Data-driven optimization for production planning with multiple demand features

Xiaoli Su and Lijun Zeng
School of Economics and Management, Fuzhou University, Fuzhou, China
Bo Shao
Department of Industrial and Systems Engineering, University of Wisconsin-Madison, Madison, WI, USA, and
Binlong Lin
School of Economics and Management, Fuzhou University, Fuzhou, China

Received 24 April 2023
Revised 2 August 2023
Accepted 6 September 2023

Abstract

Purpose – The production planning problem with fine-grained information has hardly been considered in practice. The purpose of this study is to investigate the data-driven production planning problem when a manufacturer can observe historical demand data with high-dimensional mixed-frequency features, which provides fine-grained information.

Design/methodology/approach – In this study, a two-step data-driven optimization model is proposed to examine production planning with the exploitation of mixed-frequency demand data is proposed. First, an Unrestricted MIxed DAta Sampling approach is proposed, which imposes Group LASSO Penalty (GP-U-MIDAS). The use of high frequency of massive demand information is analytically justified to significantly improve the predictive ability without sacrificing goodness-of-fit. Then, integrated with the GP-U-MIDAS approach, the authors develop a multiperiod production planning model with a rolling cycle. The performance is evaluated by forecasting outcomes, production planning decisions, service levels and total cost.

Findings – Numerical results show that the key variables influencing market demand can be completely recognized through the GP-U-MIDAS approach; in particular, the selected accuracy of crucial features exceeds 92%. Furthermore, the proposed approach performs well regarding both in-sample fitting and out-of-sample forecasting throughout most of the horizons. Taking the total cost and service level obtained under the actual demand as the benchmark, the mean values of both the service level and total cost differences are reduced. The mean deviations of the service level and total cost are reduced to less than 2.4%. This indicates that when faced with fluctuating demand, the manufacturer can adopt the proposed model to effectively manage total costs and experience an enhanced service level.

Originality/value – Compared with previous studies, the authors develop a two-step data-driven optimization model by directly incorporating a potentially large number of features; the model can help manufacturers effectively identify the key features of market demand, improve the accuracy of demand estimations and make informed production decisions. Moreover, demand forecasting and optimal production decisions behave robustly with shifting demand and different cost structures, which can provide manufacturers an excellent method for solving production planning problems under demand uncertainty.

Keywords Big data, Data-driven optimization, Production planning, Mixed-frequency data

Paper type Research paper

1. Introduction

In recent years, e-commerce has provided a secure environment for manufacturers to make personalized and individualized products to satisfy consumer demands. It is estimated that global retail e-commerce sales reached $5.7 billion in 2020, comprising 19.7% of total retail and showing a 9.7% increase over 2021 (eMarketer reports, 2020). Manufacturers have access to a tremendous

The authors sincerely appreciate helpful comments from the editor and two anonymous reviewers. This research work is supported by the National Social Science Foundation of China under Grant No. 20BGL112. The authors also express sincere gratitude for valuable suggestions provided by Professor Xingxuan Zhuo.

Kybernetes

0368-492X
DOI 10.1108/K-04-2023-0690

amount of data on individual customers for identifying the unique requirements of consumers (Feng and Shanthikumar, 2018; Ban and Rudin, 2019; Khoei *et al.*, 2023). Furthermore, advanced technologies, e.g. sensing Internet of Things (IoT) devices, radio-frequency identification (RFID) chips and artificial intelligence, are implemented to collect and process real-time data about manufacturing facilities, products and customers' preferences and demands (Tambuskar *et al.*, 2023). The huge amounts of data are large in scope (Gholizadeh *et al.*, 2020; Govindan and Gholizadeh, 2021). In reality, the frequency derived from these collected real-time data is often higher than the frequency of the demand to be predicted. For example, consumers' shopping behaviors occur with distinguishing features, including different frequencies and habits; the sampling frequency of the real-time data is characterized as high-frequency, e.g. not only day-to-day/hour-to-hour but also minute-to-minute. In contrast, the traditional production planning of a manufacturer is designed quarterly, monthly, or weekly. The setting becomes particularly complicated when considering how to incorporate these huge amounts of inconsistent frequency data into production planning. Therefore, it is very necessary for manufacturers to investigate the value and pattern of these large amounts of mixed-frequency data, such as e-commerce trade data, real-time web search data and Google Trend data. Through data mining, utilizing mixed-frequency data can significantly improve the accuracy of the descriptions and predictions regarding consumer demand.

In general, traditional demand forecasting methods require the frequency of all variables to be consistent. Therefore, there are two approaches to analyze datasets with both low- and high-frequency data. The first approach is to use only the sample data that are consistent with the frequency of the demand to be predicted. This approach filters out the high-frequency data that may contain valuable demand information. For example, Cui *et al.* (2018) and Boone *et al.* (2018) use Facebook data and Google Trend data to predict the daily demand for an online retailer. The data are sampled at a high frequency, and they can reflect the preferences of consumers and improve the accuracy of predictions. In contrast, the forecasting accuracy is reduced without using these high-frequency data. The second approach converts high-frequency data into the specific frequency of the demand to be predicted, and it can impact the original structure of the collected data, deviate from the true information and affect the accuracy of the forecasting outcomes. To ensure the specific frequency, the second approach obliterates the fluctuation of high-frequency data, which can result in information distortion and inaccurate forecasting results; see Silvestrini and Veredas (2008) and Andreou *et al.* (2010). Furthermore, due to rapid changes of consumer preferences, the lifecycle of each product has been greatly reduced (Ho *et al.*, 2022). This phenomenon indicates that weekly/monthly collected data are extremely limited for the manufacturer, and the reliability of the forecasting results based on these data cannot be guaranteed. Thus, how to optimally utilize the mixed-frequency data is a critical issue to improve the accuracy of demand forecasting.

In the production and operations management field, few studies explore the value of mixed-frequency data on forecasting demands. To analyze mixed-frequency data, a classical model in finance management is MIxed DAta Sampling (MIDAS), which can not only utilize valuable demand information contained in high-frequency data but also avoid information distortion caused by data preprocessing. In addition, it is unrealistic for manufacturers to directly input a huge amount of data into the forecasting model, because the huge amount of data also results in an increased number of variables and their ambiguous relationships (Kieslich *et al.*, 2021; Shahin *et al.*, 2021). This is a challenging issue for manufacturers. Therefore, we aim to investigate how past demand observations and a potentially large number of features, e.g. attributes and explanatory variables, related to demand can be effectively utilized to enhance forecasting accuracy.

Motivated by the aforementioned practical issues, we aim to study the following research questions within the literature at the interface of production planning and finance: (1) How should a manufacturer efficiently utilize high-dimensional feature data to solve the multiperiod production planning problem? (2) What is the value of incorporating high-dimensional feature

data in optimizing production planning decisions? (3) How can production planning decisions integrated with a high-dimensional feature dataset be effectively implemented in practice?

To answer these interesting questions, we develop a two-step data-driven optimization model to solve the production planning problem by exploiting the mixed frequency of massive demand information. Compared with the literature, our model has the following unique features. First, we explore a holistic approach by utilizing mixed-frequency demand data, while past studies mainly use the datasets with a consistent frequency or rely on low-frequency processing approaches. Second, we consider a unrestricted MIDAS approach by imposing group LASSO penalty constraints (i.e. GP-U-MIDAS), which is an efficient way of retaining the integrity of the data information and filtering the interfering variables. Through this approach, the key factors influencing demand can be completely distinguished from the massive collection of possible influencing variables, and the goodness-of-fit performance can be effectively improved for both in-sample fitting and out-of-sample forecasting. Third, through numerical simulation, the production planning model integrating the GP-U-MIDAS approach and the rolling cycle can effectively minimize the manufacturer's total cost, while improving the service level.

Considering the impact of demand uncertainty, our modeling and numerical results are helpful for manufacturers to make optimal production planning decisions. Through several numerical examples, we find that (1) compared to the linear univariate model (e.g. AutoRegressive, AR) and the multivariate linear mixed data model (e.g. Unrestricted MIxed DAta Sampling, U-MIDAS), the multivariate linear mixed data model with group LASSO penalty constraints (GP-U-MIDAS) is an effective way to fully excavate and utilize high-dimensional mixed-frequency information, which can avoid overfitting and underfitting with shifting demand. (2) Based on the accurate extraction of crucial features and the accurate estimation of market demand, the data-driven production planning model with the GP-U-MIDAS method provides more informed and robust production decisions with shifting demand and different cost structures. (3) The performance of the data-driven production planning models is further evaluated regarding various aspects, including production planning decisions, service levels and operational costs. Our numerical results also indicate that the data-driven production planning model integrated with the GP-U-MIDAS approach is an excellent choice compared to using the U-MIDAS approach and the AR approach.

The remainder of this study is organized as follows. In Section 2, we review the related literature. Based on an integrated high-dimensional mixed-frequency approach, the demand forecasting model is introduced in Section 3. A data-driven multiperiod production planning model under demand uncertainty is developed in Section 4. In Section 5, numerical examples are presented to validate the data-driven solutions and managerial implications are addressed to assist manufacturers in making production decisions. We also provide discussions and outlook in Section 6.

2. Literature review

Our study aims to investigate the effects of featured-demand data on operations planning, which is related to the following three streams of literature: (1) demand forecasting; (2) the production planning problem; and (3) the high-dimensional mixed-frequency approach.

2.1 Demand forecasting

The first stream of literature related to our work is on demand learning and forecasting. To fully utilize the massive data, operations managers are facing a significant challenge in converting the data into valuable information to enhance the accuracy of demand forecasting.

In general, there are two classes of techniques in demand forecasting. First, forecasting methods are developed based on time series, including exponential smoothing, regression models and Bayesian models. For this type of method, historical demand is the key input to predict future demand; their relationship is described through functional forms (Petropoulosa *et al.*, 2014; Feng and Shanthikumar, 2018; Petropoulos *et al.*, 2013; Bai, 2022). Second, statistical techniques are applied to learn, extract and select features of historical data to improve the accuracy of demand learning and forecasting. For example, the weather index is used to predict electricity consumption (Carbonneau *et al.*, 2008). The retail demand forecast is conducted based on demographic information (Feng *et al.*, 2013). Facebook data are also used to predict daily demand for an online retailer (Cui *et al.*, 2018).

