International Journal of Forecasting 40 (2024) 430–456

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Forecast reconciliation: A review

George Athanasopoulos [a,*], Rob J. Hyndman [b], Nikolaos Kourentzes [c], Anastasios Panagiotelis [d]

[a] *Monash University, VIC 3145, Australia*
[b] *Monash University, VIC 3800, Australia*
[c] *University of Skövde, 541 28 Skövde, Sweden*
[d] *The University of Sydney, NSW 2006, Australia*

ARTICLE INFO

Keywords:
Aggregation
Coherence
Cross-temporal
Hierarchical time series
Grouped time series
Temporal aggregation

ABSTRACT

Collections of time series formed via aggregation are prevalent in many fields. These are commonly referred to as hierarchical time series and may be constructed cross-sectionally across different variables, temporally by aggregating a single series at different frequencies, or even generalised beyond aggregation as time series that respect linear constraints. When forecasting such time series, a desirable condition is for forecasts to be coherent: to respect the constraints. The past decades have seen substantial growth in this field with the development of reconciliation methods that ensure coherent forecasts and improve forecast accuracy. This paper serves as a comprehensive review of forecast reconciliation and an entry point for researchers and practitioners dealing with hierarchical time series. The scope of the article includes perspectives on forecast reconciliation from machine learning, Bayesian statistics and probabilistic forecasting, as well as applications in economics, energy, tourism, retail demand and demography.

1. Introduction

In time series forecasting, aggregation occurs in a variety of settings. For example, regional-level tourism demand aggregates to national tourism demand; total revenue from the sale of individual stock keeping units aggregates to total revenue from all stock keeping units; the Gross Domestic Product of an economy is an aggregate of individual components; time series data measured at a quarterly frequency can be aggregated to data at an annual frequency. While hierarchical time series will be defined more formally in Section 2, the notion of hierarchical forecasting can be understood via the simple example where there is a time series X, a time series Y and a time series $Z = X + Y$, and we are interested in forecasts of X, Y and Z.

In practice, it is important to acknowledge that our variables X, Y and Z may be forecast in isolation from one another. This may occur when each forecast is obtained using a different time series model or when forecasts are produced by separate organisational silos (see Chase, 2013). In such cases, it will typically be the case that adding the forecast of X to the forecast of Y will not be equal to the forecast of Z. Indeed, even where forecasts for X, Y and Z are produced jointly, it is not typically the case that forecasts aggregate in the correct fashion[1]. This leads to two fundamental questions facing the forecaster of hierarchical time series:

Question 1: How best to adjust forecasts to agree with the known aggregation structure?

* Corresponding author.
E-mail addresses: george.athanasopoulos@monash.edu (G. Athanasopoulos), rob.hyndman@monash.edu (R.J. Hyndman), nikolaos@kourentzes.com (N. Kourentzes), anastasios.panagiotelis@sydney.edu.au (A. Panagiotelis).

[1] Some rare and necessarily restrictive exceptions are discussed in Section 3.9.

https://doi.org/10.1016/j.ijforecast.2023.10.010

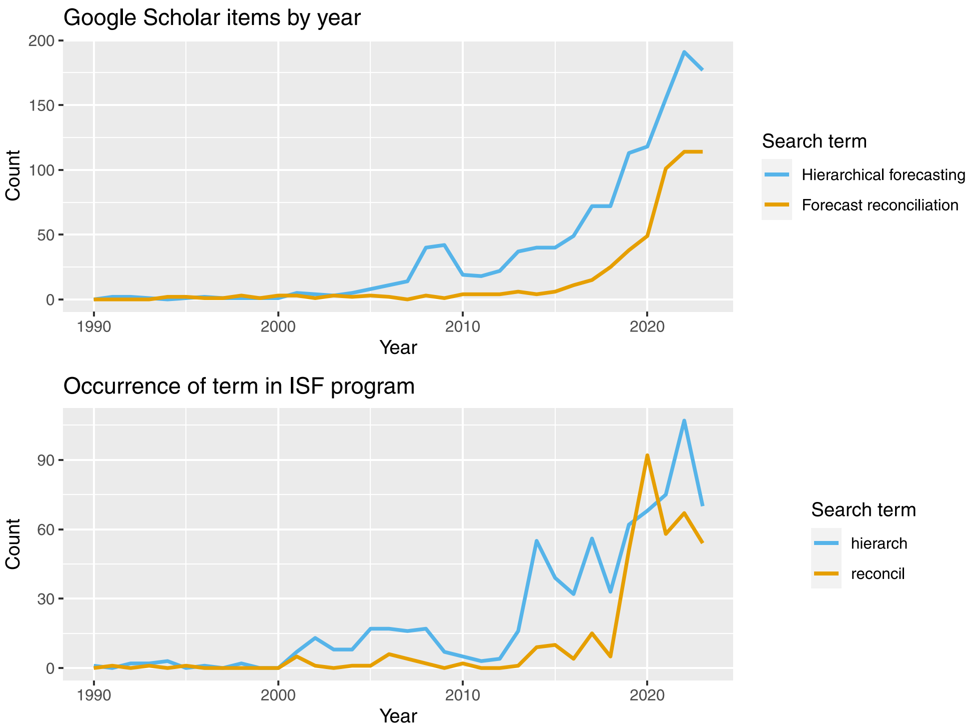

Fig. 1. Search term results in Google Scholar and the book of abstracts for the International Symposium on Forecasting during ISFs (1990–2023). The Google Scholar search is as of October 13, 2023.

and

Question 2: Does adjusting forecasts in this manner lead to improvements in forecast accuracy?

