Feature-based intermittent demand forecast combinations: accuracy and inventory implications

Li Li[a], Yanfei Kang[a], Fotios Petropoulos[b] and Feng Li[c]

[a]School of Economics and Management, Beihang University, Beijing, China; [b]School of Management, University of Bath, UK; [c]School of Statistics and Mathematics, Central University of Finance and Economics, Beijing, China

ARTICLE HISTORY

Compiled November 20, 2022

ABSTRACT

Intermittent demand forecasting is a ubiquitous and challenging problem in production systems and supply chain management. In recent years, there has been a growing focus on developing forecasting approaches for intermittent demand from academic and practical perspectives. However, limited attention has been given to forecast combination methods, which have achieved competitive performance in forecasting fast-moving time series. The current study examines the empirical outcomes of some existing forecast combination methods and proposes a generalized feature-based framework for intermittent demand forecasting. The proposed framework has been shown to improve the accuracy of point and quantile forecasts based on two real data sets. Further, some analysis of features, forecasting pools and computational efficiency is also provided. The findings indicate the intelligibility and flexibility of the proposed approach in intermittent demand forecasting and offer insights regarding inventory decisions.

KEYWORDS

Intermittent demand forecasting; Forecast combinations; Time series features; Diversity; Empirical evaluation

CONTACT Feng Li. Email: feng.li@cufe.edu.cn

1. Introduction

Intermittent demand with several periods of zero demand is ubiquitous in practice. Over half of inventory consists of spare parts, in which demand is typically intermittent (Nikolopoulos 2021). Given the high purchase and shortage costs associated with intermittent demand applications, accurate forecasts could be coupled with improved inventory management in the field of manufacturing (Jiang, Huang, and Liu 2021), aerospace (Wang and Petropoulos 2016), retailing (Sillanpää and Liesiö 2018) and so on (Balugani et al. 2019; Babai et al. 2019).

What makes intermittent demand challenging to forecast is that there are two sources of uncertainty: the sporadic demand occurrence, and the demand arrival timing. Seminal work on intermittent demand forecasting by Croston (1972) proposed to separately forecast the sizes of demand and the inter-demand intervals. Then some scholars followed this idea and put forward some developments. For example, Syntetos-Boylan Approximation (SBA) proposed by Syntetos and Boylan (2005) delivered approximately unbiased estimates and constituted the benchmark in subsequently proposed methodologies for intermittent demand forecasting.

Syntetos, Boylan, and Croston (2005) proposed a categorization of demand patterns to facilitate the selection of Croston's method (1972) and SBA (Syntetos and Boylan 2005). A classification rule was expressed in terms of the average inter-demand interval and the squared coefficient of variation of demand sizes (Syntetos, Boylan, and Croston 2005). Kostenko and Hyndman (2006) developed the SBC categorization scheme (Syntetos, Boylan, and Croston 2005) and suggested a simple and more accurate rule, which has been widely used in the research of intermittent demand (Petropoulos and Kourentzes 2015; Spiliotis et al. 2021).

However, Croston's method (1972) and SBA update demand sizes and intervals, which leads to inapplicability in periods of zero demand when considering inventory obsolescence. To overcome this shortcoming, Teunter, Syntetos, and Babai (2011) proposed a new method called Teunter-Syntetos-Babai (TSB) to update the demand probability instead of the demand interval. TSB has been proved to have good empirical performance for the demands within linear and sudden obsolescence (Babai, Syntetos, and Teunter 2014).

The aforementioned forecasting methods for intermittent demand are all parametric methods, which estimate the parameters of a specific distribution. Instead,

non-parametric intermittent demand methods directly estimate empirical distribution based on past data, with no need for any assumption of a standard probability distribution. The bootstrapping methods, and the overlapping and non-overlapping aggregation methods dominate the research field of non-parametric intermittent demand forecasting (Willemain, Smart, and Schwarz 2004; Hasni et al. 2019a,b; Boylan and Syntetos 2021; Boylan and Babai 2016).

In particular, temporal aggregation is a promising approach to intermittent demand forecasting, in which a lower-frequency time series can be aggregated to a higher-frequency time series. Latent characteristics of the demand, such as trend and seasonality, appear at higher levels of aggregation. Nikolopoulos et al. (2011) first introduced temporal aggregation to intermittent demand forecasting and proposed the Aggregate-Disaggregate Intermittent Demand Approach (ADIDA). To tackle the challenge of determining the optimal aggregation level, Petropoulos and Kourentzes (2015) considered combinations of forecasts from multiple temporal aggregation levels simultaneously. This approach is called the Intermittent Multiple Aggregation Prediction Algorithm (IMAPA). The overall results of their work suggested that combinations of forecasts from different frequencies led to improved forecasting performance.

Recently, some attention has been paid to applying machine learning approaches to improve forecasting accuracy for intermittent demand, such as neural networks (Lolli et al. 2017), support vector machines (Kaya and Turkyilmaz 2018; Jiang, Huang, and Liu 2021), and so on.

Despite that intermittent demand forecasting has obtained some research achievements in recent decades (Nikolopoulos et al. 2011; Petropoulos and Kourentzes 2015; Kourentzes and Athanasopoulos 2021), there is still much scope for improvements (Nikolopoulos 2021). For example, limited attention has been given to combination schemes for intermittent demand forecasting. The literature indicates that forecast combination can improve forecast accuracy in modeling fast-moving time series (Bates and Granger 1969; De Menezes, Bunn, and Taylor 2000; Petropoulos et al. 2022; Li, Petropoulos, and Kang 2022). In this study, we aim to examine whether the forecast combination improves intermittent demand forecasts. The main contributions of our work are: (1) providing a discussion and comparison of forecast combination methods in the context of intermittent demand forecasting, (2) developing a feature-based combination framework for intermittent demand, which can determine optimal combi-

nation weights evaluated by the given error measure, and (3) improving the accuracy of both point and quantile forecasts to support real inventory decisions.

The rest of the paper is organized as follows. Section 2 reviews a series of forecast combination methods discussed in this work. Section 3 proposes a generalized forecast combination framework for intermittent demand. In Section 4, we apply our framework to two real datasets and present results based on point forecasts and quantile forecasts. Section 5 concludes the paper.

2. A review of forecast combinations

Combining forecasts from different methods or models has achieved satisfactory results in practice. Wang et al. (2022a) provided an up-to-date review of forecast combinations including combining point forecasts and combining probabilistic forecasts. The Simple Average (SA) has been proved to be a hard-to-beat forecast combination method (Clemen 1989; Stock and Watson 2004; Lichtendahl Jr and Winkler 2020), which simply combines forecasts with an equal weight of $1/M$ (M is the number of forecasting methods to be combined). Clemen (1989) reviewed over two hundred articles and concluded that SA should be used as a benchmark when proposing more complex weighting schemes. Palm and Zellner (1992) emphasized that SA could reduce the variance of forecasts and avoid the uncertainty of weight estimation. The phenomenon that SA outperforms more complicated combination methods is referred to as "forecast combination puzzle" (Stock and Watson 2004; Smith and Wallis 2009; Claeskens et al. 2016).

Because SA is sensitive to extreme values, some attention has been paid to other more robust combination schemes, including the median and trimmed means (Stock and Watson 2004; Lichtendahl Jr and Winkler 2020; Petropoulos and Svetunkov 2020). Jose and Winkler (2008) studied two mean-based methods, trimmed and Winsorized means, and verified their improved combined forecasts. The simple combination schemes based on the mean and median are easy to calculate and avoid parameter estimation errors. However, there is still no consensus on which of the mean and the median of individual forecasts performs better.

In the field of intermittent demand forecasting, forecast combination methods have been largely overlooked. To the best of our knowledge, only SA has been applied to

improving intermittent demand forecasting (Petropoulos and Kourentzes 2015). Recently, the organizers of the M5 competition (Makridakis, Spiliotis, and Assimakopoulos 2021) used SA as the combination benchmark, such as the average of exponential smoothing (ES) and ARIMA. The M5 competition focused on sales forecasts involving a mass of intermittent time series. As shown in M5 results, combinations performed better or equally well with the individual methods that they consist of (Makridakis, Spiliotis, and Assimakopoulos 2022).

To further exploit the value of forecast combinations, a handful of research has focused on finding optimal weights for combining different forecasting models over the past half-century. The seminal work by Bates and Granger (1969) proposed the idea of weighted forecast combinations. Newbold and Granger (1974) continued this stream of research and investigated more forecasting models and multiple forecast horizons. In their work, a weighted combination can be expressed as a linear function such that

$$\hat{y}^{c}_{T+1} = \sum_{i=1}^{M} w_{i,T+1}\hat{y}_{i,T+1} = \mathbf{w}'_{T+1}\hat{\mathbf{y}}_{T+1},$$

where $\hat{\mathbf{y}}_{T+1}$ is the column vector of forecasts at time $T+1$ generated from M forecasting models, and $\mathbf{w}$ is the column vector of weights.