The aforementioned studies focus on analyzing demand information at the aggregate level. In contrast, in the era of big data, we consider how to effectively use of multisource data and efficiently extract the key features of demand information simultaneously. We propose a holistic approach to identify individualized features and then discuss how managers may use the valuable information to predict demand and make reasonable production plans.

2.2 The production planning problem

The second stream of literature regards the production planning problem under demand uncertainty. The production planning problem was proposed by Wagner and Whitin (1958) in 1958. They developed a W–W model to analyze the single-item production planning problem under dynamic demand. Since then, this classic problem has attracted extensive attention from scholars, and it has been applied to solve other operations and supply chain management issues, e.g. production capacity planning, inventory management, backlogging, remanufacturing, lost sales, demand and production time windows and perishability (Sox and Muckstadt, 1997; Koca *et al.*, 2015; Lee and Çetinkaya, 2001; Chu *et al.*, 2013; Gordon and Pistikopoulos, 2022).

Traditionally, three main approaches are used to describe demand uncertainty with various assumptions. First, demand is taken as a random variable, and its distribution is assumed to be known (Gebennini *et al.*, 2009; Perakis and Zaretsky, 2008; Englberger *et al.*, 2016; De Armas and Laguna, 2020). Second, demand is taken as a random variable with limited information; that is, full demand information is unavailable, but partial random demand information (i.e. support, mean, median, symmetry and variance) is known (Goli *et al.*, 2019; Jabbarzadeh *et al.*, 2019). Third, demand is described as fuzzy numbers and interval numbers that can reflect demand fluctuation in a certain range (Sodhi, 2005; Wang and Tang, 2009; Tirkolaee *et al.*, 2019; Darvishi *et al.*, 2020; Demirhan *et al.*, 2020).

In our paper, we incorporate a potentially large number of features directly in a production planning problem. Most importantly, we consider the mixed frequency of massive demand information, including both the inherent information of the historical demand data and the external information of multisource features. This issue is more challenging in production and operations management.

2.3 The high-dimensional mixed-frequency approach

Our work is also related to research on high-dimensional and mixed-frequency data processing, which has been used to predict macroeconomic conditions (Sheen and Wang, 2021), stock market changes (Zhu *et al.*, 2022) and oil prices (de Medeiros *et al.*, 2022).

When facing different data frequencies, both the weighted average and moving average methods are commonly used to convert high-frequency data into low-frequency data (Silvestrini and Veredas, 2008; Andreou *et al.*, 2010; Wang *et al.*, 2022). The drawback of these methods is that the inherent information of the high-frequency data can be filtered out, which leads to inaccurate forecasting results. To effectively address the mixed-frequency, the

MIDAS model was developed by Ghysels *et al.* (2004) and Andreou and Ghysels (2021). By including a polynomial weight constraint function, this model can reduce the number of estimated parameters, which makes this method more flexible and practical in applications such as macroeconomic forecasting (Clements and Galvão, 2017), water quality prediction (Penev *et al.*, 2014) and energy price analysis (Baumeister *et al.*, 2015).

In contrast, in our work, an integrated GP-U-MIDAS approach is developed to extract the key features of massive demand information (e.g. dimension and frequency). We also evaluate the performance of the production planning model integrated with the GP-U-MIDAS approach, especially for processing high-dimensional and mixed-frequency demand information.

2.4 Research gaps

Table 1 summarizes the key features of extant work and our work, especially the main features of the relevant articles on production planning problems with demand uncertainty. Specifically, the contributions of our study are compared with other work in terms of multiple aspects: problem formulation, demand forecasting, information types and information utilization paradigm.

As shown in Table 1, few papers focusing on the production planning issue consider all three parts of demand forecasting, information types and information utilization paradigms, which may incur suboptimal production decisions with large amounts of inconsistent frequency data. Thus, manufacturers must develop a comprehensive data-driven multiperiod production planning model that can address all three parts, especially obtaining informed production decisions under demand uncertainty. In addition, we incorporate mixed-frequency data into the production planning problem to effectively utilize a potentially large number of features. Through numerical simulation, we directly disclose the relationship between the utilization of mixed-frequency data and the accuracy of production decisions. Meanwhile, we examine the value of potential features of market demand to improve the accuracy of demand forecasting.

3. Demand forecasting model

Accurate demand forecasting is a major puzzle in operations and supply chain management (Feng and Shanthikumar, 2018); in particular, demand forecasting is a prerequisite for making production plans. In the era of big data, firms can access a massive amount of high-frequency data, e.g. web search/traffic data, Google Trend data, Facebook data and online clickstream data, collected through various intelligent devices and e-platforms. It is a challenging issue for firms to convert these valuable data into dynamic information to improve forecasting outcomes.

By fully exploiting the features of massive demand information, we aim to optimize the data processing method to improve the accuracy of market demand forecasting. Specifically, the GP-U-MIDAS model is used in our study to forecast market demand. Given the group effects generated via the frequency alignment operation in the U-MIDAS model, the GP-U-MIDAS model proposed in our study integrates a group penalized function, e.g. group LASSO, into the U-MIDAS regression framework to improve the accuracy of demand forecasting, especially for group selection and parameter estimation.

In general, the market demand for a product is affected by various factors, e.g. product quality and customer preferences, with mixed frequency. We denote $\{d_t, t \in Z\}$ as the low frequency demand time series with a low-frequency lag operator B, i.e. $Bd_t = d_{t-1}$ and the high-frequency factors denoted as $\{x_\tau^{(i)}, \tau \in Z\}$ with a high-frequency lag operator L, i.e. $Lx_\tau^{(i)} = x_{\tau-1}$, where $i = 1, 2, \ldots, k$ represent different high-frequency variables and m_i denotes

Table 1.
Review of the relevant papers on production decisions with demand uncertainty

Research	Problem formulation		Demand forecasting		Information types		Information utilization paradigm	
	Limited data	Massive data	Univariate	Multivariate	Consistent frequency	Inconsistent frequency	Unfiltered	Filtered
Petropoulosa *et al.* (2014)			√		√		√	
Goli *et al.* (2019)	√				√		√	
Darvishi *et al.* (2020)	√				√		√	
Andreou *et al.* (2010)				√	√		√	
Clements and Galvão (2017)				√		√	√	
Andreou and Ghysels (2021)				√		√	√	
Our work		√		√		√		√

Source(s): Table by authors

the frequency mismatch between d_t and $x_\tau^{(i)}$. In the case of $m_i = 1$, both demand d_t and factor $x_\tau^{(i)}$ are observed at the common frequency. When $m_i > 1$, d_t and $x_\tau^{(i)}$ are observed at mixed frequency. In these circumstances, for each low-frequency period from t to $(t+1)$, we can observe m_i high-frequency variables $x_\tau^{(i)}$ at $\tau = tm_i, tm_i - 1, \ldots, (t-1)m_i - 1$, i.e. $\tau = tm_i - j$ with $j = 0, 1, \ldots, m_i - 1$. In practice, the demand d_t is affected not only by the current market value $\{x_\tau^{(i)}\}_{i=1}^{k}$ but also by the multiperiod lags of factors $(x_{tm_i}^{(i)}, x_{tm_i-1}^{(i)}, \ldots, x_{tm_i-l_i}^{(i)})$, where l_i denotes the maximum lag order of $x_\tau^{(i)}$.

To predict the low-frequency demand d_t using the high-frequency factors $\{x_\tau^{(i)}\}_{i=1}^{k}$, it is necessary to normalize the differences in frequency, i.e. the process of frequency alignment and then to transform high-frequency $x_\tau^{(i)}$ into a low-frequency matrix $\boldsymbol{X}_\tau^{(i)}$, as follows:

$$\boldsymbol{X}_\tau^{(i)} = \begin{bmatrix} x_{um_i}^{(i)} & x_{um_i-1}^{(i)} & \cdots & x_{um_i-l_i}^{(i)} \\ x_{(u+1)m_i}^{(i)} & x_{(u+1)m_i-1}^{(i)} & \cdots & x_{(u+1)m_i-l_i}^{(i)} \\ \vdots & \vdots & \ddots & \vdots \\ x_{tm_i}^{(i)} & x_{tm_i-1}^{(i)} & \cdots & x_{tm_i-l_i}^{(i)} \\ \vdots & \vdots & \ddots & \vdots \\ x_{Tm_i}^{(i)} & x_{Tm_i-1}^{(i)} & \cdots & x_{Tm_i-l_i}^{(i)} \end{bmatrix}, \tag{1}$$

where $t = u, u+1, \ldots, T$, T is the length of the demand time series and u is a minimum integer that satisfies the condition $um_i - l_i \geq 0$.

Based on the results of the frequency alignment process, we use a U-MIDAS regression to perform multi-step ahead demand forecasting; thus, we have

$$d_{t+h} = \sum_{i=1}^{k} \sum_{j=0}^{l_i} \beta_j^{(i)} x_{tm_i-j}^{(i)} + \varepsilon_{t+h}, \tag{2}$$

where $\sum_{j=0}^{l_i} x_{tm_i-j}^{(i)} = (x_{tm_i}^{(i)}, x_{tm_i-1}^{(i)}, \ldots, x_{tm_i-l_i}^{(i)})$ and $\{\beta_j^{(i)}\}_{j=0}^{l_i}$ are corresponding coefficients, and the random error ε_t is assumed to be white noise. In addition, h denotes the h-step ahead forecasts, and when $h = 0$, Eq. (2) shows the same period relationship between y_t and $\{x_\tau^{(i)}\}_{i=1}^{k}$.

