Demand Forecasting in Supply Chains: A Review of Aggregation and Hierarchical Approaches

M. Zied Babai[a,*], John Edward Boylan[b], Bahman Rostami-Tabar[c]

[a] *Kedge Business School, Talence, 33405, France*

[b] *Centre for Marketing Analytics and Forecasting, Lancaster University, LA1 4YX, UK*

[c] *Cardiff Business School, Cardiff University, Cardiff, CF24 4YX, UK*

Abstract

Demand forecasts are the basis of most decisions in supply chain management. The granularity of these decisions, either at the time level or the product level, lead to different forecast requirements. For example, inventory replenishment decisions require forecasts at the individual SKU level over lead time, whereas forecasts at higher levels, over longer horizons, are required for supply chain strategic decisions, such as the location of new distribution or production centres. The most accurate forecasts are not always obtained from data at the 'natural' level of aggregation. In some cases, forecast accuracy may be improved by aggregating data or forecasts at lower levels, or disaggregating data or forecasts at higher levels, or by combining forecasts at multiple levels of aggregation. Temporal and cross-sectional aggregation approaches are well established in the academic literature. More recently, it has been argued that these two approaches do not make the fullest use of data available at the different hierarchical levels of the supply chain. Therefore, consideration of forecasting hierarchies (over time and other dimensions), and combinations of forecasts across hierarchical levels, have been recommended. This paper provides a comprehensive literature review of research dealing with aggregation and hierarchical forecasting in supply chains, based on a systematic search in the Scopus and Web of Science databases. The review enables the identification of major research gaps and the presentation of an agenda for further research.

Keywords: Supply Chain, forecasting, aggregation, hierarchies, combination

*Correspondence: M.Z. Babai, Kedge Business School, Talence, 33405, France, Tel.: +33-(0)5 56 84 53 61
Email address: mohamed-zied.babai@kedgebs.com (M. Zied Babai)

1. Introduction

Forecasts are required to support most of the decisions in managing the supply chain. Two of the main dimensions that characterise the granularity of the decisions are the time and the product. For the product dimension, decisions go from the individual SKU level, such as in inventory control, up to the level where all SKUs are considered, such as in aggregate capacity planning. For the time dimension, operational decisions are made at the daily or weekly levels, whereas tactical and strategic decisions are made at monthly and yearly levels. At the strategic level, supply chain managers deal more and more with an uncertain capacity and face increasing market and technological changes, which pushes them to consider the assortment of all products offered to those markets when managing capacity and deciding about distribution channels (online, store or omnichannel). At the tactical level, decisions are taken about product assortments, and production and warehousing capacities. Finally, at the operational level, decisions are made to control activities such as inventory replenishment, production schedules, transportation plans, workforce rostering and after-sales services.

1.1. Framework

Traditionally, planning processes at different levels have been conducted independently. It is now recognised that there are advantages of co-ordinating logistical plans through a 'Sales and Operations Planning' process (Harwell 2015, Lapide 2016, Fildes et al. 2019). This is intended to ensure coherence between strategic, tactical and operational plans and decisions across long-, medium- and short-term horizons, and determines why forecasts are needed, as indicated in the leftmost column of Figure 1. This figure aims to give a unifying framework for the topics covered in this review. The arrows going from left to right show how planning and decision processes should inform forecast requirements and different ways in which forecasts may be implemented. The bottom arrow, going from right to left, indicates that the final forecasts should inform the planning and decision processes.

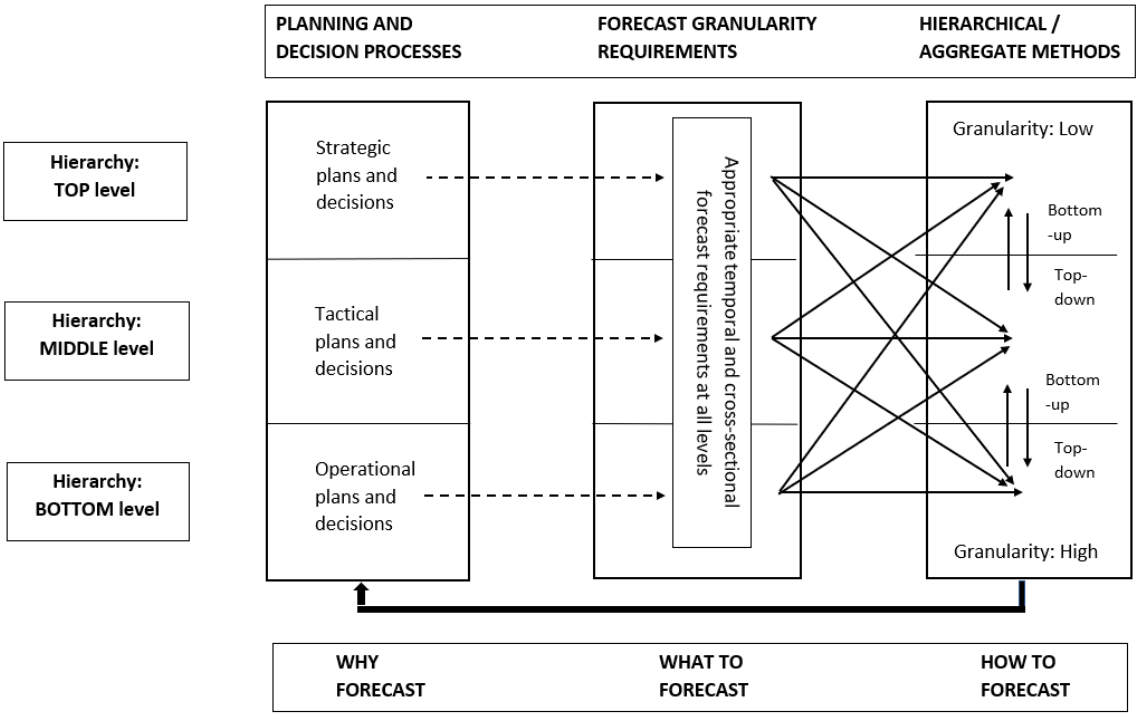

Figure 1: Framework for aggregation and hierarchical approaches in demand forecasting

Figure 1 Alt Text: A box with three levels of the planning and decisions processes (strategic, tactical and operational), leading to a box with the forecast granularity requirements, leading to a final box with the hierarchical and aggregation methods

The granularity of the forecasting requirements, both cross-sectionally and temporally, should be determined by the planning and decision requirements. These links are shown by the dashed arrows between the first and second columns of Figure 1. For example, as discussed earlier, an inventory replenishment decision requires short-term forecasting over lead time at the individual SKU level. Considerations such as these address the question of what to forecast.

It has also become evident that forecasting methods can be conducted at all levels simultaneously. Forecasting at different levels does not have to rely only on data at the level of interest. The question of how to forecast can be addressed by methods that utilise data or forecasts at any level of the hierarchy (shown by the arrows between the second and third columns in Figure 1). This change in perspective opens up opportunities for different approaches to forecasting, by combining forecasts over different levels of a hierarchy.

Alternatively, forecasts can be derived using bottom-up or top- down methods, shown by the vertical arrows in Figure 1 and discussed further in the next sub-section.

1.2. Aggregation Approaches

Different aggregation approaches have been used to deal with the historical demand and forecasting data granularity. Temporal and cross-sectional aggregation approaches represent the oldest approaches. These approaches were presented as good alternatives to manage the demand and to reduce the degree of uncertainty through "risk-pooling" (Chen et al. 2007, Chen and Blue 2010). Temporal aggregation refers to aggregation across time, whereas cross-sectional aggregation refers to aggregation across a dimension (e.g products) in a particular time period (Syntetos et al. 2016). There has been a considerable body of literature dealing with forecasting by aggregation, extending back to the 1950s, mainly in the economics and finance literature, e.g. Theil (1954), Quenouille (1958) and Amemiya and Wu (1972). Aggregation approaches started to attract the attention of supply chain forecasting researchers in the 2000s, with a stream of literature dealing with the bottom-up (BU) and top-down (TD) approaches. The TD approach can be viewed as a demand aggregation approach (associated with a disaggregation mechanism of the forecasts), whereas BU can be viewed as a forecast aggregation approach. Two temporal aggregation approaches have been considered in the supply chain forecasting literature: aggregation with blocking and aggregation with resampling (or bootstrapping). In the former, overlapping or non-overlapping blocks of consecutive time periods are used to aggregate the demand. In the latter, demand over random (not necessarily consecutive) time periods are aggregated. The choice of aggregation approach (or combinations of forecasts across the hierarchies) should be driven by accuracy considerations, subject to the constraints of the data available to the forecaster. For example, if transactional time-stamped data is available, then there is complete freedom to employ any level of temporal aggregation. On the other hand, if the finest level of granularity available is one week, then additional data would need to be collected to allow the application of forecasting methods based on daily patterns.

1.3. Previous Reviews

Over recent years, there have been a few review papers in the literature dealing with topics related to our paper, such as forecasting for inventory planning, supply chain fore-

casting and the bullwhip effect (Syntetos et al. 2009, 2016, Wang and Disney 2016). The most recent and the closest to our topic is by Syntetos et al. (2016), which identified gaps between theory and practice in supply chain forecasting. They proposed a framework based on three dimensions of the supply chain: length, depth and time. Temporal and cross-sectional aggregation were identified as important aspects of the "time" and "depth" dimensions, respectively, within the proposed framework. In the last five years, the field has moved on rapidly, particularly with regard to hierarchical forecasting. Recent developments are covered in the current review.

1.4. Aims of this Review

This paper seeks to achieve three aims relating to demand forecasting in supply chains. Firstly, informed by the framework in Figure 1, we bring together the most recent work on aggregation and hierarchical forecasting, based on a systematic review. Compared to a traditional 'narrative' literature review, a systematic literature search enables a more transparent, rigorous and comprehensive review (Meza-Peralta et al. 2020). We emphasise the importance of improving both accuracy and inventory performance. These common objectives can be achieved by data aggregation (prior to forecasting), or by forecast aggregation, or combining using hierarchical approaches. Although the means differ, the objective is the same - namely to improve performance at a given level of the hierarchy.

Secondly, we aim to identify open research questions in aggregation and hierarchical forecasting. Specific gaps in research are identified at the end of each of the major sections of this review. We bring these findings together at the end of the paper, leading to our presentation of an agenda for further work in this area.

Thirdly, we aim to generate debate on themes that cut across both aggregation-based and hierarchical forecasting. These themes relate to the requirements of such forecasts and their evaluation.

The paper is structured as follows. The following section presents the methodology used to conduct the systematic review. Section 3 reviews the literature that deals with temporal aggregation, covering research on blocking and resampling procedures. In Section 4, we discuss research related to cross-sectional aggregation, with a focus on the top-down and bottom-up approaches. Section 5 addresses the latest advances in forecasting hier-

archies and combinations. In Section 6, we discuss the practical implications of recent developments. We close the paper with conclusions and identified gaps in the literature.

2. Review methodology

We performed a systematic review of the literature to (i) identify all published supply chain forecasting research articles that deal with aggregation and hierarchies; and (ii) qualitatively evaluate their contribution to the field of supply chain forecasting and aggregation; and iii) summarise their findings, strengths and limitations.

The Scopus and Web of Science electronic databases were used to search for all articles published between 1900 and 2021. The following search ontology was used: "["supply chain" OR "inventory" OR "spare part*" OR "intermittent demand"] AND ["forecast*"] AND ["aggregat*" OR "bootstrap*"]". This search ontology was restricted to the title, or keywords, or abstract. The keywords used in the search ontology were determined by the authors of this review by examining well-known papers in the area and identifying appropriate key words to address the three aspects of "supply chain", "forecasting" and "aggregation". It was found that many papers addressing supply chain issues do not use the term as a keyword or in the abstract, and so alternatives were provided. Similarly, some papers do address temporal aggregation using bootstrapping but do not use the term "aggregation" explicitly. The terms "combination" and "hierarchy" were considered but not used because they generated too many papers outside the scope of this review. However, the number of papers in this area is still quite modest and checks have been made to ensure that there have been no significant omissions.

Application of the search ontology yielded 673 documents. Then, we screened the documents in both Scopus and Wed of Science databases, excluding a document if:

- Source type is not "Journal"
- Document types is not "Article"
- Subject (research) area is not relevant, e.g. Physics and Astronomy, Chemistry
- Source title (Journal) is not relevant, e.g. International Journal Of Vehicle Design

- Language is not English

After these exclusions, 310 documents remained.

We also included 78 articles from other sources. These articles include methodological and background papers on temporal and cross-sectional aggregation, not necessarily related to supply chain topics. This judgment was made by the authors of this review, based on their knowledge of the literature.

Following that, we checked for duplication and removed duplicated articles. After their removal, 303 documents remained. Then, unique records were assessed for eligibility to be included in the review sample. Our inclusion criterion consists of any article that is related to supply chain forecasting by temporal or cross-sectional aggregation. The exclusion criteria were identified by the authors of the review after reading all of the abstracts and, in some cases, the full-text articles. It was agreed between the authors of this review to exclude a document from the review sample if it falls into one of the following categories:

- DP - demand planning rather than forecasting
- EI - economics of inventories, i.e. aggregate inventories in the economy rather than in a supply chain
- IP - inventory planning, primarily about optimisation rather than forecasting
- IS - information sharing and bullwhip effect
- NSC - not supply chain
- TM - using aggregation terminology but not in the sense defined in this paper as temporal, cross-sectional, hierarchies and cross-temporal.

Full-text articles were assessed, using these criteria, for eligibility by the authors of this review, independently. In case of any conflict in the judgement, the lead author of the paper made the final decision whether to include the article in the final review sample or not. This assessment led to the exclusion of 170 papers. The end result was to identify 133 papers for full review.