Granger and Ramanathan (1984) investigated some regressive approaches to obtain linear combinations. They demonstrated that the method with a constant term and unrestricted weights performed better. The combination weights can be estimated by Ordinary Least Squares (OLS). Linear combination has a long and successful history in forecasting. However, the issue related to determining the best set of forecasting models to combine is also worthy of attention. Lasso-based methods can do this trick by producing the selection and shrinkage toward zero (Tibshirani 1996). Diebold and Shin (2019) proposed a variant of Lasso, partially-egalitarian LASSO (peLASSO), which set the weights of some forecasting methods to zero and shrunk the survivors toward equality. They provided an empirical assessment to forecast Eurozone GDP growth and found that peLASSO outperformed SA and the median (Diebold and Shin 2019).

The aforementioned weighted forecast combinations need to generate multiple forecasts in the training period, which multiplies the computation time. Especially for highly intermittent demand, the covariance matrix of forecast errors is often singu-

lar (can not calculate the inversion in Bates and Granger's methods (1969)), because the obtained errors may have many zero values. Similarly, for regressive approaches, the standardizing process can not be implemented when the true values for training are always zero. Therefore, Bates and Granger's methods (1969) and regression-based methods are not applicable for highly intermittent data.

Recent studies indicate that using all time series in the dataset to estimate the combination weights show outstanding performance in forecasting fast-moving time series (e.g., Montero-Manso and Hyndman 2021; Talagala, Li, and Kang 2022; Wang et al. 2022b). One mainstream is feature-based forecast combinations. For example, Montero-Manso et al. (2020) developed FFORMA (Feature-based FORecast Model Averaging), which used 42 features to estimate the optimal combination weights based on a meta-learning algorithm.

Different feature-based combination approaches applied different time series features to improve forecasting performance (e.g., Wang, Smith-Miles, and Hyndman 2009; Petropoulos et al. 2014; Kang, Hyndman, and Li 2020; Li, Kang, and Li 2022). However, a significant characteristic of intermittent demand is that there exist a large number of zeros and irregular patterns, which makes the feature sets used in previous literature inapplicable for intermittent demand. Theodorou et al. (2021) proposed a methodological approach for feature extraction and selection to explore the representativeness of M5 dataset. On the basis of the FFORMA framework (Montero-Manso et al. 2020), Kang et al. (2022) used the diversity of forecasting models as the only feature. The diversity has proved to be a novel type of efficient feature, which can not only improve the forecast accuracy but also reduce computational complexity.

The potential of time series features and the diversity of forecasts have not been investigated when producing forecast combinations for intermittent demand. In our work, we extract a set of time series features selected for intermittent demand and calculate the diversity based on a pool of intermittent demand forecasting methods. To this end, a forecast combination framework for intermittent demand can be constructed by mapping the two types of features to the combination weights based on eXtreme Gradient Boosting (XGBoost) (Chen and Guestrin 2016). The proposed framework can be applied to both point and quantile forecast combinations.

3. Forecast combination for intermittent demand

3.1. Time series features for intermittent demand

Several studies have investigated the features of intermittent demand (Kourentzes and Petropoulos 2016; O'Hara-Wild, Hyndman, and Wang 2021; Theodorou et al. 2021). First, we consider the two most popular attributes to divide intermittent demand in the SBC classification scheme (Syntetos, Boylan, and Croston 2005; Kostenko and Hyndman 2006). Then we review the 42 features selected for exploring the feature spaces of M5 competition data (Theodorou et al. 2021). To ensure the interpretability and compute as few features as possible, we remove the features based on complex statistical methods, such as STL decomposition and Fourier transform, and take out Boolean variables with minimal information. The reserved features in Theodorou et al. (2021) are used in our work.

Therefore, we consider nine explainable time series features for intermittent demand forecasting, which are listed in Table 1. They imply the intermittency, volatility, regularity and obsolescence of intermittent demand. Given a time series $\{y_t, t = 1, 2, \cdots, T\}$, we describe the nine features as follows.

- F_1, F_2: The two features are average Inter-Demand Interval (IDI) and squared Coefficient of Variation (CV^2) to measure intermittency and demand size volatility in the SBC classification scheme (Syntetos, Boylan, and Croston 2005; Kostenko and Hyndman 2006).
- F_3: Entropy-based measures have been applied to quantify the regularity and unpredictability of time-series data (Kang, Hyndman, and Smith-Miles 2017; Theodorou et al. 2021). We use approximate entropy in this paper. A relatively small value of F_3 indicates that the demand series includes more regularity and is more forecastable.
- F_4, F_5: The two features describe the ratios of some specific values in a given time series. F_4 measures the percentage of zero values. F_5 denotes the percentage of values lying outside $[\mu_y - \sigma_y, \mu_y + \sigma_y]$, where μ_y and σ_y are the mean and standard deviation of time series $\{y_t\}$, respectively.
- F_6: This feature provides the coefficient of a linear least squares regression, which measures the linear time trend of the variances of component chunks for the target series. For monthly data in the following experiments, we set the chunk

Table 1. Description of nine time series features selected for intermittent demand.

Feature	Description	Range	Implication
F_1: IDI	Averaged inter-demand interval	$[1, \infty)$	Intermittency
F_2: CV^2	Coefficient of variation squared of non-zero demand	$[0, \infty)$	Volatility
F_3: Entropy	Approximate entropy	$(0, 1)$	Regularity
F_4: Percent.zero	The percentage of observations that are zero values	$(0, 1)$	Intermittency
F_5: Percent.beyond.sigma	The percentage of observations that are more than σ away from the mean of the time series (σ is the standard deviation of the time series)	$[0, 1)$	Volatility
F_6: Linear.chunk.var	The coefficient of linear least-squares regression for variances of component chunks in the time series	$(-\infty, \infty)$	Volatility
F_7: Change.mean.abs	The mean absolute value of consecutive changes in the demand.	$[0, \infty)$	Volatility
F_8: Ratio.last.chunk	The ratio of the sum of squares of the last chunk to the whole series	[0,1]	Obsolescence
F_9: Percent.zero.end	The percentage of consecutive zero values at the end of the time series	[0,1]	Obsolescence

length $L = 12$. Moreover for daily data $L = 10$, consistent with Theodorou et al. (2021).

- F_7: F_7 first calculates the consecutive changes in the demand, i.e., the first difference of the demand series. Then the mean absolute value of the consecutive changes is taken.

The last two features focus on the presence of recent demand to capture the obsolescence, which is a challenging problem in the field of intermittent demand (Babai, Syntetos, and Teunter 2014; Babai et al. 2019).

- F_8: F_8 calculates the sum of squares of the last chunk out of K chunks expressed as a ratio with the sum of squares over the whole series. We set $K = 4$ for the Royal Air Force (RAF) dataset and $K = 10$ for M5 competition data, so that the length of the last chunk of each series is longer than the forecasting horizon.
- F_9: F_9 computes the percentage of consecutive zero values at the end of the series, i.e., the number of consecutive zero values at the end over the length of the time series.

Based on the nine time series features tailored for intermittent demand, each target time series can be represented using a nine-dimensional vector. The feature vector can be used as the input to train the forecast combination model in the proposed framework.

3.2. Diversity for intermittent demand

In this paper, we extend Kang et al. (2022)'s work and use the diversity of the pool of methods to develop a forecasting combination method for intermittent demand. The scaled diversity between any two forecasting methods is defined as:

$$DIV_{ij} = \frac{\frac{1}{H}\sum_{h=1}^{H}\left(\hat{y}_{ih} - \hat{y}_{jh}\right)^2}{\left(\frac{1}{T}\sum_{t=1}^{T}|y_t|\right)^2}, \tag{1}$$

where H is the forecast horizon, $\hat{y}_{ih}$ is the h-th step forecast generated from the i-th forecasting model, and $\{y_t, t = 1, 2, \cdots, T\}$ is a series of observed values.

Assuming that the forecasting pool contains M methods, we apply Equation (1) to each two of them. For each target time series, we can construct a diversity vector consisting of $M(M-1)/2$ pairwise diversity measures. This vector can be viewed as the feature vector for the corresponding series in the proposed framework.

The main merits of applying forecast diversity to intermittent demand forecasting are twofold. The first aspect is simplicity in principle, as the calculation only depends on forecasting values, with no need to compute a separate set of features. The second is general applicability. The diversity can be obtained automatically from intermittent demand forecasts and comprehended quickly by forecasters without expertise. Choosing relevant features to match the actual inventory management problem may lead to feature selection bias, especially when forecasters information is inadequate. Therefore, in contrast to time series features, the diversity shows remarkably simplicity and interpretability in intermittent demand forecasting.

3.3. Evaluation metrics for intermittent forecasting

In previous studies, various forecasting evaluation metrics have been used for intermittent demand. The chosen metric of forecast errors may influence the ranked performance of the forecasting methods. Silver et al. (1998) pointed out that no single metric was universally best. Wallström and Segerstedt (2010) discussed a series of forecasting error measurements, especially for intermittent demand and split them into two categories, traditional (accuracy) and bias error measurements. As traditional measures, Wallström and Segerstedt (2010) considered mean absolute deviation (MAD),

mean square error (MSE), symmetric Mean Absolute Percentage Error (sMAPE). As bias error measures, they examined the Cumulated Forecast Error (CFE), Number Of Shortages (NOS), and Periods In Stock (PIS). Kourentzes (2014) evaluated model selection results based on two accuracy metrics. The first is the Mean Absolute Scaled Error (MASE), which was suggested to be the standard measure for the data with different scales and zero values (Hyndman and Koehler 2006). The second is the scaled Absolute Periods In Stock (sAPIS), which is a scale-independent variant of PIS.