These questions have motivated a growing and fruitful area of research, particularly over the past decade. The top panel of Fig. 1 shows the growth in Google Scholar items by the search terms “Hierarchical forecasting” and “Forecast Reconciliation” (the latter to be defined in Section 3). The bottom panel tracks the occurrence of the terms “hierarch*“[2] and ”reconcil” in the book of abstracts of the International Symposium of Forecasters, the leading conference on forecasting. Both measures, while crude, pick up on the growing interest in the topic, especially in academic circles. Further, the reference list of this paper will attest to the multidisciplinary nature of the field, with breakthroughs in hierarchical forecasting being published in top-tier journals in statistics, econometrics, operations research and machine learning.

The impact of methods for forecasting hierarchical time series has not been limited to academia, with industry also showing a strong interest. Many organisations use modern hierarchical forecast methods in practice, including Amazon, the International Monetary Fund, the Bank of New York Mellon, IBM, Huawei, H&M, and Volkswagen. Methods have been implemented in leading analytics software platforms such as SAP, SAS, ForecastPro and Fiddlehead technologies, not to mention numerous open-source packages in R and Python (see Section 7). Among the broader forecasting community, including academics and practitioners, hierarchical data have featured as part of the M5 competition (Makridakis et al., 2022; Seaman & Bowman, 2022) and the Global Energy Forecasting Competition (Hong et al., 2019).

The growth and impact of hierarchical forecasting make a review paper timely. Throughout, our focus will be on forecasting, although where there are similarities between hierarchical forecasting methods and other literature, we will discuss seminal papers (for example, in Section 3.1). Methods for dealing with hierarchical time series often involve first generating forecasts of all series in the hierarchy. This review will not focus on the models and methods used to obtain these original forecasts. Also, while we will focus in Section 6 on some specific application areas where hierarchical forecasting methods have been used extensively, the methods have been applied so widely that not every applied paper can lie within the scope of this review, and instead our focus is on papers that make important methodological contributions.

The remainder of the paper is organised as follows. Section 2 provides the basic setting for hierarchical forecasting, introducing notation and terminology and covering important historical background. Section 3 covers forecast reconciliation, the hierarchical forecasting method that has garnered the most attention over the past decade. Section 4 covers the special case of temporal aggregation of a single time series, which (despite a separate historical development) has since adopted methods from (and now merged with) cross-sectional aggregation.

[2] We acknowledge that this search term may pick up instances of the word hierarchical related to hierarchical models rather than hierarchical time series, although the use of hierarchical models in a conference primarily on time series is rare.

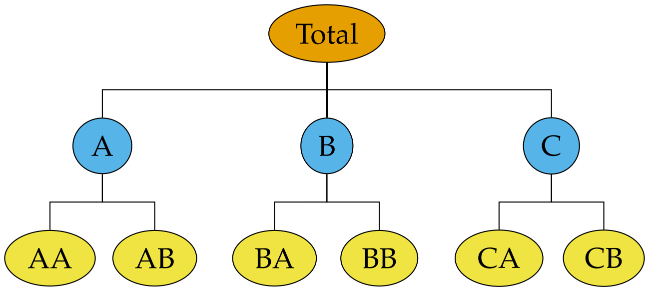

Fig. 2. The diagram shows a 2-level hierarchical tree structure.

Section 5 covers approaches to probabilistic forecasting, an area that, while previously ignored, has in recent years seen some important breakthroughs. Significant tourism, macroeconomics, energy, demography, retail and healthcare applications are covered in Section 6. Finally, in Section 8, we look to the future of the field and point to some open questions in hierarchical forecasting.

2. The setting

2.1. Hierarchical and grouped time series

We define a hierarchical time series as a multivariate time series, $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T$, that adheres to some known linear constraints.[3] For example, Fig. 2 shows a 2-level hierarchical structure. Let $y_{\text{Tot},t}$ be the total (level 0) of all series at time t; and let $y_{i,t}$ be the value of the time series at node i and time t. Let $\boldsymbol{y}_t \in \mathbb{R}^n$ be a vector comprising observations at time t of all time series in the hierarchy, and $\boldsymbol{b}_t \in \mathbb{R}^{n_b}$ ($n_b < n$) be a vector comprising the observations at time t of only the most disaggregated *bottom-level* series. The remaining $n_a = n - n_b$ aggregated series can be written as

$$\boldsymbol{a}_t = \boldsymbol{A}\boldsymbol{b}_t,$$

for an appropriate $n_a \times n_b$ *aggregation matrix* $\boldsymbol{A}$, and the full set of time series can be written for all t as

$$\boldsymbol{y}_t = \boldsymbol{S}\boldsymbol{b}_t,$$

where $\boldsymbol{y}_t = \begin{bmatrix} \boldsymbol{a}_t \\ \boldsymbol{b}_t \end{bmatrix}$ and $\boldsymbol{S} = \begin{bmatrix} \boldsymbol{A} \\ \boldsymbol{I}_{n_b} \end{bmatrix}$ is the $n \times n_b$ *summing* or *structural* matrix.

For example, for the hierarchical structure of Fig. 2, $n = 10$, $n_b = 6$, $n_a = 4$, $\boldsymbol{b}_t = [y_{AA,t}, y_{AB,t}, y_{BA,t}, y_{BB,t}, y_{CA,t}, y_{CB,t}]'$, $\boldsymbol{a}_t = [y_{Tot,t}, y_{A,t}, y_{B,t}, y_{C,t}]'$ and

$$\boldsymbol{A} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}.$$

The aggregation matrix, $\boldsymbol{A}$, describes how the bottom-level series aggregate to the series above. Hence, the columns of $\boldsymbol{S}$ span the linear subspace of $\mathbb{R}^n$ for which the linear constraints hold. We refer to this as the *coherent subspace* and denote it by $\mathfrak{s}$. We refer to the property that data adhere to these linear constraints as *coherence*.