Figure 2 illustrates the systematic literature search and its result.

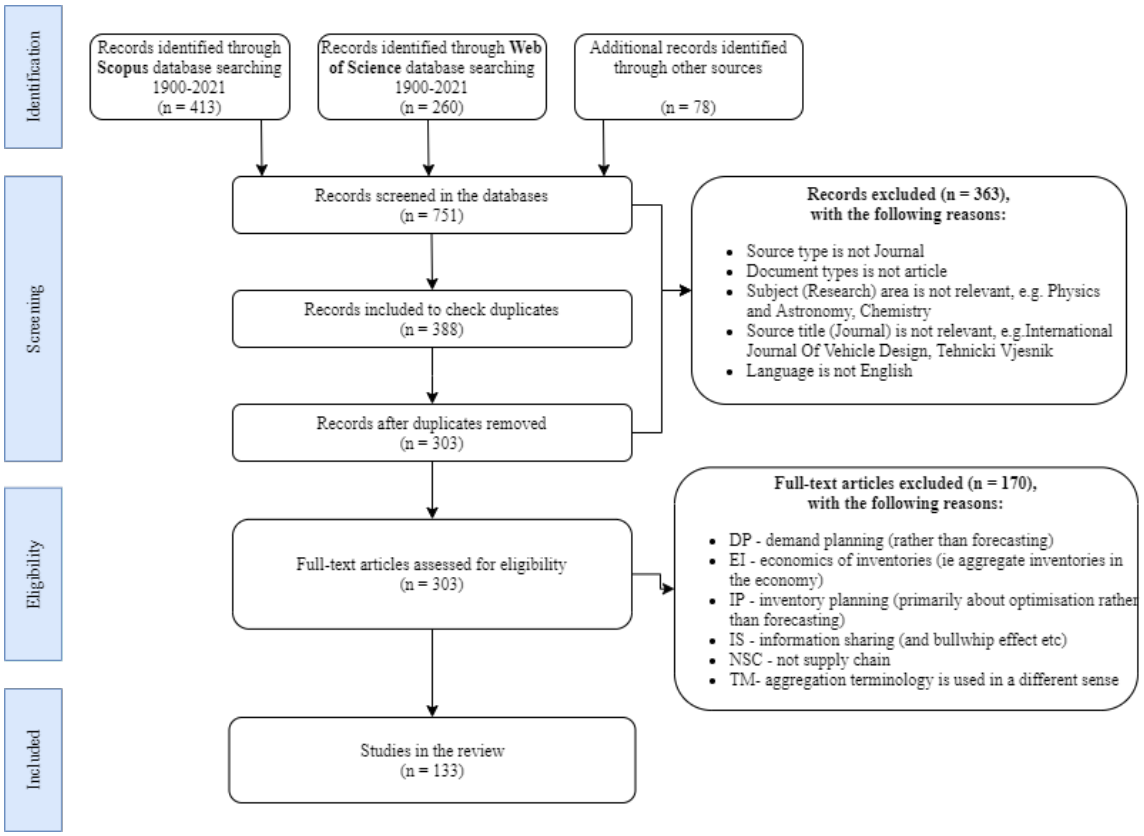

Figure 2: Systematic literature review flow chart

Figure 2 Alt Text: Four levels of the systematic literature review are shown: Identification, screening, eligibility and included where the first level shows the initial number of papers considered (n=413) and the last level shows the final number of studies included in the review (n=133)

For the purpose of checking the systematic literature review process and facilitating future reviews on the same topic, a file with the final list of papers and the excluded ones is provided by the authors upon request.

Our review of the papers included in the final list generated by the systematic literature review revealed three main topics: i) temporal aggregation, ii) cross-sectional aggregation and iii) hierarchical reconciliation and combination. These are the themes of the following sections.

3. Temporal aggregation

Temporal aggregation, or aggregation across time, refers to the process of deriving a low frequency demand series from a high frequency demand series, e.g. weekly demand or monthly demand being derived from daily demand (Nikolopoulos et al. 2011, Petropoulos et al. 2016). In supply chain forecasting, it is usually performed on demand data directly. It can also be performed on demand forecasts. For example, Verstaete et al. (2019) aggregated the estimated sales for the entire selling period.

The aggregation of demand over time can be done using blocking or resampling procedures. In the former, demands are aggregated from blocks of consecutive periods whereas, in the latter, demands are aggregated from randomly resampled periods. In the following subsections, we present each aggregation approach in more detail, and we provide a review of the related literature.

3.1. Aggregation with blocking

Two blocking procedures are commonly considered for temporal aggregation: non-overlapping aggregation and overlapping aggregation. In the former, the demand is divided into consecutive non-overlapping bucket times, where the length of the bucket time is the same and equal to the aggregation level. For example, based on the daily demand over 28 days, a weekly aggregated demand is obtained, which consists of four demands over four blocks of seven days. Non-overlapping temporal aggregation has the advantage of retaining auto-correlation structures in the demand. However, the main disadvantage is the fact that only few blocks are obtained if demand history is short or the aggregation level is long. Moreover, if the history length is not a multiple of the aggregation level, then some of oldest data are discarded.

In overlapping aggregation, the demand is divided into consecutive overlapping blocks with a moving block over time where the block's size is equal to the aggregation level. At each time period, the block is moved one period ahead, so the oldest observation is dropped and the newest is included. The main advantage of the overlapping temporal aggregation approach is that more blocks are available than in the case of the non-overlapping approach. However, in this approach there is a correlation that is induced between blocks, even if it is not present in the original disaggregated demand. Also, un-

der this approach, less 'weight' is given to the most recent observations, as they appear in fewer blocks.

3.1.1. Theoretical foundational work on aggregation with blocking

The foundational work on temporal aggregation with blocking started in the 1970s. It analysed the impact of temporal aggregation on the characteristics of time series when they are modelled as autoregressive integrated moving average (ARIMA) processes. The research started with the work of Amemiya and Wu (1972) for the non-overlapping aggregation case. It was shown that, if a time series follows a p-th order autoregressive process, ARIMA(p,0,0), then the non-overlapping aggregates follow a mixed autoregressive moving average (ARIMA) model of order $(p, 0, q*)$ where $q* = [\frac{(p+1)(m-1)}{m}]$ and $[x]$ denotes the integer part of the real number x. This result was generalised by Weiss (1984) and it was shown that the temporal aggregation of an ARIMA(p, d, q) process follows an ARIMA(p, d, r) process where $r = [\frac{(p+d+1)(m-1)+q}{m}]$. Wei (1978) studied the aggregation effect on univariate multiplicative seasonal time series models. It was shown that, for an ARIMA process of order $(p, d, q)\times(P, D, Q)s$, the corresponding aggregate process is an ARIMA$(p, d, r)\times(P, D, Q)s^*$ if s is a multiple of m (where $r = [\frac{(p+d+1)(m-1)+q}{m}]$ and $s^* = s/m$) and an ARIMA$(P + p, D + d, r)$ if m is a multiple of s. Brewer (1973) also presented a generalisation of the results for ARIMA models with exogenous variables, i.e. ARIMAX models. It is shown that the temporally aggregated ARIMAX model is an ARIMAX. For overlapping temporal aggregation, Hotta et al. (1992) have shown that that the temporal aggregation of an ARIMA(p, d, q) process follows an ARIMA(P, d, Q) process where $P \leq p$ and $Q \leq q + m - 1$. Mohammadipour and Boylan (2012) examined the case of integer autoregressive moving average, INARMA(p, q), processes. They showed that the aggregation of an INARMA(p, q) process over a forecast horizon results in an INARMA(p, q) process with the same INAR and INMA parameters but with a different innovation parameter. Theoretical results of the impact of temporal aggregation on the characteristics of time series are summarised in Table 1.

	Disaggregate demand process	Aggregated process (over m periods)
Non-overlapping aggregation	ARIMA(p,d,q)	ARIMA(p,d,r) where $r = [\frac{(p+d+1)(m-1)+q}{m}]$
	ARIMA$(p,d,q)\times(P,D,Q)s$	ARIMA$(p,d,r)\times(P,D,Q)s^*$ if $s = k*m$ where $r = [\frac{(p+d+1)(m-1)+q}{m}]$ and $s^* = s/m$ ARIMA$(P+p, D+d, r)$ if $m = k*s$
Overlapping aggregation	ARIMA(p,d,q)	ARIMA(P,d,Q) where $P \leq p$ and $Q \leq q+m-1$
	INARMA(p,q) with innovation term $Z_t = Po(\lambda)$	INARMA(p,q) with innovation term $Z_t = Po(m*\lambda)$

Table 1: Impact of temporal aggregation on the characteristics of time series

The above cited research has built the basis for a considerable literature that includes analytical and empirical research on forecast accuracy and inventory performance of temporal aggregation with blocking. In the following two sections, we review both analytical empirical research on aggregation with blocking from the perspectives of forecast accuracy and inventory performance.

3.1.2. Forecast accuracy evidence of aggregation with blocking

Most of the research on forecast accuracy of temporal aggregation with blocking is empirical in nature. Nikolopoulos et al. (2011) was among the first studies in the supply chain literature that empirically analysed the effect of non-overlapping blocks temporal aggregation on demand forecast accuracy. Based on the intermittent demand data of 5000 SKUs from the Royal Air Force (RAF, UK), the authors showed the potential benefits of the aggregation-disaggregation approach (referred to as the ADIDA approach) when it is used with the Naïve and Syntetos-Boylan Approximation (acronym SBA to be used hereafter) forecasting methods. Different aggregation block lengths were tested, from two months up to 24 months and the disaggregation process was done using equal weights. Forecast accuracy was measured with several scaled and relative error metrics including the mean absolute scaled error (MASE) and the relative geometric root mean squared error (RGRMSE). The empirical investigation indicated the potential benefit of identifying an optimal aggregation level.

Spithourakis et al. (2011) extended the work by Nikolopoulos et al. (2011) to empirically investigate the performance of the non-overlapping aggregation approach for fast-moving demand. They used data related to 1428 monthly time series from the M3-Competition. The study revealed the considerable reduction of the symmetric MAPE using the aggregation approach when it is associated with forecasting methods such as naïve, single exponential smoothing and the Theta method. Spithourakis et al. (2014) presented the ADIDA framework as a multi-rate signal processing system and they proposed some mathematical properties of each block of the system. Building on the ADIDA framework, Fu and Chien (2019) proposed a method that integrates temporal aggregation and machine learning techniques to forecast the intermittent demands of electronics components. Based on demand histories of 265 products from a worldwide leading electronics distributor, they empirically showed, based on Root Mean Squared Error, Mean Absolute Error and MASE, that the proposed method is more accurate than benchmark forecasting methods for intermittent demand as well as machine learning methods. Jin et al. (2015) empirically analysed the impact of non-overlapping temporal aggregation when order data or point of sales (POS) are used to generate forecasts. Their empirical investigation was based on weekly and monthly order and POS data over two years from a large US consumer packaged goods manufacturer. The forecast accuracy was measured using the mean absolute deviation (MAD). They first empirically confirmed the findings of Rostami-Tabar et al. (2013) on the direct impact of autocorrelation in the data on the relationship between the aggregation approach and forecast accuracy. They also showed that when order data are used, the aggregated approach significantly improves the forecast accuracy, whereas the opposite effect occurs when POS data are used.

Analytical research on forecast accuracy of temporal aggregation with blocking is relatively scarce. In the case of intermittent demand, Mohammadipour and Boylan (2012) analysed the impact of non-overlapping aggregation when demand follows integer autoregressive moving average (INARMA) processes. They showed, through a simulation based on theoretically generated demand data and empirically by means of data of two real demand data series from the automotive and aeronautics industries, that in most cases, forecasting using a temporally aggregated process leads to lower mean square errors (MSE) compared to the cumulative h-step-ahead forecasting method. The outper-

formance is pronounced when the autoregressive parameter is high. The comparative performance is reversed in the case of an INARMA(1,1) demand process with small autoregressive and moving average parameters and short length of forecast horizon.

In the case of fast moving demand, temporal aggregation with blocking was analytically studied by Rostami-Tabar et al. (2013) and Rostami-Tabar et al. (2014). They considered auto-regressive moving average (ARMA) demand processes and the single exponential smoothing forecasting method to derive the MSE of the non-overlapping aggregation approach. They numerically and empirically showed that the aggregation approach usually outperforms non-aggregation for negatively auto-correlated demand and the outperformance is pronounced for high aggregation levels. More recently, Rostami-Tabar et al. (2021) compared, numerically and empirically, using monthly time series of the M4-competition dataset, the MSE of the overlapping and non-overlapping temporal aggregation approaches when forecasting finite auto-correlated demand. They showed that the aggregation approach is preferred to non-aggregation when forecasting negatively auto-correlated series. Moreover, they provided evidence of the outperformance of the overlapping aggregation approach compared to the non-overlapping one for short time series and similar performance between the two approaches when the demand history becomes long.

3.1.3. Inventory performance evidence of aggregation with blocking

Porras and Dekker (2008) is the first empirical work published in the literature to study the inventory performance of the overlapping temporal aggregation when it is used to estimate the empirical distribution of lead-time demand. Based on the case of a Dutch petrochemical complex, the authors compared the performance of overlapping temporal aggregation against the resampling approach proposed by Willemain et al. (2004) when a reorder point policy is used. The empirical study revealed that the former approach overall yielded considerable cost savings compared to the latter. It was also shown that the overlapping aggregation approach can lead to low achieved service levels. Van Wingerden et al. (2014) extended the method of Porras and Dekker (2008) in two ways. Firstly, the window is placed at random over L consecutive periods (for a fixed number of times). Secondly, they allow the window size, L, to vary, by sampling from the realised lead times. The new method, called 'Empirical Plus', was tested empirically using 6000 parts

from three companies. The evaluation was not conducted using forecast accuracy measures, but with the inventory metrics of fill rates and holding costs. The researchers found that, for most parts, the new method did not perform as well as SBA, but Empirical Plus was better than SBA for parts with infrequent demands and low variability in demand sizes. Zhu et al. (2017) tackled the service level under-achievement issue in Porras and Dekker (2008). They combined the aggregation approach with extreme value theory (EVT) to improve the modelling of the tail of lead-time demand. They conducted an empirical investigation using datasets composed of 5549 spare parts from the automotive and aeronautics industries. They showed that the proposed approach (called 'empirical-EVT') gets closer to the target cycle service level (CSL) than the basic overlapping aggregation approach.