However, Kolassa (2016) explored traditional accuracy measures and argued that measures such as MAD, MASE and MAPE are unsuitable for intermittent demand. A flat zero forecast is frequently "best" for the measure of MAE when the demand is highly intermittent, because zero is the conditional median of the demand. Therefore, especially for intermittent demand, an MAE-minimizing method, which is the conditional median, prefers a lower forecast than the MSE-minimizing method, which is the expectation. In recent M5 competition with much intermittent demand, the accuracy of point forecasts is required to be evaluated using Root Mean Squared Scaled Error (RMSSE) (Makridakis, Spiliotis, and Assimakopoulos 2021).

Kolassa (2020) emphasized that different error measures reward different point forecasts, and different measures should not be applied to a single point forecast (Petropoulos et al. 2022). In our work, we use RMSSE (Makridakis, Spiliotis, and Assimakopoulos 2021) to measure the performance of point forecasts, which can be obtained as:

$$RMSSE = \sqrt{\frac{1}{H}\frac{\sum_{h=1}^{H}\left(y_{T+h}-\hat{y}_{T+h}\right)^2}{\frac{1}{T-1}\sum_{t=2}^{T}\left(y_t-y_{t-1}\right)^2}}, \tag{2}$$

where H is the forecasting horizon. $\hat{y}_{T+h}$ is the h-th step forecast generated from a series of observed values $\{y_t, t = 1, 2, \cdots, T\}$, and y_{T+h} is the true value.

3.4. Generalized forecast combination framework

The quality of forecast combination has been demonstrated to depend on the individual forecasts as well as the diversity between forecasts (Lemke and Gabrys 2010; Kourentzes, Barrow, and Petropoulos 2019; Kang et al. 2022). Therefore, defining an appropriate forecasting pool is one of the most crucial steps in the forecast combination process. Firstly, we define a broad pool for intermittent demand forecasting.

The pool includes traditional forecasting models, which are Naive, seasonal Naive (sNaive), Simple Exponential Smoothing (SES), Moving Averages (MA), AutoRegressive Integrated Moving Average (ARIMA), ExponenTial Smoothing (ETS), and intermittent demand forecasting methods, which are Crostons method (CRO), optimized Crostons method (optCro), SBA, TSB, ADIDA, IMAPA. The 12 forecasting methods in the pool are considered as statistical benchmarks in the M5 competition (Makridakis, Spiliotis, and Assimakopoulos 2021). In contrast to CRO with fixed smoothing parameters, the parameters in optCro are optimized to allow for more flexibility. Implementations for these methods exist in the **forecast** (Hyndman et al. 2020) and **tsintermittent** (Kourentzes and Petropoulos 2016) packages in R. Then the pooling methods (Kourentzes, Barrow, and Petropoulos 2019; Lichtendahl Jr and Winkler 2020; Diebold and Shin 2019) can be applied to reduce the number of forecasting methods and further improve the quality of the forecasting pool. We study the effect of three popular pooling algorithms in Section 4.

In the proposed forecast combination framework, we build an XGBoost model to learn the relationship between features and combination weights. This approach transforms the combination problem into a classification problem by setting the best forecasting method as the target class for each time series. The two types of features in Section 3.1 and Section 3.2 are all valid inputs for the forecast combination model. We name the approach based on the nine time series features in Section 3.1 as Feature-based Intermittent DEmand forecasting (FIDE). The diversity-based method is called DIVersity-based Intermittent DEmand forecasting (DIVIDE).

Then, given a forecast error metric, the optimization objectives for the FIDE and DIVIDE are

$$\arg\min_{w_F} \sum_{n=1}^{N}\sum_{i=1}^{M} w(F_n)_i \times \text{error}_{n,i}, \text{ and} \tag{3a}$$

$$\arg\min_{w_D} \sum_{n=1}^{N}\sum_{i=1}^{M} w(D_n)_i \times \text{error}_{n,i}, \tag{3b}$$

respectively, where F_n is the feature vector, D_n is the diversity vector of the n-th time series, N is the number of time series, and M is the number of forecasting methods. $\text{error}_{n,i}$ is the forecast error of the i-th method for the n-th time series. RMSSE is used as the error measure of point forecasts in this paper. RMSSE focuses

on the expectation, which is consistent with the candidate methods in the forecasting pool. Once the model has been trained, weights can be produced for a new series to generate the combined forecast. The process can be implemented based on the R package **M4metalearning** by Montero-Manso et al. (2020).

Based on the forecasting pool for intermittent demand, we put forward a generalized forecast combination framework containing FIDE and DIVIDE. The flowchart of the proposed framework is presented in Figure 1. In the training phase, we generate forecasts based on the methods in the intermittent demand forecasting pool and calculate errors required in the objective function. In FIDE, we compute the features selected for intermittent demand and learn the relationship between the features and combination weights by Equation (3a). In DIVIDE, the combination model can be obtained based on the diversity of different forecast methods (see Equation (3b)), where the pairwise diversity values of the methods in the pool are used as time series features. Therefore, DIVIDE can be viewed as a special case of FIDE. In the forecasting phase, we calculate the features or the diversity for the new time series, and get the combination weights through the pre-trained XGBoost model. Finally, we utilize the optimal weights to average the forecasts from different methods in the pool and achieve the combined forecast results.

To implement the proposed framework, the time series need to be divided into three periods. Let H be the forecast horizon and T be the length of data. The first $T-H$ observations are used for training the forecast combination model. Then the $T-H$ observations are split into $T-H-H$ for training the forecasting methods in the pool and H for computing the accuracy of each method. The final H observations are used to evaluate the forecasting results.

The merits of the proposed framework include: (1) using a diverse forecasting pool, consisting of intermittent demand forecasting methods and traditional time series forecasting models, (2) considering a customizable objective function, which can be modified to an error measure of interest in real inventory management, (3) selecting intelligible time series features especially for intermittent demand, and (4) using the diversity as another form of features, whose calculation only depends on forecasting values.

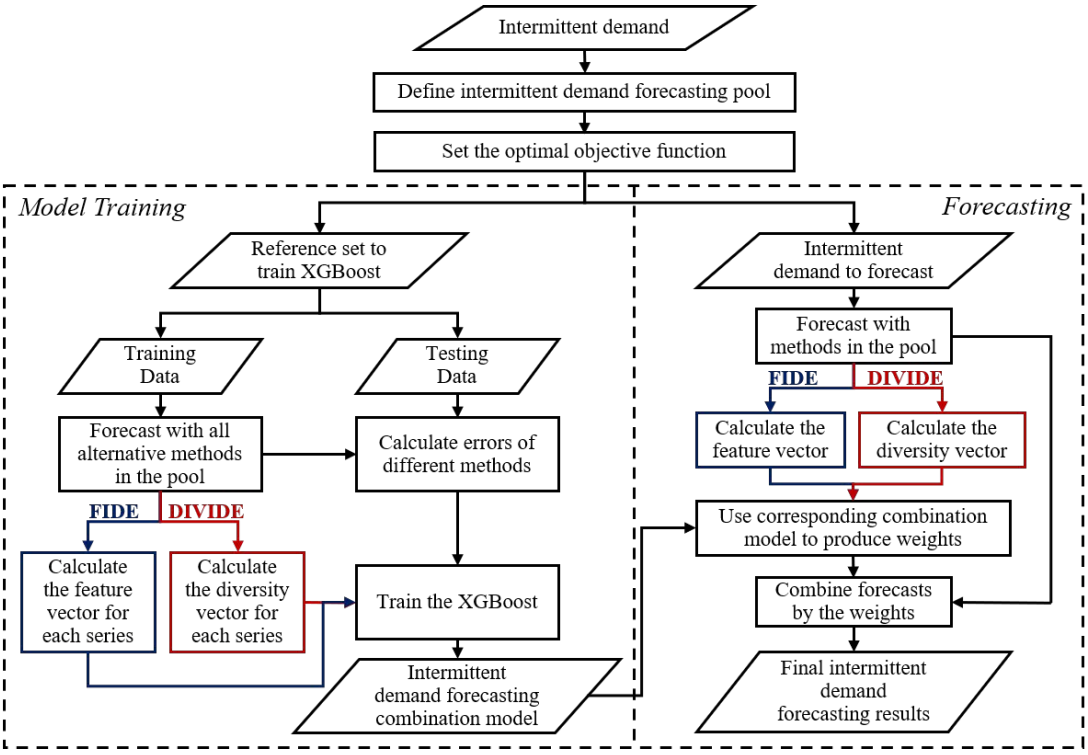

Figure 1. Caption: The flowchart of the proposed forecast combination framework for intermittent demand.
Figure 1. Alt Text: The flowchart contains two phases of model training and forecasting. The blue line refers to FIDE and the red line denotes DIVIDE.