Fig. 3 shows a *grouped* structure in which the bottom-level series are aggregated by attributes of interest that are crossed, in contrast to the *hierarchical* (nested) structure shown in Fig. 2. For this grouped example, the bottom-level series $\boldsymbol{b}_t = (y_{\text{AX},t}, y_{\text{AY},t}, y_{\text{BX},t}, y_{\text{BY},t}, y_{\text{CX},t}, y_{\text{CY},t})'$, aggregate into $y_{\text{A},t}$, $y_{\text{B},t}$ and $y_{\text{C},t}$, and also into $y_{\text{X},t}$ and $y_{\text{Y},t}$. Hence, in contrast to hierarchical time series, grouped time series do not naturally aggregate (or disaggregate) in a unique manner. However, for simplicity, we mean hierarchical and grouped structures when referring to hierarchical time series. We will highlight the difference when it is important to do so.

2.2. Other representations

The *structural representation*, based on the summing matrix $\boldsymbol{S}$ in the form shown above, is not the only way to write the constraints for the time series $\boldsymbol{y}_t$.

First, the ordering of the series within $\boldsymbol{y}_t$ is arbitrary, and there is no requirement for the bottom-level series to appear below the aggregated series. An alternative order is sometimes more convenient, and then the rows of $\boldsymbol{S}$ can be permuted to match the order of $\boldsymbol{y}_t$.

The coherent structure can also be expressed via a *constraint matrix* such that

$$\boldsymbol{C}\boldsymbol{y}_t = \boldsymbol{0}.$$

If we start with the structural representation shown above, then we can write $\boldsymbol{C} = [\boldsymbol{I}_{n_a} \quad -\boldsymbol{A}]$. Working with this zero-constrained representation is often more convenient than the structural representation. We can simply start with a general constraint matrix $\boldsymbol{C}$ that may not be of full rank without defining an aggregation or summing matrix (Di Fonzo & Girolimetto, 2023a).

Each of these representations has been used in the forecast reconciliation literature, and we will return to them in subsequent sections.

There is no requirement for the $\boldsymbol{S}$, $\boldsymbol{A}$ or $\boldsymbol{C}$ matrices to contain only 0s and $\pm$1s. They can include any real values, specifying linear constraints that apply to the available time series (Athanasopoulos et al., 2020). For an example of this, see Li and Hyndman (2021) where the elements of the $\boldsymbol{S}$ matrix are proportions. Furthermore, this brings to light the full generality of the so-called hierarchical time series. The methods discussed below can be applied to problems that need not be hierarchical and need not be time series (see Girolimetto & Di Fonzo, 2023).

[3] This paper follows the recommendations of Hyndman (2022).

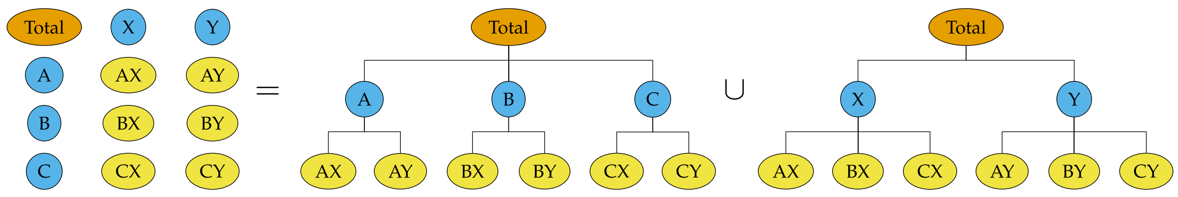

Fig. 3. The diagram shows that a 2-level grouped structure can be considered the union of two hierarchical tree structures with common top and bottom level series.

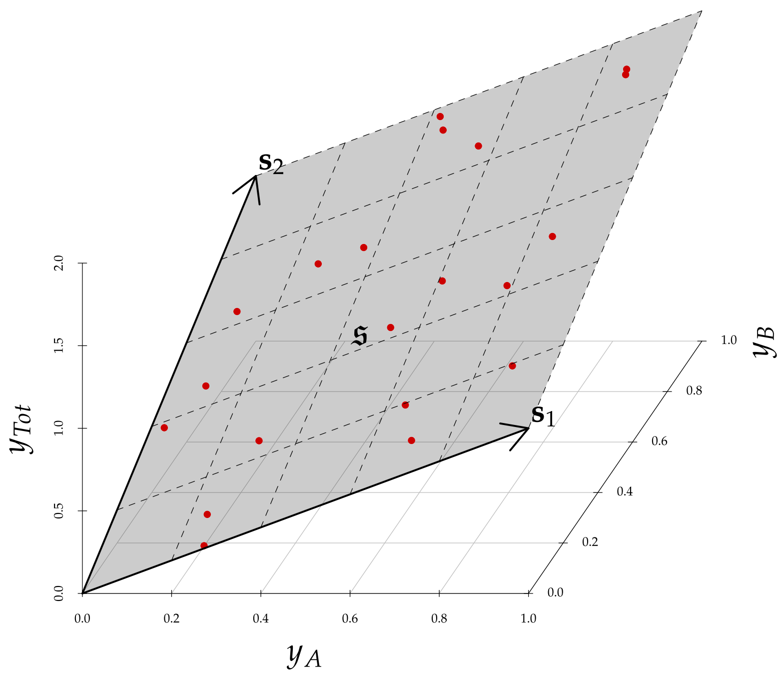

Fig. 4. Depiction of a three dimensional hierarchy with $y_{\text{Tot}} = y_{\text{A}} + y_{\text{B}}$. The gray coloured two dimensional plane depicts the coherent subspace $\mathfrak{s}$ where $\vec{s}_1 = (1, 1, 0)'$ and $\vec{s}_2 = (1, 0, 1)'$ are basis vectors that span $\mathfrak{s}$. The red points in $\mathfrak{s}$ represent realisations or coherent forecasts. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