Babai et al. (2012) concluded that most of the literature dealing with temporal aggregation focused only on the forecasting accuracy of the aggregation approaches without assessing their economic effects. To address this issue, they conducted an empirical investigation using 4815 SKUs from the RAF to compare the inventory performance (expressed through inventory holding costs and achieved cycle service levels) of forecasting using a non-aggregated approach and a non-overlapping aggregation. They demonstrated that aggregation leads to higher service-cost efficiency than the non-aggregation approach for high target CSLs.

Boylan and Babai (2016) was the first paper to conduct a theoretical analysis of the accuracy of the overlapping and non-overlapping aggregation approaches. This work focused on the estimation of the cumulative distribution function (CDF) of demand when demand is independent and identically distributed (i.i.d.). They showed that both approaches lead to unbiased estimates and derived variance expressions of the CDF estimators for each approach. They also provided evidence, numerically and empirically, that the overlapping approach often leads to a better estimate than the non-overlapping one. However, the latter can outperform the former when the demand history is very short. The analysis of both approaches under an order-up-to-level inventory control policy revealed that when the aggregation level increases the overlapping approach leads to a reduction in backorders when the target cycle service level is high.

3.2. *Aggregation with resampling*

3.2.1. *Foundational work*

Quenouille (1958) and Tukey (1958) proposed jackknife estimation of a parameter, by systematically omitting each observation from a dataset, calculating estimates, and then averaging these estimates. The jackknife can be used to estimate the bias and variance of an estimator, and its confidence intervals. This approach inspired Efron (1979) to develop bootstrap estimation, whereby observations are resampled with replacement, and each observation has an equal probability of being selected. The bootstrap has been applied in many different domains and has become a staple method in statistical science.

3.2.2. *Lead-Time Demand Resampling*

Bookbinder and Lordahl (1989) were the first to apply bootstrapping methods to the estimation of inventory reorder levels. It is assumed that there is a record of previous lead time demands (LTDs). Bootstrap samples, of the same size, are generated by sampling with replacement. The resulting bootstrap distribution over the bootstrap samples is used to estimate the parameter of interest. Bookbinder and Lordahl (1989) focused on the estimation of the p-th fractile of the demand distribution, assuming that the LTD is a stationary random variable and the demand for each SKU is independent of all others. Using synthetic data, they found that the bootstrap can provide acceptable estimates for two-point and bimodal distributions, unlike the normal distribution. No empirical evaluations were undertaken.

Lordahl and Bookbinder (1994) proposed a weighted average of two order statistics to estimate the reorder point. Their simulation analyses focused on synthetic data, generated from a discrete two-point distribution, and bimodal mixtures of normal distributions that are approximately symmetric, negatively skewed and positively skewed. They found the bootstrap method to perform well in terms of inventory service, without an undue increase in inventory costs. However, the effect of autocorrelated demand was not investigated.

Independently from Bookbinder and Lordahl (1989), Wang and Rao (1992) also proposed using bootstrapping methods to estimate the lead-time demand distribution and reorder points for an inventory control system. In their analyses, they recognised that both de-

mand and lead times may be stochastic, and that demand may be autocorrelated. The issue of serial independence of demand observations had been examined by Ray (1980), who concluded that the assumption of independence will be conservative against run-outs when there is negative autocorrelation, but will be inadequate when there is positive autocorrelation. Bootstrapping lead-time demands mitigates this problem, because autocorrelation of demands over individual periods within lead time is captured, although autocorrelation of successive lead times is not. Wang and Rao (1992) investigated the performance of the bootstrap method, using synthetic data, for a range of auto-regressive parameters in an AR(1) demand process and mean values for geometrically distributed lead times, with demand assumed to be normally distributed. The results for the bootstrap estimates were favourable in terms of bias and standard error.

3.2.3. Block resampling

Rather than resampling demands from previous lead times, an alternative approach is to resample demands from previous blocks of time, of fixed length, which may not have coincided with the time elapsing between ordering and receipt. Hall (1985) proposed resampling non-overlapping and overlapping blocks, in the context of spatial data. For univariate time series data, proposals for non-overlapping blocks (Carlstein 1986) and overlapping blocks (Künsch 1989) were made. The latter approach is also known as the Moving Blocks Bootstrap. Park and Willemain (1999) commented on the problem of the dependence structure near the block endpoints, which affects both approaches. They proposed a 'Threshold Bootstrap' method to address this problem. Monte Carlo simulation experiments showed that, if well calibrated, the Threshold Bootstrap can perform better than the Moving Blocks Bootstrap in estimating the standard error of the sample mean for a variety of auto-regressive moving average time series.

3.2.4. Resampling demands from individual periods

Fricker and Goodhart (2000) examined the demand distributions of the Marine Expeditionary Force. They found that direct sampling of lead-time demands was infeasible. Because of a lack of historical data on the inventory position, they could not directly resample lead-time demands. They also recognised that it was possible to design a resampling scheme that made more efficient use of the available historical data, for independent and identically distributed time series. They proposed random resampling of demand from

individual periods (with replacement). For a fixed lead time of L periods, the resampling is done L times, to give the first resampled lead-time demand. This process is repeated many times, and yields estimates of the Cumulative Distribution Function of lead-time demand, and the associated quantiles required for inventory reorder levels.

Willemain et al. (2004) took this idea further, in the context of intermittent demand, by including a Markov chain structure for the resampling of demand occurrence, which takes into account the historical conditional probabilities of demand occurrence, given demand occurrence (or non-occurrence) in the previous period. Their method was not extended to take into account autocorrelation in the series of non-zero demands, nor the cross-correlations between demand intervals and non-zero demand sizes. This is not merely of academic interest: Willemain et al. (1994) had found indications of such correlations in a previous study, subsequently confirmed for a significant minority of aircraft spare parts from the US Defense Logistics Agency (Altay et al. 2012).

Willemain et al. (2004) also introduced a 'jittering' procedure, whereby resampled demand values are adjusted by adding an amount calculated as the product of a standard normal variable and the square root of the demand value. This allows for the generation of plausible values that have not been previously observed. Rego and Mesquita (2015) identified a bias introduced by this procedure and proposed an alternative 'jittering' procedure, which significantly reduces the bias, but does not eliminate it entirely.

A US Patent was granted (Willemain and Smart 2001) for software including the intermittent demand resampling procedure of Willemain et al. (2004). The features and benefits of the new package were presented to professional practitioners as well as to academics, as summarised by Smith and Babai (2011), and the software has continued to be used by commercial organisations over the last 20 years.

3.2.5. Resampling demand size and demand intervals

Zhou and Viswanathan (2011) proposed an alternative resampling approach. Instead of resampling demands (including zeroes) from individual periods, demand sizes (non-zeroes only) and demand intervals are resampled separately. This method has the advantage that it does not impose a demand interval distribution, although it does assume that successive intervals are independent. Its disadvantage is that there may be few demand

intervals to resample if the demand is highly intermittent. Hasni, Babai, Aguir and Jemai (2019) have numerically and empirically evaluated the inventory performance of this method and compared it to that of Willemain et al. (2004). They found that the former outperforms the latter in terms of inventory cost reduction for moderately intermittent demand data and long lead times.

3.2.6. Resampling immediately after demand occurrence

Teunter and Duncan (2009) proposed an adaptation to simple resampling of demands from previous individual periods. In their adaptation, the first resample is taken from non-zero demands, with the remaining resamples over the lead time being taken from all previous demands. This restricts attention to those review cycles with some demand and discounts those cycles without demand.

Teunter and Duncan's modification may be applied to the resampling schemes of Willemain et al. (2004) and Zhou and Viswanathan (2011). It gives the distribution of lead-time demand, conditional upon that demand being non-zero. The empirical higher inventory efficiency of this modified method, when compared to the standard resampling method, was shown by Hasni, Aguir, Babai and Jemai (2019).

3.2.7. Resampling dependent on elapsed time since last demand occurrence

Pennings et al. (2017) introduced a variant of the Willemain et al. (2004) method, based on approximating the probability of a demand occurrence using an empirical distribution of demand occurrences over the lead time (conditional on elapsed time since the last demand occurrence). The authors also developed a parametric method taking this into account. In their empirical analyses of five datasets, they found their parametric approach to be better than their non-parametric methods in terms of of forecast accuracy (Geometric Mean Absolute Error). Mixed results were obtained for inventory performance.

3.3. Empirical evidence

Willemain et al. (2004) compared the lead-time demand distributions generated by their method with that predicted by Croston's method coupled with a normal distribution. The comparison was undertaken on over 28,000 SKUs from a variety of sectors. It was conducted using the Probability Integral Transformation (PIT) technique, recommended

by Willemain et al. (2004) and subsequently extended by Kolassa (2016). Willemain et al. (2004) found their method to perform better than Croston according to the PIT measure.

Porras and Dekker (2008) compared the overlapping blocks method with the approach advocated by Willemain et al. (2004), based on an empirical analysis of spare parts from a Dutch petrochemical complex. They examined the inventory cost implications of the two methods, finding that the overlapping blocks method produced lower costs, with both methods attaining a 90% fill rate.

Rego and Mesquita (2015) analysed over 10,000 SKUs from a Brazilian automotive manufacturer. They compared an adapted form of the method by Zhou and Viswanathan (2011) with parametric forecasting methods, including the Syntetos-Boylan Approximation (Syntetos and Boylan 2005). The evaluation was based on a trade-off between inventory costs and fill rates, for each of the quadrants identified by Syntetos et al. (2005). They found a clear preference for the method of Zhou and Viswanathan (2011) for 'lumpy' demand, for the Syntetos-Boylan Approximation for 'erratic' demand, with the results being less clear for the 'smooth' and 'intermittent' categories.

Syntetos et al. (2015) examined over 4000 series from the US jewellery and over 3000 series from the electronics sector. They compared the approach of Willemain et al. (2004) with Croston's method and the Syntetos-Boylan Approximation. Evaluations of inventories showed modest gains in cycle service levels for the method of Willemain et al. (2004) for the jewellery data, but with the opposite result for the electronics dataset.

Hasni, Babai, Aguir and Jemai (2019) undertook a direct comparison of the bootstrapping methods of Willemain et al. (2004) and Zhou and Viswanathan (2011). The inventory results were favourable for SBA, but with the advantage over the bootstrapping methods diminishing as the backordering costs increased. Babai et al. (2020) compared both of these bootstrapping methods with neural network (NN) methods, finding that the NN approaches could achieve better inventory efficiency that the bootstrapping methods. Hasni, Aguir, Babai and Jemai (2019) proposed a modification of these two bootstrapping methods where the lead-time demand is adjusted by considering that a demand occurs in the first period of each lead-time bucket. A service driven inventory system was considered with two objective service measures: the cycle service level and the fill

rate. They provided empirical evidence that the proposed adjusted methods result in a higher service-cost efficiency compared to the original methods.

Overall, the empirical results are quite mixed, with no clear 'winner' emerging from the published studies. The idea of comparing results based on data features (such as by Rego and Mesquita (2015)) seems more promising. However, a different categorisation method may be needed than that proposed by Syntetos et al. (2005) which was designed for comparing different parametric methods, and not for comparing bootstrapping methods.

3.4. Gaps of research

Research on temporal aggregation in the supply chain forecasting literature has seen many important developments over the years. However, research has been predominantly empirical rather than analytical and based on forecast accuracy evaluations with less interest in supply chain performance.

In the particular context of intermittent demand, INARMA process modelling has been used to bring a foundational framework to develop knowledge on temporal aggregation (with blocking) for this type of demand. However, with the exception of the work by Mohammadipour and Boylan (2012) which focused only on the forecast accuracy of overlapping aggregation, there are no analytical developments in this area. Inspired from the rich supply chain forecasting literature based on the ARIMA framework, the INARMA modelling should be considered to make important developments in analysing forecasting by temporal aggregation of intermittent demand (both overlapping and non-overlapping) and the implications on supply chain performance.

It is also worth noting that most of the temporal aggregation research has been built on the non-overlapping aggregation assumption and little research was dedicated to overlapping aggregation. As shown earlier in the paper, under the ARIMA framework, the characteristics of demand when aggregated with an overlapping approach have not been fully identified yet, which makes the analytical developments less advanced.

Further, despite the rich literature relating to temporal aggregation with resampling, there has been a lack of theoretical research on the resampling methods that are commonly discussed in the literature. For example, the resampling approach developed by

Willemain et al. (2004) has been implemented in commercial demand forecasting software used by many companies without examination of its theoretical foundations. This opens up an avenue for further research to analyse the theoretical properties of such methods.

Finally, although most of the research discussed in this paper deals with point forecasts, it should be noted that a stream in the forecasting literature has been developed to deal with prediction intervals and distributions. This is very important from the supply chain forecasting perspective since the determination of safety stocks and inventory policies parameters rely on such forecasts. This is in line with the research related to aggregation with resampling, where the relevant literature develops interesting approaches to forecast lead-time distributions. However, no research has been conducted on temporal aggregation with blocking aiming at improving the forecasting performance under prediction intervals and distributions, which constitutes a big gap in the supply chain forecasting literature.