4. Empirical evaluation

4.1. Real dataset

The proposed methods are applied to two real datasets. The first RAF dataset has been previously investigated in the literature (Kourentzes and Athanasopoulos 2021; Petropoulos and Kourentzes 2015; Teunter and Duncan 2009). It contains 5000 monthly time series, with 84 observations each. Moreover, the second is M5 competition data, involving the unit sales of 3049 products sold by Walmart in the USA between 2011-01-29 and 2016-06-19 (1969 days). The dataset was organized in the form of hierarchical time series in M5 competition (Makridakis, Spiliotis, and Assimakopoulos 2021). We only consider the bottom level, i.e., 30,490 product-store unit sales in this paper.

In the following experiment, we examine three forecast horizons of 3, 6 and 12 months ahead for RAF dataset and 28-day-ahead forecasts for M5 data as required in M5 competition (Makridakis, Spiliotis, and Assimakopoulos 2021). The final obser-

vations of the horizon length are used to evaluate the forecasting performance. The seasonal periods are 12 and 7 for monthly and daily demand, respectively. Moreover, we preprocess the data before forecasting, removing the initial zero values and making the first non-zero demand as the initial value (Theodorou et al. 2021). This is due to a lack of information that the initial zeros mean demands or sales are zero, or the product has not been in stores yet.

Following the SBC scheme (Kostenko and Hyndman 2006), the time series can be divided into four categories based on IDI and CV^2. Figure 2 describes the distributions of RAF and M5 datasets, respectively. The boundaries of different categories in Figure 2 are IDI= 4/3, $CV^2 = 0.5$ (Kostenko and Hyndman 2006). The equal sign is placed on the less-than sign when classifying, e.g., if IDI> 4/3 and CV^2 <= 0.5, the time series is "intermittent". As shown in Figure 2, the RAF dataset exhibits high intermittence and contains 2729 intermittent, and 2271 lumpy series. While the M5 data has a wider-ranging distribution, including 22,206 intermittent, 5359 lumpy, 897 erratic, and 2028 smooth series.

4.2. Point forecasting

We compare our methods with individual models, SA, Median and FFORMA. Other combination methods reviewed in Section 2, such as Bates and Granger (1969)'s original forecast combinations and regression-based methods (Granger and Ramanathan 1984; Diebold and Shin 2019), are omitted here, which are not applicable for highly intermittent data. We present the forecasting accuracy of different methods based RAF dataset and M5 competition data in Table 2 and Table 3 respectively.

As shown in Table 2, the best individual method is ADIDA for all forecast horizons based on RMSSE. The simple combination methods (SA and Median) can not beat the best individual method. While the proposed methods based on intermittent demand features and the diversity consistently outperform others. FIDE shows obvious superiority compared with FFORMA using 42 time series features and offers the best forecasting results overall. Therefore, our chosen features are more appropriate to describe intermittent demand compared with the time series features in FFORMA designed for fast-moving data. Furthermore, the improved forecast accuracy of DIVIDE indicates that the diversity is a simple and efficient tool for intermittent demand forecasting combinations.

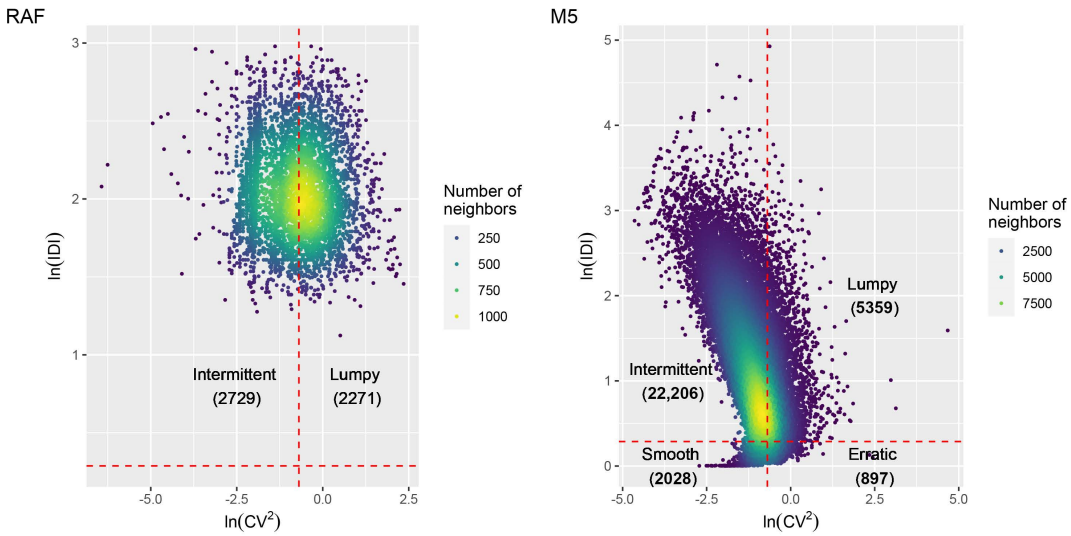

Figure 2. Caption: The density scatterplots of $\ln(\mathsf{CV}^2)$ and $\ln(\mathsf{IDI})$, for RAF (left) and M5 (right) datasets.

Figure 2. Alt Text: The x-axis and y-axis show the natural logarithmic transform of CV^2 and that of IDI, respectively, for RAF (left) and M5 (right) datasets. To present the overall distribution of data, each point is colored by the number of neighboring points. Red lines indicate the boundaries of different categories ($\mathsf{IDI} = 4/3$, $\mathsf{CV}^2 = 0.5$) (Kostenko and Hyndman 2006). In the left panel, the RAF dataset contains 2729 intermittent, and 2271 lumpy series. In the right panel, the M5 data includes 22,206 intermittent, 5359 lumpy, 897 erratic, and 2028 smooth series.

The forecasting results of M5 competition data in Table 3 are organized by SBC classification scheme (Kostenko and Hyndman 2006). For each column (a subset of data), the combination model is optimized respectively. The last three rows show the RMSSE of the top three winning methods in the M5 competition for comparison, e.g., "M5-w1" denotes the first ranked method. The results of last three rows in Table 3 are calculated based on corresponding M5 submissions. It should be noted that the M5 competition took the hierarchy of data into consideration and used weighted RMSSE to rank participants. The weights put more emphasis on the series that account for higher monetary sales (Makridakis, Spiliotis, and Assimakopoulos 2021). Therefore, comparing these methods in absolute terms is not entirely fair, as they were obtained from different application contexts and optimization objectives.

As shown in Table 3, the best individual method is IMAPA for all classifications, which is inconsistent with the RAF data. The results emphasize the risk of choosing a single forecasting method and elicit the necessity of forecast combination. Our proposed methods achieve the competitive forecasting performance based on RMSSE when compared with the top three ranked methods in the M5 competition. Based on different classifications of M5 data, the performance of the proposed methods exhibits significant differences. The proposed FIDE and DIVIDE outperform FFORMA for the intermittent and lumpy data, which is consistent with the RAF dataset. The forecasting results based on the two datasets provide good evidence for the superiority of the proposed framework in intermittent demand forecasting. However, for the erratic and smooth data, our methods perform slightly worse than FFORMA. We acknowledge the limitations of the features used in the proposed framework, which are more applicable to intermittent demand.

The features and diversity in the proposed framework have been proven to improve the accuracy of intermittent demand forecasting. In the following experiments, we continue to provide a sensitivity analysis of multiple features, including diversity viewed as another form of features. We investigate the relationship between RMSSE and the number of features used in the proposed FIDE and DIVIDE, as shown in Figure 3. In FIDE, we set the feature number to be 3, 6 and 9(all). While in DIVIDE, the feature number varies across 5, 10, 20, 30, 40, 50 and 66(all). We use two ways to alter the number of features. One is to select features in the order of feature importance in XGBoost model, and the other is by random feature selection. The importance of each

Table 2. Forecasting accuracy (RMSSE) of different methods based on the RAF dataset. For each forecasting horizon (column), the smallest value is marked in **bold**. The last column is the average rank of all methods based on the three horizons.

Method	$H = 3$	$H = 6$	$H = 12$	Average rank
Naive	0.493	0.552	0.658	13.0
sNaive	0.619	0.764	0.911	17.0
SES	0.466	0.540	0.641	8.3
MA	0.461	0.551	0.644	9.0
ARIMA	0.490	0.558	0.616	12.0
ETS	0.483	0.552	0.615	9.2
CRO	0.500	0.564	0.616	14.0
optCro	0.500	0.564	0.616	14.0
SBA	0.499	0.562	0.615	12.2
TSB	0.487	0.554	0.612	10.0
ADIDA	0.464	0.539	0.606	5.0
IMAPA	0.478	0.545	0.603	5.8
SA	0.485	0.552	0.618	11.0
Median	0.478	0.546	0.605	6.5
FFORMA	0.413	0.491	0.583	3.0
FIDE	0.369	**0.461**	**0.562**	**1.3**
DIVIDE	**0.359**	0.462	0.563	1.7

Table 3. Forecasting accuracy (RMSSE) of different methods based on the M5 competition dataset. The last three rows show the RMSSE of the top three winning methods in the M5 competition for comparison. The results for different classifications and the whole data set are presented respectively. For each column, the smallest value is marked in **bold** (without including the last three rows).