Since $\boldsymbol{X}_\tau^{(i)} = (x_{tm_i}^{(i)}, x_{tm_i-1}^{(i)}, \ldots, x_{tm_i-l_i}^{(i)})$ is generated from the high-frequency factor $x_\tau^{(i)}$ using the frequency alignment approach, the estimation of Eq. (2) is transformed into a parameter proliferation problem; that is, the number of covariates is significantly increased from k to $\sum_{i=1}^{k}(l_i + 1)$. This results in a challenging issue in parameter estimation, the curse of dimensionality. This problem is addressed through extensions of the variable selection method toward U-MIDAS regression. However, the frequency alignment vector $(x_{tm_i}^{(i)}, x_{tm_i-1}^{(i)}, \ldots, x_{tm_i-l_i}^{(i)})$ should be treated as a complete group structure in the variable selection procedures. Because the covariates are in $\{x_{tm_i}^{(i)}, x_{tm_i-1}^{(i)}, \ldots, x_{tm_i-l_i}^{(i)}\}$, all these variables belong to the initial high-frequency factor $x_\tau^{(i)}$, and both have substantial impacts on market demand; thus, $x_\tau^{(i)}$ is a significant factor in demand forecasting.

K

According to the discussion above, we next apply group selection methods, such as group LASSO, to analyze the issue of variable selection in a group manner. By integrating the group LASSO penalty into Eq. (2), we can formulate the objective function of the GP-U-MIDAS model as follows:

$$\min Q\left(\widehat{\boldsymbol{\beta}}\right) = \frac{1}{2}\left\| \boldsymbol{d} - \sum_{i=1}^{k} \boldsymbol{X}^{(i)}\boldsymbol{\beta}^{(i)} \right\|^2 + \sum_{i=1}^{k} \lambda \left\| \boldsymbol{\beta}^{(i)} \right\|, \tag{3}$$

where $\boldsymbol{X}^{(i)} = (x^{(i)}_{tm_i}, x^{(i)}_{tm_i-1}, \ldots, x^{(i)}_{tm_i-l_i})$ is a $T \times (l_i + 1)$-dimensional design matrix formed by the factors in the i-th group and $\boldsymbol{\beta}^{(i)} = (\beta^{(i)}_0, \beta^{(i)}_1, \ldots, \beta^{(i)}_{l_i})'$ denotes a $(l_i + 1)$ parameters vector corresponding to $\boldsymbol{X}^{(i)}_{\tau}$.

In addition, λ is the regularization parameter applied to the L_1-norm ($\|\cdot\|$) of $\boldsymbol{\beta}^{(i)}$, and it impacts the strength of shrinkage and group selection results. Generally, the larger the value of λ is, the greater the strength of the regularization, and the greater the sparsity of the model. Meanwhile, considering the situation definition in the time series data, we use the older observation to forecast a new one. The classic K-fold cross-validation method cannot be used to solve this problem; thus, we use the Hyndmans time series cross-validation method.

Furthermore, to estimate the parameters $\{\boldsymbol{\beta}^{(i)}\}_{i=1}^{k}$ of the GP-U-MIDAS model in Eq. (3), we can utilize the group coordinate descent (GCD) algorithm proposed by Breheny and Huang (2015). The GCD algorithm is an efficient approach for fitting models with grouped penalties, and it optimizes the objective function shown in Eq. (3) with respect to a single group at a time, iteratively cycling through the groups until convergence is reached. Specifically, let the solution to Eq. (3) be denoted as $\widehat{\boldsymbol{\beta}}(\boldsymbol{z}_i, \lambda)$ and the closed form of the solution be as follows:

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{z}_i, \lambda) = S(\boldsymbol{z}_i, \lambda), \tag{4}$$

where $\boldsymbol{z}_i = \boldsymbol{X}^{(i)'}(\boldsymbol{d} - \boldsymbol{X}^{-(i)}\boldsymbol{\beta}^{-(i)})$ is the least squares solution, $\boldsymbol{X}^{-(i)}$ denotes the remaining portion of the design matrix with $\boldsymbol{X}^{(i)}$ being excluded, and $\boldsymbol{\beta}^{-(i)}$ denotes the associated parameters. Moreover,

$$S(\boldsymbol{z}_i, \lambda) = S(\|\boldsymbol{z}_i\|, \lambda)\frac{\boldsymbol{z}_i}{\|\boldsymbol{z}_i\|}, \tag{5}$$

where

$$S(\|\boldsymbol{z}_i\|, \lambda) = \begin{cases} \|\boldsymbol{z}_i\| - \lambda, & \text{if } \|\boldsymbol{z}_i\| > \lambda \\ 0, & \text{if } \|\boldsymbol{z}_i\| \le \lambda \\ \|\boldsymbol{z}_i\| + \lambda, & \text{if } \|\boldsymbol{z}_i\| < -\lambda \end{cases}. \tag{6}$$

Without loss of generality, the group covariates $\{\boldsymbol{X}^{(i)}\}_{i=1}^{k}$ are assumed to be orthonormal. Let $\boldsymbol{r} = \boldsymbol{d} - \sum_{i=1}^{k}\boldsymbol{X}^{(i)}\boldsymbol{\beta}^{(i)}$ and s denote the residuals and the number of iterations, respectively. Given the initial parameters $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}^{(1)}_0, \boldsymbol{\beta}^{(2)}_0, \ldots, \boldsymbol{\beta}^{(k)}_0)'$, we can implement the GCD algorithm as follows:

Algorithm 1: The GCD algorithm

Step 1: Set $s = 0$. Initialize vector of residuals $\boldsymbol{r} = \boldsymbol{d} - \sum_{i=1}^{k} \boldsymbol{X}^{(i)}\boldsymbol{\beta}^{(i)}$.

Step 2: For $i = 1, 2, \ldots, k$, execute the following calculations

(a) calculate $\boldsymbol{z}_i = \boldsymbol{X}^{(i)\prime}\boldsymbol{r} + \boldsymbol{\beta}_s^{(i)}$;

(b) update $\boldsymbol{\beta}_{s+1}^{(i)} \leftarrow \hat{\boldsymbol{\beta}}(\boldsymbol{z}_i, \lambda)$;

(c) update $\boldsymbol{r} \leftarrow \boldsymbol{r} - \boldsymbol{X}^{(i)\prime}(\boldsymbol{\beta}_{s+1}^{(i)} - \boldsymbol{\beta}_s^{(i)})$.

Step 3: Update $s \leftarrow s + 1$.

Step 4: Given the convergence criteria, repeat steps 2-3 until convergence, and output $\left\{\boldsymbol{\beta}^{(i)}\right\}_{i=1}^{k}$ is an estimator $\hat{\boldsymbol{\beta}}$ of the GP-U-MIDAS model.

In this section, a data-driven model is proposed to study the demand forecasting issue by exploiting various frequencies of massive demand information. The GP-U-MIDAS model is proposed to optimize the data processing method to improve the accuracy of demand forecasting. In the following section, the proposed algorithm is integrated into a production planning model.

4. Data-driven production planning model

In this section, we use the production planning model as the workhouse to show how big data can change operations planning. Specifically, we consider the cost minimization problem of a manufacturer who coordinates internal production, outsourcing and inventory to satisfy uncertain market demand. The market capacity is infinite, and the delivery lead time at the spot market is zero (Mandl and Minner, 2023). The manufacturer can access demand data from history or a market study. The manufacturer aims to decide the production and outsourcing quantities that minimize the operations cost. However, the true demand information is difficult to forecast. In what follows, we introduce the market features into the production planning model, and we also demonstrate how the manufacturer can use available demand data to enhance the aggregate forecast and production planning. That is, we aim to study how to reconcile valuable demand information and enable data-driven production planning. The specific decision-making process is shown in Figure 1.

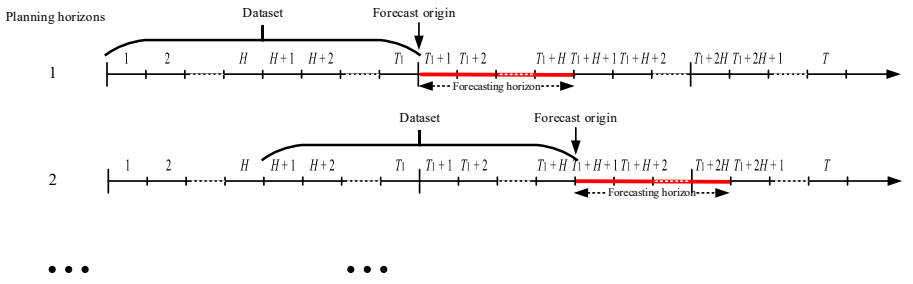


Figure 1.
Schematic diagram for multiple horizon production planning

Source(s): Figure by authors

First, at the beginning of the first planning horizon, the manufacturer uses the dataset containing T_1 periods of historical demand data and corresponding high-frequency data to predict the demand of H periods in the planning period based on the GP-U-MIDAS model introduced in Section 3. Second, based on the predicted demand, the manufacturer makes the production plan. Thirdly, the manufacturer implements the production plan and then collects H periods of historical demand data and corresponding high-frequency data to update the dataset. Finally, the manufacturer can repeat the above steps to conduct production planning in subsequent periods.

We consider a capacitated production planning problem with outsourcing (CPP-O) over a finite planning horizon $[t_0, t_0+H]$, where t_0 represents the beginning of the planning horizon. The production capacity is assumed to be constant, while outsourcing is incapacitated. The manufacturer obtains the predicted demand for the next H periods based on the GP-U-MIDAS model introduced in Section 3 and then decides on internal production and the forward sourcing quantity for delivering over the planning horizon $[t_0+1, t_0+H]$. Therefore, the objective of the manufacturer is to coordinate the internal production, outsourcing, and inventory, that can satisfy uncertain market demand with the minimum operations costs. We assume that the initial inventory level in one period is zero and that the demand of a period cannot be backlogged. The summary of notation is shown in Table 2.

