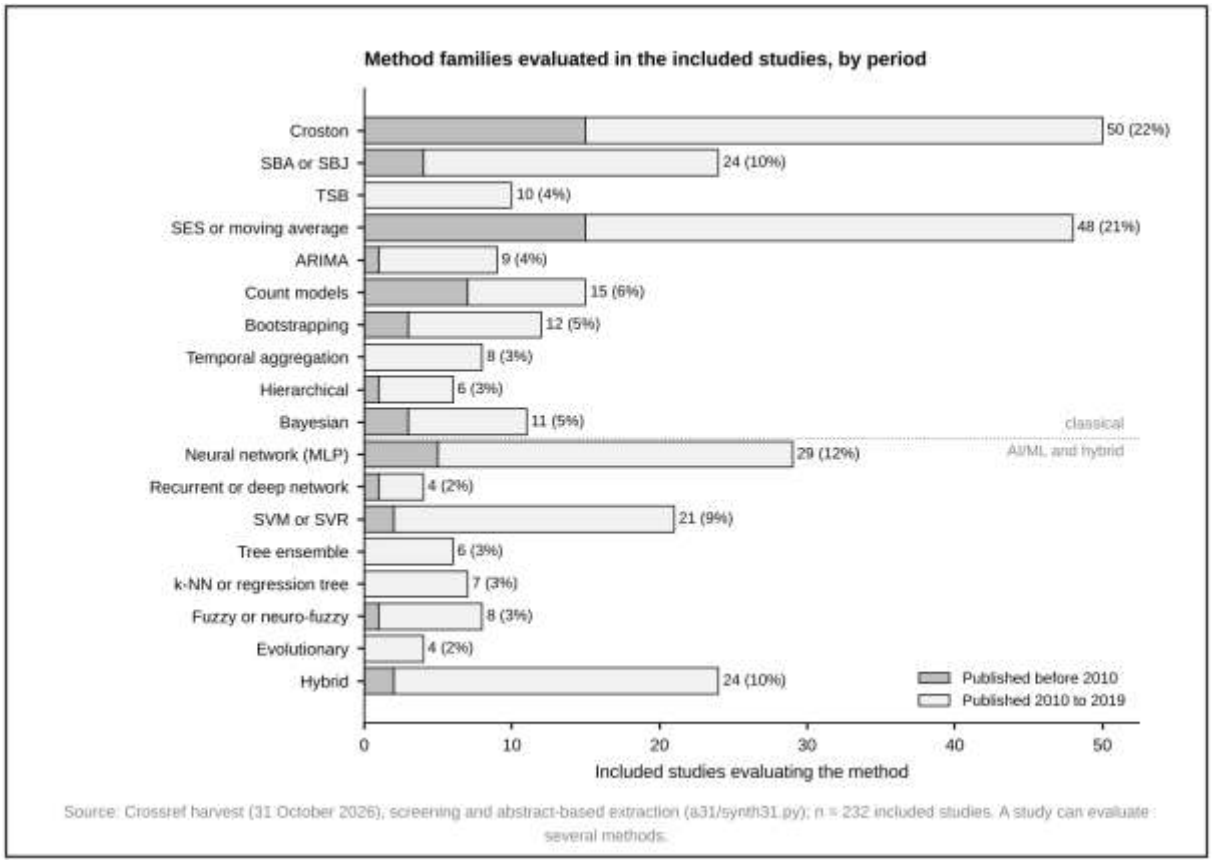
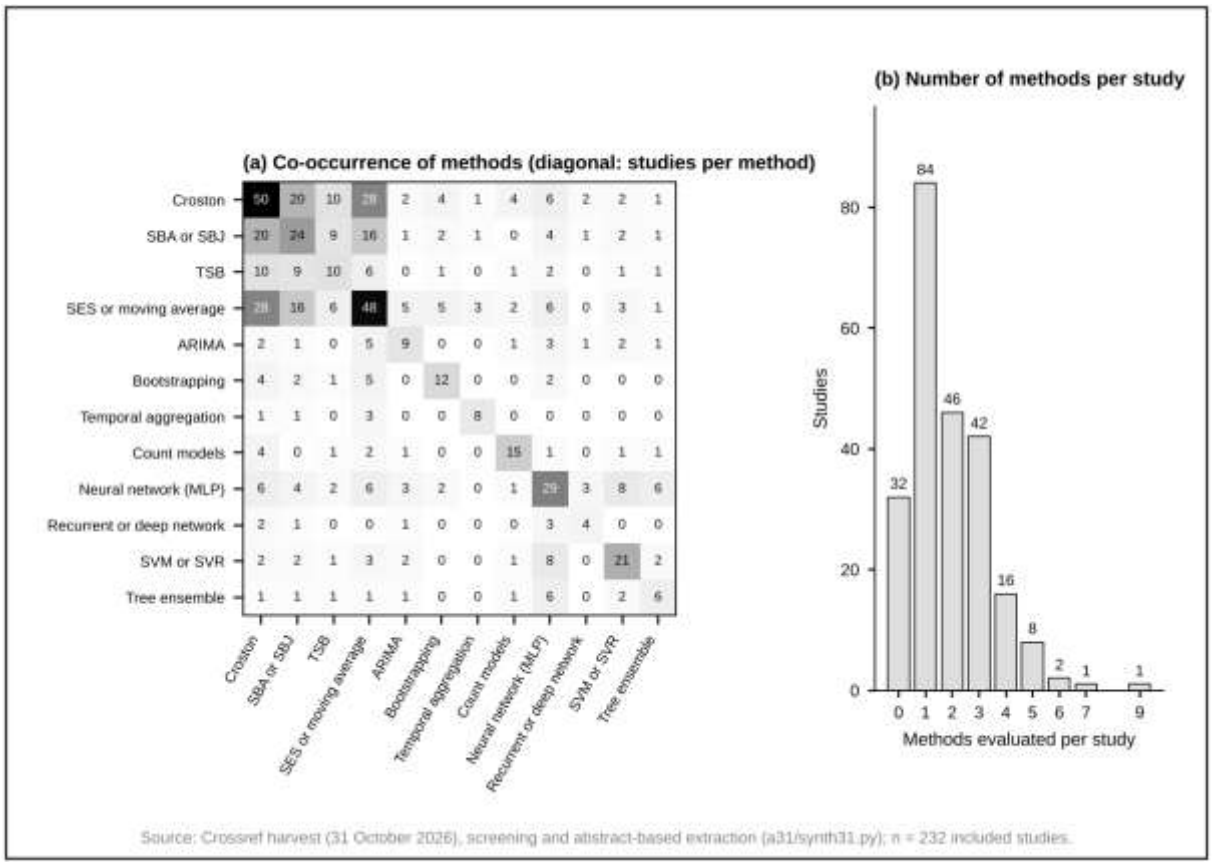


Figure 6 shows how methods were combined within studies. The median study evaluated 2 of the coded method categories; 116 studies evaluated a single category or none that could be coded from the abstract, and 28 evaluated four or more. The co-occurrence matrix in Figure 6(a) shows that Croston's method, SBA and simple exponential smoothing are evaluated together far more often than any AI method is evaluated against them: of the 29 neural-network studies, 6 included Croston's method as a comparator and 4 included SBA or SBJ, while 6 compared the network with simple exponential smoothing or a moving average. Overall, an AI or machine-learning method was compared with at least one classical method in the same study in 23 studies (10%); this is the comparative evidence base on which the benchmark synthesis of RQ3 rests.

Figure 6: Method combinations: (a) co-occurrence of methods within studies, (b) number of coded method categories per study



Evaluation practice: accuracy metrics and cost measures

Tables 9 and 10 and Figure 7 describe how the included studies evaluated their forecasts. On the accuracy side (Table 9), the most frequently reported measures were MSE or RMSE (12 studies), ME or bias (10) and MASE (6). MAPE or sMAPE, which are undefined or unstable on series with zeros (Hyndman & Koehler, 2006; Kim & Kim, 2016), were used by 5 studies and were the only accuracy measure in 3 (1%). Scale-free measures recommended for intermittent series, MASE and other scaled errors, appeared in 7 studies, bias measures (ME or the bias of cumulative errors) in 10, PIS or CFE (Wallström & Segerstedt, 2010) in 1, Theil's U in 0 and relative or percent-better measures (Syntetos & Boylan, 2005; Syntetos et al., 2010) in 2. The evaluation design was stated as rolling origin in 0 studies and as fixed origin in 3, and the forecast horizon was stated in 7 (3%); the remainder did not report these elements in the abstract, which does not mean that they were absent from the full text but does mean that a reader of the abstract cannot judge the strength of the evidence (Tashman, 2000).

Table 9: Accuracy metrics and evaluation design reported in the included studies

Accuracy metric or evaluation element	Studies	Share
MAE or MAD	5	2%
MSE or RMSE	12	5%
MAPE or sMAPE	5	2%
MASE	6	3%
RMSSE or other scaled error	1	0%
ME or bias	10	4%
PIS or CFE	1	0%
Theil's U	0	0%
Relative or percent-better measures	2	1%

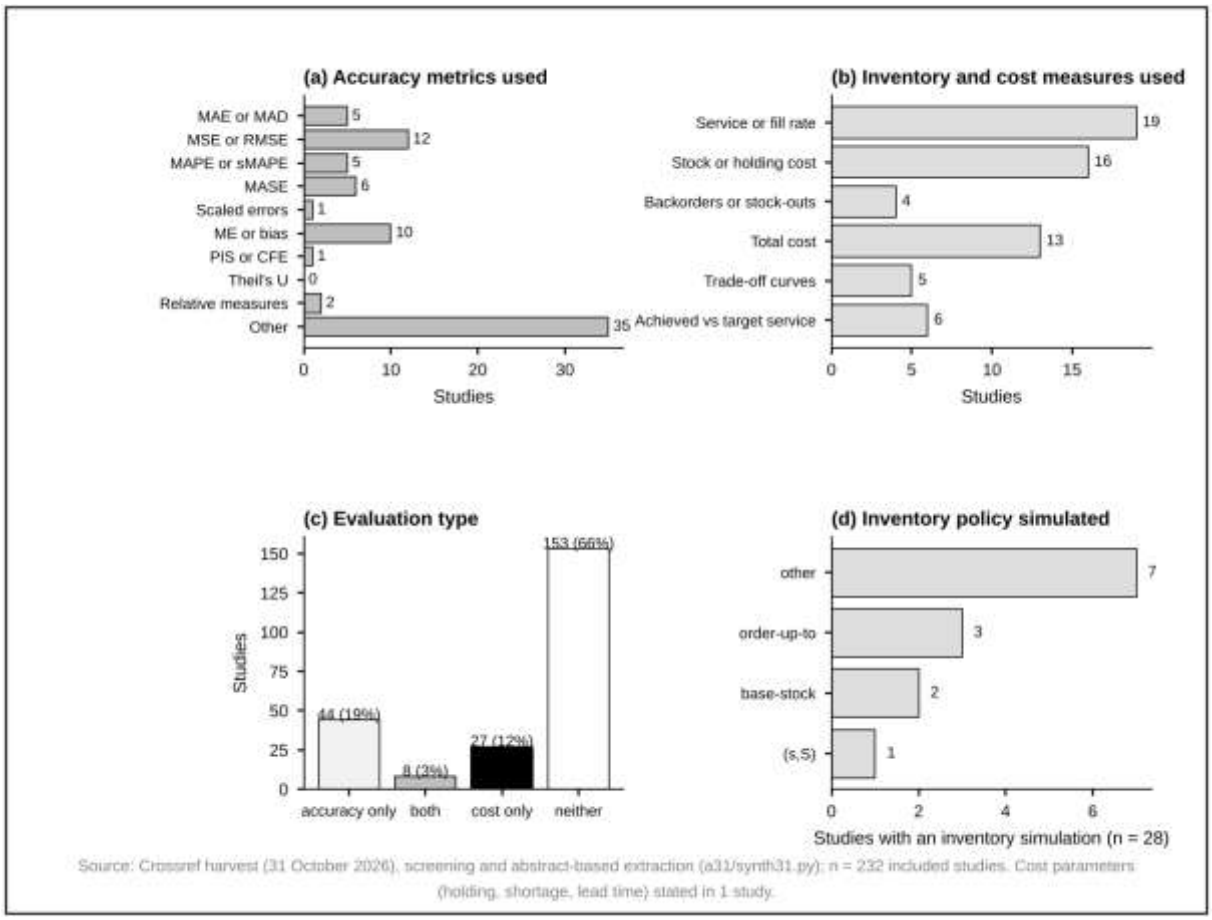
Accuracy metric or evaluation element	Studies	Share
Other	35	15%
None reported	140	60%
MAPE or sMAPE as the only accuracy measure	3	1%
Evaluation design stated: rolling origin	0	0%
Evaluation design stated: fixed origin	3	1%
Forecast horizon stated	7	3%

On the inventory and cost side (Table 10), 35 of the 232 studies (15%) evaluated the forecasts with at least one inventory or cost measure. The most frequent were service level or fill rate (19 studies) and stock or holding cost (16); total cost was reported in 13 studies, trade-off curves of stock against service in 5 and achieved against target service in 6. An inventory simulation that feeds the forecasts into a replenishment policy was performed in 28 studies (12%). A replenishment policy was named in the abstract of 13 studies: order-up-to policies in 3, (s,S) in 1, (s,Q) in 0, base-stock in 2 and other policies in 7. The cost parameters that determine the result (holding rate, shortage or backorder penalty, lead time) were stated in 1 study (<1%). Classified by evaluation type (Figure 7(c)), 44 studies (19%) evaluated accuracy only, 8 (3%) evaluated both accuracy and inventory or cost performance, 27 (12%) evaluated inventory or cost performance only and 153 reported neither in the abstract (reviews and methodological papers for the most part). The share of studies with a cost evaluation was 11% before 2010 and 16% from 2010 onwards, and 15% among the 196 original, case and simulation studies that could have performed one.