4. Cross-sectional aggregation

Cross-sectional aggregation, also known as hierarchical or contemporaneous aggregation, refers to the aggregation across a number of SKUs at a specific time period. Existing approaches to cross-sectional aggregation include the bottom-up, the top-down and the middle-out approach.

4.1. Top-down, Bottom-up and Middle-out

The bottom-up (BU) approach is based on forecasting each series at the bottom-level, and then aggregating these forecasts to produce forecasts for all the series to the group level (if a forecast at the aggregate level is required). Top-down (TD) consists in forecasting directly at the group level (after aggregating the demand) and then disaggregating these forecasts down to the bottom level (if a forecast at the disaggregate level is required). The middle-out (MO) approach combines bottom-up and top-down approaches. A middle level is first chosen and then the BU approach (or the TD approach) is used to generate coherent forecasts for the series above (or below) the middle level by aggregating (or by disaggregating) the middle level forecasts. The three cross-sectional aggregation approaches are illustrated in Figure 3.

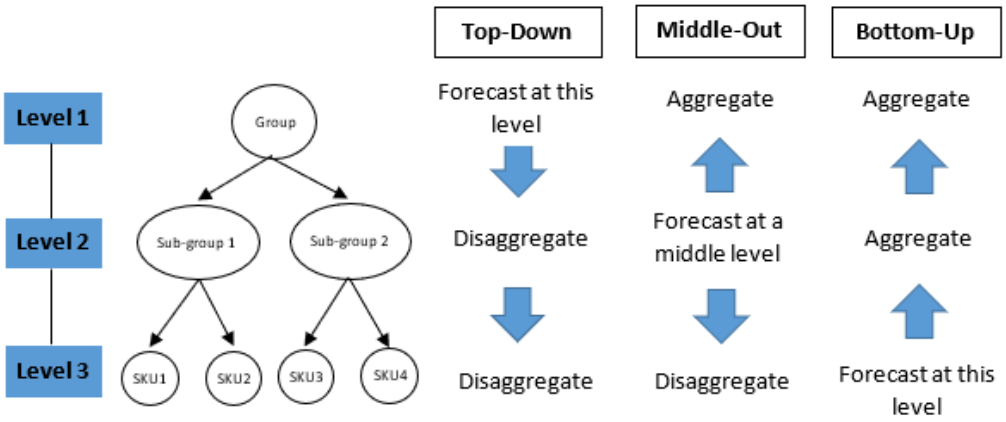

Figure 3: Illustration of the bottom-up, top-down and middle-out approaches

Figure 3 Alt Text: Three levels of a hierarchy are shown and in each level the corresponding aggregation and disaggregation processes are indicated under the Top-Down, Middle-Out and Bottom-Up approaches

4.1.1. Theoretical foundational work on cross-sectional aggregation

The foundational work on cross-sectional aggregation can be divided into two streams. The first stream in the literature deals with the performance analysis of TD and BU approaches. This stream of the literature started with the work of Theil (1954) and was mainly conducted in the economics domain (Shlifer and Wolff 1979, Lütkepohl 2011). There whave been disagreements in that literature on the outperformance of the top-down or the bottom-up forecasting approach. Theil (1954) and Grunfeld and Griliches (1960) argued that the TD approach is more efficient and more accurate for stable demands whereas Orcutt et al. (1968), Edwards and Orcutt (1969), Dangerfield and Morris (1992), Zellner and Tobias (1998) and Weatherford et al. (2001)), among others, argued that BU is preferred when there are differences across time series. As stated by Schwarzkopf et al. (1988), this disagreement is mainly due to type of the generating data process, the forecasting method and the considered forecast accuracy method.

The second stream analysed the impact of cross-sectional aggregation on the characteristics of time series when they are modelled with ARIMA processes. Granger and Morris (1976) showed that the cross-sectional aggrgation of the demand of N subaggregate SKUs following ARMA(p_i,q_i) is an ARMA(x,y) demand where $x \leq \sum_{i=1}^{N}(p_i)$ and $y \leq max(x - p_i + q_i)$. Particularly, aggregating two subaggregate ARMA(1,1) processes with the same parameters leads to the same ARMA(1,1) process. It was also shown by Harvey (1993) that the cross-sectional aggregation of the demand of two subaggre-

gate SKUs following AR(1) processes with parameters ϕ_1 and ϕ_2 is an AR(1) process if $\phi_1 = \phi_2$, an AR(2) process if $\phi_1 = -\phi_2$, and an ARMA(2,1) otherwise. Another example is the work by Silvestrini and Veredas (2008), which showed that the cross-sectional aggregation of two ARIMA(0,1,1) processes is an ARIMA(0,1,1) process. These results on the impact of cross-sectional aggregation on the characteristics of some processes are summarised in Table 2.

	Subaggregate demand processes	Cross-sectional aggregated process
N products	ARMA(p_i,q_i)	ARMA(x,y) where $x \leq \sum_{i=1}^{N}(p_i)$ and $y \leq max(x - p_i + q_i)$
2 products	ARMA(1,1)	ARMA(1,1)
	AR(1) with parameters ϕ_1 and ϕ_2	AR(1) if $\phi_1 = \phi_2$
		AR(2) if $\phi_1 = -\phi_2$
		ARMA(2,1) otherwise
	ARIMA(0,1,1)	ARIMA(0,1,1)

Table 2: Impact of cross-sectional aggregation on the characteristics of time series

4.1.2. Performance of cross-sectional aggregation

In the supply chain forecasting literature, the work by Zotteri et al. (2005) is among the first that empirically analysed the performance of TD and BU approaches. They used sales data from a food retailer to show that both TD and BU can lead to substantial improvements in the mean absolute percentage error (MAPE) and that the choice of the best level of aggregation depends on the underlying demand generation process. Zotteri and Kalchschmidt (2007) compared the MSE of TD and BU forecasting approaches where the forecasting is made at the SKU/store level or at an aggregated level for a set of geographical locations of SKUs/stores. They assumed stationary and non-correlated (over time and across SKUs) demand that is estimated using the minimum MSE method. They concluded that BU should be used only in cases of low demand variability and small size chains. Viswanathan et al. (2008) compared the performance of TD and BU through simulation experiments on subaggregate SKUs with intermittent demand. The forecasts under the BU approach were generated using Croston's method whereas, under TD, the forecasts were calculated using SES (since the degree of intermittence of the aggregate

demand was low). The forecast accuracy was measured by means of the mean absolute deviation (MAD) and the inventory performance was reflected through the total inventory cost including the holding and the shortage costs). The simulation results showed that when the variability of demand intervals of the subaggregate SKUs is low, BU leads to more accurate forecasts than TD (in forecasting at the aggregate level), but when this variability increases, the relative performance of TD improves and it becomes better than the latter under a high number of SKUs. Moreover, if one aggregates a high number of subaggregate SKUs having their demand intervals and demand sizes highly variable, TD is the best forecasting method. They also showed that under TD, SES outperforms Croston in most cases when forecasting the aggregate demand. It should be noted that there is a considerable body of research that analysed the performance of cross-sectional aggregation in terms of forecast accuracy or inventory performance, without referring to the TD or BU approach. This research includes Razi et al. (2004), Zhou et al. (2007), Strijbosch et al. (2008), Murray et al. (2018*a*) and Murray et al. (2018*b*), Villegas et al. (2018) and Narayanan et al. (2019).

Inspired by the BU and TD approaches, Li and Lim (2018) proposed a method to forecast intermittent demand at the store level for a fashion retail in Singapore. The method, referred to as the greedy aggregation-decomposition method, is composed of three-steps. The first step of the method consists in forecasting the daily demand by using a modification of the Holt-Winters method after cross-sectionally aggregating the demand of all SKUs. The second step consists in forecasting the demand size and interval for each SKU by using SES as in Croston's method. The last step is to allocate the total demand to each SKU at each store based on size and interval forecasts generated in the first steps, instead of using the popular proportional allocation method as in the TD approach discussed in Gross and Sohl (1990). The method was assessed using MAE and measures related to MASE. The outperformance of the proposed method was shown when compared to some benchmark methods commonly used in the literature for intermittent demand forecasting.

Analytical research on cross-sectional aggregation is relatively scarce. Widiarta et al. (2007) is among the first research works that analytically evaluated and compared, by means of the MSE, the performance of TD and BU approaches for autocorrelated de-

mands in the supply chain. They assumed an autoregressive AR(1) demand process that is forecasted at the SKU level using SES. They showed that if the lag-1 autocorrelation of the demand for at least one of the SKUs in the family is higher than 1/3, the BU approach leads to lower variance of forecast error than TD. Widiarta et al. (2008) extended the previous work when the demand of all the subaggregate SKUs follow an MA(1) process. They showed that the performance of the TD and BU approaches is the same if the smoothing constants used for forecasting the subaggregate SKUs and the aggregate family demand are set to the optimum value or equal to each other. Under the same demand process and forecasting method, Widiarta et al. (2009) additionally showed, by means of simulation, that when the correlation parameters of the subaggregate SKUs are negative, TD outperforms BU in terms of variance of forecast error. When the parameters have different signs, BU performs better TD when the correlation between the components is negative, whereas when the correlation between the components is positive, TD becomes the preferred approach.

Sbrana and Silvestrini (2013) and Rostami-Tabar et al. (2015) analytically derived the MSE expression of BU and TD approaches when forecasting aggregate and subaggregate demand in the case of a non-stationary demand process. They assumed that the sub-aggregate demand follows an Integrated Moving Average of order one (i.e., ARIMA(0,1,1)) demand process and it is forecasted using SES. By means of numerical experiments, they showed that when the moving average parameter for all the subaggregate SKUs or the smoothing constant used for these SKUs are identical, the performance of BU and TD is the same. Moreover, Rostami-Tabar et al. (2015) demonstrated, based on an empirical investigation using data of a European superstore, that when the demands of the subaggregate SKUs are highly autocorrelated, the performance of BU and TD is also the same for all autocorrelation values, according to ratios of variances. Their investigation revealed that, at the aggregate level, BU is preferable to TD when the cross-correlation between the sub-aggregate SKUs is positive and low or takes negative values. These findings confirm the early simulation based results shown by Fliedner (1999) on the benefit of demand cross-sectional aggregation of highly negatively cross-correlated subaggregate items. At the sub-aggregate level, BU outperforms TD when the smoothing

constant is set to its optimal value for both approaches, regardless of the cross-correlation, the disaggregation weight or the values of the process parameters.

Kremer et al. (2016) assessed how judgment affects the relative accuracy of the BU approach and the direct forecasting at the group level (referred to as top-direct) approach. They provided evidence that BU outperforms top-direct, using an MAE measure, if the subaggregate items are affected similarly by short- and long-term shocks, e.g., products that are affected similarly by general market growth ("change") and weather effects ("noise").

4.2. Seasonal group indices

An important special case of cross-sectional aggregation relates to the forecasting of seasonal time series. These are ubiquitous in retailing but accurate seasonal estimation using classical methods requires long demand histories (Hyndman and Kostenko 2007). In practice, organisations are often obliged to contend with short demand histories, because the product is relatively new to the market, or because the entire history is not available on an Enterprise Resource Planning system.

Classical methods for forecasting seasonal demand, such as the Holt-Winters method, rely only on a product's own demand history. Methods like these are sometimes known as 'Individual Seasonal Indices' (ISI) methods. They will not produce accurate forecasts if there are short demand histories or even for longer histories if the data are very noisy. However, there is an opportunity to generate more accurate forecasts if the individual series is part of a group of seasonally homogeneous series (for example, across locations, or across products).

There are two basic methods of seasonal aggregation, to form 'Group Seasonal Indices' (GSI). One approach is to sum demands across a seasonal group and then to estimate the seasonal indices (for all series) from the aggregate series (Withycombe 1989). This is known as the 'Withycombe Group Seasonal Index' (WGSI). The other approach is to calculate individual seasonal indices for each item in the group and use this average for all items in the group (Dalhart 1974). This is known as the 'Dalhart Group Seasonal Index' (DGSI). Both of these approaches address seasonality for non-trended data.

Chen and Boylan (2007) commented that, instead of applying the ISI or a GSI method for all products, one can choose between ISI or GSI for each product and, further, choose between WGSI and DGSI. They found that WGSI is more accurate than ISI. in terms of mean square error, if the coefficient of variation of the deseasonalised individual series is greater than the coefficient of variation of the deseasonalised aggregate series. These results show that more noisy time series can 'borrow strength' from other series with homogeneous seasonality, but less noisy series may 'borrow weakness' even if the seasonal patterns are homogeneous.

The above studies are restricted to non-trended data. Dekker et al. (2004) and Ouwehand et al. (2005) proposed an adaptation of the Hot-Winters method, whereby level and trend estimates are updated at the level of the individual series, but seasonal indices are updated using aggregated data across a product family. Empirical results on data from food and electrotechnical wholesalers showed that the adapted method was more accurate than the classical Holt-Winters approach, based on an assessment of MAD, MSE and symmetric MAPE measures.

Most research work on seasonal aggregation has assumed that the groupings are given. For example, an organisation's standard product groupings could be used. This approach has limitations because homogeneity in product features is not always associated with seasonal homogeneity (Zotteri et al. 2005). Boylan et al. (2014) proposed a k-means clustering method, based on theoretical linkages to the MSE criterion. Testing on empirical data from a lighting company showed that the approach may be used with confidence if a company lacks a grouping method.

4.3. Gaps of research

Despite the huge literature dealing with the comparative performance of BU and TD approaches, there is still a lack of simple theoretical rules and indications on which approach should be used in general and complex situations.