2.3. Coherent forecasts

When forecasting hierarchical time series, we require the forecasts to adhere to the same linear constraints as the data, i.e., to aggregate in the same manner or to follow the same linear constraints. We define a set of h-step-ahead point forecasts $\tilde{\boldsymbol{y}}_{t+h|t} \in \mathbb{R}^n$ as *coherent* if $\tilde{\boldsymbol{y}}_{t+h|t} \in \mathfrak{s}$. Fig. 4 presents an example of the simplest possible hierarchy for which $\boldsymbol{y}_t \in \mathbb{R}^3$, $\boldsymbol{b}_t \in \mathbb{R}^2$ and $y_{Tot,t} = y_{A,t} + y_{B,t}$. The coherent subspace is shown as a grey 2-dimensional plane within a 3-dimensional space. Note that the columns of $\boldsymbol{S}$, $\vec{s}_1 = (1, 1, 0)'$ and $\vec{s}_2 = (1, 0, 1)'$, span the coherent subspace; i.e., $\mathfrak{s} = \text{span}(\boldsymbol{S})$. The red points in $\mathfrak{s}$ represent realisations or coherent forecasts.

Pritularga et al. (2021) note that this definition of coherence implicitly assumes that the data measurement occurs at a given level, typically the lowest. In practice, this may not be the case, and due to measurement errors, different data collection methodologies, or otherwise, there may be discrepancies in the coherence. Therefore, they propose to add to the aggregation a statistical discrepancy term $\boldsymbol{\delta}_t$:

$$\boldsymbol{y}_t = \boldsymbol{S}\boldsymbol{b}_t + \boldsymbol{\delta}_t.$$

Equivalently, this can be expressed as a slack term in the coherence constraints. When the data collection is done perfectly, this term is zero. Athanasopoulos et al. (2020) and Kourentzes et al. (2021) provide examples where a time series is included in the hierarchy to deal with this discrepancy.

2.4. Single-level approaches

Traditionally, forecasts of hierarchical time series have involved selecting one level of aggregation, generating forecasts for that level, and then linearly combining these to obtain a set of coherent forecasts for the rest of the structure. These methods are usually classified as bottom-up, top-down or middle-out. Bottom-up approaches require generating forecasts at the bottom level and then aggregating these up. Generating forecasts at the most disaggregate level implies no information is lost due to aggregation. On the other hand, bottom-level data can be very noisy or even intermittent, and hence more challenging to forecast. Top-down approaches require forecasts for only one time series at the most aggregate level and then disaggregating these down. Forecasting at the most aggregate level implies a large loss of information, and it can also be challenging to disaggregate these forecasts down. The disaggregation becomes even more challenging when the structure is grouped, as the disaggregation paths are not unique. Further, Hyndman et al. (2011) and Panagiotelis et al. (2021) show that top-down approaches can introduce bias, even if the forecasts for the top level are unbiased. Middle-out approaches require forecasts at some intermediate level and then aggregating these up and disaggregating them down. In general, single-level approaches are limited to using information from a single level and potentially ignoring valuable information from all other levels.

Another consideration comes from the number of forecasting models used in the hierarchy. In the top-down case, all predictions in the hierarchy are anchored to a single forecasting model at the top level (although forecasts at other levels may be used to compute the proportions for disaggregation). Similarly, in bottom-up, all forecasts are anchored to the bottom level. This introduces modelling and estimation risks, where the few forecasts used to populate the rest of the hierarchy may be poor quality. For example, Kourentzes et al. (2017) show that even with full knowledge of the data-generating process, estimation errors can substantially reduce the quality of the resulting forecasts on other levels of the hierarchy.

This section concentrates on implementing single-level approaches for generating point forecasts. Related approaches for generating coherent probabilistic forecasts are discussed in Section 5. Methods in the purely temporal setting are reviewed in Section 4.1.

The vast majority of the literature, prior to the introduction of the concept of forecast reconciliation, almost exclusively focused on comparing bottom-up and top-down methods. Orcutt et al. (1968) and Edwards and Orcutt (1969) are from the early works arguing that information loss is substantial and, therefore, it is important to work with the most disaggregate data available. Kinney (1971) found that disaggregated earnings data by market segments resulted in more accurate forecasts than when firm-level data were used. Building on this result, Collins (1976) compared segmented econometric models with aggregate models for a group of 96 firms and found that the segmented models produced more accurate forecasts for both sales and profit. Dunn et al. (1976) show that forecasts aggregated from a lower level for modelling telephone demand are more accurate than the top-down method, although the comparison was based on only nine series. Shlifer and Wolff (1979) concluded that the bottom-up method is preferable under some conditions on the hierarchy structure and the forecast horizon.