Table 10: Inventory and cost measures, inventory simulation, replenishment policies and cost parameters reported in the included studies

Inventory or cost element	Studies	Share
Service level or fill rate	19	8%
Stock or holding cost	16	7%
Backorders or stock-outs	4	2%
Total cost	13	6%
Trade-off curves	5	2%
Achieved versus target service	6	3%
None reported	159	69%
Any inventory or cost measure	35	15%
Inventory simulation performed	28	12%
Policy: not reported	160	69%
Policy: none	59	25%
Policy: other	7	3%
Policy: order-up-to	3	1%
Policy: base-stock	2	1%
Policy: (s,S)	1	0%
Cost parameters stated (holding, shortage, lead time)	1	0%
Evaluation type: accuracy only	44	19%
Evaluation type: both accuracy and cost	8	3%
Evaluation type: cost only	27	12%

Figure 7: Evaluation practice: (a) accuracy metrics, (b) inventory and cost measures, (c) evaluation type, (d) replenishment policies named



Accuracy against cost: the literature-based benchmark (RQ3)

Table 11 and Figure 8 assemble the comparative evidence. Of the 23 studies that compared at least one AI or machine-learning method with at least one classical method, the family reported best under accuracy metrics was AI or machine learning in 14 (61%), classical in 2 (9%), a hybrid in 4 and mixed or without a clear winner in 1; 2 of these studies reported no accuracy verdict in the abstract. Under inventory or cost measures the picture is thinner because fewer of the comparative studies performed such an evaluation: the family reported best was AI or machine learning in 3 studies, classical in 0, a hybrid in 0 and mixed in 0, while 20 of the 23 comparative studies (87%) reported no inventory or cost verdict at all. This asymmetry is the central empirical result of the review: the literature that claims accuracy gains for AI methods on intermittent series is, for the most part, a literature that has not asked whether those gains survive the inventory stage.

Table 11: Comparative evidence on AI and machine-learning methods against classical methods: family reported best under accuracy metrics and under inventory or cost measures

Comparative evidence	Under accuracy metrics	Under inventory or cost measures	Share
Studies comparing AI/ML with classical methods	23		
Family reported best: AI/ML	14 (61%)	3 (13%)	
Family reported best: classical	2 (9%)	0 (0%)	
Family reported best: hybrid	4 (17%)	0 (0%)	
Family reported best: mixed or no clear winner	1 (4%)	0 (0%)	
Family reported best: not applicable (no such evaluation)	2 (9%)	20 (87%)	

Comparative evidence	Under accuracy metrics	Under inventory or cost measures	Share
Studies reporting both verdicts	3		
Rankings disagree (of those reporting both)	1		33%
Best under accuracy: AI/ML; best under cost: AI/ML 2			
Best under accuracy: classical; best under cost: AI/ML 1			
Best under accuracy: hybrid; best under cost:			
Best under accuracy: mixed or no clear winner; best under cost:			

Only 3 studies reported both an accuracy verdict and an inventory or cost verdict for an AI-against-classical comparison. In 1 of them (1 of 3) the two rankings disagreed: the method that was best under the accuracy metric was not the method that produced the lowest cost or the best service, or the cost evaluation found no material difference where the accuracy metric had found one. The cross-tabulation in Table 11 and Figure 8(b) shows the pattern of agreement: the diagonal cells (same family best under both criteria) hold 2 studies, and the off-diagonal cells 1. These numbers are small, and the review does not treat them as an estimate of the frequency of disagreement in general; what they establish is that disagreement between the two criteria is a documented and recurring phenomenon, as the earlier stock-control studies of the Croston family had already shown for classical methods (Syntetos & Boylan, 2006; Teunter et al., 2011; Syntetos et al., 2010), and that no study in the corpus has demonstrated an AI advantage that is both accuracy-based and cost-based on a wholesale dataset.

Figure 8: Benchmark synthesis: (a) family reported best, by evaluation type, (b) agreement between the accuracy-based and the cost-based verdict

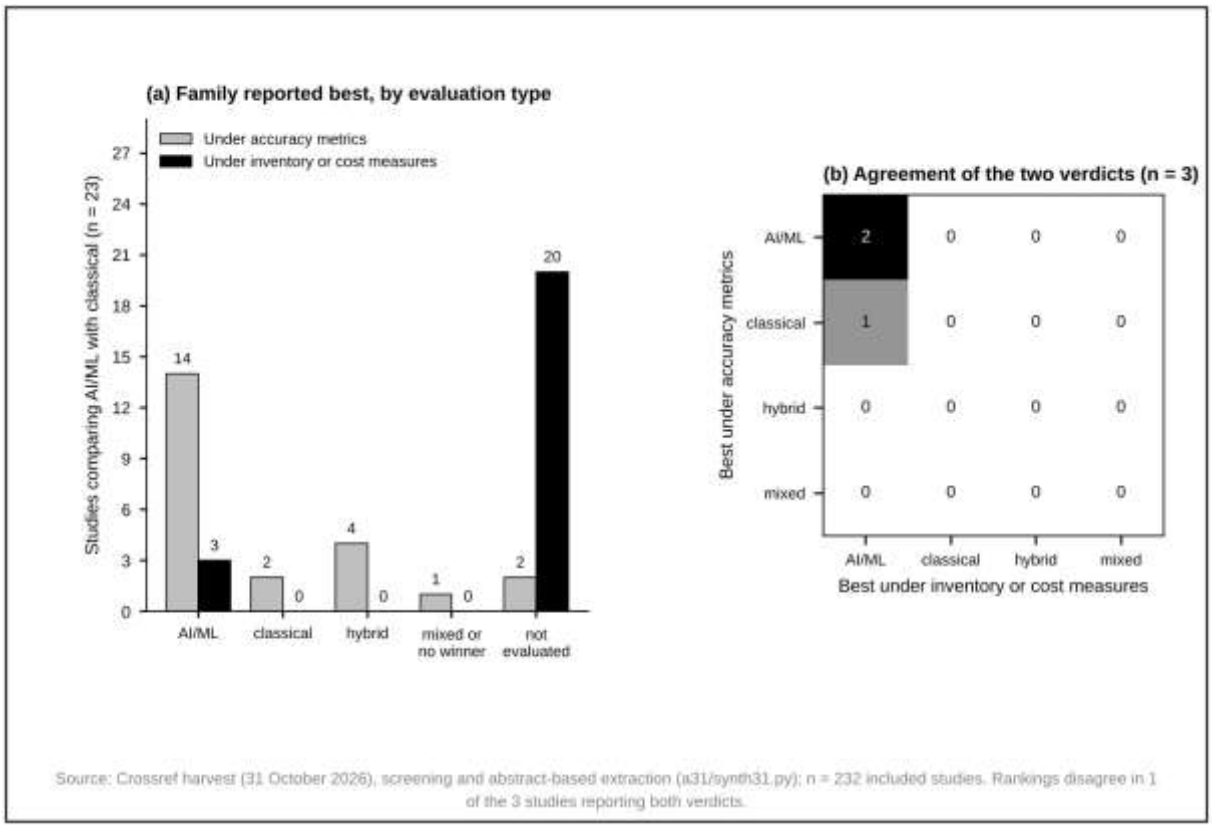


Table 12 lists every study with a cost-based verdict, including classical-only comparisons, with the data, the number of series, the policy and the method reported best. Across all 31 studies with such a verdict, the winning family was classical in 18 and AI/ML in 8. The methods named as best are heterogeneous:

no single method was named in more than 2 studies, and Croston variants, bootstrapping methods and hybrid and support vector schemes all appear among the winners. Two features of Table 12 matter for the protocol of Section 5. First, the magnitudes reported are heterogeneous in unit (percentage of holding cost, service-level points, stock-days, total cost index) and in baseline, so no pooled estimate is attempted; the review reports the direction and the study's own number. Second, the winners are conditional on the policy and the cost parameters that the study assumed, which Table 10 showed to be stated in only 1 study; a method that wins under an order-up-to policy with a 95 per cent target may not win under an (s,Q) policy with a shortage penalty, and the literature does not yet permit a reader to separate the effect of the method from the effect of the setting.

Table 12: Studies reporting a cost-based verdict: data, number of series, replenishment policy, method reported best under inventory or cost measures and reported magnitude

Study	Domain	Data	Series	Policy	Best method under cost measures	Family	Reported magnitude
Gardner (1990)	wholesale or distribution	NR	NR	NR	NR	mixed	NR
Eaves and Kingsman (2004)	aerospace or defence	UK Royal Air Force consumable spare part	NR	NR	Approximation method (Syntetos-Boylan modific	classical	significant reductions in value of stock-holdings required to attain a
Syntetos and Boylan (2006)	NR	empirical dataset of 3,000 SKUs (company	500-4999	NR	authors' new intermittent demand method	classical	NR
Syntetos and Boylan (2008)	aerospace or defence	Royal Air Force (UK) spare parts, 753 in	500-4999	NR	NR (adjusted forecasts improve parametric sto	classical	NR
Teunter and Duncan (2009)	aerospace or defence	UK Royal Air Force spare parts (large da	NR	NR	Croston's method (and variant) and bootstrapp	classical	NR
Syntetos et al. (2009a)	aerospace or defence	Royal Air Force (UK) large demand datase	NR	order-up-to	proposed modified forecasting/order-up-to app	classical	NR
Mitchell (2010)	NR	NR	NR	NR	NR	mixed	all combinations replaced demanded units an average of 10 weeks before
Bala (2010)	spare or service parts	NR	NR	NR	purchase-driven decision-tree forecasting mod	AI/ML	NR
Teunter et al. (2010)	aerospace or defence	Royal Air Force (RAF) data	NR	order-up-to	proposed compound binomial order-up-to method	classical	NR

Study	Domain	Data	Series	Policy	Best method under cost measures	Family	Reported magnitude
Babai et al. (2012)	aerospace or defence	Royal Air Force (UK) large demand datase	NR	NR	temporal aggregation approach	classical	NR
Rahman and Sarker (2012)	other	US Census Bureau time series of a season	<50	other	Bayesian model	classical	NR
Kourentzes (2013)	NR	NR	NR	NR	neural networks (on service level); best Cros	AI/ML	NR
Moon (2013)	aerospace or defence	South Korean navy spare parts demand	NR	NR	ANN (classification error, inventory cost); l	AI/ML	NR
Song et al. (2013)	utilities or transport	port critical spare parts, real case stu	NR	NR	PSO-LSSVM	AI/ML	NR
Kourentzes (2014)	automotive	3,000 automotive spare-parts series of 2	500-4999	order-up-to	MAR and MSR cost functions for optimising Cro	classical	MAR/MSR give smoothing parameters mostly 0.05-0.2 versus higher unstab
Verhagen and Curran (2014)	aerospace or defence	aircraft maintenance operator, 9 LRUs	<50	NR	Homogeneous Poisson Process model	classical	NR
Song et al. (2014)	utilities or transport	port critical spare parts, real case stu	NR	NR	PSO-LSSVM	AI/ML	NR
Lindsey and Pavur (2014)	spare or service parts	simulated spare parts with periodic repl	NR	NR	Bayesian model (lower inventory cost only whe	mixed	NR
Petropoulos et al. (2016)	spare or service parts	two spare-parts datasets (automotive and	5000+	NR	proposed volume-based aggregation framework	classical	NR
Lei et al. (2017)	utilities or transport	State Grid Corporation of China procurem	NR	NR	multiple temporal aggregation-disaggregation	classical	NR
Solis et al. (2017)	NR	NR	NR	NR	bias-corrected Croston (exact and approximate	mixed	NR
Mobarakeh et al. (2017)	aerospace or defence	Dassault Aviation non-contract business-	NR	NR	sliding-window bootstrapping with Mean functi	classical	proposed method outperformed the others for 75% of spare parts; 'signi