Method	Intermittent	Lumpy	Erratic	Smooth	All
Naive	0.888	0.835	0.917	1.001	0.887
sNaive	0.892	0.840	0.929	1.024	0.892
SES	0.835	0.794	0.870	0.925	0.835
MA	0.859	0.807	0.884	0.960	0.857
ARIMA	0.820	0.782	0.842	0.898	0.819
ETS	0.822	0.790	0.868	0.920	0.824
CRO	0.814	0.782	0.860	0.930	0.817
optCro	0.810	0.782	0.845	0.850	0.809
SBA	0.811	0.784	0.841	0.852	0.810
TSB	0.809	0.780	0.838	0.848	0.807
ADIDA	0.823	0.790	0.870	0.925	0.826
IMAPA	0.806	0.775	0.832	0.842	0.804
SA	0.813	0.773	0.816	0.865	0.809
Median	0.819	0.784	0.850	0.901	0.819
FFORMA	0.809	0.747	**0.739**	**0.775**	0.801
FIDE	0.792	0.745	0.746	0.783	0.783
DIVIDE	**0.790**	**0.741**	0.743	0.776	**0.779**
M5-w1	0.785	0.734	0.743	0.774	0.774
M5-w2	0.799	0.747	0.767	0.797	0.789
M5-w3	0.784	0.731	0.730	0.769	0.772

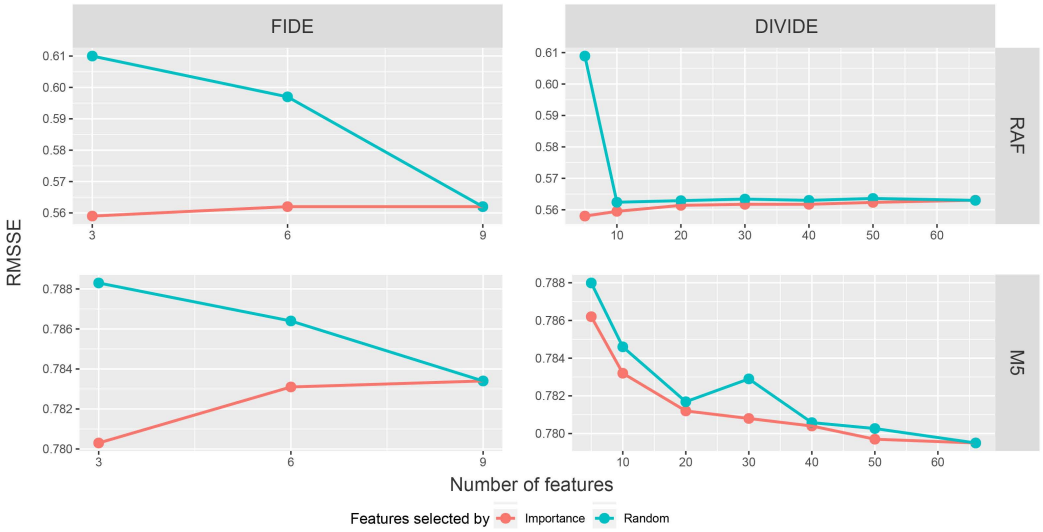

Figure 3. Caption: The relationship between RMSSE and the number of features used in the proposed FIDE (left) and DIVIDE (right), for RAF (top) and M5 (bottom) datasets.
Figure 3. Alt Text: The results based on RAF dataset (indicatively for $H = 12$) are presented on the top panel, and the bottom panel shows the results from M5 competition data. The features in FIDE and DIVIDE are selected based on two methods. One is to select features in order of importance in XGBoost model (red lines), and the other is by random selection (blue lines).

feature to the FIDE or DIVIDE is measured by the gain of features in the XGBoost model (Chen and Guestrin 2016).

As shown in Figure 3, with the increase of the feature number, RMSSE of FIDE decreases when selecting features randomly (blue lines), but increases when selecting features in order of importance (red lines). The findings emphasize the importance of choosing appropriate features as input in the proposed FIDE. For RAF dataset, the three most important features are Percent.zero.end (F_9), Ratio.last.chunk (F_8) and Linear.chunk.var (F_6). While for M5 competition data, the top three features are Percent.zero.end (F_9), IDI (F_1) and Ratio.last.chunk (F_8). Thus, the features to capture the recent demand are more critical for constructing the combination model in FIDE. However, the relationship between RMSSE and the feature number in DIVIDE is markedly different from that of FIDE. As the number of features increases, the overall trend of RMSSE is downward, though there is a non-significant increase when considering the importance of features for RAF dataset. Therefore, we recommend applying the whole diversity in the proposed framework.

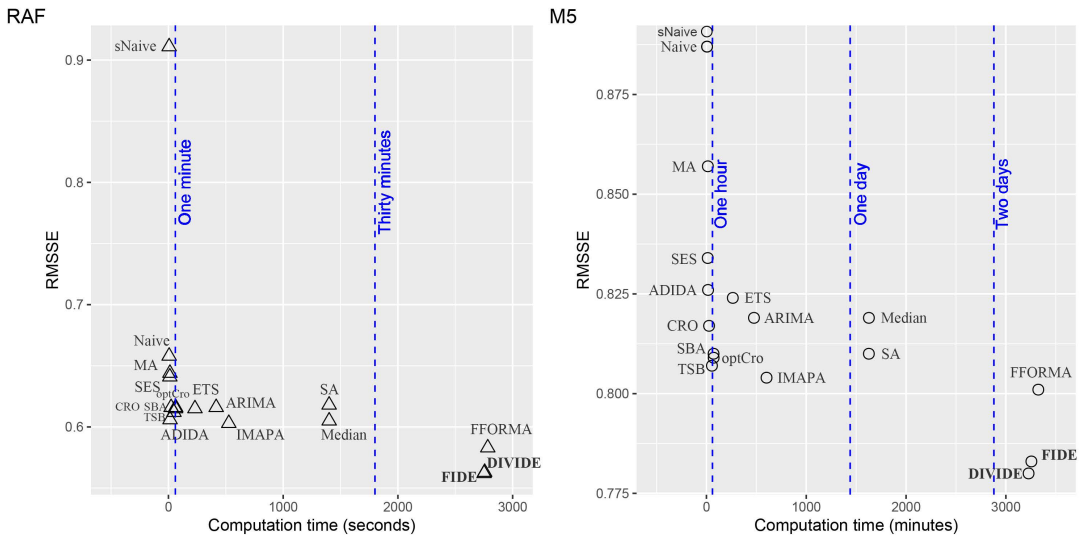

Figure 4. Caption: The relationship between RMSSE and computation time, for RAF (left) and M5 (right) datasets.
Figure 4. Alt Text: Different forecasting methods are marked on the picture. The computation time is obtained by using a Microsoft Windows 10 desktop with 8 cores and 16 logical processors at 3.59 GHz, 16 GB RAM.

To analyze the efficiency of the examined forecasting methods, we investigate the relationship between RMSSE and computation time based on RAF and M5 datasets. The results are computed indicatively for RAF dataset when $H = 12$. As shown in Figure 4, the time consumption of our methods mainly comes from the individual forecasting methods, which is the limitation of forecast combinations. Simple combination schemes, such as SA and median, can save nearly half the time, but they perform worse than the best individual method for intermittent demand. The proposed framework generates forecasts in the training and forecasting periods, increasing the time consumption. However, the process of model training is in the off-line phase in real applications. Therefore, the increased computational time of training does not affect the forecasting efficiency. Compared with FFORMA, our methods based on features for intermittent demand and diversity are more computationally efficient, especially for the M5 dataset with large amounts of long time series. Based on these findings, decision makers can consider the trade-off between accuracy and computational cost in actual inventory management.

Reducing the number of forecasting methods used in the proposed framework can significantly save computational time. For instance, the computation time of our meth-

ods can be halved by removing the three most time-consuming methods in the forecasting pool. In the following experiment, we aim to investigate the potential of shrinking the forecasting pool to further improve the accuracy of the proposed framework. We call this process pooling, deriving from Kourentzes, Barrow, and Petropoulos (2019)'s research.

We study the effect of three pooling algorithms based on the RAF (indicatively for $h = 12$) and M5 dataset, which are forecast islands proposed by Kourentzes, Barrow, and Petropoulos (2019), a screened method from Lichtendahl Jr and Winkler (2020), and a Lasso-based method by Diebold and Shin (2019). The forecast islands (Kourentzes, Barrow, and Petropoulos 2019) remove some poorly performing models from the pool, which is shortened to "Islands". It conducts $C' = \{0, \Delta C\}$ for a series of ordered forecasts based on a criterion of forecasting performance C and includes all forecasts until $C' \geq T$. $T = \text{Q3} + 1.5\text{IQR}$ is related to the outlier of the boxplot, where Q3 is the third quartile and IQR is the interquartile range. The screened method (Lichtendahl Jr and Winkler 2020), shortened to "Screened", screens out forecasting models with highly correlated errors (correlation coefficient is over 0.95). The Lasso-based method (Diebold and Shin 2019), "Lasso" for short, sets the regression coefficients of some forecasts to zero via a standard Lasso software, i.e., R package **glmnet** (Friedman, Hastie, and Tibshirani 2010), and the survivors form the final forecast pool.