Schwarzkopf et al. (1988) looked at the bias and robustness of the two methods and concluded that the bottom-up method is better except when there are missing or unreliable data at the lowest levels. Dangerfield and Morris (1992) construct 15,000 artificial 2-level hierarchies using the M-competition data with two series at the bottom level. They found that bottom-up forecasts were more accurate, especially when the two bottom-level series were highly correlated. Zellner and Tobias (2000) used annual GDP growth rates from 18 countries and found that disaggregation provided better forecasts, results in line with earlier perspectives expressed by Espasa (1994). Another comparison is that of Wanke and Saliby (2007), who compare the two approaches for safety inventory levels. Wan et al. (2013) analyse aggregate versus disaggregate forecasts for international arrivals into Hong Kong considering alternative bottom-up approaches, arguing that these take advantage of the heterogeneity across the disaggregate series and show that the traditional bottom-up approach is more accurate compared to directly forecasting at the aggregate level.

Fewer studies find clear evidence and argue for a top-down approach in contrast to a bottom-up approach. Grunfeld and Griliches (1960) argue that disaggregated data are error-prone and that top-down forecasts may be more accurate. Strijbosch et al. (2008) in an estimation setting show the superiority of a top-down estimator. Huddlestone et al. (2015) find evidence of using a top-down method when forecasting noisy geographic time series in many applications. Athanasopoulos et al. (2009) propose two new top-down approaches based on forecast proportions rather than historical proportions, which show promising performance. These approaches can lead to negative weights, perhaps stretching the definition of disaggregation. Gross and Sohl (1990) and Athanasopoulos et al. (2009) provide a summary of top-down approaches.

Fliedner and Mabert (1992) experiment with how different groupings (based on clustering) of time series affect the forecast accuracy of traditional approaches. Fliedner (1999) argues that strong positive or negative correlation of sub-aggregate series enhances forecast accuracy of the aggregate series whether a bottom-up or a direct approach is used for forecasting the aggregate and vice versa (low correlation in the bottom-level series diminishes forecast accuracy of the aggregate series). He further concludes that direct forecasts of an aggregate variable are more accurate than a bottom-up approach. Lütkepohl (1984a) and Ilmakunnas (1990) show that it might be preferable to forecast the aggregate variable directly rather than using a bottom-up approach. Hubrich (2005) also concludes using a bottom-up approach to forecast inflation for the Euro area. He attributes this to shocks affecting components of inflation in a similar way over the evaluation period, and therefore, forecast bias is increased when aggregating subcomponent forecasts. Kremer et al. (2016) examine bottom-up versus direct forecasts for the aggregate through a behavioural lens and argue for advantages and disadvantages of a bottom-up judgemental approach, which depend to a large degree on the underlying correlation structure at the bottom level.

Zotteri et al. (2005) and Zotteri and Kalchschmidt (2007) argue that forecast accuracy is highly correlated to the choice of aggregation level, which depends on the underlying data generation process. Widiarta et al. (2008) compare top-down to bottom-up in a restricted simulation setting and find that the difference in forecast accuracy between these is insignificant when the correlation between the sub-aggregate components is small or moderate. Sbrana and Silvestrini (2013) extend these results and conclude that neither the top-down nor the bottom-up approach should be preferred a priori in any empirical analysis. Williams and Waller (2011) compare top-down to bottom-up demand forecasts. They conclude that the superiority of the methods depends on whether shared weekly point-of-sale data are used.

Fliedner (2001) also reviewed these approaches and discussed their advantages and disadvantages. He notes that different forecasting methods may be better suited to different aggregation levels, affecting the choice of which level to use for forecasting.

Kahn (1998) highlighted the need for a method that would enjoy and combine the good features of single-level approaches. A simple solution of averaging top-down and bottom-up forecasts was suggested by Lapide (1998) and later built upon by Rehman et al. (2023). The call was taken up in full by Hyndman et al. (2011) and Athanasopoulos et al. (2009), who introduced the concept of forecast reconciliation.

3. Forecast reconciliation

3.1. Least squares reconciliation outside of forecasting

The concept of least squares reconciliation has appeared in several contexts outside the forecasting domain. As far back as the 1940s, reconciliation procedures were used for national economic accounts. The national economic account is disaggregated into production, income and outlay, and capital transactions, further disaggregated by various factors. The aim is to coherently estimate the national account for all disaggregated and aggregated levels. Stone et al. (1942) formulated the problem using simultaneous linear equations (similar to the zero-constrained form). Stone (1961) proposed a constrained estimation approach to balancing national accounts, where the constrained estimates are a weighted linear combination of initial estimates. This underpinned the work for which Richard Stone later won the 1984 Nobel Prize in Economics. Byron (1978, 1979) formalised and extended Stone's work using more computationally efficient procedures. Suppose the national accounts are expressed as a vector $\boldsymbol{y}$, which need to satisfy the constraint $\boldsymbol{C}\boldsymbol{y} = \boldsymbol{0}$, and let the original (incoherent) account estimates be denoted by $\hat{\boldsymbol{y}}$. Then the reconciled estimates $\tilde{\boldsymbol{y}}$ are found by solving the constrained generalised least squares (GLS) problem

$$\tilde{\boldsymbol{y}} = \arg\min_{\boldsymbol{y}} (\boldsymbol{y} - \hat{\boldsymbol{y}})'\boldsymbol{W}^{-1}(\boldsymbol{y} - \hat{\boldsymbol{y}}), \qquad s.t.\ \boldsymbol{C}\boldsymbol{y} = \boldsymbol{0}.$$

Assuming $\boldsymbol{C}$ is full rank, Byron (1978, 1979) provide the solution $\tilde{\boldsymbol{y}} = \boldsymbol{M}\hat{\boldsymbol{y}}$, where

$$\boldsymbol{M} = \boldsymbol{I} - \boldsymbol{W}\boldsymbol{C}'(\boldsymbol{C}\boldsymbol{W}\boldsymbol{C}')^{-1}\boldsymbol{C}$$

is a projection matrix, and $\boldsymbol{W}$ is a positive definite matrix. See also Weale (1992), Smith et al. (1998) Bikker et al. (2013) for a more modern treatment of this approach.