Study	Domain	Data	Series	Policy	Best method under cost measures	Family	Reported magnitude
Sulistyo and Sutrisno (2018)	spare or service parts	oil and gas company spare parts	NR	NR	artificial neural network	AI/ML	ANN MASE 0.761; average inventory reduced 11.9%; service level increas
Boukhtouta and Jentsch (2018)	aerospace or defence	Canadian Armed Forces spare parts (40 se	<50	NR	SVM	AI/ML	NR
Kitti-udomporn et al. (2018)	spare or service parts	case company purchasing shelf-life-limit	NR	other	NR (lowest-MAD method among the five not name	classical	total inventory cost reduced by 19.08% versus current purchasing
Babajanivalashedi et al. (2018)	aerospace or defence	business aircraft manufacturer spare par	NR	NR	neural network (best for 98% of demand); rand	AI/ML	neural network best for 98% of demand, random forest for 2% of parts (
Hasni et al. (2019b)	spare or service parts	spare parts demand data, more than 9,000	5000+	NR	adjusted bootstrapping methods	classical	NR
Kalaya et al. (2019)	pharmaceutical or healthcare	community hospital Mae Hong Son, 2 vital	<50	other	Croston and TSB lowest MSE; proposed Exponent	classical	NR
Babai et al. (2019)	spare or service parts	simulated series with linear and sudden	NR	NR	new modified SBA/TSB method (outperforms Cros	classical	NR (no numbers in abstract)
Hasni et al. (2019c)	NR	generated demand distributions plus empi	NR	NR	parametric approach for highly intermittent d	mixed	NR
Van der Auweraer and Boute (2019)	spare or service parts	NR	NR	base-stock	proposed service-maintenance-information meth	classical	NR

Evidence gaps and reporting quality (RQ4)

Table 13 and Figure 9 map the evidence by method family and by application domain against the evaluation type. The Croston-type family has the most complete evidence, with 3 studies evaluating both accuracy and cost and 5 cost only, against 20 accuracy only. The neural-network family (MLP, recurrent and deep) has 3 studies with both evaluations against 6 with accuracy only, and the SVR, tree and nearest-neighbour family 2 against 8. The gap by domain follows from the first finding of RQ1: the wholesale or distribution domain contributes 0 studies with both evaluations, 1 with accuracy only and 1 with cost only. Demand classification schemes, which the protocol places at the start of any benchmark, were used in 32 studies: the ADI-CV2 scheme (Syntetos et al., 2005; Kostenko & Hyndman, 2006) in 5, ABC-XYZ or similar value-variability schemes in 2 and other schemes in 25; the level of intermittence of the data (share of zero periods, average inter-demand interval or squared coefficient

of variation) was reported in 1 study (<1%).

Table 13: Evidence map: demand classification, intermittence reporting and the distribution of studies by method family and evaluation type

Data characteristic or evidence cell	Studies	Share
Demand classification used: not reported	174	75%
Demand classification used: none	26	11%
Demand classification used: other	25	11%
Demand classification used: ADI-CV2	5	2%
Demand classification used: ABC-XYZ	2	1%
Intermittence level reported (share of zeros, ADI, CV2)	1	0%
Evaluation type: neither	153	66%
Evaluation type: accuracy only	44	19%
Evaluation type: cost only	27	12%
Evaluation type: both	8	3%
Family Croston-type (Croston, SBA, TSB): accuracy only / both / cost only / neither	20 / 3 / 5 / 26	
Family Smoothing, ARIMA, count models: accuracy only / both / cost only / neither	24 / 5 / 9 / 33	
Family Bootstrapping and aggregation: accuracy only / both / cost only / neither	4 / 2 / 7 / 12	
Family Neural networks (MLP, RNN, deep): accuracy only / both / cost only / neither	6 / 3 / 0 / 21	
Family SVR, trees, k-NN: accuracy only / both / cost only / neither	8 / 2 / 3 / 16	
Family Fuzzy, evolutionary, hybrid: accuracy only / both / cost only / neither	6 / 0 / 2 / 22	

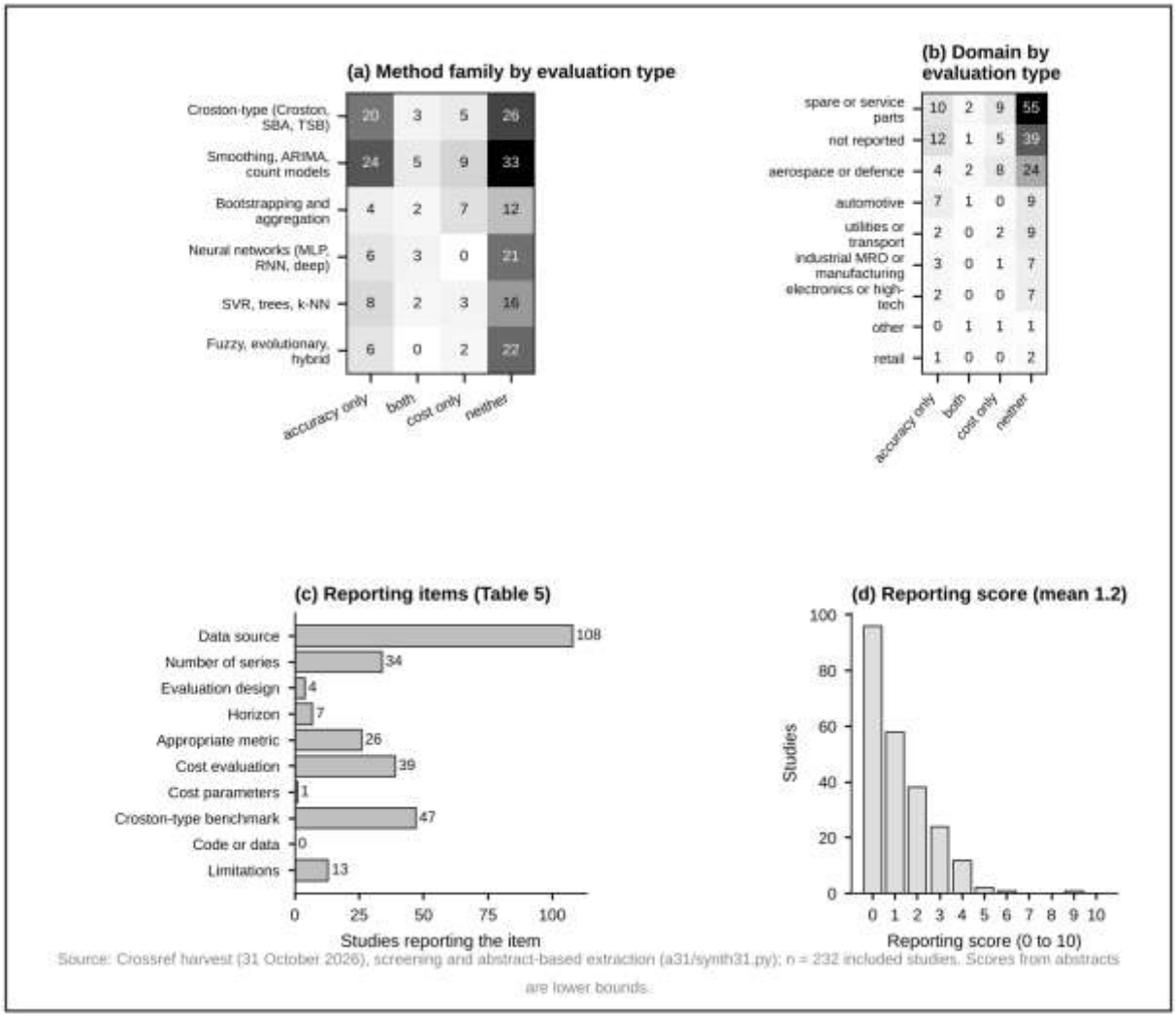
Table 14 and Figure 9(c) and (d) report the ten-item reporting score. The items most often satisfied were data source described (108 studies, 47%) and Croston-type benchmark included (47, 20%); the items least often satisfied were code or data availability (0 studies) and cost parameters stated (1).replace('_', ' ')}} (1). The mean score was 1.2 of 10 across all studies and 1.5 among the studies whose extraction rests on an abstract or full text; 50 studies (22%) were scored from the title alone and necessarily score at or near zero. Limitations were stated in the abstract of 13 studies (6%). These scores are lower bounds, because abstracts omit much that full texts contain; their value lies in the comparison between items rather than in their level. On that comparison, the two items that the cost-based benchmarking protocol of Section 5 depends on most, the performance of an inventory or cost evaluation (item 6, 39 studies) and the statement of the cost parameters (item 7, 1 study), are among the least often reported, and the availability of code or data (item 9) is reported by 0 studies, so the comparisons of Table 12 cannot be reproduced from the published record alone.

Table 14: Reporting-quality items satisfied by the included studies (abstract-based scoring)

Reporting item	Studies reporting	Share
1 Data source described	108	47%
2 Number of series stated	34	15%
3 Evaluation design stated	4	2%
4 Forecast horizon stated	7	3%
5 Intermittent-appropriate accuracy metric	26	11%

Reporting item	Studies reporting	Share
6 Inventory or cost evaluation performed	39	17%
7 Cost parameters stated	1	0%
8 Croston-type benchmark included	47	20%
9 Code or data availability	0	0%
10 Limitations stated	13	6%
Mean reporting score (all studies)	1.2	
Mean reporting score (abstract-based studies)	1.5	
Studies decided on the title alone	50	22%

Figure 9: Evidence gaps and reporting: (a) evidence map by method family and evaluation type, (b) evidence map by domain and evaluation type, (c) reporting items satisfied, (d) distribution of the reporting score



A COST-BASED BENCHMARKING PROTOCOL FOR WHOLESALE DISTRIBUTION

From the evidence to the protocol

The framework of Figure 1 organised the coding; this section turns the coded evidence into a benchmarking protocol that a wholesale distributor can run on its own data, stating for each step what the literature supports, what it leaves open and what the inventory-forecasting literature prescribes (Table 15). Four findings shape the protocol. First, the evidence base is large at the level of methods and

thin at the level of cost: 15% of the included studies evaluate forecasts with an inventory or cost measure, and <1% state the cost parameters that such an evaluation needs. Second, where both evaluation types are reported, the family judged best differs in 1 of 3 of the comparisons, so that an accuracy-based benchmark is not a reliable guide to a cost-based decision. Third, the Croston-type estimators remain the reference methods and are included as benchmarks in 20% of the studies, so that a benchmark without them is not comparable with the literature. Fourth, wholesale distribution is studied by name in 5 studies, and the characteristics that matter for it (assortment scale, service commitments, cost structure) are reported unevenly, so that the protocol must create its own evidence rather than transfer it.

Table 15: Steps of the cost-based benchmarking protocol, with the evidence of the review that supports each step, the open questions it leaves and the prescription adopted

Protocol step	Evidence of the review	Open question	Prescription adopted
1 Data and classification	Intermittence level reported by 0% of studies; ADI-CV2 scheme used by 2%; method ranking depends on the class	Class cut-offs for a given policy and cost structure	Classify every SKU by ADI and CV2; stratify the benchmark by class; weight by class share
2 Reference set	Croston-type benchmark included by 20% of studies; SBA and TSB established as references	Parameter optimisation on intermittent series	SES, Croston, SBA, TSB and a bootstrap; parameters fitted on the training window
3 Candidate set	AI/ML evaluated in 25% of studies; neural networks dominant; tree ensembles and deep networks late and rare	Cross-learning design; computational cost	MLP, tree ensemble, RNN and hybrids tuned on the training window; cross-learning and cost recorded
4 Rolling-origin accuracy	Evaluation design stated by 2%; horizon by 3%; MAPE-type measures alone in 1%	Window lengths for short histories	Rolling origins over one demand cycle; horizon = lead time plus review period; MASE, bias, PIS
5 Inventory simulation	Inventory or cost evaluation in 15%; cost parameters stated in 0%; rankings differ in 1 of the 3 studies reporting both	Policy and cost structure of the distributor	Order-up-to policy with stated lead time, target fill rate, holding and shortage cost; service, stock, cost and trade-off curves per class
6 Decision and monitoring	Monitoring after adoption rarely reported; obsolescence addressed by TSB-type methods	Retraining frequency; reclassification triggers	Method per class; monthly bias and service charts; reclassification; benchmark repeated on change
Reporting checklist	Mean reporting score 1.2 of 10; code or data availability 0%	Full-text reporting unknown	Ten items of Table 5 as the minimum content of a benchmark report

Step 1, data and classification, fixes the benchmark dataset: a demand history per SKU and location at the period length of the review cycle, with lost sales and returns separated where the system records them, and the ADI and CV2 of every series computed and the assortment categorised into smooth, erratic, intermittent and lumpy classes (Syntetos et al., 2005; Kostenko & Hyndman, 2006; Heinecke et al., 2013). The classification is the stratification variable of the benchmark: methods are compared within classes, because the evidence shows that the ranking depends on the class, and the class shares of the assortment weight the results (Boylan et al., 2008; Petropoulos et al., 2014; Lengu et al., 2014). The review found that <1% of the studies report the level of intermittence of their data and 2% use the ADI-CV2 scheme; the protocol requires both, so that the distributor's results can be placed against the literature.