It is also worth stating that most of the research dealing with the analysis of performance of the cross-sectional aggregation approaches in the context of intermittent demand has been empirical in nature and simulation-based. Similar to the case of temporal aggregation, the INARMA modelling represent an interesting framework to model such demand

patterns, which should be further considered in the literature to strengthen the findings in this area with some theoretical properties of the cross-sectional aggregation approaches.

Finally, another gap in the research dealing with cross-sectional aggregation consists in the focus of the relevant research on forecast accuracy and the limits in using other utility functions when evaluating the effectiveness of the considered approaches. The inventory service level/cost is an obvious utility function that should be considered at the bottom levels of the hierarchy but other utility functions (e.g. finance, marketing) could be used at different upper levels of the hierarchy. Evaluating the performance of cross-sectional aggregation approaches under different utility functions for different levels of the hierarchy is lacking in the existing literature.

5. Hierarchical reconciliation and combination

Temporal and cross-sectional aggregations, discussed in previous sections, are limited in the sense that they do not fully utilise all the information available at various levels of aggregation. In this section, we discuss some approaches designed to overcome this limitation. These approaches include hierarchical and grouped time series reconciliation, multiple temporal aggregation, and temporal and cross-temporal hierarchies.

5.1. Hierarchical and grouped time series reconciliation

In supply chains, a collection of time series can be represented as a hierarchical or grouped time series structure. (For example, the total demand for a retail item can be disaggregated into demand on each regional warehouse, and further disaggregated by demand on each retail outlet.) These categories are nested within the larger group categories and the resulting time series of nested categories are refereed to as "hierarchical time series" (Hyndman et al. 2011). An alternative aggregation structure is grouped time series where the collection of time series can be grouped together in a number of non-hierarchical ways. For example, a supply chain manager might be interested in attributes such as product family, customer type, price range, etc. Such attributes do not naturally disaggregate in a unique hierarchical manner as they are not nested (Hyndman and Athanasopoulos 2021). In supply chains, one may have more complex structures including both hierarchical and grouped time series. For example, it would be natural for the

supply chain manager to be interested in demand by product family, customer type and also by geographic locations.

The traditional methods discussed in Section 4 (BU, TD and MO) have some limitations. They only use base forecasts from a single level of aggregation which have either been aggregated or disaggregated to obtain forecasts at other levels. Hyndman et al. (2011) proposed the optimal combination (OC) approach as an alternative which uses the information available at all levels of the structure. This approach first generates forecasts at each node of the hierarchy separately and then combines and reconciles all forecasts, in order to produce coherent forecasts. That is, forecasts can add up in a way that is consistent with the aggregation structure of the hierarchy or group that defines the collection of time series. For example, considering the demand for a retail item, forecasts of demand for the item in regional stores should add up to demand forecasts for regional warehouses, which should in turn add up to give a demand forecast at the total level. Following the development of the optimal combination approach, several extensions have been proposed that focus on the theoretical advancement of forecast reconciliations on both points (Hyndman et al. 2016, Wickramasuriya et al. 2019, Panagiotelis et al. 2021) and probabilistic forecasts (Taieb et al. 2017, Jeon et al. 2019, Taieb et al. 2020, Panagiotelis et al. 2020). While these hierarchical time series reconciliation and combination approaches have wide applications, there are a limited number of studies that investigate the application in the real supply chains. Mircetic et al. (2021) used real time series of a supply chain distribution network from a European brewery company to assess the performance of optimal reconciliation proposed by Wickramasuriya et al. (2019) and Hyndman and Athanasopoulos (2021) against the BU and TD approaches using an empirical investigation. The dataset available for the purpose of this research consisted of weekly time series for 56 SKUs for the period from 2012 to 2015. These time series were then grouped based on: i) regions (marketing regions); ii) distribution centres; iii) wholesalers; and iv) product types. The ETS models from the forecast package in R were used to produce the out-of-sample base forecasts for 52-steps-ahead (one year ahead), which is required for planning. They reported forecast accuracy using Root Mean Square Scaled Error (RMSSE) and showed that the forecast performance of BU and OC evaluated across the whole structure is not statistically different. They also examined the point

forecast combination of BU and OC instead of using them individually. They showed that combining the forecasts of OC and BU produce consistently more accurate forecasts through all nodes of the supply chain grouped structure.

There are three studies (Abolghasemi et al. 2019, 2020, Spiliotis et al. 2020) that use a dataset containing sales of 55 products for 120 weeks, obtained from a food manufacturing company in Australia. Each product forms a hierarchy with three levels: 12 distribution centers at the bottom, two retailers in the middle and the total, giving 660 product-location combinations at the bottom level. Retailers at the middle level of the hierarchy have different sales patterns, while the bottom level series have a similar sales pattern to the middle-level series.

Abolghasemi et al. (2019) investigated the hierarchical forecasting problem of sales time series in the presence of promotion on a three-level structure including top, middle and bottom levels. They used a middle-out (MO) approach to generate forecasts at all levels. Forecasts are first generated at a middle level and then the middle-level forecasts are aggregated to the top level and disaggregated to the lower levels. They proposed using machine learning (ML) models including artificial neural networks (ANN), extreme gradient boosting (XGBoost), and support vector regression (SVR) for dynamic hierarchical forecasting where the time series dynamics may change due to promotion. These models estimate the proportions of lower-level time series from the upper level. They also compared the proposed approaches with various variations of optimal combination, BU, TD and MO approaches. They used ARIMAX with price as the explanatory variable to generate base forecasts of four-step-ahead and eight-step-ahead averages. The symmetric MAPE measure was used to evaluate the forecast accuracy. The results showed that the performance of the considered approaches depends on the forecasting horizon and the level of the hierarchy. At the bottom level, the XGBoost outperforms the other ML and statistical models. For the top level, they showed that the best forecasts across the entire horizon are generated with the TD and the top-down forecasted proportions (TDFP) model.

From the existing hierarchical approaches, the accuracy of an optimal combination approach with minimum trace was shown superior in many empirical studies over other

alternatives (Wickramasuriya et al. 2019). However, there are still some circumstances where this method may fail, which are summarised in Abolghasemi et al. (2020). In fact, there is no single approach that generates accurate forecasts across all levels of different hierarchical and/or grouped time series structures. The suitability of approaches may depend on the characteristics of the time series and the structure of the hierarchy. Abolghasemi et al. (2020) examined the selection of suitable approaches, based on time series features, and used Machine Learning (ML) for classification. They compared the proposed approach against BU, TD and MO approaches. To generate forecasts for 4 weeks ahead, they used a regression model with ARMA errors (Reg-ARMA), where product prices are used as a predictor variable. They used the MASE and RMSSE error metrics to examine the performance of hierarchical approaches. The results indicated that, on average, the proposed approach was the most accurate hierarchical forecasting method. A detailed analysis showed that the TD approach outperforms the proposed approach at the top level. They recommended to expand model selection to reconciliation method selection when dealing with forecasting hierarchical and grouped time series. While the hierarchical combination approaches explored in the literature are generally linear in nature, Spiliotis et al. (2020) proposed a non-linear approaches to the problem of hierarchical forecast reconciliation. They used Random Forests (RF) and XGBoost (XGB) methods to derive the combination weights for the forecasts across the various aggregation levels. These methods have been shown to perform well in time series contexts and cross-learning. They used ARIMA to estimate the base forecasts and to act as a benchmark against BU, TD and OC. They evaluated the forecasting performance of the hierarchical forecasting methods using MASE, RMSSE and absolute mean scaled error (AMSE). They showed that ML reconciliation approaches were superior to existing, linear ones, in terms of forecast accuracy.

The above studies have examined the application of the OC approach and its extensions in the supply chain. The overall conclusion is that using information across the hierarchy improved forecast accuracy, compared to a situation when separate levels are used to generate forecast requirements. Moreover, using a combination of hierarchical approaches or multiple approaches instead of using a single approach for the entire hierarchy can improve accuracy. However, it is not easy to draw concrete conclusions on when

each approach provides more accuracy. Datasets used in these studies have weekly granularity, so it may not be appropriate to generalise results to other granularities such as sub-daily, daily and monthly. More studies with other time granularities in the supply chain should be considered.

The M5 forecasting competition (Makridakis et al. 2020) was organised as an online contest to predict the sales of thousands of products from a US retailer (Walmart). It is the biggest, so far, of a series of forecasting competitions organised since 1982 by Professor Makridakis, aimed at enhancing forecasting methodology and practice (Makridakis et al. 1993). The purpose of the M5 forecasting competition was to compare the empirical accuracy of forecasts (up to 28 days ahead) using a wide range of forecasting methods in a hierarchical supply chain with grouped time series, thereby allowing assessment of methods based on aggregation and hierarchies. The dataset contains 42,840 daily time series of sales data in total. It has a structure with the SKU items at the bottom level and aggregation based on three states in the US, store, department and product categories. In addition to the sales time series, it also includes the exogenous variables of promotions, price, and special events for the bottom level series. Both point forecasts and prediction intervals are generated at all levels for 28 days ahead. Results of the M5 competition (Makridakis et al. 2021, 2020) show that ML approaches such as LightGBM outperform statistical models in forecasting hierarchical retail sales. Moreover, results indicate that using exogenous variables improves forecast accuracy, according to RMSSE for point estimates. The M5 competition is the most comprehensive experiment related to hierarchical forecasting in supply chains so far. There is a special issue under preparation for the *International Journal of Forecasting* that will be dedicated to the competition.

There are some other studies that proposed hierarchical forecasting approaches that are more specific to the context of supply chains. These approaches also used all information available in a hierarchy. Nenova and May (2016) used an empirical approach to create a model to forecast the optimal forecast aggregation technique for a data set with two levels of hierarchies: bottom and top. The approach establishes a relationship between correlation of time series at the bottom level and the outperforming aggregation. Therefore, it is possible that various approaches are used in the forecasting process instead of using one approach. They developed an analytical model to choose an expected optimal

forecast consolidation strategy. They showed an accuracy gain from using the proposed procedure, as opposed to using the same strategy (e.g. BU or TD) for all datasets. (Accuracy of hierarchical approaches was reported using MAE and RMSE.) The paper does not explicitly describe what values of correlation favour the BU or TD approach. Also, their approach was not compared with the OC approach in terms of performance or computational time.

Pennings and Van Dalen (2017) proposed an integrated hierarchical forecasting approach to forecast the demand of products at different hierarchical aggregation levels. It first generates forecasts at all levels and then incorporates available information. The generated forecasts are already reconciled and add up in the same manner as data in the hierarchy. Therefore, this approach avoids ex-post revising of forecasts, as is done in the OC approach. Two different datasets from food and personal care sectors were used to evaluate forecasting and inventory performance. The results demonstrate that forecast accuracy and inventory performance can be substantially improved with respect to the BU, TD and the optimal combination approaches. (Forecast accuracy based on MAPE.)

Huber et al. (2017) proposed a decision support system to provide hierarchical forecasts at different organisational levels, based on point-of-sales data of multiple items. The approach identifies clusters of items that are used to extend the hierarchy based on intraday sales patterns. They used univariate and multivariate ARIMA models to forecast time series. They evaluated the proposed approach in the context of demand forecasting for an industrialised bakery. The dataset comprises point-of-sales data over 18 months of 16 articles that are sold in six stores and two regions and articles can be grouped into two categories. The clustering approach in hierarchical forecasting seems to outperform traditional BU and TD approaches, based on forecast evaluations using MAPE and RMSE. Li and Lim (2018) proposed a greedy aggregation–decomposition approach to forecast intermittent demand in a hierarchical structure of a fashion retailer. The proposed approach utilises both forecasts at top and bottom aggregation levels, unlike the traditional BU and TD approaches. The performance of the approach was compared against popular intermittent demand forecasting including Croston, SBA, TSB as well as temporal aggregation approach such as MAPA and ADIDA, used in each level separately. Using a real database of the SKU-store-day demand over two years provided by a retailer, they showed that the

proposed method that combined information in the hierarchy outperformed other existing intermittent demand forecasting methods. The revised mean absolute scaled error (RMASE), MAE and MASE were used to evaluate forecast accuracy.

There is no consistent agreement among studies to determine the conditions under which each hierarchical reconciliation and or combination approach works better. The performance of these approaches generally depends on the characteristics of time series, forecasting horizon, level of aggregation and the structure of the hierarchy.

5.2. Temporal hierarchies and Multiple temporal aggregation

Similar to hierarchical and grouped time series in cross-sectional aggregation, one can also create coherent forecasts in temporal hierarchies, or benefit from obtaining different information at multiple levels of aggregation in non-overlapping temporal aggregation. The idea here is to exploit the information available at various levels of temporal aggregation instead of using only one single optimal temporal aggregation level (Rostami-Tabar et al. 2013, Kourentzes et al. 2017).

Combinations of forecasts at different levels of temporal aggregation were evaluated empirically by Moon et al. (2012). They tested direct methods, TD methods and combination methods on a sample of 300 items with lumpy demand patterns from the South Korean Navy. Monthly, yearly and quarterly aggregations were compared and it was found that, overall, the best year-ahead forecasting method was a simple (unweighted) combination of the forecast for quarterly aggregated data (adjusted for linear trend) at group level and a forecast of monthly aggregated data at the item level. This evaluation was conducted considering both forecast accuracy and inventory costs. Mean absolute deviation and RMSE were used to evaluate the forecasting performance.

Further, Moon et al. (2013) examined features leading to the outperformance of direct methods or methods based on group level time series. The performance was evaluated according to a measure based on absolute error to mean demand ratios. It was found that the correlation between demands for different items, the variability in demand volume, and the equipment group had the greatest influence on relative forecasting performance. A logistic regression classification model was found to be marginally superior to the method based on group level time series.