The production planning model is formulated as a mixed-integer programming model.

$$\min F = \min \sum_{t=t_0+1}^{t_0+H} (c_s y_t + c_m X_t + c_t^o L_t + c_h I_t + c_l S_t) \tag{7}$$

s.t.

$$I_t = I_{t-1} + X_t + L_t + S_t - d_t \quad t = t_0+1, t_0+2, ..., t_0+H \tag{8}$$

$$0 \le X_t \le M y_t \quad t = t_0+1, t_0+2, ..., t_0+H \tag{9}$$

$$0 \le S_t \le d_t \quad t = t_0+1, t_0+2, ..., t_0+H \tag{10}$$

$$I_{t_0} = I_{t_0+H} = 0 \tag{11}$$

$$I_t, L_t \ge 0 \quad t = t_0+1, t_0+2, ..., t_0+H \tag{12}$$

Notation	Description
t	The index of each forecasting period, $t \in \{t_0+1, t_0+2, ..., t_0+H\}$
d_t	The predicted demand in period t
M	The maximum production capacity
c_s	The setup cost
c_m	The unit production cost
c_t^o	The unit outsourcing cost in period t
c_h	The unit holding cost
c_l	The shortage cost
I_t	The on-hand inventory in period t
S_t	The shortage quantity in period t
X_t	The realized production quantity in period t
L_t	The outsourcing quantity in period t
y_t	The indicator of production activity, $y_t \in \{1, 0\}$

Table 2. Summary of notations

Source(s): Table by authors

$$y_t = \{0,1\} \quad t = t_0 + 1, t_0 + 2, ..., t_0 + H \tag{13}$$

On the right-hand side of Eq. (7), the first term is the production setup cost in period t; the second term is the production cost in period t; the third term is the outsourcing cost in period t; and the fourth and fifth terms are the holding cost and shortage cost, respectively, in period t. This model can be considered a variant of the data-driven production planning model since it relies on predictive analysis of the demand model, specifically the GP-U-MIDAS model, using a substantial volume of real-time and featured demand information. Next, we conduct numerical experiments to validate the models proposed in this section.

5. Numerical examples

In this section, numerical experiments are conducted in two steps to validate the outcomes of the theoretical analysis of demand forecasting and data-driven production planning. In Section 5.1, we examine the effectiveness of the forecasting algorithm and analyze the impacts of features of massive demand on the accuracy of forecasting. Then, in Section 5.2, based on the forecasting outcomes (integrated with the multiple frequencies of massive demand), we further analyze the data-driven production planning model.

5.1 Demand forecasting analysis

5.1.1 Data generating process. To assess the predictive and variable selection performance of the proposed model in Section 4, we simulate the data from the following data generating process.

Generate the high-frequency factors $x_\tau^{(i)}$. Without loss of generality, we consider a weekly/monthly mixed dataset, with a low-frequency sample size of $T = 300$ and frequency mismatch $m_i = 4$. Considering the differences between high-frequency factors $x_\tau^{(i)}$, we adopt the following methods to generate a sample with size $T \times m_i = 1200$. Specifically, we use a normal distribution with mean u_T and a constant coefficient of variation ρ to generate a sample of $x_\tau^{(1)}$ and use a first-order autoregressive process with standardized normal distribution to generate a sample of $x_\tau^{(i)}$, $i = 2, \cdots, 10$.

High-frequency factors $x_\tau^{(1)}$ can be divided into T groups, and each group obeys a normal distribution with different mean u_T and a constant coefficient of variation ρ. We set u_T equal to the arithmetic sequence of interval [100,300], and $\rho = 1$.

For high-frequency factors $x_\tau^{(i)}$:

$$x_\tau^{(i)} = \rho_0^{(i)} + \rho_1^{(i)} x_{\tau-1}^{(i)} + \varepsilon_\tau^{(i)} \tag{14}$$

where $\varepsilon_\tau^{(i)} \sim_{i.i.d.} N(0,1)$ for all $i = 2, 3, \cdots, 10$, and $\tau = 1, 2, \cdots, 1200$ denotes the high-frequency sampling; $\rho_0^{(i)}$ is the intercept for $x_\tau^{(i)}$, and we set $\boldsymbol{\rho}_0 = (\rho_0^{(2)}, \rho_0^{(3)}, \cdots, \rho_0^{(10)})' = (90, 80, ..., 10)'$; $\rho_1^{(i)}$ is the persistence parameter for $x_\tau^{(i)}$, and we set $\boldsymbol{\rho}_1 = (\rho_1^{(2)}, \rho_1^{(3)}, \cdots, \rho_1^{(10)})' = (0.5, 0.5, \cdots, 0.5)'$.

Generate the low-frequency demand d_t. Let the target of interest d_t be driven by $l^{(i)}$ autoregressive lags augmented with high-frequency factors $x_\tau^{(i)}$; we simulate d_t as follows:

$$d_t = \beta_0 + \sum_{i=1}^{k} \beta^{(i)} \sum_{j=0}^{l^{(i)}} w(\boldsymbol{\delta}; j) x_{tm_i - j}^{(i)} + u_t \tag{15}$$

where $u_t \sim_{i.i.d.} N(0,1)$, we set the intercept $\beta_0 = 1$ and regression coefficients $\boldsymbol{\beta} = (\beta^{(1)}, \beta^{(2)}, \cdots, \beta^{(10)})' = (5,0,0,0,0,3,0,0,0,1)'$. In the numerical simulation procedure, we add seven noisy factors that are generated in the same way as the relevant factors, but the coefficients are equal to zero. In addition, we rely on a commonly used weighting scheme in the mixed data analysis literature, namely, the exponential Almon weighting function $w(\boldsymbol{\delta}; j)$, which is defined as $w(\boldsymbol{\delta}; j) = \frac{\exp\{\delta_1 j + \delta_2 j^2\}}{\sum_{j=0}^{l^{(i)}} \exp\{\delta_1 j + \delta_2 j^2\}}$, for $\boldsymbol{\delta} = (\delta_1, \delta_2) = (0, -0.5)$, and j ($j = 0, 1, \cdots, l^{(i)}$) denotes the $j-$ th high-frequency lag for factor $x_\tau^{(i)}$.

5.1.2 Experimental design process. Depending on the data generating process, we further process the design of numerical experiments as follows.

Step 1. Data splitting. To examine the predictive performance of the proposed model in multiple horizon demand forecasts, we employ a rolling forecasting scheme, and the schematic diagram is shown in Figure 2. The first $T_1 = 240$ in-sample dataset is used as the initial estimation period, and the window size T_1 is fixed for later use. At each step, we perform a sequence of multiple horizon forecasts of d_{t+h}, with the forecast horizon $h = \{1, 2, 3, 4\}$. Consequently, the forecast origin of the evaluation period is $T_1 + h$ initially, and it becomes $T_1 + h + Ht_2$ in the next t_2-th case with $t_2 = 1, 2, \cdots, (T - T_1)/H$. Thus, the length of the evaluation period is $(T - T_1 - h + 1)$.

Step 2. Reference model setup. To assess the performance of GP-U-MIDAS, we develop two reference models. One is a classic autoregressive (AR) model, which utilizes only the historical demand $\{d_t\}_{t=1}^N$ to predict future demand. The other is the U-MIDAS model without variable selection, as shown in Eq. (2). The AR model is formulated as follows:

$$d_t = a_0 + \sum_{p=1}^{P} a_p d_{t-p} + \varepsilon_t \tag{16}$$

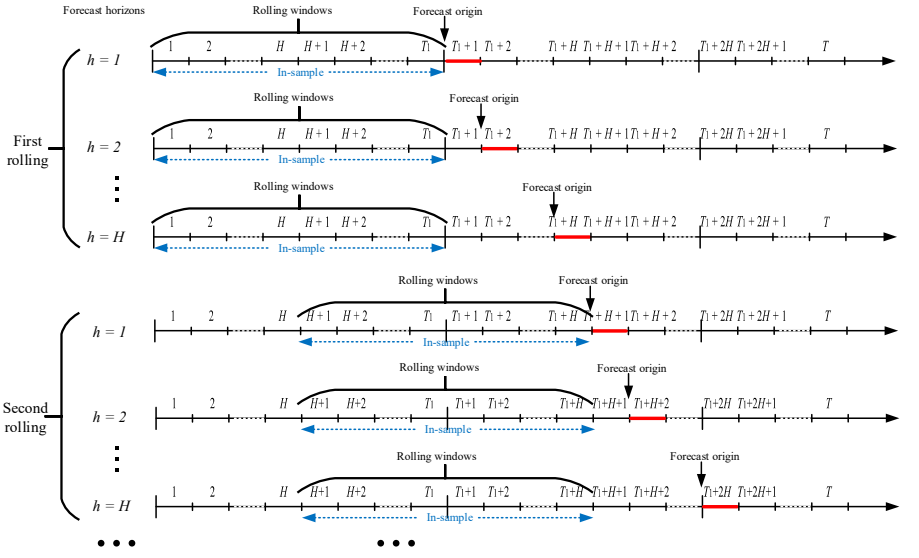


Figure 2.
Schematic diagram for multiple horizon demand forecasts

Source(s): Figure by authors

where a_p is an autoregressive coefficient, random error ε_t is described as white noise, and P describes the autoregressive maximum order.

Step 3. Model training. The maximum length of the lags $\{l^{(i)}\}_{i=1}^{10}$ included in the above models is selected based on the Bayesian information criterion (BIC), and the optimal regularization parameter λ for the GP-U-MIDAS model is determined by the Hyndmans time series cross-validation method. Specifically, we generate $N_\lambda = 20$ different values of λ at equal intervals between $\lambda_{\min} = 0.01$ and $\lambda_{max} = 0.4$.