Step 2, the reference set, consists of simple exponential smoothing, Croston, SBA, TSB and a bootstrap

of lead-time demand, with parameters optimised on the training window only (Croston, 1972; Syntetos & Boylan, 2005; Teunter et al., 2011; Willemain et al., 2004; Kourentzes, 2014). The reference set is the part of the protocol that the evidence supports best: these methods are the benchmarks of the literature, their relative merits by demand class are established, and any candidate must beat them under the cost ranking to justify its complexity (Syntetos & Boylan, 2006; Teunter & Duncan, 2009; Babai et al., 2014; Syntetos et al., 2015).

Step 3, the candidate set, adds the AI/ML methods that the evidence shows to be most often evaluated and most often reported best: multilayer perceptrons with lagged demand and elapsed-time inputs, tree ensembles on feature vectors that pool information across SKUs, and recurrent networks where the series are long enough, together with hybrids and combinations (Gutierrez et al., 2008; Kourentzes, 2013; Lolli et al., 2017; Makridakis et al., 2018a; Petropoulos & Kourentzes, 2015). The protocol requires that every candidate be tuned on the training window, that cross-learning be stated (which SKUs share a model), and that the computational cost be recorded, because the evidence rarely reports it.

Step 4, rolling-origin evaluation and accuracy metrics, follows the evaluation literature: forecasts are produced from several origins over a hold-out period that spans at least one demand cycle, for the horizon that equals the lead time plus the review period, and accuracy is summarised with scale-free and bias measures (MASE, mean error, periods in stock) rather than percentage errors (Tashman, 2000; Hyndman & Koehler, 2006; Wallström & Segerstedt, 2010; Prestwich et al., 2014; Kim & Kim, 2016). The review found that 2% of the studies state their evaluation design and 1% rely on MAPE-type measures alone; the protocol makes the design and the measures explicit so that the accuracy step is reproducible.

Step 5, inventory simulation, is the step that the evidence shows to be most often missing and that decides the benchmark. Each method's forecasts and forecast-error distribution feed a periodic-review order-up-to policy (or the distributor's actual policy), with the lead time, the review period, the target fill rate, the holding-cost rate and the shortage or backorder cost stated; the simulation reports, per method and class, the achieved service, the average stock, the holding and shortage cost, and the trade-off curve of service against stock obtained by varying the target (Gardner, 1990; Sani & Kingsman, 1997; Syntetos & Boylan, 2006; Syntetos et al., 2010; Tiacci & Saetta, 2009; Silver et al., 2016). The ranking of methods by total cost at equal service is the verdict of the benchmark; the accuracy ranking of step 4 is reported beside it and the agreement between them is recorded, because the review shows how often they differ.

Step 6, decision and monitoring, adopts a method per demand class, documents its parameters, and monitors bias and service monthly with the control-chart logic of the quality literature; SKUs are reclassified as their demand changes and the benchmark is repeated when the assortment, the lead times or the cost parameters change (Teunter et al., 2011; Babai et al., 2014; Fildes & Kingsman, 2011). The step closes the loop of Figure 1 and gives the distributor the evidence that the literature does not provide for its own assortment.

Workflow and decision logic

Figure 10 gives the protocol as a workflow across five organisational lanes. Demand planning selects the SKU sample by class and value and decides on the method per class at the end; data and IT extract and freeze the benchmark dataset; analytics run the reference and candidate methods and the rolling-origin evaluation; inventory control runs the simulation under the stated policy and computes service and cost; and management and finance set the holding rate, the shortage cost and the service targets, and approve the adoption. The gateway in the inventory-control lane is the agreement check between the cost ranking and the accuracy ranking: disagreement sends the analysis back to analytics for diagnosis (which classes, which horizons, which service levels drive the difference) before a decision is taken. The workflow puts the statistics and machine-learning work in one lane and the inventory evaluation in another so that the cost evaluation cannot be skipped: the review found it missing from most of the published comparisons.

Figure 10: Cost-based benchmarking workflow for a wholesale distributor in BPMN form: five organisational lanes, gateways and message flows

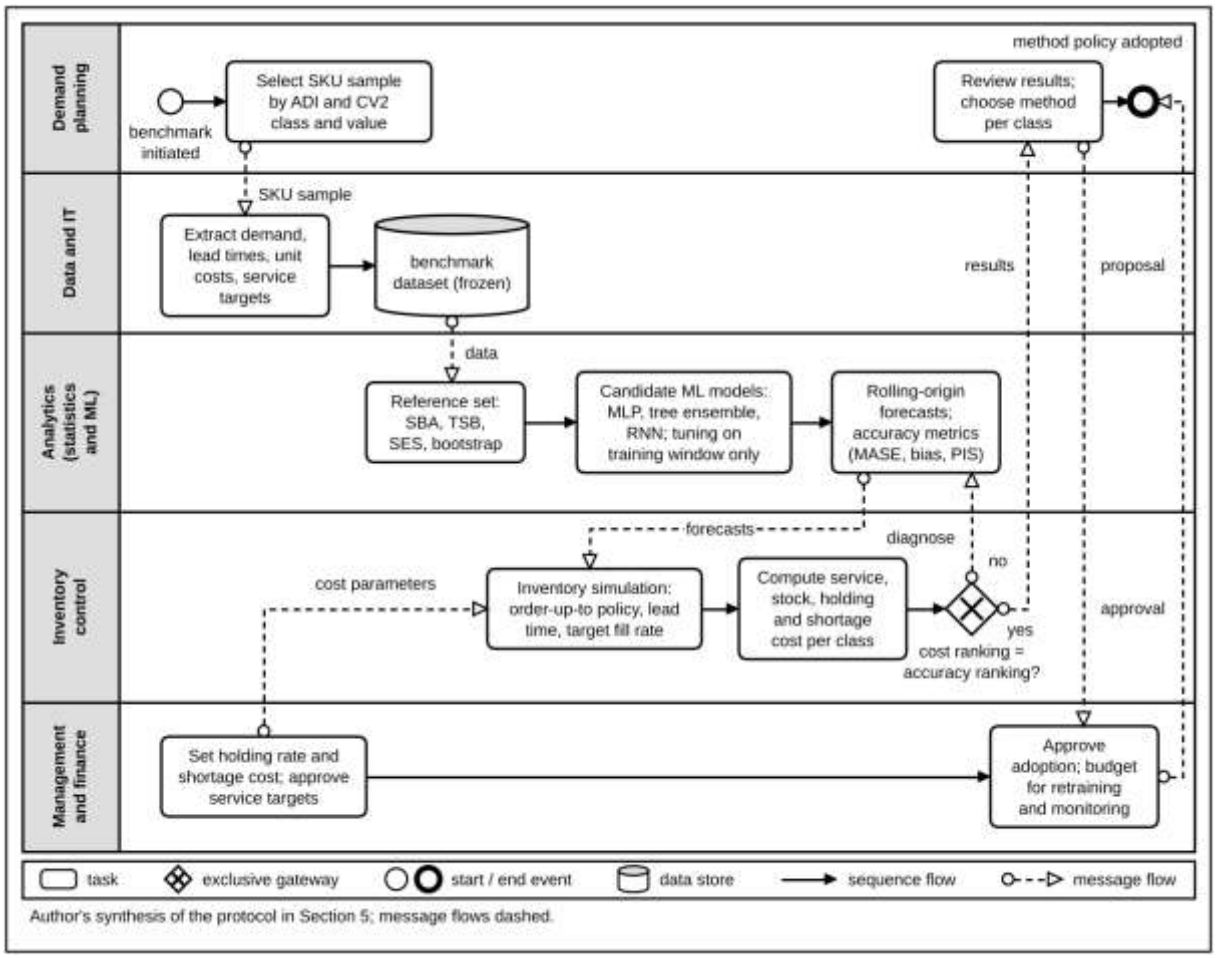
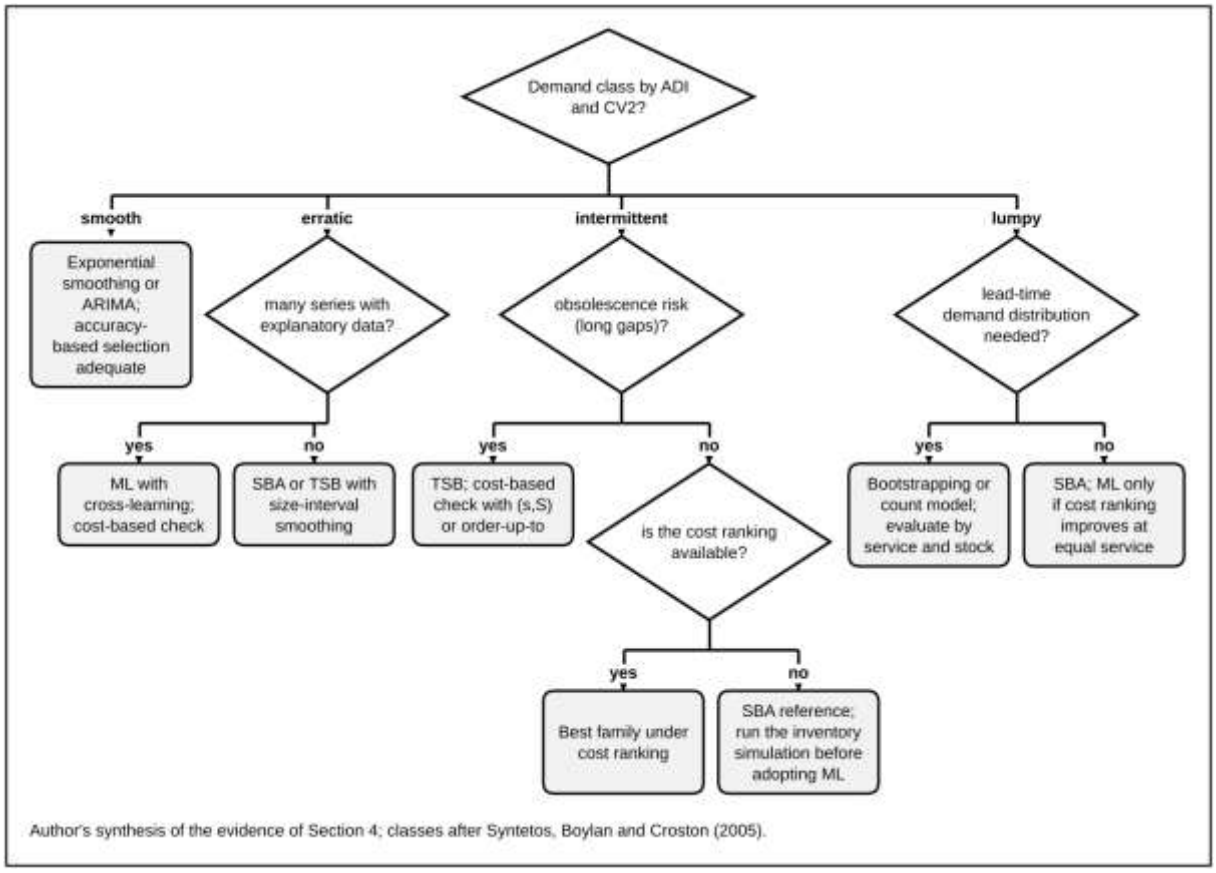


Figure 11 gives the decision logic that a demand planner applies per demand class. For smooth demand, accuracy-based selection among smoothing and ARIMA methods is adequate, because with few zero periods the accuracy and cost rankings are expected to agree more closely. For erratic demand, the question is whether there are many series with explanatory data, in which case ML with cross-learning is a candidate and a cost-based check is required; otherwise SBA or TSB with size-interval smoothing is the choice. For intermittent demand, obsolescence risk (long gaps) points to TSB and a cost-based check under the actual policy; without obsolescence risk, the planner adopts the family that the cost ranking favours if a cost ranking is available and runs the inventory simulation before adopting ML if it is not. For lumpy demand, the need for a lead-time demand distribution points to bootstrapping or a count model evaluated by service and stock, and otherwise to SBA with ML adopted only if the cost ranking improves at equal service. The tree encodes the evidence of Section 4 in the order in which the evidence is strongest: class first, policy second, family last.