The idea of multiple temporal aggregation was also explored in a paper by Kourentzes et al. (2014). They proposed the Multiple Aggregation Prediction Algorithm (MAPA) which first constructs multiple time series from the original series using non-overlapping temporal aggregation, for example creating weekly and monthly series from daily series. Then, an appropriate state-space exponential smoothing (ETS) model is fitted to each series separately and its respective time series components are forecast. Next, the time series components from each aggregation level are combined to create the final forecast. The advantage of this approach is that it is not restricted to any assumption regarding the time series process. It also benefits from forecast combination (Blanc and Setzer 2016) and reduces the uncertainty in model selection. However, there are some limitations in using MAPA: i) the forecasting model is not flexible because it uses only ETS family of models, so this is the only forecasting approach available in the framework; ii) in time series with peaks in the seasonality (i.e. within day or within week peaks), the approach might be problematic because it shrinks seasonal indices. There are few studies that investigated the application of this approach and modelling with multiple temporal aggregation level in general in supply chains. Petropoulos and Kourentzes (2014) provide empirical evidence on intermittent demand using a large data set of spare parts demand. They showed that combination across forecasts generated from multiple non-overlapping temporally aggregated series using the same single forecasting method or multiple methods improves the forecasting performance. Five different errors measures including scaled Mean Error, scaled Absolute Error, scaled Squared Error, scaled Periods in Stock and scaled Absolute Periods in Stock were used to assess the performance of the approaches.

Kourentzes and Petropoulos (2016) extended the MAPA approach to include external variables such as promotions. They examined the performance of the proposed approach using historical demand time series of cider of a popular brand in the UK. They benchmarked against the extended exponential smoothing that includes external promotional data. Their results indicated that the proposed approach outperforms all benchmarks. (Scaled Mean Error and scaled Mean Absolute Error were used to report the forecasting performance of approaches.) Barrow and Kourentzes (2016) also compared MAPA with standard forecasting methods such as ETS, ARIMA and Theta using time series of sales

of products from a major UK fast moving consumer goods manufacturer. They indicated that the forecast resulting from MAPA outperforms others in terms of forecast accuracy and bias. They used scaled mean error (sME) and scaled median error (sMdE) to measure forecast bias, and scaled mean squared error (sMSE) and scaled median squared error (sMdSE) to measure the magnitude of forecast errors.

Petropoulos et al. (2019) empirically investigated the inventory performance of MAPA using the monthly industry series of the M3 competition. They indicated that multiple temporal aggregation not only improves the forecasting performance but also generates smoother forecasts which minimises the bullwhip effect and has the best trade-off curves for inventory costs versus service levels. (Lei et al. 2016) proposed a new algorithm that combines MAPA with fuzzy Markov chain model (FMC-MAPA). Material demand data from the STATE GRID Corporation of China was used to test the forecasting accuracy of the new approach by comparing it with the exponential smoothing (ES) and fuzzy Markov chain (FMC) benchmarks. The results showed that FMC-MAPA with, an equal weight dissaggregation method, outperformed the benchmarks. They indicated that forecasts generated from the combined method are more stable and robust than the ES and FMC models, separately.

Building on the idea of optimal reconciliation proposed by Hyndman et al. (2011) and multiple aggregation levels by Kourentzes et al. (2014), Athanasopoulos et al. (2017) proposed Temporal Hierarchies Forecasting (THieF). A temporal hierarchy is created considering high frequency time series at the bottom level (e.g. hourly time series) and lower frequency time series at higher levels(e.g. daily = 24 hours and weekly=7 days). It can be created for any time series using non-overlapping temporal aggregation. THieF combines forecasts from all levels of the hierarchy. This approach overcomes the limitation of MAPA. It allows for other forecasting models to be used and it does not shrink the seasonal indices arbitrarily. Given the fact that supply chain forecasting informs decisions at multiple level of time granularity (e.g. short term, mid-term, long-term), this approach was recommended to practitioners as forecasts are based on the same information about the future. Using monthly and quarterly series of the M3 competition and weekly data of AE departments, Athanasopoulos et al. (2017) showed that forecasting with temporal hierarchies can improve forecast accuracy significantly. The forecasts are evaluated using

the Relative Mean Absolute Error (RMAE) and MASE. THieF has also some limitations: i) when constructing new series from the original series, the frequency of the new time series cannot be a fraction, it must be an integer, ii) the approach also uses a linear combination of forecasts generated at all levels to create the reconciled forecasts. Therefore, it is sub optimal at each separate level of aggregation. Moreover, the relationship between final reconciled forecast and forecasts at each level might not be linear. We should also note that both approaches, i.e. MAPA and THieF, are based on non-overlapping temporal aggregation. These approaches cannot accommodate the use of overlapping temporal aggregation.

Building on the THieF framework proposed by Athanasopoulos et al. (2017), Kourentzes and Athanasopoulos (2021) extended the idea to forecast intermittent demand series in a temporal hierarchy focusing on forecast improvement at the disaggregate intermittent demand level. They demonstrated that the proposed approach brings significant gains for both point and quantile forecasts, through an empirical investigation using a dataset of aircraft spare parts. They evaluated the forecast accuracy using four metrics, the Mean Error (ME), RMSE, the Mean Interval Score (MIS) and the Pinball loss (PIN).

5.3. Cross-temporal hierarchies

In the previous sections, temporal and cross- sectional aggregation are used separately. This means that either: i) cross-sectional aggregation is considered only at one level of temporal granularity (e.g. monthly) or ii) multiple temporal granularities (e.g. hourly, daily, weekly etc) are used, assuming a single level in the cross-sectional structure (e.g. total). Using either of the approaches separately in a supply chain provides benefits but, in practice, multiple levels of temporal granularities across the entire cross-sectional structure are required. Generating forecasts for all levels of temporal and cross-sectional granularities first requires producing forecasts for each level of temporal aggregation considering the whole cross-sectional structure. This will need some post-processing to generate forecasts for the entire levels, which are also coherent (Kourentzes and Athanasopoulos 2019). This problem has been addressed by creating a framework to generate cross-temporally coherent forecasts that supports all levels of temporal and cross-sectional aggregation. This framework allows the generation of a single version of the forecast, which is critical for supply chains to align decisions across different horizons

and across different departments. The approach has not yet been examined using real data from supply chains, but Kourentzes and Athanasopoulos (2019) indicated that it provides further forecast accuracy gains to either cross-sectional reconciliation and its variations (Hyndman et al. 2011) or temporal hierarchies forecasting (Athanasopoulos et al. 2017). This was based on an anlaysis of tourism data, using the Average Relative Mean Squared Error (AvgRelMSE) to track the forecast accuracy. More specific to supply chains, Punia et al. (2020) proposed a cross-temporal forecasting approach that generates coherent forecasts for all levels of decision making for a retailer including products, time, and channel dimensions of supply chain forecasting. Additionally, the authors investigated the suitability of TD and BU approaches in the context of online and offline retail. They showed that forecasts from the proposed framework significantly improve forecast accuracy, compared to direct forecasts at all levels of retailer decision making. The following error metrics were used to examine the forecast accuracy: average relative mean absolute error (ARMAE), average relative mean squared error (ARMSE) and the average relative mean absolute percentage error (ARMAPE). A weekly dataset consisting of ten SKUs for more than two years was used for the empirical analysis, which is rather too limited for reliable conclusions to be drawn.

5.4. Gaps of research

The idea of using available information at various levels of cross-sectional and temporal aggregation to improve forecasting performance is promising and there have been multiple theoretical developments in this area that potentially could be very useful in supply chain forecasting: i) it can improve forecast accuracy ; ii) it can reduce the risk related to model selection and uncertainty and ii) it is better aligned with multiple levels of decision making, which is essential in modern supply chains. However, there are still important gaps in this area of research.

Despite the recent developments of hierarchical and temporal hierarchies, there is a need to examine empirically the validity of these theoretical developments and how they might benefit supply chains to make better and more aligned decisions with other parts of the network. In particular, the challenging problem of investigating the benefit of both temporal and cross-sectional hierarchies beyond forecast accuracy still remains a huge gap in the literature. In fact, forecasts are required at multiple levels of the hierarchy to

inform different type of decisions in finance, logistics, marketing or transportation planning, etc. These departments might have different conflicting objective functions. Moreover, the evaluation requires the knowledge of how these functions are implemented, as well as their relevant utility functions.

Another important question is how to evaluate the performance of the entire hierarchy. Currently, forecasting performance is evaluated and reported at each level separately and the forecasting performance is averaged to report the performance across the entire structure. It is desirable that forecasting performance metrics are introduced that would be able to measure the performance of each approach in the hierarchy as a whole, rather than focusing on the performance on the top or the bottom or the middle level.

Research on reconciliation (both temporal and cross sectional) and combination of multiple levels of aggregation shows that forecasts can be improved, but the conditions for this improvement remain unclear. Generally, the structure of the data and its features play a critical role in determining the conditions under which each approach should be recommended. Very little research has examined the association between data structure and characteristics of time series and the performance of approaches. This is the case for both continuous and discrete time series. The association of time series characteristics from all levels of the hierarchy in both cross-sectional and temporal hierarchies, and not only the original series, with the performance of approaches needs to be investigated.

Although the hierarchical models utilise the information available through the historical time series, this is not the case for potentially useful exogenous variables. The potential benefit of incorporating exogenous variables in a hierarchy structure still needs to be examined. There are various types of exogenous variables that might be useful in this case: i) variables that are independent of the aggregation level, ii) variables that could be aggregated in the same way as the time series data, and iii) variables that are unique to each level. Determining when incorporating such models provides benefits to the supply chain, as well as identifying what type of exogenous variables should be used, has important implications in practice.

Obviously, point forecasts do not provide information about the uncertainty in supply chain forecasting. However, most of the articles reviewed in this study report only point

forecast accuracy. Investigating uncertainty in forecasting hierarchies in terms of prediction intervals and/or probabilistic forecast is another important gap.

Finally, the theoretical developments in this area do not support the count nature of time series. Many forecasts ultimately will be used as inputs in count numbers to other function in supply chains. Therefore, extending hierarchical, temporal and cross-temporal reconciliation to account for count time series is of practical as well as theoretical importance.

6. Practical Implications

As previous sections of this review have indicated, the aggregation and hierarchical forecasting literature has evolved considerably over recent decades. Greater emphasis on aggregation and hierarchies in academic research has been reflected by developments in commercial software. Some examples will suffice to illustrate this trend. Demand Works has introduced a feature enabling users to work at any level of aggregation, without being constrained by predefined hierarchical structures. Relex software has extended forecast visibility to any level of granularity or aggregation. Logility includes facilities to disaggregate forecasts at higher levels down to the lowest levels. SAS and SAP have adopted the 'optimal combination approach' as an alternative to the traditional top-down and bottom-up methods. Finally, Blue Yonder allows forecasts to be generated at multiple levels of the hierarchy, enabling trends and seasonality at higher levels to drive forecasts at lower levels. A survey of forecasting software that includes more details on forecasting features, such as temporal aggregation and hierarchical forecasting, is provided in Fildes et al. (2018).

These developments now mean that practitioners have a wider choice of methods that they can implement in practice. They should certainly consider using recently developed hierarchical methods, given the encouraging empirical evidence on forecast accuracy. Cross-temporal methods should also be considered. However, the empirical evidence has focussed on a relatively small number of cross-sectional series (typically less than 100 at the bottom level of the hierarchy). As the number of series grows, hierarchical methods can become computationally expensive. If this problem becomes prohibitive, then benefits may still accrue from aggregation methods. Indeed, they should always be used

as a benchmark method, to ensure that the added complexity of hierarchical methods is worthwhile. There are no comprehensive rules for the outperformance of top-down or bottom-up approaches. Simulation comparisons are recommended, not only in terms of accuracy but also with regard to inventory performance. For inventory applications, temporal aggregations should be evaluated, using the lead time as the level of aggregation.

7. Conclusions

7.1. Main findings

The research reviewed in this paper has led to important developments, including making better use of the data available over different time and hierarchical levels. There have been significant advances in the analytical modelling of aggregation, as summarised in Sections 3 and 4, and in the understanding of hierarchical forecasting, as outlined in Section 5. We summarise in Table 3 the main findings that arise from the literature with regard to the contexts where temporal and cross-sectional aggregations are beneficial. It should be noted that most of these findings hold when the performance is considered at the aggregate level. Note also that some of these findings are supported more consistently in the literature than others as discussed earlier in the paper. The corresponding sections in the paper are also indicated in Table 3.

Furthermore, it has been pointed out in this paper as one of the findings that hierarchical forecasting does not generalise the aggregation approach. This is demonstrated by the fact that forecasting methods based on temporal aggregation with overlapping blocks is not integrated within the hierarchical and combinations approaches proposed in the literature.

7.2. Main research gaps

Overall, there has been significant progress in the forecasting literature in recent years, not only in analytical developments but also in empirical research. However, our literature review has identified several research gaps, which have been discussed in the concluding sub-sections of this paper. In this section, building on these gaps, we draw together the major research themes that provide an agenda for further research. In Section 5.1, we reviewed the results of the M5 forecasting competition (Makridakis et al.

	Temporal aggregation	Cross-sectional aggregation	
		Bottom-Up	Top-Down
Intermittence	High intermittence degree (Long demand intervals) and Low variability of demand sizes [Section 3.1.3]	Low variability of demand intervals of the subaggregate SKUs [Section 4.1.2]	High variability of demand intervals and demand sizes for subaggregate SKUs (especially for a high number of aggregated SKUs) [Section 4.1.2]
Seasonality		The coefficient of variation of the deseasonalised individual series is greater than the coefficient of variation of the deseasonalised aggregate series [Section 4.2]	
Correlation	Negatively auto-correlated demand [Section 3.1.2]	Negative cross-correlation between the subaggregate SKUs (or positive very low cross correlation) and Correlation parameters of the subaggregate SKUs are with different signs [Section 4.1.2]	High positive cross-correlation between the subaggregate SKUs and Negative auto-correlation of the subaggregate SKUs [Section 4.1.2]

Table 3: Contexts where aggregation is beneficial: main findings from the literature

2020). The M5 competition included grouped time series, thereby allowing assessment of methods based on aggregation and hierarchies through the supply chain. This competition was important because it went beyond the evaluation of point forecasts, to include prediction intervals as well. These intervals are crucial for many supply chain applications and it is important that future studies also take interval forecasts into account.