Later, the same idea was applied to reconciling other time series produced by national statistics offices. Some of this work is reviewed by Dagum and Cholette (2006, chapter 11). As an example, seasonally adjusted time series require reconciliation. While the original time series data are coherent (e.g., national and state employment numbers), they become incoherent after each series is seasonally adjusted. The same least squares solution is used for this problem (Corona et al., 2021; Di Fonzo & Marini, 2011; Quenneville & Fortier, 2012).

Temporal reconciliation also interests national statistics offices, ensuring monthly or quarterly estimates sum to the annual estimates (Chow & Lin, 1971). Simultaneous least squares reconciliation of time series estimates in cross-sectional and temporal dimensions was introduced by Di Fonzo (1990), building on Rossi (1982).

Least squares reconciliation has also found its way into chemical process measurement. Chemical process data are inherently noisy, and data reconciliation methods allow adjustment of measured values to satisfy specific material and energy constraints (Romagnoli & Sanchez, 2000).

In the engineering literature, a related problem involves optimal vehicle tracking where roads provide locally linear constraints on the position of a vehicle (Simon & Chia, 2002). This line of research is summarised in Simon (2006) and Simon (2010).

3.2. First attempts at reconciliation in forecasting

To our knowledge, the earliest published work that applied least squares reconciliation in a forecasting context was the PhD thesis of Roman Ahmed (2009), working under the supervision of Rob Hyndman and George Athanasopoulos. The main methodological contributions from this thesis eventually appeared as Hyndman et al. (2011). First, they showed that all the existing bottom-up, middle-out and top-down methods could be expressed as

$$\tilde{\boldsymbol{y}}_h = \boldsymbol{S}\boldsymbol{G}_h\hat{\boldsymbol{y}}_h \tag{1}$$

for a suitably chosen $n_b \times n$ matrix $\boldsymbol{G}_h$. (We will drop the subscript h when $\boldsymbol{G}$ does not depend on the forecast horizon, h.) Here, $\boldsymbol{G}_h$ maps the base forecasts $\hat{\boldsymbol{y}}_h$ into the bottom level, and so can be thought of as a forecast combination that combines all base forecasts to form bottom-level reconciled forecasts. In the special case of bottom-up forecasting, $\boldsymbol{G} = [\boldsymbol{O}_{n_b \times n_a} \quad \boldsymbol{I}_{n_b}]$, while for top-down forecasts, the first column of $\boldsymbol{G}$ contains the proportions for each of the bottom-level series, while the remaining columns are all zero.

Hyndman et al. (2011) showed that if the base forecasts $\hat{\boldsymbol{y}}_h$ are unbiased with covariance $\boldsymbol{W}_h$, and $\boldsymbol{S}\boldsymbol{G}_h\boldsymbol{S} = \boldsymbol{S}$, then the reconciled forecasts $\tilde{\boldsymbol{y}}_h$ are also unbiased and have covariance $\boldsymbol{S}\boldsymbol{G}_h\boldsymbol{W}_h\boldsymbol{G}_h'\boldsymbol{S}'$. Notably, the condition $\boldsymbol{S}\boldsymbol{G}_h\boldsymbol{S} = \boldsymbol{S}$ is generally not satisfied for top-down methods, except for those discussed in Section 3.6.

To find the optimal matrix $\boldsymbol{G}_h$, Hyndman et al. (2011) formulated the problem as a regression of the form

$$\hat{\boldsymbol{y}}_h = \boldsymbol{S}\boldsymbol{\beta}_h + \boldsymbol{\varepsilon}_h$$

where $\boldsymbol{\varepsilon}_h$ is the reconciliation error with covariance $\boldsymbol{V}_h$. This led to the GLS solution

$$\boldsymbol{G}_h = (\boldsymbol{S}'\boldsymbol{V}_h^{-1}\boldsymbol{S})^{-1}\boldsymbol{S}'\boldsymbol{V}_h^{-1}. \tag{2}$$

The covariance $\boldsymbol{V}_h$ is unknown (and later shown to be unidentifiable by Wickramasuriya et al., 2019), but under some conditions, Hyndman et al. (2011) showed that (2) collapses to an OLS solution where $\boldsymbol{V}_h$ is replaced by an identity matrix, giving $\boldsymbol{G} = (\boldsymbol{S}'\boldsymbol{S})^{-1}\boldsymbol{S}'$.

Another key contribution of Hyndman et al. (2011) was to propose using sparse matrix algebra to speed up the computations for large time series systems greatly. Simultaneously and independently, Di Fonzo and Marini (2011) also proposed sparse matrix algebra for reconciling historical time series.

The first application of these new ideas was Athanasopoulos et al. (2009), which appeared two years earlier due to delays in publishing Hyndman et al. (2011). There, the OLS reconciliation was compared to various top-down and bottom-up methods, using some quarterly Australian tourism data disaggregated by a geographic hierarchy and purpose of travel. Variations of these Australian tourism data have since become ubiquitous for benchmarking forecast reconciliation methods.