Figure 11: Decision tree for method selection by demand class under cost-based evaluation



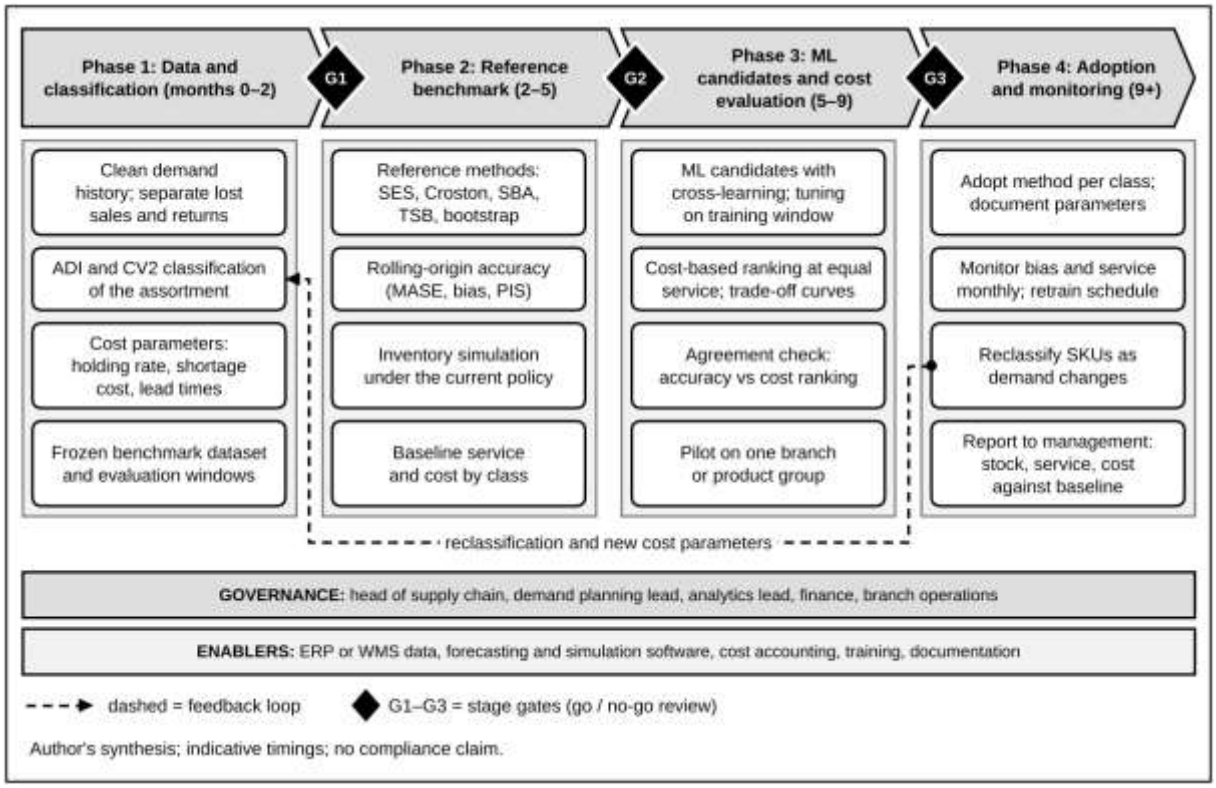
Elements specific to wholesale distribution

Four features of wholesale distribution are built into the protocol. First, scale: a distributor benchmarks thousands of SKUs, so the sample of step 1 is stratified by class and value and the methods are run in batch with stated computational cost; among the studies that state the number of series, 21 of 39 evaluate fewer than 500, so that the distributor's benchmark will be larger than the literature's and the cross-learning candidates matter more. Second, the cost structure: holding rates, shortage costs and lead times differ between product lines and suppliers, so the cost parameters of step 5 are set per product line by finance rather than taken from the literature, where they are stated in <1% of the studies. Third, the service commitment: distributors sell availability, so the trade-off curve of service against stock is the output that management reads, and the benchmark verdict is stated at the committed fill rate. Fourth, the supplier interface: the distributor's orders are its suppliers' demand, so the bias of the adopted method is monitored not only for its stock effect but for the order variability that it passes upstream, which the collaborative arrangements of the sector are designed to reduce (Lee et al., 1997; Chen et al., 2000; Cachon & Fisher, 2000; Disney & Towill, 2003).

Implementation roadmap and reporting

Figure 12 gives an implementation roadmap in four phases with three stage gates. Phase 1 builds the data foundation: cleaned demand histories, the ADI-CV2 classification of the assortment, cost parameters per product line, and a frozen benchmark dataset with evaluation windows. Phase 2 runs the reference benchmark: the classical set, rolling-origin accuracy, the inventory simulation under the current policy, and a baseline of service and cost by class. Phase 3 adds the ML candidates with cross-learning, the cost-based ranking at equal service with trade-off curves, the agreement check between accuracy and cost rankings, and a pilot on one branch or product group. Phase 4 adopts the method per class, monitors bias and service monthly with a retraining schedule, reclassifies SKUs as demand changes and reports stock, service and cost against the baseline. The dashed feedback from phase 4 to phase 1 is the learning path: reclassification and new cost parameters restart the benchmark. The timings are indicative.

Figure 12: Implementation roadmap for cost-based benchmarking of intermittent demand forecasting in a wholesale distributor: four phases, three stage gates, governance and enablers



Finally, the review supports a reporting checklist for benchmark studies, because the reporting score of Section 4.6 shows that the published evidence rarely states the items that would make a benchmark reusable: the ten items of Table 5 (data source; number of series; evaluation design; horizon; an intermittent-appropriate accuracy metric; inventory or cost evaluation; cost parameters; a Croston-type benchmark; code or data availability; limitations) are proposed as the minimum content of a benchmark report, whether published or internal. A distributor that reports its benchmark against the checklist can compare its results with the literature, and a study that reports against it adds to an evidence base that, as it stands, cannot be aggregated.

DISCUSSION

Principal findings against the research questions

The review set out to establish what the pre-2020 literature can tell a wholesale distributor about forecasting intermittent demand with AI and machine-learning methods, and to do so on the criterion that matters to the distributor, inventory cost, rather than on accuracy statistics alone. Four findings answer the research questions. First, on volume and distribution (RQ1), the literature is substantial, as the earlier narrative reviews indicated (Boylan & Syntetos, 2010; Syntetos et al., 2009b; Bacchetti & Saccani, 2012), but narrow in setting: 232 studies met the criteria, 76% of them published from 2010 onwards, concentrated in the operational research and forecasting journals and in spare-parts, maintenance and after-sales environments; wholesale or distribution settings were named in 5 studies and the datasets were small relative to a distributor's assortment. Second, on methods (RQ2), the Croston family and simple smoothing remained the reference set throughout the period, neural networks were the dominant AI family and were evaluated well before the deep-learning period, and tree ensembles, recurrent networks and deep architectures entered the literature late and in small numbers; AI and classical methods were compared head-to-head in 23 studies. Third, on evaluation (RQ3), 15% of the studies evaluated forecasts with an inventory or cost measure, cost parameters were stated in <1%, and among the comparative studies that reported both an accuracy and a cost verdict the two rankings disagreed in 1 of 3 cases; AI methods were reported best under accuracy metrics more often than classical methods, but under cost measures the comparative evidence is too sparse (3 studies

with both verdicts) to support a general ranking in either direction. Fourth, on gaps and reporting (RQ4), the evidence map is densest for Croston-type methods evaluated on accuracy and cost and sparsest for the AI families evaluated on cost; the reporting score averaged 1.2 of 10 on abstracts, with the inventory evaluation, the cost parameters and the availability of code or data the least reported items.

The disagreement finding deserves emphasis because it is easy to misread. It does not show that AI methods are worse for inventory control than classical methods, nor that accuracy metrics are useless; it shows that the accuracy-to-cost relationship is not monotone for intermittent series, which the classical stock-control literature established for the Croston family two decades ago (Syntetos & Boylan, 2006; Teunter et al., 2011; Syntetos et al., 2010) and which the AI literature has largely not tested. An estimator that minimises a squared-error criterion on a series that is zero in most periods is rewarded for forecasting close to zero; an estimator that is slightly biased upwards may cost more in stock but save much more in shortages when the policy converts the forecast into an order-up-to level (Syntetos & Boylan, 2001; Teunter et al., 2011; Prestwich et al., 2014). Which effect dominates depends on the cost ratio, the lead time and the demand class, none of which an accuracy metric knows about. The practical consequence for a distributor is that a vendor's claim of "x per cent better accuracy" for an AI forecasting module cannot be converted into a stock or service outcome without the inventory simulation that Section 5 prescribes.

Implications for wholesale distribution practice

For the U.S. wholesale sector, whose scale and distinct inventory behaviour were set out in Section 2.4 (U.S. Census Bureau, 2019; Chen et al., 2007; Rumyantsev & Netessine, 2007), the review carries three implications. The first is that the evidence base for AI-driven intermittent demand forecasting in distribution is thin, so a distributor should expect to generate its own evidence rather than inherit it: the 5 studies that name a wholesale setting, together with the 3 retail studies, are the only direct evidence, and they do not cover the range of demand classes, lead times and cost structures that a full-line distributor faces. The second is that the classical reference set is inexpensive to run and well understood, so any AI candidate should be required to beat Croston, SBA and TSB on cost measures under the distributor's own policy before it is adopted, exactly as the benchmark protocol specifies; the literature shows AI methods winning accuracy comparisons often enough that this test is worth running, while the comparative cost evidence (3 studies with both verdicts) is too thin for the test to be skipped. The third is that demand classification is the first decision, not an afterthought: the method ranking depends on the class (Syntetos et al., 2005; Boylan et al., 2008; Petropoulos et al., 2014), and a distributor that classifies its SKUs by ADI and CV2 and benchmarks within each class, weighting the result by the class share of stock value, obtains a decision that can be defended and repeated, which a single aggregate accuracy figure cannot provide.

Two further points concern implementation. The reporting checklist derived in Section 5 is as useful for a distributor's internal benchmark report as for a journal article: a report that states the data, the number of series, the evaluation design, the horizon, the metrics, the inventory simulation, the cost parameters, the classical benchmarks and the availability of the code allows the next planner to reproduce the decision, and the review found that even the published literature rarely does so. The monitoring step is also drawn from the evidence: obsolescence and the drift of intermittent patterns are addressed by estimators of the TSB type (Teunter et al., 2011; Babai et al., 2014), and a benchmark that is not repeated when the demand classes change will recommend methods for a demand that no longer exists.

Implications for research

The evidence map in Figure 9 identifies where new studies would add most. The first gap is cost-based evaluation of the AI families: 3 neural-network studies and 2 studies of support vector, tree-ensemble or nearest-neighbour methods evaluated both accuracy and cost, so a well-designed study that runs a tree ensemble, a recurrent network and a hybrid against Croston, SBA and TSB through an order-up-to policy on a large SKU set, with stated cost parameters and trade-off curves, would be among the most informative in the field. The second gap is distribution data: a benchmark on a wholesaler's SKU set, reported against the checklist, would be the first of its kind in the pre-2020 corpus and would test whether the spare-parts findings transfer to inexpensive, numerous intermittent SKUs with short lead

times. The third gap is methodological: the metrics for intermittent series are still debated (Hyndman & Koehler, 2006; Wallström & Segerstedt, 2010; Prestwich et al., 2014; Kolassa, 2016; Kim & Kim, 2016), and the review's finding that MAPE-type measures were the only accuracy measure in 3 studies shows that the recommendations of the metrics literature have not been adopted uniformly. A fourth gap is the reporting of cost parameters and code; the review could not reproduce a single comparison in Table 12 from the published record, and the field's move towards open data in the M-competitions (Makridakis et al., 2018a; Makridakis et al., 2018b) has yet to reach the intermittent demand literature of the period. Finally, the review's own method is a contribution that others can reuse: the search strings, the keyword classes, the decision log and the extraction files are supplied in the supplementary material, so an update covering 2020 onwards, when deep learning and global models became common on intermittent retail data, can be run against the same protocol.