Step 4. Performance evaluation. To assess the performance of different methods by varying selection on mixed-frequency factors, we employ five indicators. Specifically, (1) the capability of variable selection for the GP-U-MIDAS model is measured by the selected proportion (%), i.e. $(S_{sim}/T_s) \times 100\%$, where S_{sim} denotes the variable selection numbers for the GP-U-MIDAS model, and T_s denotes the number of times to repeat the simulation. (2) The fitting and forecasting capability for all models is measured by the mean absolute percentage error (MAPE), i.e. $MAPE = \sum_{t=1}^{n} \left| (d_t - \widehat{d}_t)/d_t \right| /n$, where d_t is the actual demand, $\widehat{d}_t$ is the fitted or predicted value of d_t, and n is the sample size of the estimation period or evaluation period. (3) Production decisions. We calculate the production quantity X_t, outsourcing quantity L_t, theoretical on-hand inventory I_t, theoretical shortage quantity S_t, actual inventory $I_t^{real}(I_{t-1}^{real} + X_t + L_t - d_t > 0, I_t^{real} = I_t - 1^{real} + X_t + L_t - d_t; I_{t-1}^{real} + X_t + L_t - d_t \le 0, I_t^{real} = 0)$ and actual shortage quantity $S_t^{real}(d_t - I_{t-1}^{real} - X_t - L_t > 0, S_t^{real} = d_t - I_{t-1}^{real} - X_t - L_t; d_t - I_{t-1}^{real} - X_t - L_t \le 0, S_t^{real} = 0)$. (4) Service level. The service level is the ratio of supplied products to the actual demand (i.e. $SL_t = (1 - S_t^{real}/d_t) * 100\%$). (5) Out-of-sample models. The total production cost (C_t) is calculated in each period, $C_t = \sum_{t=t_0+1}^{t_0+H} (c_s y_t + c_m X_t + c_t^o L_t + c_h I_t^{real} + c_l S_t^{real})$.

Step 5. Robustness validation. We repeat the above experimental process 100 times ($T_s = 100$) and report the detailed results of variable selection, goodness-of-fit, demand forecasting, production decisions, service level and out-of-sample cost. The Diebold Mariano (DM) test is conducted to determine whether the performances of the candidate models (i.e. AR, U-MIDAS and GP-U-MIDAS) are statistically significantly different.

5.1.3 Analysis of demand forecast. In this subsection, the performance of the GP-U-MIDAS model is summarized in terms of variable selection, as shown in Table 3.

In Table 3, the error proportion results show that $x_\tau^{(1)}$ and $x_\tau^{(6)}$, which have a major influence on demand, have an equal probability of 100% of being selected by GP-U-MIDAS. Then, $x_\tau^{(10)}$ has a probability of 92% of being sequentially selected, as it has a relatively small effect on demand. In addition, the rest of the factors $x_\tau^{(2)}$, $x_\tau^{(3)}$, $x_\tau^{(4)}$, $x_\tau^{(5)}$, $x_\tau^{(7)}$, $x_\tau^{(8)}$ and $x_\tau^{(9)}$ do not

Variables	Selected proportion (%)	Variables	Selected proportion (%)
$x_\tau^{(1)}$	100	$x_\tau^{(6)}$	100
$x_\tau^{(2)}$	14	$x_\tau^{(7)}$	13
$x_\tau^{(3)}$	14	$x_\tau^{(8)}$	13
$x_\tau^{(4)}$	14	$x_\tau^{(9)}$	13
$x_\tau^{(5)}$	14	$x_\tau^{(10)}$	92

Source(s): Table by authors

Table 3. Summary of the error proportion of variable selection results from 100 repetitions

exceed a probability of 15% of being selected because those factors have no significant impact on demand. Clearly, the results meet our needs well, and the validity of the GP-U-MIDAS model on key variable selection has been confirmed.

Next, Table 4 presents the average MAPE for demand fitting and forecasting across four forecast horizons and three different models, and Figure 3 shows the MAPE values for forecasting across all horizons.

The results in Table 4 and Figure 3 provide several insights. First, in both in-sample fitting and out-of-sample forecasting, the AR model performs less well throughout most horizons, which verifies that it is far from sufficient to consider only historical demand data. Second, as the number of high-frequency external factors grows, the goodness-of-fit of the U-MIDAS model improves remarkably, but overfitting occurs frequently and has a serious influence on the forecasting accuracy. Third, since regularization prevents overfitting in regression models, the predictive power of the GP-U-MIDAS model can be improved significantly by imposing group LASSO penalty constraints on parameters. In this case, the GP-U-MIDAS model with fewer high-frequency factors is inferior to U-MIDAS in terms of goodness-of-fit, but it outperforms the AR and U-MIDAS models to obtain the smallest MAPE for out-of-sample forecasting across most forecast horizons.

To determine whether the difference in forecast MAPE from the paired comparisons model is statistically significant, we use the Diebold-Mariano (DM) test, and the alternative hypothesis is that Model 2 is significantly more accurate than Model 1. According to the DM test results in Table 5, both the U-MIDAS and GP-U-MIDAS models significantly outperform the AR model across forecast horizons $h = \{1, 3, 4\}$ at the 1% level. In addition, the predictive accuracy of the GP-U-MIDAS model is significantly superior to that of the U-MIDAS model

Table 4.
Summary of fitted and forecast error results from 100 repetitions

	AR		U-MIDAS		GP-U-MIDAS	
Forecast horizons	Fitted MAPE	Forecast MAPE	Fitted MAPE	Forecast MAPE	Fitted MAPE	Forecast MAPE
$h = 1$	6.29%	4.46%	4.46%	4.06%	4.84%	3.72%
$h = 2$	6.29%	2.94%	4.46%	4.17%	4.80%	3.74%
$h = 3$	6.29%	5.64%	4.48%	4.09%	4.80%	3.74%
$h = 4$	6.29%	6.53%	4.44%	4.15%	4.82%	3.84%

Source(s): Table by authors

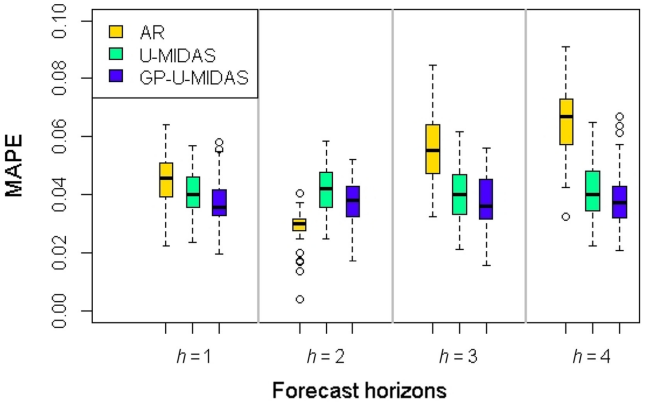


Figure 3.
The forecast error (MAPE) results from 100 repetitions

Source(s): Figure by authors

across all forecast horizons at the 1% significance level. Therefore, the DM results confirm the reliability of the conclusions.

In summary, when conducting market demand forecasting in traditional ways, we refer to only historical demand data and paint only partial pictures at best. Under these circumstances, linear univariate models, e.g. the AR, are unable to fully excavate and utilize the high-dimensional mixed-frequency information, and they are prone to underfitting. In contrast, multivariate linear mixed data models, e.g. U-MIDAS, can effectively utilize the available multisource mixed-frequency data to improve the interpretability of the time-varying demand. However, overfitting generally occurs when the number of variables is excessively large, and the predictive power of the U-MIDAS model is clearly weakened. As a remedy, incorporating the group or variable selection approach to sparse modeling, e.g. GP-U-MIDAS, our numerical results show that this approach can provide an effective way to prevent overfitting and underfitting and significantly improve the predictive ability of models without sacrificing goodness-of-fit.

5.2 Optimal production planning

Thus far, we have validated the performance of the forecasting approach, i.e. GP-U-MIDAS. In this section, we conduct numerical examples to illustrate data-driven production planning, which aims to assist the manufacturer in obtaining optimal production planning in the long term. That is, based on demand forecasting results, the manufacturer makes the optimal production plan that minimizes operational costs and maximizes service levels. We also compare the total cost when the manufacturer conducts production planning using the actual demand and the AR, U-MIDAS and GP-U-MIDAS forecasting results. We consider a dataset with 5 planning horizons, and each planning horizon (e.g. a month) consists of 4 forecasting periods (e.g. weeks); thus, the low-frequency sample set has a total of 20 periods. The basic parameters used in this example are summarized in Table 6.

5.2.1 Analysis of production planning. We first explore the performances of the models based on the case selected from 100 simulation experiments. The sample set, see Figure 4, with a low frequency is taken as the manufacturer's actual demand. Figure 4 shows three main high-frequency factors $x^{(1)}, x^{(6)}, x^{(10)}$ and the estimated values of demand for the AR,

	Model 1							
	AR				U-MIDAS			
Model 2	$h=1$	$h=2$	$h=3$	$h=4$	$h=1$	$h=2$	$h=3$	$h=4$
U-MIDAS	4.64^{***}	−12.51	11.96^{***}	17.52^{***}	–	–	–	–
GP-U-MIDAS	7.66^{***}	−9.77	14.20^{***}	19.85^{***}	5.30^{***}	6.69^{***}	5.09^{***}	4.82^{***}

Note(s): "***" indicates significance at the level of 1%, i.e. p-value < 0.01
Source(s): Table by authors

Table 5. DM test results for the predictive ability of the models

Planning horizon	1	2	3	4	5
M	1,550	1,550	1,550	1,550	1,550
c_s	1,500	1,500	1,500	1,500	1,500
c_m	10	10	10	10	10
$c^o{}_t$	20	12	18	15	25
c_h	2	2	2	2	2
c_l	30	30	30	30	30