Table 4 presents the forecasting accuracy of original FIDE and DIVIDE and those considering the three pooling algorithms. In Table 4, we find that none of the pooling methods significantly improves the forecasting performance. The proposed framework automatically reduces the weights of some methods to minimal values, which can be regarded as a generalized pooling method customized for each time series. The Islands remove the worst performing methods in the pool, but reduce the accuracy, especially for RAF dataset. For highly intermittent data, poorly performing methods, such as Naive, sNaive, make important contributions to the forecast combination. Therefore, unless the pool of forecasting methods is too large, which would render the computation process time-consuming, there is no need to add modeling complexity by implementing pooling approaches on top of our proposed framework.

Table 4. Forecasting accuracy (RMSSE) of the original FIDE and DIVIDE methods and those considering pooling algorithms based on RAF and M5 dataset. For each row (dataset), the smallest values for FIDE and DIVIDE are marked in **bold**, respectively.

	FIDE				DIVIDE			
	Original	Islands	Screened	Lasso	Original	Islands	Screened	Lasso
RAF($H = 12$)	0.562	0.600	0.562	**0.561**	0.563	0.597	0.561	**0.560**
M5	**0.783**	0.787	**0.783**	0.784	**0.779**	0.783	0.781	0.782

4.3. Quantile forecasting

Based on the improving accuracy of point forecasts, the proposed combination methods are shown to be effective in providing robust forecasts to support decisions. However, in real supply chain management, estimating the right part of the demand distribution is also necessary for determining safety stock levels, which has been largely ignored in the research (Barrow and Kourentzes 2016; Spiliotis et al. 2021). Fildes, Ma, and Kolassa (2019) reviewed retail demand forecasting and emphasized the connection of quantile, density, or volatility forecasting to inventory control.

The intermittent demand forecasting methods (such as CRO, optCro, SBA, TSB, ADIDA and IMAPA) can not directly output quantile forecasts. We apply Trapero, Cardos, and Kourentzes (2019)'s empirical approach to estimate the desired quantiles. They recommended a kernel density estimation to model the forecast error distribution. We generate quantile forecasts by adjusting the point forecasts based on the respective quantiles calculated from the empirical distribution of residual errors, as follows:

$$Q_{T+h}(u) = \hat{y}_{T+h} + \hat{q}_{|e}(u), \tag{4}$$

where T is the length of observations, $Q_{T+h}(u)$ is the probabilistic forecast for quantile u at time $T + h$, $\hat{y}_{T+h}$ is the h-th step point forecast, $\hat{q}_{|e}(u)$ is the estimated u-th quantile of the residual errors. As shown in Equation (4), we assume that the demand pattern that occurred in the past will continue in the future. The obtained forecasts are based on in-sample approximations without requiring computing multiple forecasts. The approach has been verified to perform well for the RAF and M5 datasets (Spiliotis et al. 2021; Kourentzes and Athanasopoulos 2021).

The proposed framework can be extended to quantile forecast combinations by

mapping the features to the errors of quantile forecasts. In FIDE and DIVIDE, we still compute the nine features in Section 3.1 based on historical data and the diversity of different point forecasts as shown in Section 3.2. The training and testing processes are consistent with Section 3.4. We use the Scaled Pinball Loss (SPL) function to measure the precision of the quantile forecasts, which is required in the M5 competition (Makridakis, Spiliotis, and Assimakopoulos 2021). The SPL can be obtained as follows:

$$SPL(u) = \frac{\sum_{h=1}^{H} u\left(y_{T+h} - Q_{T+h}(u)\right)\mathbf{1}\left\{Q_{T+h}(u) \leq y_{T+h}\right\} + (1-u)\left(Q_{T+h}(u) - y_{T+h}\right)\mathbf{1}\left\{Q_{T+h}(u) > y_{T+h}\right\}}{H \cdot \frac{1}{T-1}\sum_{t=2}^{T}|y_t - y_{t-1}|}, \tag{5}$$

where y_{T+h} is the actual future value of the examined time series at point $T+h$, $Q_{T+h}(u)$ is the generated forecast for quantile u, H is the forecasting horizon, T is the length of the number of historical observations, and $\mathbf{1}$ is the indicator function (being 1 if true is within the postulated interval and 0 otherwise).

The following experiment based on the RAF and M5 datasets focuses on four quantiles, i.e. $u_1 = 0.750$, $u_2 = 0.835$, $u_3 = 0.975$, and $u_4 = 0.995$. u_1 and u_2 provide a good sense of the mid-right part of the distribution, while u_3 and u_4 provide information about its right tail, which is essential for the risk of extreme outcomes. We customize the objective function by assigning the error measure to SPL based on the correlated quantiles. Therefore, we obtain different combination weights for the four quantiles, respectively. The forecasting results in Table 5 are computed based on 12-month-ahead forecasts for RAF dataset and 28-day-ahead forecasts for M5 data.

We can find in Table 5 that the performance of individual methods changes considerably based on different quantiles. The finding indicates that each method is more appropriate for estimating different parts of the distribution of the series, which echoes the weakness of choosing a single method. The proposed combination methods exhibit steady performance across the four quantiles. For RAF dataset, FIDE and DIVIDE consistently outperform the rest in Table 5. For M5 dataset, we add the top three ranked methods in the M5 competition for comparison, the results of which derive from Spiliotis et al. (2021). The proposed DIVIDE outperforms the first ranked method in M5 competition at quantiles 0.835, 0.975 and 0.995. While at quantile 0.750, our methods also provide competitive forecasting results. The improved performance of high quantile forecasts can contribute to practical inventory decisions for higher levels

Table 5. Quantile forecasting performance (SPL) of different methods based on the RAF and M5 datasets. The last three rows show the SPL of the top three winning methods in the M5 competition for comparison. The results based on four quantiles (0.750, 0.835, 0.975, and 0.995) are reported. For each column (quantile), the smallest value is marked in **bold** (without including the last three rows).

Method	RAF				M5			
	0.750	0.835	0.975	0.995	0.750	0.835	0.975	0.995
Naive	1.395	1.296	0.455	0.183	1.078	1.035	0.329	0.090
sNaive	0.844	0.778	0.353	0.211	0.636	0.557	0.181	0.063
SES	0.864	0.793	0.353	0.207	0.576	0.510	0.214	0.110
MA	0.503	0.509	0.351	0.205	0.599	0.561	0.204	0.068
ARIMA	0.740	0.685	0.354	0.237	0.586	0.514	0.213	0.109
ETS	0.741	0.687	0.355	0.238	0.579	0.509	0.215	0.113
CRO	0.404	0.447	0.348	0.192	0.580	0.519	0.171	0.053
optCro	0.406	0.448	0.349	0.192	0.566	0.510	0.169	0.053
SBA	0.405	0.448	0.349	0.192	0.566	0.509	0.169	0.053
TSB	0.406	0.449	0.348	0.192	0.566	0.510	0.169	0.053
ADIDA	0.583	0.560	0.405	0.343	0.568	0.520	0.292	0.199
IMAPA	0.404	0.455	0.367	0.217	0.561	0.504	0.170	0.056
SA	0.629	0.609	0.347	0.198	0.569	0.500	0.168	0.057
Median	0.532	0.531	0.345	0.199	0.556	0.496	0.171	0.056
FFORMA	0.401	0.450	0.343	0.188	0.523	0.471	0.156	0.051
FIDE	0.402	**0.446**	**0.340**	**0.182**	0.516	0.456	0.150	0.046
DIVIDE	**0.400**	0.449	0.341	0.184	**0.511**	**0.453**	**0.146**	**0.045**
M5-w1					0.509	0.455	0.151	0.048
M5-w2					0.610	0.492	0.157	0.055
M5-w3					0.513	0.457	0.165	0.070

of service.

5. Conclusion

This paper focuses on forecast combinations for intermittent demand. We review a handful of forecasting methods, and investigate the performance of some existing forecast combination methods for intermittent demand. We introduce time series features and diversity to propose a generalized forecast combination framework, which can automatically determine the optimal combination weights. We conduct an empirical investigation based on real-life data to analyze the forecast accuracy and gain insights related to inventory decisions.

The results of point forecasts are measured by RMSSE, which focuses on the expectation. The proposed framework notably outperforms other combination methods and the best individual method, especially for the RAF dataset with highly intermittent series. Moreover, for M5 competition data, our methods achieve a competitive

performance compared with the top three ranked methods in the M5 competition. In addition, the proposed framework can be regarded as a generalized pooling method customized for each time series by reducing the weights of some methods to minimal values. The empirical evaluation based on RAF and M5 datasets provides good evidence of the superiority and flexibility of the proposed framework. We acknowledge that our combination methods increase the computational time compared with individual methods. Decision makers should consider the trade-off between accuracy and computational cost in actual inventory management.

The proposed framework has also been applied to quantile forecast combinations, especially for high quantiles to estimate the right part of the demand distribution. We use SPL to measure the quantile forecasting performance and make it used in the optimization objective. The examined results show that our methods can provide accurate forecasts of both central tendency and high quantiles, which directly connect with the inventory decision.