Limitations

The review has limitations that bear on each finding; Table 16 lists them with their likely effect and the remedy. The first group concerns the search: one index, bibliographic fields only, relevance-ranked retrieval capped at 200 records per string, and two pages lost to index throttling. The yield analysis shows that 52 included studies were found by a single string, which indicates that coverage depends on the breadth of the string set, and the review cannot estimate what a second index would have added; studies whose titles use none of the search vocabulary (for example, a neural-network study on "slow movers" or on "sporadic sales" described only in the abstract) would be missed. The second group concerns screening and extraction: one reviewer, assisted by keyword classes and by extraction agents working from a written brief, with a 10 per cent independent re-screening to estimate consistency; and extraction from abstracts rather than full texts, so that every count in Section 4 is a lower bound and the reporting scores are conservative. The 50 studies decided on the title alone (22%) are mostly book chapters from the same handful of edited volumes; they carry little extracted information and their inclusion inflates the denominator of the coverage statistics, which is why Section 4 reports the abstract-based mean reporting score separately. The third group concerns the synthesis: vote counts weight every study equally regardless of dataset size or design, no magnitudes are pooled, and the benchmark verdicts rest on 23 comparative studies, of which only 3 report both verdicts, so the disagreement rate is a documented phenomenon rather than a population estimate. Finally, the restriction to pre-2020 literature was a design choice that defines the corpus but excludes the deep-learning and global-model results of the following years, and the Crossref issued date may in a few cases be an online-first date for a work whose version of record is dated 2020.

Table 16: Limitations of the review, their effect on the findings and what would address them

Limitation	Effect on the findings	What would address it
Single index (Crossref), bibliographic fields only	Studies whose titles lack the search vocabulary and studies not registered with a DOI are missed; abstracts were not searched	Add Scopus, Web of Science and OpenAlex when its search endpoints are available; search abstracts
Relevance-ranked retrieval capped at 200 records per string	Lower-ranked relevant records not retrieved; some pages not retrieved because of index throttling (Table 2)	Retrieve all matches with cursor paging; re-run the strings
Single reviewer with keyword- and agent-assisted screening	Screening errors not caught by a second reviewer; the 10% re-screening estimates consistency, not validity	Dual independent screening with adjudication
Abstract-based extraction, no full texts	Every count is a lower bound; reporting scores conservative; 22% of included studies decided on the title alone	Full-text extraction for the included studies
Index year may be the online-first year	A few included records may have a 2020 version of record	Check the version of record of every included study
No meta-analysis	Magnitudes cannot be pooled; vote counts weight every study equally	Standardised effect sizes from full texts where designs permit

Limitation	Effect on the findings	What would address it
Wholesale distribution studied by name in few studies	The wholesale reading rests on spare-parts and retail evidence	Benchmark studies on distributor data reported against the checklist
Pre-2020 restriction	Deep-learning and global-model results after 2019 are excluded by design	Update review covering 2020 onwards

CONCLUSION

This review synthesised 232 pre-2020 studies on forecasting intermittent demand with classical, AI and machine-learning methods, organised by the type of evaluation the studies used, and derived from them a cost-based benchmarking protocol for a wholesale distributor. The literature is recent, methodologically rich and concentrated in spare-parts settings; the Croston family remains the reference set; neural networks are the established AI family and the later tree ensembles and deep networks are under-evaluated. The decisive finding is evaluative rather than methodological: 15% of the studies judged forecasts by inventory or cost measures, cost parameters were stated in <1%, and where both an accuracy and a cost verdict were reported for an AI-against-classical comparison they disagreed in 1 of 3 cases. The literature therefore supports neither the claim that AI methods reduce inventory cost for intermittent SKUs nor its denial; it supports the conclusion that the question must be answered by inventory simulation on the distributor's own data, within demand classes, under the distributor's own policy and cost parameters, against the classical reference set, and reported against a checklist that allows the decision to be reproduced.

REFERENCES

[1]. Aditya, D., & Ahmed, H. C. H. R. (2022). Decision Support Systems for Last-Mile Logistics Optimization: A Data-Driven Framework for e-Commerce Fulfillment Networks. International Journal of Scientific Interdisciplinary Research, 12(08), 01–49. https://doi.org/10.63125/h46vp562

[2]. Adnan, S. (2024). Human Trust in AI-Assisted Project Decision Systems: A Mixed-Methods Analysis of Explainability, Adoption, and Managerial Decision Quality. Journal of Sustainable Development and Policy, 3(04), 287-332. https://doi.org/10.63125/9c139h44

[3]. Aminul Islam, M. (2024). Interpreting Cyber-Threat Indicators in Enterprise Memory Forensics: A Mixed-Methods Study of Machine Learning Performance and Explainability. Review of Applied Science and Technology, 14(12), 01–13. https://doi.org/10.63125/2p8hte92

[4]. Aminul Islam, M., & Md, A. (2023). Adaptive Class-Imbalance Learning for High-Sensitivity Fraud Detection in Large-Scale Digital Transaction Networks. International Journal of Business and Economics Insights, 13(09), 01-33. https://doi.org/10.63125/bmxwv741

[5]. Armstrong, J. S., & Collopy, F. (1992). Error measures for generalizing about forecasting methods: Empirical comparisons. International Journal of Forecasting, 8(1), 69–80. https://doi.org/10.1016/0169-2070(92)90008-W

[6]. Asheke, M. (2023). Machine Learning–Based Predictive Quality Modeling for Defect Detection and Process Optimization in Advanced Manufacturing Systems. Review of Applied Science and Technology, 13(12), 01–49. https://doi.org/10.63125/thy73b16

[7]. Asheke, M., & Md Maksudul, H. (2022). AI-Driven Demand Forecasting for Sporting-Goods and Teamwear Supply Chains Under Seasonal and Tournament Demand. International Journal of Business and Economics Insights, 2(1), 73–110. https://doi.org/10.63125/ed6by557

[8]. Asheke, M., & Mohammad Badrul, A. (2022). Development Of Functionalized Cellulose Nanofiber Adsorbents for Selective Critical-Metal Recovery from Industrial Wastewater. American Journal of Scholarly Research and Innovation, 1(02), 270–320. https://doi.org/10.63125/v6fxzq09

[9]. Ashik, M. (2024). Meta-Analysis of Risk-Based Inspection Strategies for Improving Reliability and Safety Across Critical Energy Infrastructure Systems. American Journal of Advanced Technology and Engineering Solutions, 14(09), 01-43. https://doi.org/10.63125/tgttkt09

[10]. Aviv, Y. (2001). The effect of collaborative forecasting on supply chain performance. Management Science, 47(10), 1326–1343. https://doi.org/10.1287/mnsc.47.10.1326.10260

[11]. Babai, M. Z., Ali, M. M., & Nikolopoulos, K. (2012). Impact of temporal aggregation on stock control performance of intermittent demand estimators: Empirical analysis. Omega, 40(6), 713–721. https://doi.org/10.1016/j.omega.2011.09.004

[12]. Babai, M. Z., Dallery, Y., Boubaker, S., & Kalai, R. (2019). A new method to forecast intermittent demand in the presence of inventory obsolescence. International Journal of Production Economics, 209, 30–41. https://doi.org/10.1016/j.ijpe.2018.01.026

[13]. Babai, M. Z., Syntetos, A., & Teunter, R. (2014). Intermittent demand forecasting: An empirical study on accuracy and the risk of obsolescence. International Journal of Production Economics, 157, 212–219. https://doi.org/10.1016/j.ijpe.2014.08.019

[14]. Babajanivalashedi, R., Baboli, A., Shahzad, M. K., & Tonadre, R. (2018). A predictive approach to define the best forecasting method for spare parts: A case study in business aircrafts' industry. In 2018 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM) (pp. 773–777). IEEE. https://doi.org/10.1109/ieem.2018.8607461
[15]. Bacchetti, A., & Saccani, N. (2012). Spare parts classification and demand forecasting for stock control: Investigating the gap between research and practice. Omega, 40(6), 722–737. https://doi.org/10.1016/j.omega.2011.06.008
[16]. Bala, P. K. (2010). Purchase-driven classification for improved forecasting in spare parts inventory replenishment. International Journal of Computer Applications, 10(9), 40–45. https://doi.org/10.5120/1507-2025
[17]. Bergmeir, C., & Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. Information Sciences, 191, 192–213. https://doi.org/10.1016/j.ins.2011.12.028
[18]. Boukhtouta, A., & Jentsch, P. (2018). Support vector machine for demand forecasting of Canadian armed forces spare parts. In 2018 6th International Symposium on Computational and Business Intelligence (ISCBI) (pp. 59–64). IEEE. https://doi.org/10.1109/iscbi.2018.00021
[19]. Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). Time series analysis: Forecasting and control (5th ed.). Wiley. https://www.wiley.com/en-us/Time+Series+Analysis%3A+Forecasting+and+Control%2C+5th+Edition-p-9781118675021
[20]. Boylan, J. E., & Syntetos, A. A. (2007). The accuracy of a modified Croston procedure. International Journal of Production Economics, 107(2), 511–517. https://doi.org/10.1016/j.ijpe.2006.10.005
[21]. Boylan, J. E., & Syntetos, A. A. (2010). Spare parts management: A review of forecasting research and extensions. IMA Journal of Management Mathematics, 21(3), 227–237. https://doi.org/10.1093/imaman/dpp016
[22]. Boylan, J. E., Syntetos, A. A., & Karakostas, G. C. (2008). Classification for forecasting and stock control: A case study. Journal of the Operational Research Society, 59(4), 473–481. https://doi.org/10.1057/palgrave.jors.2602312
[23]. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324
[24]. Cachon, G. P., & Fisher, M. L. (2000). Supply chain inventory management and the value of shared information. Management Science, 46(8), 1032–1048. https://doi.org/10.1287/mnsc.46.8.1032.12029
[25]. Cachon, G. P., Randall, T., & Schmidt, G. M. (2007). In search of the bullwhip effect. Manufacturing & Service Operations Management, 9(4), 457–479. https://doi.org/10.1287/msom.1060.0149
[26]. Chen, F., Drezner, Z., Ryan, J. K., & Simchi-Levi, D. (2000). Quantifying the bullwhip effect in a simple supply chain: The impact of forecasting, lead times, and information. Management Science, 46(3), 436–443. https://doi.org/10.1287/mnsc.46.3.436.12069
[27]. Chen, H., Frank, M. Z., & Wu, O. Q. (2005). What actually happened to the inventories of American companies between 1981 and 2000? Management Science, 51(7), 1015–1031. https://doi.org/10.1287/mnsc.1050.0368
[28]. Chen, H., Frank, M. Z., & Wu, O. Q. (2007). U.S. retail and wholesale inventory performance from 1981 to 2004. Manufacturing & Service Operations Management, 9(4), 430–456. https://doi.org/10.1287/msom.1060.0129
[29]. Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273–297. https://doi.org/10.1007/BF00994018
[30]. Croston, J. D. (1972). Forecasting and stock control for intermittent demands. Operational Research Quarterly, 23(3), 289–303. https://doi.org/10.1057/jors.1972.50
[31]. Davydenko, A., & Fildes, R. (2013). Measuring forecasting accuracy: The case of judgmental adjustments to SKU-level demand forecasts. International Journal of Forecasting, 29(3), 510–522. https://doi.org/10.1016/j.ijforecast.2012.09.002
[32]. Disney, S. M., & Towill, D. R. (2003). The effect of vendor managed inventory (VMI) dynamics on the bullwhip effect in supply chains. International Journal of Production Economics, 85(2), 199–215. https://doi.org/10.1016/S0925-5273(03)00110-5
[33]. Durach, C. F., Kembro, J., & Wieland, A. (2017). A new paradigm for systematic literature reviews in supply chain management. Journal of Supply Chain Management, 53(4), 67–85. https://doi.org/10.1111/jscm.12145
[34]. Eaves, A. H. C., & Kingsman, B. G. (2004). Forecasting for the ordering and stock-holding of spare parts. Journal of the Operational Research Society, 55(4), 431–437. https://doi.org/10.1057/palgrave.jors.2601697
[35]. Fildes, R., & Kingsman, B. (2011). Incorporating demand uncertainty and forecast error in supply chain planning models. Journal of the Operational Research Society, 62(3), 483–500. https://doi.org/10.1057/jors.2010.40
[36]. Fildes, R., & Petropoulos, F. (2015). Simple versus complex selection rules for forecasting many time series. Journal of Business Research, 68(8), 1692–1701. https://doi.org/10.1016/j.jbusres.2015.03.028
[37]. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. The Annals of Statistics, 29(5), 1189–1232. https://doi.org/10.1214/aos/1013203451
[38]. Gardner, E. S., Jr. (1990). Evaluating forecast performance in an inventory control system. Management Science, 36(4), 490–499. https://doi.org/10.1287/mnsc.36.4.490
[39]. Gardner, E. S., Jr. (2006). Exponential smoothing: The state of the art: Part II. International Journal of Forecasting, 22(4), 637–666. https://doi.org/10.1016/j.ijforecast.2006.03.005
[40]. Gutierrez, R. S., Solis, A. O., & Mukhopadhyay, S. (2008). Lumpy demand forecasting using neural networks. International Journal of Production Economics, 111(2), 409–420. https://doi.org/10.1016/j.ijpe.2007.01.007
[41]. Hasni, M., Aguir, M. S., Babai, M. Z., & Jemai, Z. (2019a). Spare parts demand forecasting: A review on bootstrapping methods. International Journal of Production Research, 57(15-16), 4791–4804. https://doi.org/10.1080/00207543.2018.1424375