Source(s): Table by authors

Table 6. The basic parameter settings for production planning

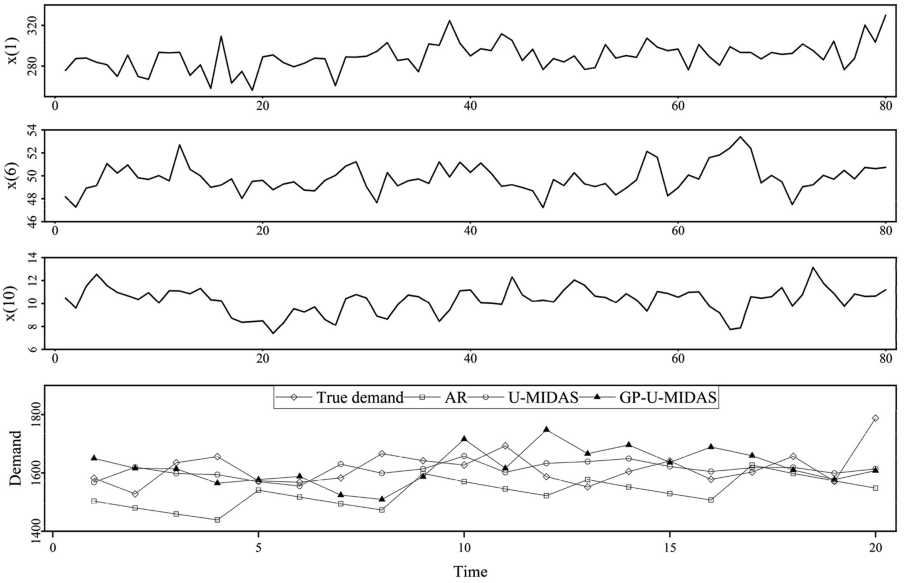


Figure 4.
The forecasting results of the AR, U-MIDAS and GP-U-MIDAS approaches

Source(s): Figure by authors

U-MIDAS and GP-U-MIDAS models. From Figure 4, we can learn that there is no obvious regularity in the variation of high-frequency factors $x^{(1)}, x^{(6)}, x^{(10)}$ in the 20 low-frequency periods, but the variation of the forecasting results using the GP-U-MIDAS approach is gentle and slight. In contrast, the variation of the forecasting results using the AR approach shows a larger fluctuation, and the pattern of the fluctuations does not match real demand changes.

For further explanation, Figure 5 shows the MAPE for each period among the AR, U-MIDAS and GP-U-MIDAS models. The figure uses a bar for the AR MAPE and lines for the U-MIDAS and GP-U-MIDAS MAPE values to show their differences. Compared with the AR model, the

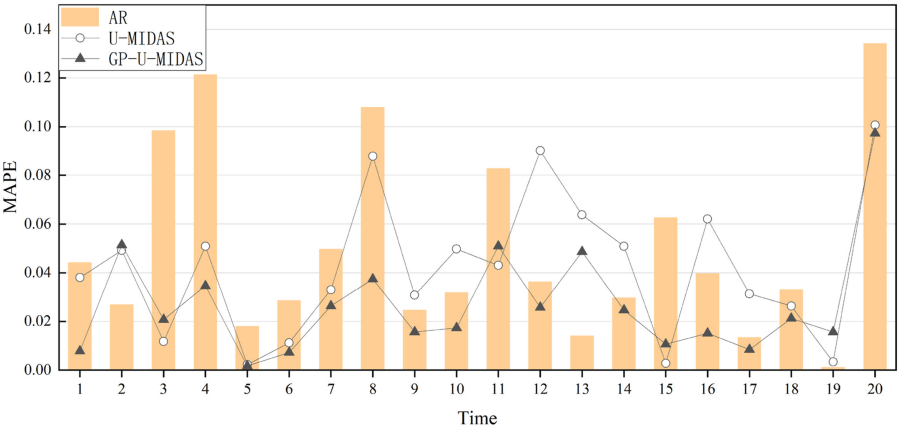


Figure 5.
MAPEs among the AR, U-MIDAS and GP-U-MIDAS approaches

Source(s): Figure by authors

GP-U-MIDAS model has a smaller error in 17 of 20 observations (85% of the time). Thus, the GP-U-MIDAS model may pick up certain microtrends from high-frequency factors resulting in better predictions. Overall, compared to the AR approach, the GP-U-MIDAS approach can effectively assist the manufacturer in obtaining more accurate demand information that is close to the actual demand by capturing complex microtrends in high-frequency factors.

According to the forecasting results obtained using the AR, U-MIDAS and GP-U-MIDAS approaches, the manufacturer's optimal production and operational decisions are determined, and the results are summarized in Table 7.

Table 7 shows that the optimal production planning outcomes obtained based on the forecasting results of the AR, U-MIDAS and GP-U-MIDAS approaches are similar to those planning outcomes under true demand. In general, to meet market demand, the production planning decisions in each period are made by utilizing available products in inventory and maximizing production capacity. In addition, demand greater than production capacity and inventory level can be satisfied through outsourcing because the unit production cost is lower than the unit outsourcing cost. Thus, the manufacturer prefers to make products itself first. On the other hand, given that the unit shortage cost is greater than the outsourcing cost, the manufacturer prefers to adopt an outsourcing strategy rather than lose market demand.

Furthermore, when the manufacturer's production capacity is smaller than the predicted demand, the manufacturer may make different production plans, outsourcing the demand that cannot be met by production in the current period or producing this part of demand in the current period and storing it. This results in two different planning outcomes because the marginal operations costs are different. For part of the demand that cannot be met by production in the current period, if the unit outsourcing cost is less than the unit production and inventory costs, the manufacturer will choose to outsource; otherwise, the manufacturer prefers to produce and store in advance.

In addition, Table 7 shows the inventory and shortage levels based on different production strategies under real demand. Compared to the theoretical shortage (holding) quantity and actual shortage (holding) quantity, there exist serious shortages using the data-driven model with the AR method, because this method underestimates market demand. The production strategy based on the U-MIDAS model leads to greater inventory. Different from these two models, the GP-U-MIDAS model can balance the shortage and inventory, which may reduce the total production cost.

5.2.2 Sensitivity analysis. To test the robustness of the proposed approach, we further evaluate the performance, i.e. service levels, of the different production planning outcomes obtained under AR, U-MIDAS and GP-U-MIDAS to that of production planning outcomes with the actual demand. The comparison results are shown in Figure 6 and Table 8.

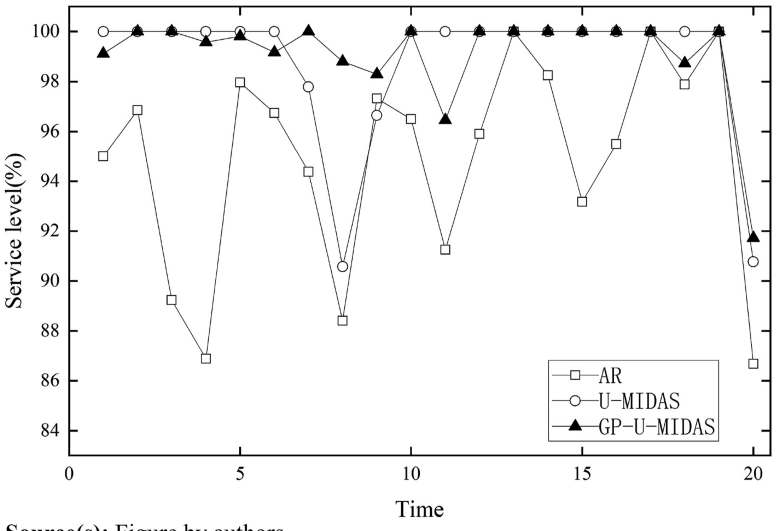
As shown in Figure 6 and Table 8, the service level using the AR forecasting approach is lower than that of the U-MIDAS and GP-U-MIDAS approaches for most periods. In addition, the service level curve of the GP-U-MIDAS approach does not change as much as that of both the AR and U-MIDAS approaches. The GP-U-MIDAS approach can maintain service a stable and high level (more than 91.72% over the whole planning period); meanwhile, the stable service level can bring positive evaluation of the manufacturer.

Table 9 presents operations cost difference in each planning horizon, where $\Delta_{AR} = (\prod_{AR} - \prod_{A})/\prod_{A}$, $\Delta_{UM} = (\prod_{UM} - \prod_{A})/\prod_{A}$ and $\Delta_{GP} = (\prod_{GP} - \prod_{A})/\prod_{A}$. In addition, the expected costs under the actual demand, AR, U-MIDAS and GP-U-MIDAS forecasting results, that is, Π_{A}, Π_{AR} Π_{UM} and $\Pi_{GP}\Delta_{GP}$, are also shown. Table 8 indicates that the total cost difference is reduced from 6.64% using the AR forecasting results to 0.87% using the GP-U-MIDAS forecasting results. The total cost difference is reduced from 6.19% using the AR forecasting results to 2.08% using the GP-U-MIDAS forecasting results. Thus, the overall performance of the GP-U-MIDAS approach is comparable to that of the planning outcomes with the actual demand.