The good performance of our proposed framework can be attributed to: (i) defining an appropriate forecasting pool on the top of the framework, which consists of intermittent demand forecasting methods and traditional time series forecasting models, (ii) applying diversity or time series features to determine the optimal combination weights automatically, and (iii) applying to both point and quantile forecasts to support inventory decisions. The diversity and the features selected for intermittent demand are all effective inputs of the proposed framework. Extracting the diversity independent of historical data makes it more flexible for intermittent demand forecasting, especially when the training set is limited in positive demands. In addition, the features in FIDE are all easily understood. The two features focusing on the presence of recent demand are proved more critical for constructing the forecast combination model. These advantages of the proposed methods lead to broad application prospects in intermittent demand forecasting.

However, we recognize the lack of a comprehensive evaluation of inventory performance in the current study. Petropoulos, Wang, and Disney (2019) combined financial, operational, and service metrics to form a holistic measure for inventory control objectives. Ducharme, Agard, and Trépanier (2021) focused on stock-out events and proposed a novel metric called Next Time Under Safety Stock. The utility measures are essential to achieve a direct link between inventory holding costs and service levels

in the production system. Such analysis needs to proceed based on restocking policies, which are not available for the RAF and M5 datasets without any background information of inventory. Future research should investigate the inventory performance of our proposed framework in the field of a specific inventory management problem. Another limitation of this paper is lacking an automatic procedure for choosing features for modeling FIDE. Several scholars have investigated selecting features automatically from a large number of features (Lubba et al. 2019; Theodorou et al. 2021). Although these approaches seem more general, they take over much computational time, and the selected features are often difficult to understand in the applications. Based on the results of our work, the nine features in FIDE are efficient and can be used as the benchmark pool of features for intermittent demand. In further research, we will study a standard procedure to select features automatically for the proposed framework, aiming to achieve both interpretability and computational efficiency.

Data Availability Statement

The RAF dataset has been used in previous literature (Teunter and Duncan 2009; Petropoulos and Kourentzes 2015; Kourentzes and Athanasopoulos 2021) and is available upon request. The M5 competition (Makridakis, Spiliotis, and Assimakopoulos 2021) data involves the unit sales of 3049 products between 2011-01-29 and 2016-06-19 (1969 days). The first 1941 observations for model training can be obtained from `https://github.com/Mcompetitions/M5-methods`; the final 28 observations are available upon request.

Acknowledgments

Yanfei Kang is supported by the National Natural Science Foundation of China (No. 72171011). Feng Li is supported by the Beijing Universities Advanced Disciplines Initiative (No. GJJ2019163) and the Emerging Interdisciplinary Project of CUFE. This research was supported by Alibaba Group through the Alibaba Innovative Research Program and the high-performance computing (HPC) resources at Beihang University.

Declaration of Interest Statement

No potential conflict of interest was reported by the authors.

References

Babai, Mohamed Zied, Yves Dallery, Selmen Boubaker, and Rim Kalai. 2019. "A New Method to Forecast Intermittent Demand in the Presence of Inventory Obsolescence." *International Journal of Production Economics* 209: 30–41.

Babai, Mohamed Zied, Aris Syntetos, and Ruud Teunter. 2014. "Intermittent Demand Forecasting: An Empirical Study on Accuracy and the Risk of Obsolescence." *International Journal of Production Economics* 157: 212–219.

Balugani, Elia, Francesco Lolli, Rita Gamberini, Bianca Rimini, and M Zied Babai. 2019. "A Periodic Inventory System of Intermittent Demand Items with Fixed Lifetimes." *International Journal of Production Research* 57 (22): 6993–7005.

Barrow, Devon K, and Nikolaos Kourentzes. 2016. "Distributions of Forecasting Errors of Forecast Combinations: Implications for Inventory Management." *International Journal of Production Economics* 177: 24–33.

Bates, John M, and Clive WJ Granger. 1969. "The Combination of Forecasts." *Journal of the Operational Research Society* 20 (4): 451–468.

Boylan, John E, and M Zied Babai. 2016. "On the Performance of Overlapping and Non-overlapping Temporal Demand Aggregation Approaches." *International Journal of Production Economics* 181: 136–144.

Boylan, John E, and Aris A Syntetos. 2021. *Intermittent Demand Forecasting: Context, Methods and Applications*. John Wiley & Sons.

Chen, Tianqi, and Carlos Guestrin. 2016. "Xgboost: A Scalable Tree Boosting System." In *Proceedings of the 22nd ACM SIGKDD International Conference on knowledge Discovery and Data Mining*, 785–794.

Claeskens, Gerda, Jan R Magnus, Andrey L Vasnev, and Wendun Wang. 2016. "The Forecast Combination Puzzle: A Simple Theoretical Explanation." *International Journal of Forecasting* 32 (3): 754–762.

Clemen, Robert T. 1989. "Combining Forecasts: A Review and Annotated Bibliography." *International Journal of Forecasting* 5 (4): 559–583.

Croston, John D. 1972. "Forecasting and Stock Control for Intermittent Demands." *Journal of the Operational Research Society* 23 (3): 289–303.

De Menezes, Lilian M, Derek W Bunn, and James W Taylor. 2000. "Review of Guidelines

for the Use of Combined Forecasts." *European Journal of Operational Research* 120 (1): 190–204.

Diebold, Francis X, and Minchul Shin. 2019. "Machine Learning for Regularized Survey Forecast Combination: Partially-egalitarian LASSO and Its Derivatives." *International Journal of Forecasting* 35 (4): 1679–1691.

Ducharme, Corey, Bruno Agard, and Martin Trépanier. 2021. "Forecasting a Customer's Next Time Under Safety Stock." *International Journal of Production Economics* 234: 108044.

Fildes, Robert, Shaohui Ma, and Stephan Kolassa. 2019. "Retail Forecasting: Research and Practice." *International Journal of Forecasting* (in press).

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. 2010. "Regularization Paths for Generalized Linear Models via Coordinate Descent." *Journal of Statistical Software* 33 (1): 1–22.

Granger, Clive WJ, and Ramu Ramanathan. 1984. "Improved Methods of Combining Forecasts." *Journal of Forecasting* 3 (2): 197–204.

Hasni, Marwa, MS Aguir, Mohamed Zied Babai, and Zied Jemai. 2019a. "On the Performance of Adjusted Bootstrapping Methods for Intermittent Demand Forecasting." *International Journal of Production Economics* 216: 145–153.

Hasni, Marwa, MS Aguir, Mohamed Zied Babai, and Zied Jemai. 2019b. "Spare Parts Demand Forecasting: A Review on Bootstrapping Methods." *International Journal of Production Research* 57 (15-16): 4791–4804.

Hyndman, Rob, George Athanasopoulos, Christoph Bergmeir, Gabriel Caceres, Leanne Chhay, Mitchell O'Hara-Wild, Fotios Petropoulos, Slava Razbash, Earo Wang, and Farah Yasmeen. 2020. *forecast: Forecasting Functions for Time Series and Linear Models*. R package version 8.12, http://pkg.robjhyndman.com/forecast.

Hyndman, Rob J, and Anne B Koehler. 2006. "Another Look at Measures of Forecast Accuracy." *International Journal of Forecasting* 22 (4): 679–688.

Jiang, Peng, Yibin Huang, and Xiao Liu. 2021. "Intermittent Demand Forecasting for Spare Parts in the Heavy-duty Vehicle Industry: A Support Vector Machine Model." *International Journal of Production Research* 59 (24): 7423–7440.

Jose, Victor Richmond R, and Robert L Winkler. 2008. "Simple Robust Averages of Forecasts: Some Empirical Results." *International Journal of Forecasting* 24 (1): 163–169.

Kang, Yanfei, Wei Cao, Fotios Petropoulos, and Feng Li. 2022. "Forecast with Forecasts: Diversity Matters." *European Journal of Operational Research* 301 (1): 180–190.

Kang, Yanfei, Rob J Hyndman, and Feng Li. 2020. "GRATIS: GeneRAting TIme Series with Diverse and Controllable Characteristics." *Statistical Analysis and Data Mining: The ASA Data Science Journal* 13 (4): 354–376.

Kang, Yanfei, Rob J Hyndman, and Kate Smith-Miles. 2017. "Visualising Forecasting Algorithm Performance Using Time Series Instance Spaces." *International Journal of Forecasting* 33 (2): 345–358.

Kaya, Gamze Ogcu, and Ali Turkyilmaz. 2018. "Intermittent Demand Forecasting Using Data Mining Techniques." *Applied Computer Science* 14 (2): 38–47.

Kolassa, Stephan. 2016. "Evaluating Predictive Count Data Distributions in Retail Sales Forecasting." *International Journal of Forecasting* 32 (3): 788–803.

Kolassa, Stephan. 2020. "Why the "Best" Point Forecast Depends on the Error or Accuracy Measure." *International Journal of Forecasting* 36 (1): 208–211.

Kostenko, Audrey V, and Rob J Hyndman. 2006. "A Note on the Categorization of Demand Patterns." *Journal of the Operational Research Society* 57 (10): 1256–1257.