The `hts` R package (Hyndman et al., 2010) implementing the OLS reconciliation method appeared on CRAN in 2010. The method quickly became popular in business and industry long before the methodological paper appeared.

An early explanation of the method intended for practitioners appeared as Hyndman and Athanasopoulos (2014), while the ideas made their way into an undergraduate textbook in Hyndman and Athanasopoulos (2018) and Hyndman and Athanasopoulos (2021).

3.3. Scaled reconciliation methods

One obvious drawback of the OLS approach is that it weights all series equally, whether they are aggregates or disaggregates, and whether their base forecasts are good or bad. An early recognition of this issue is in Kourentzes et al. (2014), who treat the aggregate time series to bring all series on the same scale. The same issue prompted Hyndman et al. (2016) to propose a weighted least squares (WLS) solution, where the series are weighted by the inverse variances of the base forecasts, later referred to as "variance scaling". If $\boldsymbol{W}_h$ is the covariance matrix $\mathrm{Var}(\boldsymbol{y}_{T+h|h} - \hat{\boldsymbol{y}}_h)$, then the WLS solution is $\boldsymbol{G}_h = (\boldsymbol{S}'\boldsymbol{\Lambda}_h^{-1}\boldsymbol{S})^{-1}\boldsymbol{S}'\boldsymbol{\Lambda}_h^{-1}$ and $\boldsymbol{\Lambda} = \mathrm{diag}(\boldsymbol{W}_h)$. Hyndman et al. (2016) was also the first forecasting paper to note that the methods applied to grouped time series and strictly hierarchical structures.

3.4. Minimum trace reconciliation

Wickramasuriya et al. (2019) provided theoretical insights into the problem by taking an optimisation rather than a regression approach. They formulated the problem as minimizing the trace (MinT) of the covariance matrix $\mathrm{Var}(\boldsymbol{y}_{T+h|h} - \tilde{\boldsymbol{y}}_h)$, equal to the sum of the variances of all the reconciled forecasts, and showed that the solution is given by

$$\boldsymbol{G}_h = (\boldsymbol{S}'\boldsymbol{W}_h^{-1}\boldsymbol{S})^{-1}\boldsymbol{S}'\boldsymbol{W}_h^{-1},$$

where $\boldsymbol{W}_h$ is the covariance matrix $\mathrm{Var}(\boldsymbol{y}_{T+h|h} - \hat{\boldsymbol{y}}_h)$. Equivalently, $\tilde{\boldsymbol{y}}_h = \boldsymbol{M}_h\hat{\boldsymbol{y}}_h$, where

$$\boldsymbol{M}_h = \boldsymbol{S}(\boldsymbol{S}'\boldsymbol{W}_h^{-1}\boldsymbol{S})^{-1}\boldsymbol{S}'\boldsymbol{W}_h^{-1}.$$

This MinT solution is equivalent to GLS and has the WLS and OLS solutions as special cases. Wickramasuriya et al. (2019) also showed that there is an equivalent and more computationally efficient solution given by

$$\boldsymbol{M}_h = \boldsymbol{I}_n - \boldsymbol{W}_h\boldsymbol{C}'(\boldsymbol{C}\boldsymbol{W}_h\boldsymbol{C}')^{-1}\boldsymbol{C}$$

where $\boldsymbol{C} = [\boldsymbol{I}_{n_a} \quad -\boldsymbol{A}]$, which matches the earlier work of Byron (1978, 1979) for reconciling national accounts (although derived from a different perspective). This inverts an $n_a \times n_a$ matrix rather than an $n \times n$ matrix.

A difficulty with the MinT solution is in estimating the covariance matrix, $\boldsymbol{W}_h$, especially for $h > 1$. The sample covariance matrix of the base models' residuals estimates $\boldsymbol{W}_1$. Still, it is often a very poor estimate and may be singular, especially when $n > T$ (which is true for many real hierarchical time series). Wickramasuriya et al. (2019) proposed a shrinkage estimator of $\boldsymbol{W}_1$, where the off-diagonal elements are shrunk towards zero, and suggested approximating $\boldsymbol{W}_h$ as a scalar multiple of $\boldsymbol{W}_1$. The scalar multiple cancels when computing $\boldsymbol{G}_h$, so it does not need to be estimated when computing point forecasts.

Wickramasuriya et al. (2019) also discuss a simple alternative approach to finding $\boldsymbol{W}_h$, first proposed by Athanasopoulos et al. (2017), based only on the structure of the hierarchy, and not on a statistical estimate. In this "structural scaling" approximation, $\boldsymbol{W}_h = \boldsymbol{\Lambda} \propto \mathrm{diag}(\boldsymbol{S}\boldsymbol{1}_{n_b})$, where $\boldsymbol{1}_{n_b}$ is an n_b-vector of 1s. That is, $\boldsymbol{W}_h$ is a diagonal matrix with entries proportional to the row sums of $\boldsymbol{S}$. This is the covariance matrix that would arise if all the most disaggregated series were uncorrelated with each other and had the same forecast variance.