Table 7.
The optimal production and operational decisions

Planning horizon	Forecast horizon	Actual demand					AR						U-MIDAS						GP-U-MIDAS						
		d_t	X_t	L_t	I_t	S_t	$\widehat{d}_t^{AR}$	X_t	L_t	I_t	S_t	I_t^{real}	S_t^{real}	$\widehat{d}_t^{UM}$	X_t	L_t	I_t	S_t	I_t^{real}	S_t^{real}	$\widehat{d}_t^{GP}$	X_t	L_t	I_t	S_t	I_t^{real}	S_t^{real}
1	1	1,582	1,550	32	0	0	1,503	1,503	0	0	0	0	79	1,568	1,550	100	0	0	68	0	1,650	1,550	18	0	0	0	14
	2	1,528	1,550	0	22	0	1,480	1,480	0	0	0	0	48	1,620	1,550	66	0	0	156	0	1,616	1,550	70	0	0	92	0
	3	1,635	1,550	63	0	0	1,459	1,459	0	0	0	0	176	1,598	1,550	64	0	0	135	0	1,614	1,550	48	0	0	55	0
	4	1,656	1,550	106	0	0	1,439	1,439	0	0	0	0	217	1,594	1,550	15	0	0	44	0	1,565	1,550	44	0	0	0	7
2	5	1,573	1,550	23	0	0	1,541	1,541	0	0	0	0	32	1,570	1,533	0	0	0	4	0	1,577	1,550	20	0	0	0	2
	6	1,568	1,550	18	0	0	1,517	1,517	0	0	0	0	51	1,555	1,550	38	0	0	24	0	1,588	1,550	5	0	0	0	13
	7	1,583	1,550	33	0	0	1,494	1,494	0	0	0	0	89	1,630	1,524	0	0	0	0	35	1,524	1,550	80	0	0	47	0
	8	1,666	1,550	116	0	0	1,473	1,473	0	0	0	0	193	1,599	1,509	0	0	0	0	157	1,509	1,550	49	0	0	0	20
3	9	1,642	1,550	92	0	0	1,598	1,550	48	0	0	0	44	1,614	1,550	37	0	0	0	55	1,587	1,550	64	0	0	0	28
	10	1,627	1,550	77	0	0	1,570	1,550	20	0	0	0	57	1,658	1,550	166	0	0	89	0	1,716	1,550	108	0	0	31	0
	11	1,693	1,550	143	0	0	1,545	1,545	0	0	0	0	148	1,602	1,550	66	0	0	12	0	1,616	1,550	52	0	0	0	60
	12	1,587	1,550	37	0	0	1,522	1,522	0	0	0	0	65	1,633	1,550	198	0	0	173	0	1,748	1,550	83	0	0	46	0
4	13	1,552	1,550	2	0	0	1,577	1,550	27	0	0	25	0	1,639	1,550	0	57	0	171	0	1,666	1,550	43	0	0	87	0
	14	1,605	1,550	55	0	0	1,552	1,550	2	0	0	0	28	1,649	1,550	89	0	0	205	0	1,696	1,550	99	0	0	131	0
	15	1,641	1,550	91	0	0	1,529	1,529	0	0	0	0	112	1,622	1,550	86	0	0	200	0	1,636	1,550	72	0	0	112	0
	16	1,578	1,550	28	0	0	1,507	1,507	0	0	0	0	71	1,605	1,550	139	0	0	311	0	1,689	1,550	55	0	0	139	0
5	17	1,603	1,550	53	0	0	1,627	1,550	77	0	0	24	0	1,618	1,493	0	145	0	201	0	1,659	1,550	0	71	0	86	0
	18	1,657	1,550	107	0	0	1,598	1,550	48	0	0	0	35	1,619	1,550	0	85	0	94	0	1,610	1,550	0	2	0	0	21
	19	1,571	1,550	21	0	0	1,573	1,550	23	0	0	2	0	1,599	1,550	0	58	0	73	0	1,577	1,550	47	0	0	26	0
	20	1,788	1,550	238	0	0	1,548	1,548	0	0	0	0	238	1,614	1,550	0	0	0	0	165	1,608	1,550	64	0	0	0	148

Note(s): The predicted demand $\widehat{d}_t^{AR}$, $\widehat{d}_t^{UM}$ and $\widehat{d}_t^{GP}$ are obtained based on the AR model, U-MIDAS model and GP-U-MIDAS model, respectively
Source(s): Table by authors



Source(s): Figure by authors

Figure 6. The service levels among AR, U-MIDAS and GP-U-MIDAS approaches

	AR	U-MIDAS	GP-U-MIDAS
The minimum service level	86.69%	90.58%	91.72%
The average service level	94.90%	98.79%	99.08%

Source(s): Table by authors

Table 8. The service levels among AR, U-MIDAS and GP-U-MIDAS approaches

Planning horizon	1	2	3	4	5	Total cost
The actual demand	72064	70280	74282	70640	78475	365741
AR	80410	77200	78314	74175	79922	390021
Δ_{AR}	11.58%	9.85%	5.43%	5.00%	1.84%	6.64%
U-MIDAS	73706	73432	78604	74484	73116	373342
Δ_{UM}	2.28%	4.48%	5.82%	5.44%	−6.83%	2.08%
GP-U-MIDAS	72524	71022	76320	72973	76069	368908
Δ_{GP}	0.64%	1.06%	2.74%	3.30%	−3.07%	0.87%

Source(s): Table by authors

Table 9. The total operational costs in each planning horizon

Furthermore, we conduct 100 simulations to evaluate the robustness of the three forecasting approaches. The mean and the corresponding 95% confidence intervals of the service level stability and the total cost difference with different approaches are recorded in the repetitive computing procedure. Service level stability is measured by the standard deviation of the service level. Based on the results in Table 10, the service level stability and the total cost difference using the AR approach have the largest deviation compared to the results obtained under the actual demand. In particular, the mean values of service level stability and total cost difference are reduced from 3.60 to 5.08% using the AR approach to 2.15 and 2.36% using the GP-U-MIDAS approach. A notable feature of the comparison results is that these mean values

are relatively small (less than 3%). That is, the manufacturer can save operations costs using the data-driven production planning model under the GP-U-MIDAS approach, and this method can generate excellent results that are very close to those results obtained under the actual demand.

Table 10. Comparison results regarding service level stability and total cost difference

Approaches	Service level stability		Total cost difference	
	Mean	95% confidence interval	Mean	95% confidence interval
AR	3.60%	[3.45%,3.75%]	5.08%	[4.81%,5.36%]
U-MIDAS	2.49%	[2.33%,2.64%]	2.76%	[2.50%,3.01%]
GP-U-MIDAS	2.15%	[2.02%,2.29%]	2.36%	[2.12%,2.59%]

Source(s): Table by authors

In this section, several numerical examples are presented to evaluate the performance of the data-driven production planning model integrated with the GP-U-MIDAS approach from two aspects. In Section 5.1, experiments on market demand forecasting are designed and conducted in several ways. We find that under certain scenarios, linear univariate models, e.g. AR, cannot utilize high-dimensional mixed-frequency information. These models also lead to underfitting. In contrast, multisource mixed-frequency data can be utilized by multivariate linear mixed data models, e.g. U-MIDAS; however, these models result in overfitting. Thus, we incorporate a group selection approach to sparse modeling, e.g. GP-U-MIDAS, which is an effective way to avoid overfitting and underfitting. In Section 5.2, the performance of the data-driven production planning models is further evaluated in terms of multiple aspects, including planning decisions (production quantity and outsourcing quantity), service levels and operational costs. For the cost-saving manufacturer, the numerical results indicate that the data-driven production planning model integrated with the GP-U-MIDAS approach is an excellent choice compared to using the U-MIDAS approach and the AR approach.

5.3 Managerial insight

Through the aforementioned numerical analysis, there are practical insights of our work that can assist manufacturers in making informed production decisions utilizing mixed-frequency data. Moreover, our work presents practical guidance regarding how optimization integrated with a data-driven approach can be applied to optimize production planning decisions. The practical insights of our work are summarized as follows.

First, we propose an efficient way to process historical data in terms of different frequencies, i.e. identify and utilize the feature-demand dataset. This approach could utilize various types of data and is widely applied in different industries, e.g. the retail and apparel industries.

Second, the external variable of market demand plays a significant role in improving the accuracy of demand estimations and production decisions. The crucial variables can be accurately extracted to help operations managers become more aware of what factors influence market demand. Operations managers also need to identify and optimize the most valuable parameters of the demand features.

Third, operations managers need to pay more attention to underfitting and overfitting when forecasting results perform well in-sample but not out-of-sample, which incurs misleading errors in demand prediction and production decisions.

Finally, our proposed data-driven production planning model can efficiently solve a realistic issue, i.e. the capacitated production planning problem with outsourcing over a finite planning horizon.

6. Conclusions and outlook

In this paper, we investigate the production planning problem when the operations manager has massive amounts of historical demand data on demands and demand features. Specifically, we study the feature value of massive demand information to improve the accuracy of a firm's demand forecasting; we show how incorporating predicted demand information with the production planning results in efficient planning improvements.

In terms of production and operations management, our work contributes to the literature by taking a "big-data" perspective to investigate a manufacturer's response to process a large amount of relevant information. Our study also advances the production research literature by incorporating the existing focus on historical demand information to understand the impacts of features related to demand. Our model contributes to forecasting by proposing a new way to handle high-dimensional feature data. Specifically, the data-driven production planning problem is solved based on the estimation of demand.

To fully utilize the features of massive demand information (i.e. the different frequencies of demand information), the GP-U-MIDAS model is proposed to optimize the data processing method and to improve the demand forecasting accuracy. In our work, the forecasting model can eliminate the insignificant impacts incurred by some high-frequency features and retain the key influential features; thus, the predictive capability of the model can be improved. This result can benefit substantial operations decisions in a variety of situations. Our analytical results are in line with the literature. For example, the safety stock and the inventory cost can be effectively reduced based on accurate sales and demand forecasting (Chen and Lee, 2009).

Regarding future research, there are many potential directions, and we discuss only a few. First, investigating how sourcing decisions can be solved with demand-feature data is an open question. Second, social media data could be incorporated to learn more information about customers' preferences. Finally, another interesting extension is to jointly consider unpredicted supply or production disruptions in multiproduct manufacturing systems. The application of our models in a realistic setting requires further investigation.

References

Andreou, E. and Ghysels, E. (2021), "Predicting the VIX and the volatility risk premium: the role of short-run funding spreads volatility factors", *Journal of Econometrics*, Vol. 220 No. 2, pp. 366-398.

Andreou, E., Ghysels, E. and Kourtellos, A. (2010), "Regression models with mixed sampling frequencies", *Journal of Econometrics*, Vol. 158 No. 2, pp. 246-261.

Bai, B. (2022), "Acquiring supply chain agility through information technology capability: the role of demand forecasting in retail industry", *Kybernetes*, Vol. ahead-of-print No. ahead-of-print, doi: 10.1108/K-09-2021-0853.

Ban, G.Y. and Rudin, C. (2019), "The big data newsvendor: practical insights from machine learning", *Operations Research*, Vol. 67 No. 2, pp. 90-108.

Baumeister, C., Guérin, P. and Kilia, L. (2015), "Do high-frequency financial data help forecast oil prices? The MIDAS touch at work", *International Journal of Forecasting*, Vol. 31 No. 2, pp. 238-252.