Research on hierarchical forecasting and combinations has mainly focused on statistical forecasting methods and, more precisely, extrapolative techniques. Further research is required to integrate other forecasting approaches. Judgmental forecasting is used extensively in demand forecasting in supply chains but, except for the work by Kremer et al. (2016), no research has looked at judgmental forecasting when dealing with aggregation and hierarchies. Furthermore, the M5 competition (Makridakis et al. 2021, 2020) has provided evidence on the potential benefit of using machine learning techniques and incorporating exogenous variables when forecasting hierarchies and combining forecasts. Hence, further research into this would appear to be merited.

From the theoretical perspective, the INARMA demand process modelling has been shown to provide a good framework to deal with intermittent demand within supply chains and count series in general (as discussed in Section 3.1). So far, this modelling framework has been employed for temporal aggregation but it would also be an interesting framework to develop a theoretical basis for the research on hierarchical forecasts and combinations.

7.3. Open debates

In the hierarchical literature, there is an emphasis on strict coherence, so that there are no discrepancies in the sum of forecasts at a lower level matching the corresponding forecast at a higher level. The emphasis is based on the assumption that strict coherence of forecasts will contribute towards coherence in decision making. This is an issue that warrants further debate. Is strict coherence necessary for enhanced decision making? Could a weaker definition of coherence lead to any benefits in forecasting or supply chain performance?

The impact of judgemental forecasting has already been mentioned as a research gap. Empirical and laboratory-based research on judgement in forecasting is well established. Questions of how such research should be conducted, based on judgmental amendments to aggregation or hierarchical forecasts, have not yet been debated.

Another issue that is unresolved is the assessment of forecasting accuracy across the whole supply chain. Accuracy measures for a single series, or for multiple series at the same level, have been discussed extensively. Measuring accuracy across the chain has received less attention. From an accuracy-implication perspective, the total inventory cost (across all levels) and the inventory service (at the customer level) would seem to be paramount. So, how should accuracy measures be weighted across the different levels?

In conclusion, aggregation and hierarchical forecasting have seen major advances in recent years. These advances have not been confined to the pages of academic journals. They have been made available to practitioners through the availability of commercial and open-source software. It is hoped that this review will stimulate wider use of the methods surveyed in this paper, debate on the questions raised above, and further research on the remaining gaps in our understanding of this important subject.

Data Availability Statement (DAS)

The data related to the final list of papers (results of the systematic literature review) can be freely available by the authors upon request.

References

Abolghasemi, M., Hyndman, R. J., Tarr, G. and Bergmeir, C. (2019), 'Machine learning applications in time series hierarchical forecasting', *arXiv preprint arXiv:1912.00370* .

Abolghasemi, M., Hyndman, R., Spiliotis, E. and Bergmeir, C. (2020), 'Model selection in reconciling hierarchical time series', *arXiv preprint: 2010.10742* .

Altay, N., Litteral, L. and Rudisill, F. (2012), 'Effects of correlation on intermittent demand forecasting and stock control', *International Journal of Production Economics* **135**, 275–283.

Amemiya, T. and Wu, R. Y. (1972), 'The effect of aggregation on prediction in the autoregressive model', *Journal of the American Statistical Association* **67**(339), 628–632.

Athanasopoulos, G., Hyndman, R. J., Kourentzes, N. and Petropoulos, F. (2017), 'Forecasting with temporal hierarchies', *European Journal of Operational Research* **262**(1), 60–74.

Babai, M. Z., Ali, M. M. and Nikolopoulos, K. (2012), 'Impact of temporal aggregation on stock control performance of intermittent demand estimators: Empirical analysis', *Omega* **40**(6), 713–721.

Babai, M. Z., Tsadiras, A. and Papadopoulos, C. (2020), *IMA Journal of Management Mathematics* **31**, 281–305.

Barrow, D. K. and Kourentzes, N. (2016), 'Distributions of forecasting errors of forecast combinations: implications for inventory management', *International Journal of Production Economics* **177**, 24–33.

Blanc, S. M. and Setzer, T. (2016), 'When to choose the simple average in forecast combination', *Journal of Business Research* **69**(10), 3951–3962.

Bookbinder, J. H. and Lordahl, A. E. (1989), 'Estimation of inventory re-order levels using the bootstrap statistical procedure', *IIE Transactions* **21**, 302–312.

Boylan, J. E. and Babai, M. Z. (2016), 'On the performance of overlapping and non-overlapping temporal demand aggregation approaches', *International Journal of Production Economics* **181**, 136–144.

Boylan, J. E., Chen, H., Mohammadipour, M. and Syntetos, A. A. (2014), 'Formation of seasonal groups and application of seasonal indices', *Journal of the Operational Research Society* **65**, 227–241.

Brewer, K. R. (1973), 'Some consequences of temporal aggregation and systematic sampling for arma and armax models', *Journal of Econometrics* **1**(2), 133–154.

Carlstein, E. (1986), 'The use of subseries methods for estimating the variance of a general statistic from a stationary time series', *Annals of Statistics* **14**, 1711–1719.

Chen, A. and Blue, J. (2010), 'Performance analysis of demand planning approaches for aggregating, forecasting and disaggregating interrelated demands', *International Journal of Production Economics* **128**(2), 586–602.

Chen, A., Hsu, C.-H. and Blue, J. (2007), 'Demand planning approaches to aggregating and forecasting interrelated demands for safety stock and backup capacity planning', *International journal of production Research* **45**(10), 2269–2294.

Chen, H. and Boylan, J. E. (2007), 'Use of individual and group seasonal indices in subaggregate demand forecasting', *Journal of the Operational Research Society* **58**, 1660–1671.

Dalhart, G. (1974), Class seasonality - a new approach, *in* 'American Production and Inventory Control Society COnference Proceedings', APICS.

Dangerfield, B. J. and Morris, J. S. (1992), 'Top-down or bottom-up: Aggregate versus disaggregate extrapolations', *International journal of forecasting* **8**(2), 233–241.

Dekker, M., Van Donselaar, K. and Ouwehand, P. (2004), 'How to use aggregation and combined forecasting to improve seasonal demand forecasts', *International Journal of Production Economics* **90**, 151–167.

Edwards, J. B. and Orcutt, G. H. (1969), 'Should aggregation prior to estimation be the rule?', *The Review of Economics and Statistics* pp. 409–420.

Efron, B. (1979), 'Bootstrap methods. another look at the jackknife.', *Annals of Statistics* **7**, 1–26.

Fildes, R., Ma, S. and Kolassa, S. (2019), 'Retail forecasting: Research and practice', *International Journal of Forecasting* .

Fildes, R., Schaer, O. and Svetunkov, I. (2018), 'Software survey: Forecasting 2018', *INFORMS ORMS-Today* **45**(3), 47–51.

Fliedner, G. (1999), 'An investigation of aggregate variable time series forecast strategies with specific subaggregate time series statistical correlation', *Computers & operations research* **26**(10-11), 1133–1149.

Fricker, R. and Goodhart, C. (2000), 'Applying a bootstrap approach for setting reorder points in military supply systems', *Naval Research Logistics* **47**, 459–478.

Fu, W. and Chien, C.-F. (2019), 'Unison data-driven intermittent demand forecast framework to empower supply chain resilience and an empirical study in electronics distribution', *Computers & Industrial Engineering* **135**, 940–949.

Granger, C. W. J. and Morris, M. J. (1976), 'Time series modelling and interpretation', *Journal of the Royal Statistical Society: Series A (General)* **139**(2), 246–257.

Gross, C. W. and Sohl, J. E. (1990), 'Disaggregation methods to expedite product line forecasting', *Journal of forecasting* **9**(3), 233–254.

Grunfeld, Y. and Griliches, Z. (1960), 'Is aggregation necessarily bad?', *The Review of Economics and Statistics* pp. 1–13.

Hall, P. (1985), 'Resampling a coverage pattern', *Stochastic Processes and their Applications* **20**, 231–246.

Harvey, A. C. (1993), 'Time series models'.

Harwell, J. (2015), *Sales and operations planning in the retail industry*, In M. Gilliland, L. Tashman, U. Sglavo (Eds.), Business forecasting: Practical problems and solutions (pp. 363–372). New Jersey.

Hasni, M., Aguir, M., Babai, M. and Jemai, Z. (2019), 'On the performance of adjusted bootstrapping methods for intermittent demand forecasting', *International Journal of Production Economics* **216**, 145–153.

Hasni, M., Babai, M. Z., Aguir, M. and Jemai, Z. (2019), 'An investigation on bootstrapping forecasting methods for intermittent demands', *International Journal of Production Economics* **209**, 20–29.

Hotta, L. K., Morettin, P. A. and Pereira, P. L. V. (1992), 'The effect of overlapping aggregation on time series models: an application to the unemployment rate in brazil', *Brazilian Review of Econometrics* **12**(2), 223–241.

Huber, J., Gossmann, A. and Stuckenschmidt, H. (2017), 'Cluster-based hierarchical demand forecasting for perishable goods', *Expert systems with applications* **76**, 140–151.

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

Hyndman, R. J. and Athanasopoulos, G. (2021), *Forecasting: principles and practice*, 3rd edition, OTexts: Melbourne, Australia.
URL: *OTexts.com/fpp3*

Hyndman, R. J., Lee, A. J. and Wang, E. (2016), 'Fast computation of reconciled forecasts for hierarchical and grouped time series', *Computational statistics & data analysis* **97**, 16–32.

Hyndman, R. and Kostenko, A. (2007), 'Minimum sample size requirements for seasonal forecasting models', *Foresight: the International Journal of Applied Forecasting* **6**, 12–15.

Jeon, J., Panagiotelis, A. and Petropoulos, F. (2019), 'Probabilistic forecast reconciliation with applications to wind power and electric load', *European Journal of Operational Research* **279**(2), 364–379.

Jin, Y. Williams, B. D., Tokar, T. and Waller, M. A. (2015), 'Forecasting with temporally aggregated demand signals in a retail supply chain', *Journal of Business Logistics* **36**(2), 199–211.

Kolassa, S. (2016), 'Evaluating predictive count distributions in retail sales forecasting', *International Journal of Forecasting* **32**, 788–803.

Kourentzes, N. and Athanasopoulos, G. (2019), 'Cross-temporal coherent forecasts for australian tourism', *Annals of Tourism Research* **75**, 393–409.

Kourentzes, N. and Athanasopoulos, G. (2021), 'Elucidate structure in intermittent demand series', *European Journal of Operational Research* **288**(1), 141–152.

Kourentzes, N. and Petropoulos, F. (2016), 'Forecasting with multivariate temporal aggregation: The case of promotional modelling', *International Journal of Production Economics* **181**, 145–153.

Kourentzes, N., Petropoulos, F. and Trapero, J. R. (2014), 'Improving forecasting by estimating time series structural components across multiple frequencies', *International Journal of Forecasting* **30**(2), 291–302.

Kourentzes, N., Rostami-Tabar, B. and Barrow, D. K. (2017), 'Demand forecasting by temporal aggregation: using optimal or multiple aggregation levels?', *Journal of Business Research* **78**, 1–9.

Kremer, M., Siemsen, E. and Thomas, D. J. (2016), 'The sum and its parts: Judgmental hierarchical forecasting', *Management Science* **62**(9), 2745–2764.

Künsch, H. (1989), 'The jackknife and the bootstrap for general stationary observations', *Annals of Statistics* **17**, 1217–1241.

Lapide, L. (2016), 'Retail omnichannel needs better forecasting & planning', *The Journal of Business Forecasting* **35**(3), 12.

Lei, M., Li, S. and Tan, Q. (2016), 'Intermittent demand forecasting with fuzzy markov chain and multi aggregation prediction algorithm', *Journal of Intelligent & Fuzzy Systems* **31**(6), 2911–2918.

Li, C. and Lim, A. (2018), 'A greedy aggregation–decomposition method for intermittent demand forecasting in fashion retailing', *European Journal of Operational Research* **269**(3), 860–869.

Lordahl, A. E. and Bookbinder, J. H. (1994), 'Order-statistic calculation, costs, and service in an (s,Q) inventory system', *Naval Research Logistics* **41**, 81–97.

Lütkepohl, H. (2011), 'Forecasting aggregated time series variables: A survey', *OECD Journal: Journal of Business Cycle Measurement and Analysis* **2010**(2), 1–26.

Makridakis, S., Chatfield, C., Hibon, M., Lawrence, M., Mills, T., Ord, K. and Simmons, L. F. (1993), 'The M2-competition: A real-time judgmentally based forecasting study', *International Journal of Forecasting* **9**(1), 5–22.

Makridakis, S., Spiliotis, E. and Assimakopoulos, V. (2020), 'The M5 accuracy competition: Results, findings and conclusions', *Working paper, https://bit.ly/3iKWwj9* .

Makridakis, S., Spiliotis, E., Assimakopoulos, V., Chen, Z., Gaba, A., Tsetlin, I. and Winkler, R. (2021), 'The M5 uncertainty competition: Results, findings and conclusions', *Working paper, https://bit.ly/3ruWT5F* .

Meza-Peralta, K., Gonzalez-Feliu, J., Montoya-Torres, J. R. and Khodadad-Saryazdi, A. (2020), 'A unified typology of urban logistics spaces as interfaces for freight transport: A systematic literature review', *Supply Chain Forum: An International Journal* **21**(4), 274–289.