Kourentzes, Nikolaos. 2014. "On Intermittent Demand Model Optimisation and Selection." *International Journal of Production Economics* 156: 180–190.

Kourentzes, Nikolaos, and George Athanasopoulos. 2021. "Elucidate Structure in Intermittent Demand Series." *European Journal of Operational Research* 288 (1): 141–152.

Kourentzes, Nikolaos, Devon Barrow, and Fotios Petropoulos. 2019. "Another Look at Forecast Selection and Combination: Evidence from Forecast Pooling." *International Journal of Production Economics* 209: 226–235.

Kourentzes, Nikolaos, and Fotios Petropoulos. 2016. *tsintermittent: Intermittent Time Series Forecasting*. R package version 1.9, https://CRAN.R-project.org/package=tsintermittent.

Lemke, Christiane, and Bogdan Gabrys. 2010. "Meta-learning for Time Series Forecasting and Forecast Combination." *Neurocomputing* 73 (10-12): 2006–2016.

Li, Li, Yanfei Kang, and Feng Li. 2022. "Bayesian forecast combination using time-varying features." *International Journal of Forecasting* (in press). https://doi.org/10.1016/j.ijforecast.2022.06.002.

Li, Xixi, Fotios Petropoulos, and Yanfei Kang. 2022. "Improving Forecasting by Subsampling Seasonal Time Series." *International Journal of Production Research* 1–17. https://doi.org/10.1080/00207543.2021.2022800.

Lichtendahl Jr, Kenneth C, and Robert L Winkler. 2020. "Why Do Some Combinations Perform Better than Others?" *International Journal of Forecasting* 36 (1): 142–149.

Lolli, Francesco, Rita Gamberini, A Regattieri, Elia Balugani, T Gatos, and S Gucci. 2017. "Single-hidden Layer Neural Networks for Forecasting Intermittent Demand." *International Journal of Production Economics* 183: 116–128.

Lubba, Carl H, Sarab S Sethi, Philip Knaute, Simon R Schultz, Ben D Fulcher, and Nick S Jones. 2019. "catch22: Canonical Time-series Characteristics." *Data Mining and Knowledge*

Discovery 33 (6): 1821–1852.

Makridakis, Spyros, Evangelos Spiliotis, and Vassilios Assimakopoulos. 2021. "The M5 Competition: Background, Organization, and Implementation." *International Journal of Forecasting* (in press). https://doi.org/10.1016/j.ijforecast.2021.07.007.

Makridakis, Spyros, Evangelos Spiliotis, and Vassilios Assimakopoulos. 2022. "M5 Accuracy Competition: Results, Findings, and Conclusions." *International Journal of Forecasting* (in press). https://doi.org/10.1016/j.ijforecast.2021.11.013.

Montero-Manso, Pablo, George Athanasopoulos, Rob J Hyndman, and Thiyanga S Talagala. 2020. "FFORMA: Feature-based Forecast Model Averaging." *International Journal of Forecasting* 36 (1): 86–92.

Montero-Manso, Pablo, and Rob J Hyndman. 2021. "Principles and Algorithms for Forecasting Groups of Time Series: Locality and Globality." *International Journal of Forecasting* 37 (4): 1632–1653.

Newbold, Paul, and Clive WJ Granger. 1974. "Experience with Forecasting Univariate Time Series and the Combination of forecasts." *Journal of the Royal Statistical Society: Series A (General)* 137 (2): 131–146.

Nikolopoulos, Konstantinos. 2021. "We Need to Talk about Intermittent Demand Forecasting." *European Journal of Operational Research* 291 (2): 549–559.

Nikolopoulos, Konstantinos, Aris A Syntetos, John E Boylan, Fotios Petropoulos, and Vassilis Assimakopoulos. 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.

O'Hara-Wild, Mitchell, Rob Hyndman, and Earo Wang. 2021. *feasts: Feature Extraction and Statistics for Time Series*. R package version 0.2.2, https://CRAN.R-project.org/package=feasts.

Palm, Franz C, and Arnold Zellner. 1992. "To Combine or not to Combine? Issues of Combining Forecasts." *Journal of Forecasting* 11 (8): 687–701.

Petropoulos, Fotios, Daniele Apiletti, Vassilios Assimakopoulos, Mohamed Zied Babai, Devon K Barrow, Souhaib Ben Taieb, Christoph Bergmeir, et al. 2022. "Forecasting: Theory and Practice." *International Journal of Forecasting* 38 (3): 705–871.

Petropoulos, Fotios, and Nikolaos Kourentzes. 2015. "Forecast Combinations for Intermittent Demand." *Journal of the Operational Research Society* 66 (6): 914–924.

Petropoulos, Fotios, Spyros Makridakis, Vassilios Assimakopoulos, and Konstantinos Nikolopoulos. 2014. "'Horses for Courses' in Demand Forecasting." *European Journal of Operational Research* 237 (1): 152–163.

Petropoulos, Fotios, and Ivan Svetunkov. 2020. "A Simple Combination of Univariate Models."

International Journal of Forecasting 36 (1): 110–115.

Petropoulos, Fotios, Xun Wang, and Stephen M Disney. 2019. "The Inventory Performance of Forecasting Methods: Evidence from the M3 Competition Data." *International Journal of Forecasting* 35 (1): 251–265.

Sillanpää, Ville, and Juuso Liesiö. 2018. "Forecasting Replenishment Orders in Retail: Value of Modelling Low and Intermittent Consumer Demand with Distributions." *International Journal of Production Research* 56 (12): 4168–4185.

Silver, Edward Allen, David F Pyke, Rein Peterson, et al. 1998. *Inventory Management and Production Planning and Scheduling*. Vol. 3. Wiley New York.

Smith, Jeremy, and Kenneth F Wallis. 2009. "A Simple Explanation of the Forecast Combination Puzzle." *Oxford Bulletin of Economics and Statistics* 71 (3): 331–355.

Spiliotis, Evangelos, Spyros Makridakis, Anastasios Kaltsounis, and Vassilios Assimakopoulos. 2021. "Product Sales Probabilistic Forecasting: An Empirical Evaluation Using the M5 Competition Data." *International Journal of Production Economics* 240: 108237.

Stock, James H, and Mark W Watson. 2004. "Combination Forecasts of Output Growth in a Seven-country Data Set." *Journal of Forecasting* 23 (6): 405–430.

Syntetos, Aris A, and John E Boylan. 2005. "The Accuracy of Intermittent Demand Estimates." *International Journal of Forecasting* 21 (2): 303–314.

Syntetos, Aris A, John E Boylan, and JD Croston. 2005. "On the Categorization of Demand Patterns." *Journal of the Operational Research Society* 56 (5): 495–503.

Talagala, Thiyanga S, Feng Li, and Yanfei Kang. 2022. "FFORMPP: Feature-based Forecast Model Performance Prediction." *International Journal of Forecasting* 38 (3): 920–943.

Teunter, Ruud H, and Laura Duncan. 2009. "Forecasting Intermittent Demand: A Comparative Study." *Journal of the Operational Research Society* 60 (3): 321–329.

Teunter, Ruud H, Aris A Syntetos, and M Zied Babai. 2011. "Intermittent Demand: Linking Forecasting to Inventory Obsolescence." *European Journal of Operational Research* 214 (3): 606–615.

Theodorou, Evangelos, Shengjie Wang, Yanfei Kang, Evangelos Spiliotis, Spyros Makridakis, and Vassilios Assimakopoulos. 2021. "Exploring the Representativeness of the M5 Competition Data." *International Journal of Forecasting* (in press). https://doi.org/10.1016/j.ijforecast.2021.07.006.

Tibshirani, Robert. 1996. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society: Series B (Methodological)* 58 (1): 267–288.

Trapero, Juan R, Manuel Cardos, and Nikolaos Kourentzes. 2019. "Empirical Safety Stock Estimation Based on Kernel and GARCH Models." *Omega* 84: 199–211.

Wallström, Peter, and Anders Segerstedt. 2010. "Evaluation of Forecasting Error Measure-

ments and Techniques for Intermittent Demand." *International Journal of Production Economics* 128 (2): 625–636.

Wang, Xiaoqian, Rob J Hyndman, Feng Li, and Yanfei Kang. 2022a. "Forecast Combinations: An over 50-year Review." *arXiv preprint arXiv:2205.04216* .

Wang, Xiaoqian, Yanfei Kang, Fotios Petropoulos, and Feng Li. 2022b. "The Uncertainty Estimation of Feature-based Forecast Combinations." *Journal of the Operational Research Society* 73 (5): 979–993.

Wang, Xiaozhe, Kate Smith-Miles, and Rob Hyndman. 2009. "Rule Induction for Forecasting Method Selection: Meta-learning the Characteristics of Univariate Time Series." *Neurocomputing* 72 (10-12): 2581–2594.

Wang, Xun, and Fotios Petropoulos. 2016. "To Select or to Combine? The Inventory Performance of Model and Expert Forecasts." *International Journal of Production Research* 54 (17): 5271–5282.

Willemain, Thomas R, Charles N Smart, and Henry F Schwarz. 2004. "A New Approach to Forecasting Intermittent Demand for Service Parts Inventories." *International Journal of Forecasting* 20 (3): 375–387.