Boone, T., Ganesha, R., Hicks, R.L. and Sanders, N.R. (2018), "Can google trends improve your sales forecast?", *Production and Operations Management*, Vol. 27 No. 10, pp. 1770-1774.

Breheny, P. and Huang, J. (2015), "Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors", *Statistics and Computing*, Vol. 25 No. 2, pp. 173-187.

Carbonneau, R., Laframboise, K. and Vahidov, R. (2008), "Application of machine learning techniques for supply chain demand forecasting", *European Journal of Operational Research*, Vol. 184 No. 3, pp. 1140-1154.

Chen, L. and Lee, H.L. (2009), "Information sharing and order variability control under a generalized demand model", *Management Science*, Vol. 55 No. 5, pp. 781-797.

Chu, C., Chu, F., Zhong, J. and Yang, S. (2013), "A polynomial algorithm for a lot-sizing problem with backlogging outsourcing and limited inventory", *Computers and Industrial Engineering*, Vol. 647 No. 1, pp. 200-210.

Clements, M.P. and Galvão, A.B. (2017), "Model and survey estimates of the term structure of US macroeconomic uncertainty", *International Journal of Forecasting*, Vol. 33 No. 3, pp. 591-604.

Cui, R., Gallino, S., Moreno, A. and Zhang, D.J. (2018), "The operational value of social media information", *Production and Operations Management*, Vol. 27 No. 10, pp. 1749-1769.

Darvishi, F., Yaghin, R.G. and Sadeghi, A. (2020), "Integrated fabric procurement and multi-site apparel production planning with cross-docking: a hybrid fuzzy-robust stochastic programming approach", *Applied Soft Computing*, Vol. 92, pp. 106-267.

De Armas, J. and Laguna, M. (2020), "Parallel machine, capacitated lot-sizing and scheduling for the pipe-insulation industry", *International Journal of Production Research*, Vol. 58 No. 3, pp. 800-817.

de Medeiros, R.K., da Nóbrega Besarria, C., de Jesus, D.P. and de Albuquerquemello, V.P. (2022), "Forecasting oil prices: new approaches", *Energy*, Vol. 238, 121968.

Demirhan, C.D., Boukouvala, F., Kim, K., Song, H., Tso, W.W., Floudas, C.A. and Pistikopoulos, E.N. (2020), "An integrated data-driven modeling and global optimization approach for multi-period nonlinear production planning problems", *Computers and Chemical Engineering*, Vol. 141, 107007.

Englberger, J., Herrmann, F. and Manitz, M. (2016), "Two-stage stochastic master production scheduling under demand uncertainty in a rolling planning environment", *International Journal of Production Research*, Vol. 54 No. 20, pp. 6192-6215.

Feng, Q. and Shanthikumar, G. (2018), "How research in production and operations management may evolve in the era of big data", *Production and Operations Management*, Vol. 27 No. 9, pp. 1670-1684.

Feng, Q., Luo, S. and Zhang, D. (2013), "Integrating dynamic pricing and replenishment decisions under supply capacity uncertainty", *Manufacturing and Service Operations Management*, Vol. 16 No. 1, pp. 149-160.

Gebennini, E., Gamberini, R. and Manzini, R. (2009), "An integrated production-distribution model for the dynamic location and allocation problem with safety stock optimization", *International Journal of Production Economics*, Vol. 122 No. 1, pp. 286-304.

Gholizadeh, H., Fazlollahtabar, H. and Khalilzadeh, M. (2020), "A robust fuzzy stochastic programming for sustainable procurement and logistics under hybrid uncertainty using big data", *Journal of Cleaner Production*, Vol. 258, 120640.

Ghysels, E., Santa-Clara, P. and Valkanov, R. (2004), "The MIDAS touch: mixed data sampling regression models", working paper, Anderson School of Management, UCLA, Los Angeles, 22 June.

Goli, A., Tirkolaee, E.B., Malmir, B., Bian, G.-B. and Sangaiah, A.K. (2019), "A multi-objective invasive weed optimization algorithm for robust aggregate production planning under uncertain seasonal demand", *Computing*, Vol. 101 No. 6, pp. 499-529.

Gordon, C.A.K. and Pistikopoulos, E.N. (2022), "Data-driven and safety-aware holistic production planning", *Journal of Loss Prevention in the Process Industries*, Vol. 77, 104754.

Govindan, K. and Gholizadeh, H. (2021), "Robust network design for sustainable-resilient reverse logistics network using big data: a case study of end-of-life vehicles", *Transportation Research Part E: Logistics and Transportation Review*, Vol. 149, 102279.

Ho, G.T.S., Choy, S.K., Tong, P.H. and Tang, V. (2022), "A forecasting analytics model for assessing forecast error in e-fulfilment performance", *Industrial Management and Data Systems*, Vol. 122 No. 11, pp. 2583-2608.

Jabbarzadeh, A., Haughton, M. and Pourmehdi, F. (2019), "A robust optimization model for efficient and green supply chain planning with postponement strategy", *International Journal of Production Economics*, Vol. 214, pp. 266-283.

Khoei, M.A., Aria, S.S., Gholizadeh, H., Goh, M. and Cheikhrouhou, N. (2023), "Big data-driven optimization for sustainable reverse logistics network design", *Journal of Ambient Intelligence and Humanized Computing*, Vol. 14, pp. 10867-10882.

Kieslich, C.A., Alimirzaei, F., Song, H., Do, M. and Hall, P.D. (2021), "Data-driven prediction of antiviral peptides based on periodicities of amino acid properties", *Computer Aided Chemical Engineering*, Vol. 50, pp. 2019-2024.

Koca, E., Yaman, H. and Selim Aktürk, M. (2015), "Stochastic lot sizing problem with controllable processing times", *Omega*, Vol. 53, pp. 1-10.

Lee, C.Y. and Çetinkaya, S. (2001), "A dynamic lot-sizing model with demand time windows", *Management Science*, Vol. 47 No. 10, pp. 1384-1395.

Mandl, C. and Minner, S. (2023), "Data-driven optimization for commodity procurement under price uncertainty", *Manufacturing and Service Operations Management*, Vol. 25 No. 2, pp. 371-390.

Penev, S., Leonte, D., Lazarov, Z. and Mann, R.A. (2014), "Applications of MIDAS regression in analysing trends in water quality", *Journal of Hydrology*, Vol. 511, pp. 151-159.

Perakis, G. and Zaretsky, M. (2008), "Multiperiod models with capacities in competitive supply chain", *Production and Operations Management*, Vol. 17 No. 4, pp. 439-454.

Petropoulos, F., Nikolopoulos, K., Spithourakis, G.P. and Assimakopoulos, V. (2013), "Empirical heuristics for improving intermittent demand forecasting", *Industrial Management and Data Systems*, Vol. 113 No. 5, pp. 683-696.

Petropoulosa, F., Makridakisb, S., Assimakopoulosc, V. and Nikolopoulosd, K. (2014), "'Horses for Courses' in demand forecasting", *European Journal of Operational Research*, Vol. 237 No. 1, pp. 152-163.

Shahin, M., Saeidi, S., Shah, S.A., Kaushik, M., Sharma, M., Pious, S. and Draheim, D. (2021), "Cluster-based association rule mining for an intersection accident dataset", in Umer, R., Nadeem, M., Tareen, M.A., Rehman, A.U. and Hussain, S.M. (Eds), *2021 International Conference on Computing, Electronic and Electrical Engineering (ICE CUBE 2021)*, IEEE, Quetta, pp.1-6.

Sheen, J. and Wang, B.Z. (2021), "Measuring macroeconomic disagreement–a mixed frequency approach", *Journal of Economic Behavior and Organization*, Vol. 189, pp. 547-566.

Silvestrini, A. and Veredas, D. (2008), "Temporal aggregation of univariate and multivariate time series models: a survey", *Journal of Economic Surveys*, Vol. 22 No. 3, pp. 458-497.

Sodhi, M.S. (2005), "Managing demand risk in tactical supply chain planning for a global consumer electronics company", *Production and Operations Management*, Vol. 14 No. 1, pp. 69-79.

Sox, C.R. and Muckstadt, J.A. (1997), "Optimization-based planning for the stochastic lot-scheduling problem", *IIE Transactions*, Vol. 29 No. 5, pp. 349-357.

Tambuskar, D.P., Jain, P. and Narwane, V.S. (2023), "An exploration into the factors influencing the implementation of big data analytics in sustainable supply chain management", *Kybernetes*, Vol. ahead-of-print No. ahead-of-print, doi: 10.1108/K-07-2022-1057.

Tirkolaee, E.B., Goli, A. and Weber, G.W. (2019), "Multi-objective aggregate production planning model considering overtime and outsourcing options under fuzzy seasonal demand", in Hamrol, A., Kujawińska, A. and Barraza, M.F.S. (Eds), *Advances in Manufacturing II: Volume 2-Production Engineering and Management*, Springer, Cham, pp. 81-96.

Wagner, H.M. and Whitin, T.M. (1958), "Dynamic version of the economic lot size model", *Management Science*, Vol. 5 No. 1, pp. 89-96.

Wang, X. and Tang, W. (2009), "Optimal production run length in deteriorating production processes with fuzzy elapsed time", *Computers and Industrial Engineering*, Vol. 56 No. 4, pp. 1627-1632.

Wang, W., Zhang, Z., Wang, L., Zhang, X. and Zhang, Z. (2022), "Mixed-frequency data-driven forecasting the important economies' performance in a smart city: a novel RUMIDAS-SVR model", *Industrial Management and Data Systems*, Vol. 122 No. 10, pp. 2175-2198.

Zhu, S., Wu, X., He, Z. and He, Y. (2022), "Mixed frequency domain spillover effect of international economic policy uncertainty on stock market", *Kybernetes*, Vol. 51 No. 2, pp. 876-895.

Corresponding author
Binlong Lin can be contacted at: 210720098@fzu.edu.cn