Mircetic, D., Rostami-Tabar, B., Nikolicic, S. and Maslaric, M. (2021), 'Forecasting hierarchical time series in supply chains: an empirical investigation', *International Journal of Production Research* pp. 1–20.

Mohammadipour, M. and Boylan, J. E. (2012), 'Forecast horizon aggregation in integer autoregressive moving average (inarma) models', *Omega* **40**(6), 703–712.

Moon, S., Hicks, C. and Simpson, A. (2012), 'The development of a novel hierarchical forecasting method for predicting spare parts demand in the south korean navy - a case study', *International Journal of Production Economics* **140**, 794–802.

Moon, S., Simpson, A. and Hicks, C. (2013), 'The development of a classification model for predicting the performance of forecasting methods for naval spare parts demand', *International Journal of Production Economics* **143**, 449–453.

Murray, P. W., Agard, B. and Barajas, M. A. (2018*a*), 'Asact-data preparation for forecasting: A method to substitute transaction data for unavailable product consumption data', *International Journal of Production Economics* **203**, 264–275.

Murray, P. W., Agard, B. and Barajas, M. A. (2018*b*), 'Forecast of individual customer's demand from a large and noisy dataset', *Computers & industrial engineering* **118**, 33–43.

Narayanan, P., Verhagen, W. J. and Dhanisetty, V. V. (2019), 'Identifying strategic maintenance capacity for accidental damage occurrence in aircraft operations', *Journal of Management Analytics* **6**(1), 30–48.

Nenova, Z. D. and May, J. H. (2016), 'Determining an optimal hierarchical forecasting model based on the characteristics of the data set, technical note', *Journal of Operations Management* **44**, 62–68.

Nikolopoulos, K., Syntetos, A. A., Boylan, J. E., Petropoulos, F. and Assimakopoulos, V. (2011), 'An aggregate–disaggregate intermittent demand approach (adida) to forecasting: an empirical proposition and analysis', *Journal of the Operational Research Society* **62**(3), 544–554.

Orcutt, G. H., Watts, H. W. and Edwards, J. B. (1968), 'Data aggregation and information loss', *The American Economic Review* **58**(4), 773–787.

Ouwehand, P., Van Donselaar, K. H. and de Kok, A. (2005), The impact of forecasting horizon when forecasting with group seasonal indices, Technical report.

Panagiotelis, A., Athanasopoulos, G., Gamakumara, P. and Hyndman, R. J. (2021), 'Forecast reconciliation: A geometric view with new insights on bias correction', *International Journal of Forecasting* **37**(1), 343–359.

Panagiotelis, A., Gamakumara, P., Athanasopoulos, G., Hyndman, R. et al. (2020), Probabilistic forecast reconciliation: Properties, evaluation and score optimisation, Technical report, Monash University, Department of Econometrics and Business Statistics.

Park, D. and Willemain, T. R. (1999), 'The threshold bootstrap and threshold jackknife', *Computational statistics & data analysis* **31**(2), 187–202.

Pennings, C. L. and Van Dalen, J. (2017), 'Integrated hierarchical forecasting', *European Journal of Operational Research* **263**(2), 412–418.

Pennings, C., van Dalen, J. and van der Laan, E. (2017), 'Exploiting elapsed time for managing intermittent demand for spare parts', *European Journal of Operational Research* **258**, 958–969.

Petropoulos, F. and Kourentzes, N. (2014), 'Forecast combinations for intermittent demand', *Journal of the Operational Research Society* **66**(6), 914–924.

Petropoulos, F., Kourentzes, N. and Nikolopoulos, K. (2016), 'Another look at estimators for intermittent demand', *International Journal of Production Economics* **181**, 154–161.

Petropoulos, F., Wang, X. and Disney, S. M. (2019), 'The inventory performance of forecasting methods: Evidence from the m3 competition data', *International Journal of Forecasting* **35**(1), 251–265.

Porras, E. and Dekker, R. (2008), 'An inventory control system for spare parts at a refinery: An empirical comparison of different re-order point methods', *European Journal of Operational Research* **184**(1), 101–132.

Punia, S., Singh, S. P. and Madaan, J. K. (2020), 'A cross-temporal hierarchical framework and deep learning for supply chain forecasting', *Computers & Industrial Engineering* **149**, 106796.

Quenouille, M. (1958), 'Discrete autoregressive schemes with varying time-intervals', *Metrika* **1**(1), 21–27.

Ray, W. (1980), 'The significance of correlated demands and variable lead times for stock control policies', *Journal of the Operational Research Society* **31**, 187–190.

Razi, M., Kurtulus, I. and Smith, C. (2004), 'Development and evaluation of an inventory model for low-demand spare parts', *International Journal of Industrial Engineering-Theory Applications and Practice* **11**(1), 90–98.

Rego, J. and Mesquita, M. (2015), 'Demand forecasting and inventory control: A simulation study on automotive spare parts', *International Journal of Production Economics* **161**, 1–16.

Rostami-Tabar, B., Babai, M. and Syntetos, A. (2021), 'To aggregate or not to aggregate: Forecasting of finite autocorrelated demand', *Journal of Economic Surveys-arXiv: 2103.16310* .

Rostami-Tabar, B., Babai, M. Z., Ducq, Y. and Syntetos, A. (2015), 'Non-stationary demand forecasting by cross-sectional aggregation', *International Journal of Production Economics* **170**, 297–309.

Rostami-Tabar, B., Babai, M. Z., Syntetos, A. and Ducq, Y. (2013), 'Demand forecasting by temporal aggregation', *Naval Research Logistics (NRL)* **60**(6), 479–498.

Rostami-Tabar, B., Babai, M. Z., Syntetos, A. and Ducq, Y. (2014), 'A note on the forecast performance of temporal aggregation', *Naval Research Logistics (NRL)* **61**(7), 489–500.

Sbrana, G. and Silvestrini, A. (2013), 'Forecasting aggregate demand: analytical comparison of top-down and bottom-up approaches in a multivariate exponential smoothing framework', *International Journal of Production Economics* **146**(1), 185–198.

Schwarzkopf, A. B., Tersine, R. J. and Morris, J. S. (1988), 'Top-down versus bottom-up forecasting strategies', *The International Journal Of Production Research* **26**(11), 1833–1843.

Shlifer, E. and Wolff, R. W. (1979), 'Aggregation and proration in forecasting', *Management Science* **25**(6), 594–603.

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

Smith, M. and Babai, M. Z. (2011), A review of bootstrapping for spare parts forecasting, *in* N. Altay and L. Litteral, eds, 'Service parts management: demand forecasting and inventory control', Springer, London, chapter 6, pp. 125–141.

Spiliotis, E., Abolghasemi, M., Hyndman, R. J., Petropoulos, F. and Assimakopoulos, V. (2020), 'Hierarchical forecast reconciliation with machine learning', *arXiv preprint arXiv:2006.02043* .

Spithourakis, G. P., Petropoulos, F., Babai, M. Z., Nikolopoulos, K. and Assimakopoulos, V. (2011), Improving the performance of popular supply chain forecasting techniques, *in* 'Supply Chain Forum: an international journal', Vol. 12, Taylor & Francis, pp. 16–25.

Spithourakis, G. P., Petropoulos, F., Nikolopoulos, K. and Assimakopoulos, V. (2014), 'A systemic view of the adida framework', *IMA Journal of Management Mathematics* **25**(2), 125–137.

Strijbosch, L., Heuts, R. M. and Moors, J. J. (2008), 'Hierarchical estimation as a basis for hierarchical forecasting', *IMA Journal of Management Mathematics* **19**(2), 193–205.

Syntetos, A. A., Babai, M. Z. and Gardner, E. (2015), 'Forecasting intermittent inventory demands: Simple parametric methods vs. bootstrapping', *Journal of Business Research* **68**, 1746–1752.

Syntetos, A. A., Babai, Z., Boylan, J. E., Kolassa, S. and Nikolopoulos, K. (2016), 'Supply chain forecasting: Theory, practice, their gap and the future', *European Journal of Operational Research* **252**(1), 1–26.

Syntetos, A. A. and Boylan, J. E. (2005), 'The accuracy of intermittent demand estimates', *International Journal of Forecasting* **21**, 303–314.

Syntetos, A. A., Boylan, J. E. and Croston, J. D. (2005), 'On the categorization of demand patterns', *Journal of the Operational Research Society* **56**, 495–503.

Syntetos, A. A., Boylan, J. E. and Disney, S. M. (2009), 'Forecasting for inventory planning: a 50-year review', *Journal of the Operational Research Society* **60**(sup1), S149–S160.

Taieb, S. B., Taylor, J. W. and Hyndman, R. J. (2017), Coherent probabilistic forecasts for hierarchical time series, *in* 'International Conference on Machine Learning', PMLR, pp. 3348–3357.

Taieb, S. B., Taylor, J. W. and Hyndman, R. J. (2020), 'Hierarchical probabilistic forecasting of electricity demand with smart meter data', *Journal of the American Statistical Association* pp. 1–17.

Teunter, R. and Duncan, L. (2009), 'Forecasting intermittent demand: A comparative study', *Journal of the Operational Research Society* **60**, 321–329.

Theil, H. (1954), 'Linear aggregation of economic relations'.

Tukey, J. (1958), 'Bias and confidence in not-quite large samples. (abstract)', *Annals of Mathematical Statistics* **29**, 614.

Van Wingerden, E., Basten, R., Dekker, R. and Rustenberg, W. (2014), 'More grip on inventory control through improved forecasting: A comparative study at three companies', *International Journal of Production Economics* **157**, 220–237.

Verstaete, G., Aghezzaf, E.-H. and Desmet, B. (2019), 'A data-driven framework for predicting weather impact on high-volume low-margin retail products', *Journal of Retail and Consumer Services* **48**, 169–177.

Villegas, M. A., Pedregal, D. J. and Trapero, J. R. (2018), 'A support vector machine for model selection in demand forecasting applications', *Computers & industrial engineering* **121**, 1–7.

Viswanathan, S., Widiarta, H. and Piplani, R. (2008), 'Forecasting aggregate time series with intermittent subaggregate components: top-down versus bottom-up forecasting', *IMA Journal of Management Mathematics* **19**(3), 275–287.

Wang, M.-C. and Rao, S. (1992), 'Estimating reorder points and other management science applications by bootstrap procedure', *European Journal of Operational Research* **56**, 332–342.

Wang, X. and Disney, S. M. (2016), 'The bullwhip effect: Progress, trends and directions', *European Journal of Operational Research* **250**(3), 691–701.

Weatherford, L. R., Kimes, S. E. and Scott, D. A. (2001), 'Forecasting for hotel revenue management: Testing aggregation against disaggregation', *Cornell hotel and restaurant administration quarterly* **42**(4), 53–64.

Wei, W. W. (1978), Some consequences of temporal aggregation in seasonal time series models, *in* 'Seasonal analysis of economic time series', NBER, pp. 433–448.

Weiss, A. A. (1984), 'Systematic sampling and temporal aggregation in time series models', *Journal of Econometrics* **26**(3), 271–281.

Wickramasuriya, S. L., Athanasopoulos, G. and Hyndman, R. J. (2019), 'Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization', *Journal of the American Statistical Association* **114**(526), 804–819.

Widiarta, H., Viswanathan, S. and Piplani, R. (2007), 'On the effectiveness of top-down strategy for forecasting autoregressive demands', *Naval Research Logistics (NRL)* **54**(2), 176–188.

Widiarta, H., Viswanathan, S. and Piplani, R. (2008), 'Forecasting item-level demands: an analytical evaluation of top–down versus bottom–up forecasting in a production-planning framework', *IMA Journal of Management Mathematics* **19**(2), 207–218.

Widiarta, H., Viswanathan, S. and Piplani, R. (2009), 'Forecasting aggregate demand: An analytical evaluation of top-down versus bottom-up forecasting in a production planning framework', *International Journal of Production Economics* **118**(1), 87–94.

Willemain, T. and Smart, C. (2001), 'System and method for forecasting intermittent demand'.

Willemain, T., Smart, C., Schocker, J. and DeSautels, P. (1994), 'Forecasting intermittent demand in manufacturing', *International Journal of Forecasting* **10**, 529–538.

Willemain, T., Smart, C. and Schwarz, H. (2004), 'A new approach of forecasting intermittent demand for service parts inventories', *International Journal of Forecasting* **20**, 375–387.

Withycombe, R. (1989), 'Forecasting with combined seasonal indexes', *International Journal of Forecasting* **5**, 547–552.

Zellner, A. and Tobias, J. (1998), A note on aggregation, disaggregation and forecasting performance, Technical report.

Zhou, C. and Viswanathan, S. (2011), 'Comparison of a new bootstrapping method with parametric approaches for safety stock determination in service parts inventory systems', *International Journal of Production Economics* **133**, 481–485.

Zhou, S., Jackson, P., Roundy, R. O. and Zhang, R. Q. (2007), 'The evolution of family level sales forecasts into product level forecasts: Modeling and estimation', *IIE Transactions* **39**(9), 831–843.

Zhu, S., Dekker, R., Van Jaarsveld, W., Renjie, R. W. and Koning, A. J. (2017), 'An improved method for forecasting spare parts demand using extreme value theory', *European Journal of Operational Research* **261**(1), 169–181.

Zotteri, G. and Kalchschmidt, M. (2007), 'A model for selecting the appropriate level of aggregation in forecasting processes', *International Journal of Production Economics* **108**(1-2), 74–83.

Zotteri, G., Kalchschmidt, M. and Caniato, F. (2005), 'The impact of aggregation level on forecasting performance', *International Journal of Production Economics* **93**, 479–491.