[42]. Hasni, M., Aguir, M., Babaï, M. Z., & Jemaï, Z. (2019b). On the performance of adjusted bootstrapping methods for intermittent demand forecasting. International Journal of Production Economics, 216, 145–153. https://doi.org/10.1016/j.ijpe.2019.04.005
[43]. Hasni, M., Babai, M. Z., Aguir, M. S., & Jemai, Z. (2019c). An investigation on bootstrapping forecasting methods for intermittent demands. International Journal of Production Economics, 209, 20–29. https://doi.org/10.1016/j.ijpe.2018.03.001
[44]. Heinecke, G., Syntetos, A. A., & Wang, W. (2013). Forecasting-based SKU classification. International Journal of Production Economics, 143(2), 455–462. https://doi.org/10.1016/j.ijpe.2011.11.020
[45]. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735
[46]. Hua, Z. S., Zhang, B., Yang, J., & Tan, D. S. (2007). A new approach of forecasting intermittent demand for spare parts inventories in the process industries. Journal of the Operational Research Society, 58(1), 52–61. https://doi.org/10.1057/palgrave.jors.2602119
[47]. Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. International Journal of Forecasting, 22(4), 679–688. https://doi.org/10.1016/j.ijforecast.2006.03.001
[48]. Hyndman, R. J., Ahmed, R. A., Athanasopoulos, G., & Shang, H. L. (2011). Optimal combination forecasts for hierarchical time series. Computational Statistics & Data Analysis, 55(9), 2579–2589. https://doi.org/10.1016/j.csda.2011.03.006
[49]. Hyndman, R. J., Koehler, A. B., Snyder, R. D., & Grose, S. (2002). A state space framework for automatic forecasting using exponential smoothing methods. International Journal of Forecasting, 18(3), 439–454. https://doi.org/10.1016/S0169-2070(01)00110-8
[50]. Johnston, F. R., & Boylan, J. E. (1996). Forecasting for items with intermittent demand. Journal of the Operational Research Society, 47(1), 113–121. https://doi.org/10.1057/jors.1996.10
[51]. Kalaya, P., Termsuksawad, P., & Wasusri, T. (2019). Forecasting lumpy demand for planning inventory: The case of community hospitals in Thailand. In 2019 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM) (pp. 686–690). IEEE. https://doi.org/10.1109/ieem44572.2019.8978541
[52]. Kerkkänen, A., Korpela, J., & Huiskonen, J. (2009). Demand forecasting errors in industrial context: Measurement and impacts. International Journal of Production Economics, 118(1), 43–48. https://doi.org/10.1016/j.ijpe.2008.08.008
[53]. Kim, S., & Kim, H. (2016). A new metric of absolute percentage error for intermittent demand forecasts. International Journal of Forecasting, 32(3), 669–679. https://doi.org/10.1016/j.ijforecast.2015.12.003
[54]. Kitti-udomporn, S., Sukchareonpong, P., & Witchakul, S. (2018). Forecasting and purchasing planning for shelf life-limited spare parts. In 2018 5th International Conference on Business and Industrial Research (ICBIR) (pp. 264–269). IEEE. https://doi.org/10.1109/icbir.2018.8391204
[55]. Kolassa, S. (2016). Evaluating predictive count data distributions in retail sales forecasting. International Journal of Forecasting, 32(3), 788–803. https://doi.org/10.1016/j.ijforecast.2015.12.004
[56]. Kostenko, A. V., & Hyndman, R. J. (2006). A note on the categorization of demand patterns. Journal of the Operational Research Society, 57(10), 1256–1257. https://doi.org/10.1057/palgrave.jors.2602211
[57]. Kourentzes, N. (2013). Intermittent demand forecasts with neural networks. International Journal of Production Economics, 143(1), 198–206. https://doi.org/10.1016/j.ijpe.2013.01.009
[58]. Kourentzes, N. (2014). On intermittent demand model optimisation and selection. International Journal of Production Economics, 156, 180–190. https://doi.org/10.1016/j.ijpe.2014.06.007
[59]. Kourentzes, N., Petropoulos, F., & Trapero, J. R. (2014). Improving forecasting by estimating time series structural components across multiple frequencies. International Journal of Forecasting, 30(2), 291–302. https://doi.org/10.1016/j.ijforecast.2013.09.006
[60]. Kynatun, M., & Saba, A. (2024). Machine Learning Enhanced Intrusion Detection and Anomaly Classification for MIS-Driven Enterprise Network Security. International Journal of Scientific Interdisciplinary Research, 14(09), 01–41. https://doi.org/10.63125/xk6kpc72
[61]. Lee, H. L., Padmanabhan, V., & Whang, S. (1997). Information distortion in a supply chain: The bullwhip effect. Management Science, 43(4), 546–558. https://doi.org/10.1287/mnsc.43.4.546
[62]. Lei, M., Yin, Z., Li, S., & Tan, Q. (2017). Intermittent demand forecasting and inventory control with multiple temporal and cross-sectional aggregation and disaggregation methods. In 2017 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD) (pp. 1969–1977). IEEE. https://doi.org/10.1109/fskd.2017.8393070
[63]. Lengu, D., Syntetos, A. A., & Babai, M. Z. (2014). Spare parts management: Linking distributional assumptions to demand classification. European Journal of Operational Research, 235(3), 624–635. https://doi.org/10.1016/j.ejor.2013.12.043
[64]. Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P. A., Clarke, M., Devereaux, P. J., Kleijnen, J., & Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: Explanation and elaboration. PLoS Medicine, 6(7), Article e1000100. https://doi.org/10.1371/journal.pmed.1000100
[65]. Lindsey, M., & Pavur, R. (2014). Evaluating a Bayesian approach to forecasting stocking spare parts that require periodic replenishment. In Advances in Business and Management Forecasting (pp. 111–128). Emerald Group Publishing Limited. https://doi.org/10.1108/s1477-407020140000010018

[66]. Lolli, F., Gamberini, R., Regattieri, A., Balugani, E., Gatos, T., & Gucci, S. (2017). Single-hidden layer neural networks for forecasting intermittent demand. International Journal of Production Economics, 183, 116–128. https://doi.org/10.1016/j.ijpe.2016.10.021

[67]. Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018a). The M4 Competition: Results, findings, conclusion and way forward. International Journal of Forecasting, 34(4), 802–808. https://doi.org/10.1016/j.ijforecast.2018.06.001

[68]. Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018b). Statistical and machine learning forecasting methods: Concerns and ways forward. PLOS ONE, 13(3), Article e0194889. https://doi.org/10.1371/journal.pone.0194889

[69]. Mansura Akter, E., & Sumaiya, M. (2022). Antimicrobial Resistance Profiles of Acinetobacter Baumannii and Vibrio Cholerae in Hospital Wastewater and Adjacent Household Water Sources. American Journal of Data Science and Analytics, 3(09), 01-55. https://doi.org/10.63125/gbervt14

[70]. Md Hasinur, R. (2024). Human in the Loop Artificial Intelligence for Loan Document Verification and Credit Risk Reporting: A Bibliometric and Science Mapping Analysis. American Journal of Interdisciplinary Studies, 14(06), 41-71. https://doi.org/10.63125/8x359j95

[71]. Md Imran, H., & Jannatul, F. (2022). Molecular and Precision Rehabilitation Molecular Mechanisms Linking Mitochondrial Dysfunction, Inflammation, and Muscle Degeneration with Functional Decline and Rehabilitation Response in Neuromuscular Disease. American Journal of Interdisciplinary Studies, 12(08), 01-44. https://doi.org/10.63125/0ym5bh79

[72]. Md Maksudul, H. (2024). Supplier-Risk Scoring and Procurement Resilience for Distributed Athletic-Apparel and Equipment Supply Networks. American Journal of Advanced Technology and Engineering Solutions, 14(09), 44-90. https://doi.org/10.63125/a9fmgj50

[73]. Md Maksudul, H., & Md Tazul Islam, C. (2022). Machine Learning-Based Cash-Flow Forecasting for Improving Financial Planning in U.S. Small and Mid-Sized Businesses. International Journal of Scientific Interdisciplinary Research, 12(12), 01–38. https://doi.org/10.63125/w3pgjx97

[74]. Md Shahid, N. (2023). Digital Twin-Based Network Reliability Modeling for LAN And WAN Infrastructure Resilience. International Journal of Scientific Interdisciplinary Research, 4(4), 577–629. https://doi.org/10.63125/ahjvxj14

[75]. Md Shahid, N. (2024). Predictive Analytics Models for Network Traffic Optimization in Distributed Enterprise Environments. Review of Applied Science and Technology, 3(01), 367-419. https://doi.org/10.63125/qm6cyf84

[76]. Md Tazul Islam, C. (2024). Explainable AI For Financial and Operational Decision Support: Improving Transparency and Managerial Trust. American Journal of Advanced Technology and Engineering Solutions, 14(12), 01-38. https://doi.org/10.63125/kwfvzw68

[77]. Md Tazul Islam, C., & Ahmed, H. C. H. R. (2023). Predictive Operations Analytics for Resource Allocation and Service-Capacity Planning in Data-Driven Organizations. Review of Applied Science and Technology, 13(02), 01–41. https://doi.org/10.63125/z3sm3089

[78]. Md. Mainuddin, F. (2024). Quantitative Structural Retrofit Assessment Models for Strengthening Existing Steel Buildings Under Increased Load Demands. Review of Applied Science and Technology, 3(04), 325–366. https://doi.org/10.63125/yyqnte84

[79]. Md. Morshedul, I., & Aminul Islam, M. (2022). A Quantitative Evaluation of Explainable Ensemble Learning for Probabilistic Risk Estimation and Decision Reliability in Enterprise Analytics. American Journal of Data Science and Analytics, 3(02), 01-38. https://doi.org/10.63125/k22shk74

[80]. Min, H. (2010). Artificial intelligence in supply chain management: Theory and applications. International Journal of Logistics Research and Applications, 13(1), 13–39. https://doi.org/10.1080/13675560902736537

[81]. Mitchell. (2010). On replenishing items with seasonal intermittent demand. American Journal of Economics and Business Administration, 2(1), 90–102. https://doi.org/10.3844/ajebasp.2010.90.102

[82]. Mobarakeh, N. A., Shahzad, M. K., Baboli, A., & Tonadre, R. (2017). Improved forecasts for uncertain and unpredictable spare parts demand in business aircraft's with bootstrap method. IFAC-PapersOnLine, 50(1), 15241–15246. https://doi.org/10.1016/j.ifacol.2017.08.2379

[83]. Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & The PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. PLoS Medicine, 6(7), Article e1000097. https://doi.org/10.1371/journal.pmed.1000097

[84]. Moon, S. (2013). Predicting the performance of forecasting strategies for naval spare parts demand: A machine learning approach. Management Science and Financial Engineering, 19(1), 1–10. https://doi.org/10.7737/msfe.2013.19.1.001

[85]. Mukhopadhyay, S., Solis, A. O., & Gutierrez, R. S. (2012). The accuracy of non-traditional versus traditional methods of forecasting lumpy demand. Journal of Forecasting, 31(8), 721–735. https://doi.org/10.1002/for.1242

[86]. Nikolopoulos, K., Syntetos, A. A., Boylan, J. E., Petropoulos, F., & Assimakopoulos, V. (2011). An aggregate-disaggregate intermittent demand approach (ADIDA) to forecasting: An empirical proposition and analysis. Journal of the Operational Research Society, 62(3), 544–554. https://doi.org/10.1057/jors.2010.32

[87]. Office of Management and Budget. (2017). North American Industry Classification System: 2017 NAICS manual, Sector 42: Wholesale trade. U.S. Census Bureau. https://www.census.gov/naics/reference_files_tools/2017_NAICS_Manual.pdf

[88]. Pennings, C. L. P., van Dalen, J., & van der Laan, E. A. (2017). Exploiting elapsed time for managing intermittent demand for spare parts. European Journal of Operational Research, 258(3), 958–969. https://doi.org/10.1016/j.ejor.2016.09.017

[89]. Petropoulos, F., & Kourentzes, N. (2015). Forecast combinations for intermittent demand. Journal of the Operational Research Society, 66(6), 914–924. https://doi.org/10.1057/jors.2014.62
[90]. Petropoulos, F., Kourentzes, N., & Nikolopoulos, K. (2016). Another look at estimators for intermittent demand. International Journal of Production Economics, 181, 154–161. https://doi.org/10.1016/j.ijpe.2016.04.017
[91]. Petropoulos, F., Makridakis, S., Assimakopoulos, V., & Nikolopoulos, K. (2014). 'Horses for courses' in demand forecasting. European Journal of Operational Research, 237(1), 152–163. https://doi.org/10.1016/j.ejor.2014.02.036
[92]. Petticrew, M., & Roberts, H. (2006). Systematic reviews in the social sciences: A practical guide. Blackwell. https://doi.org/10.1002/9780470754887
[93]. Prestwich, S. D., Rossi, R., Tarim, S. A., & Hnich, B. (2014). Mean-based error measures for intermittent demand forecasting. International Journal of Production Research, 52(22), 6782–6791. https://doi.org/10.1080/00207543.2014.917771
[94]. Rahman, M. A., & Sarker, B. R. (2012). A Bayesian approach to forecast intermittent demand for seasonal products. International Journal of Industrial and Systems Engineering, 11(1/2), 137. https://doi.org/10.1504/ijise.2012.046660
[95]. Rostami-Tabar, B., Babai, M. Z., Syntetos, A., & Ducq, Y. (2013). Demand forecasting by temporal aggregation. Naval Research Logistics (NRL), 60(6), 479–498. https://doi.org/10.1002/nav.21546
[96]. Rumyantsev, S., & Netessine, S. (2007). What can be learned from classical inventory models? A cross-industry exploratory investigation. Manufacturing & Service Operations Management, 9(4), 409–429. https://doi.org/10.1287/msom.1070.0166
[97]. Sadia, A., & Md. Mosheur, R. (2023). Consumer Trust, Perceived Authenticity, and Personalized Digital Marketing: A Quantitative Decision-Science Study. International Journal of Scientific Interdisciplinary Research, 4(2), 252–295. https://doi.org/10.63125/cajkqt60
[98]. Salinas, D., Flunkert, V., & Gasthaus, J. (2017). DeepAR: Probabilistic forecasting with autoregressive recurrent networks (arXiv:1704.04110). arXiv. https://doi.org/10.48550/arXiv.1704.04110
[99]. Sanders, N. R., & Graman, G. A. (2009). Quantifying costs of forecast errors: A case study of the warehouse environment. Omega, 37(1), 116–125. https://doi.org/10.1016/j.omega.2006.10.004
[100]. Sani, B., & Kingsman, B. G. (1997). Selecting the best periodic inventory control and demand forecasting methods for low demand items. Journal of the Operational Research Society, 48(7), 700–713. https://doi.org/10.1057/palgrave.jors.2600418
[101]. Sari, K. (2008). On the benefits of CPFR and VMI: A comparative simulation study. International Journal of Production Economics, 113(2), 575–586. https://doi.org/10.1016/j.ijpe.2007.10.021
[102]. Seuring, S., & Müller, M. (2008). From a literature review to a conceptual framework for sustainable supply chain management. Journal of Cleaner Production, 16(15), 1699–1710. https://doi.org/10.1016/j.jclepro.2008.04.020
[103]. Shenstone, L., & Hyndman, R. J. (2005). Stochastic models underlying Croston's method for intermittent demand forecasting. Journal of Forecasting, 24(6), 389–402. https://doi.org/10.1002/for.963
[104]. Silver, E. A., Pyke, D. F., & Thomas, D. J. (2016). Inventory and production management in supply chains (4th ed.). CRC Press. https://doi.org/10.1201/9781315374406
[105]. Snyder, R. D., Ord, J. K., & Beaumont, A. (2012). Forecasting the intermittent demand for slow-moving inventories: A modelling approach. International Journal of Forecasting, 28(2), 485–496. https://doi.org/10.1016/j.ijforecast.2011.03.009
[106]. Solis, A. O., Longo, F., Mukhopadhyay, S., & Nicoletti, L. (2017). An empirical investigation of comparative performance of approximate and exact corrections of the bias in croston's method in forecasting lumpy demand. International Journal of Simulation and Process Modelling, 12(6), 535. https://doi.org/10.1504/ijspm.2017.089637
[107]. Song, Z. J., Fu, Z., Wang, H., & Hou, G. B. (2013). Demand forecasting model of port critical spare parts based on PSO-LSSVM. Applied Mechanics and Materials, 433-435, 545–549. https://doi.org/10.4028/www.scientific.net/amm.433-435.545
[108]. Song, Z., Fu, Z., Wang, H., & Hou, G. (2014). Demand forecasting model of port critical spare parts. TELKOMNIKA Indonesian Journal of Electrical Engineering, 12(5). https://doi.org/10.11591/telkomnika.v12i5.3238
[109]. Stabak, D., & Md Shahid, N. (2022). Empirical Assessment of Uncertainty-Aware Real-Time Decision-Making for Reliable and Resilient Cyber-Physical Systems. International Journal of Scientific Interdisciplinary Research, 1(01), 432–477. https://doi.org/10.63125/q3r4kk16
[110]. Sulistyo, S. R., & Sutrisno, A. J. (2018). Lumpy demand forecasting using linear exponential smoothing, artificial neural network, and bootstrap. Angkasa: Jurnal Ilmiah Bidang Teknologi, 10(2), 107. https://doi.org/10.28989/angkasa.v10i2.362
[111]. Sumaiya, M. (2024). Hospital Wastewater as A Reservoir of Antimicrobial Resistance Genes: Implications for Public Health and Genomic Surveillance in The United States. Review of Applied Science and Technology, 3(04), 448–508. https://doi.org/10.63125/ahvas976
[112]. Syed Feroz, A., & Md Mahamud Pervaz, P. (2022). Root-Cause Analysis and FMEA-Driven Reliability Improvement in Gas Production Facilities: A Multi-Year Empirical Study. International Journal of Scientific Interdisciplinary Research, 12(10), 01–53. https://doi.org/10.63125/70msp878
[113]. Syed Feroz, A., & Shabbir, A. (2022). Finite Element-Based Thermomechanical Optimization of Pressure Components for Extending Industrial Boiler Service Life. Journal of Sustainable Development and Policy, 12(12), 01-51. https://doi.org/10.63125/x24pht28
[114]. Syntetos, A. A., & Boylan, J. E. (2001). On the bias of intermittent demand estimates. International Journal of Production Economics, 71(1-3), 457–466. https://doi.org/10.1016/S0925-5273(00)00143-2

[115]. Syntetos, A. A., & Boylan, J. E. (2005). The accuracy of intermittent demand estimates. International Journal of Forecasting, 21(2), 303–314. https://doi.org/10.1016/j.ijforecast.2004.10.001
[116]. Syntetos, A. A., & Boylan, J. E. (2006). On the stock control performance of intermittent demand estimators. International Journal of Production Economics, 103(1), 36–47. https://doi.org/10.1016/j.ijpe.2005.04.004
[117]. Syntetos, A. A., & Boylan, J. E. (2008). Demand forecasting adjustments for service-level achievement. IMA Journal of Management Mathematics, 19(2), 175–192. https://doi.org/10.1093/imaman/dpm034
[118]. Syntetos, A. A., Babai, M. Z., & Gardner, E. S., Jr. (2015). Forecasting intermittent inventory demands: Simple parametric methods vs. bootstrapping. Journal of Business Research, 68(8), 1746–1752. https://doi.org/10.1016/j.jbusres.2015.03.034
[119]. Syntetos, A. A., Babai, M. Z., Dallery, Y., & Teunter, R. (2009a). Periodic control of intermittent demand items: theory and empirical analysis. Journal of the Operational Research Society, 60(5), 611–618. https://doi.org/10.1057/palgrave.jors.2602593
[120]. Syntetos, A. A., Boylan, J. E., & Croston, J. D. (2005). On the categorization of demand patterns. Journal of the Operational Research Society, 56(5), 495–503. https://doi.org/10.1057/palgrave.jors.2601841
[121]. Syntetos, A. A., Boylan, J. E., & Disney, S. M. (2009b). Forecasting for inventory planning: A 50-year review. Journal of the Operational Research Society, 60(S1), S149–S160. https://doi.org/10.1057/jors.2008.173
[122]. Syntetos, A. A., Nikolopoulos, K., & Boylan, J. E. (2010). Judging the judges through accuracy-implication metrics: The case of inventory forecasting. International Journal of Forecasting, 26(1), 134–143. https://doi.org/10.1016/j.ijforecast.2009.05.016
[123]. Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: An analysis and review. International Journal of Forecasting, 16(4), 437–450. https://doi.org/10.1016/S0169-2070(00)00065-0
[124]. Teunter, R. H., & Duncan, L. (2009). Forecasting intermittent demand: A comparative study. Journal of the Operational Research Society, 60(3), 321–329. https://doi.org/10.1057/palgrave.jors.2602569
[125]. Teunter, R. H., Syntetos, A. A., & Babai, M. Z. (2010). Determining order-up-to levels under periodic review for compound binomial (intermittent) demand. European Journal of Operational Research, 203(3), 619–624. https://doi.org/10.1016/j.ejor.2009.09.013
[126]. Teunter, R. H., Syntetos, A. A., & Babai, M. Z. (2011). Intermittent demand: Linking forecasting to inventory obsolescence. European Journal of Operational Research, 214(3), 606–615. https://doi.org/10.1016/j.ejor.2011.05.018
[127]. Tiacci, L., & Saetta, S. (2009). An approach to evaluate the impact of interaction between demand forecasting method and stock control policy on the inventory system performances. International Journal of Production Economics, 118(1), 63–71. https://doi.org/10.1016/j.ijpe.2008.08.010
[128]. Tranfield, D., Denyer, D., & Smart, P. (2003). Towards a methodology for developing evidence-informed management knowledge by means of systematic review. British Journal of Management, 14(3), 207–222. https://doi.org/10.1111/1467-8551.00375
[129]. Turrini, L., & Meissner, J. (2019). Spare parts inventory management: New evidence from distribution fitting. European Journal of Operational Research, 273(1), 118–130. https://doi.org/10.1016/j.ejor.2017.09.039
[130]. U.S. Census Bureau. (2019). Annual Wholesale Trade Survey (AWTS). https://www.census.gov/programs-surveys/awts.html
[131]. Van der Auweraer, S., & Boute, R. (2019). Forecasting spare part demand using service maintenance information. International Journal of Production Economics, 213, 138–149. https://doi.org/10.1016/j.ijpe.2019.03.015
[132]. Van der Auweraer, S., Boute, R. N., & Syntetos, A. A. (2019). Forecasting spare part demand with installed base information: A review. International Journal of Forecasting, 35(1), 181–196. https://doi.org/10.1016/j.ijforecast.2018.09.002
[133]. Verhagen, W. J. C., & Curran, R. (2014). Stochastic forecasting of lumpy-distributed aircraft spare parts demand. In Moving Integrated Product Development to Service Clouds in the Global Economy (series: Advances in Transdisciplinary Engineering). IOS Press. https://doi.org/10.3233/978-1-61499-440-4-706
[134]. Wallström, P., & Segerstedt, A. (2010). Evaluation of forecasting error measurements and techniques for intermittent demand. International Journal of Production Economics, 128(2), 625–636. https://doi.org/10.1016/j.ijpe.2010.07.013
[135]. Webster, J., & Watson, R. T. (2002). Analyzing the past to prepare for the future: Writing a literature review. MIS Quarterly, 26(2), xiii–xxiii. https://doi.org/10.2307/4132319
[136]. Willemain, T. R., Smart, C. N., & Schwarz, H. F. (2004). A new approach to forecasting intermittent demand for service parts inventories. International Journal of Forecasting, 20(3), 375–387. https://doi.org/10.1016/S0169-2070(03)00013-X
[137]. Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing, 50, 159–175. https://doi.org/10.1016/S0925-2312(01)00702-0
[138]. Zhang, G., Patuwo, B. E., & Hu, M. Y. (1998). Forecasting with artificial neural networks: The state of the art. International Journal of Forecasting, 14(1), 35–62. https://doi.org/10.1016/S0169-2070(97)00044-7