Observe that (1) implies a combination of forecasts in $\boldsymbol{G}_h\hat{\boldsymbol{y}}_h$. Motivated by this, Pritularga et al. (2021) investigated the implications of estimation uncertainty in forecast reconciliation. They showed that uncertainties in both the forecasting models generating $\hat{\boldsymbol{y}}_h$ and in $\boldsymbol{G}_h$ will influence the quality of the reconciled forecasts. As $\boldsymbol{S}$ is fixed, the approximation of $\boldsymbol{W}_h$ carries the reconciliation uncertainty in $\boldsymbol{G}_h$, where each element that needs to be estimated can potentially increase the forecast error of the reconciled forecasts. This can help explain the surprising performance of "structural scaling" (Athanasopoulos et al., 2017), where all elements of $\boldsymbol{G}$ are fixed, and the volatile performance of MinT, especially for short time series, where estimation errors in $\boldsymbol{W}_1$ can dominate. This is similar to the forecast combination puzzle (Claeskens et al., 2016), where the performance of forecast combinations is hindered by imprecise estimation of covariances. By introducing various approximations of increasing complexity of $\boldsymbol{W}_1$, Pritularga et al. (2021) demonstrate that this choice can have a significant effect on the quality of the reconciled forecasts, especially when it comes to having consistent performance over different forecast origins, a key element of trustworthiness in forecasting (Spavound & Kourentzes, 2022). From the investigated approximations, retaining only the block-diagonal structure of the shrinkage estimator of $\boldsymbol{W}_1$ performed well in various situations.

3.5. Other optimisation approaches

van Erven and Cugliari (2015) took a game-theoretic approach to forecast reconciliation and chose to find the solution to the minimax problem

$$V = \min_{\tilde{\boldsymbol{y}} \in \mathfrak{s}} \max_{\boldsymbol{y} \in \mathfrak{s}} \left\{ \ell(\boldsymbol{y}, \tilde{\boldsymbol{y}}) - \ell(\boldsymbol{y}, \hat{\boldsymbol{y}}) \right\},$$

where ℓ is a loss function, and $\mathfrak{s}$ is the coherent subspace. They demonstrate that $V \leq 0$ so that the reconciled forecasts are guaranteed to have smaller losses than the base forecasts. They further show that when ℓ is L_2 loss, the minimax solution is equivalent to solving the constrained least squares problem, where the reconciled and base forecasts are as close as possible subject to the reconciled forecasts being coherent, leading to the closed form solution of (2) for $\boldsymbol{G}_h$.

Panagiotelis et al. (2021) unify, and in certain cases generalise, the results of van Erven and Cugliari (2015) and Wickramasuriya et al. (2019), providing a geometric intuition. In particular, they consider a loss function of the form $\ell(\boldsymbol{y}, \breve{\boldsymbol{y}}) = (\boldsymbol{y} - \breve{\boldsymbol{y}})' \boldsymbol{\Psi} (\boldsymbol{y} - \breve{\boldsymbol{y}})$, where $\boldsymbol{\Psi}$ can be any symmetric positive definite matrix, $\breve{\boldsymbol{y}}$ is either $\hat{\boldsymbol{y}}$ or $\tilde{\boldsymbol{y}}$ and derive two main results. The first is that the reconciled forecast $\tilde{\boldsymbol{y}} = \boldsymbol{S}(\boldsymbol{S}' \boldsymbol{\Psi} \boldsymbol{S})^{-1} \boldsymbol{S}' \boldsymbol{\Psi} \hat{\boldsymbol{y}}$ will always improve upon the base forecast in the sense that $\ell(\boldsymbol{y}, \tilde{\boldsymbol{y}}) \leq \ell(\boldsymbol{y}, \hat{\boldsymbol{y}})$, where the strict inequality holds if $\hat{\boldsymbol{y}}$ is incoherent. This generalises the result of van Erven and Cugliari (2015) to non-diagonal $\boldsymbol{\Psi}$

The second is that the MinT solution $\tilde{\boldsymbol{y}} = \boldsymbol{S}(\boldsymbol{S}' \boldsymbol{W}_h^{-1} \boldsymbol{S})^{-1} \boldsymbol{S}' \boldsymbol{W}_h^{-1} \hat{\boldsymbol{y}}$ will optimise loss in expectation for any choice of $\boldsymbol{\Psi}$ (generalising the result of Wickramasuriya et al., 2019). Note that the second result does not consider the estimation uncertainty in $\boldsymbol{W}_h$ (Pritularga et al., 2021), which can be especially prominent for short time series, explaining cases in the literature where this result seems to be violated.

If we are willing to drop the unbiased condition and allow both base and reconciled forecasts to be biased, a different least squares solution emerges, as shown by Ben Taieb and Koo (2019). They use an expanding window approach, applying the base forecasting method iteratively to the training data, computing $\hat{\boldsymbol{y}}_{t+h|t}$, the h-step-ahead base forecast of $\boldsymbol{y}_{t+h}$ based on training data $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_t$, for $t = T_1, \ldots, T - h$. They consider the regularised empirical risk minimisation problem

$$\min_{\boldsymbol{G}} L_T(\boldsymbol{G}),$$

where

$$L_T(\boldsymbol{G}) = \frac{1}{Nn} \| \boldsymbol{Y} - \hat{\boldsymbol{Y}} \boldsymbol{G}' \boldsymbol{S}' \|_F + \lambda \| \mathrm{vec} \boldsymbol{G} \|_1,$$

$N = T - T_1 - h + 1$, $\| \cdot \|_F$ is the Frobenius norm, $\boldsymbol{Y} = [\boldsymbol{y}_{T_1+h}, \ldots, \boldsymbol{y}_T]'$, $\hat{\boldsymbol{Y}} = [\hat{\boldsymbol{y}}_{T_1+h|T_1}, \ldots, \hat{\boldsymbol{y}}_{T|T-h}]'$, and λ is a regularisation parameter. The first term contains the errors of the reconciled forecasts, while the second shrinks the elements of $\boldsymbol{G}$ to zero, providing some regularisation of the amount of reconciliation involved. When $\lambda = 0$, they show that its solution is

$$\hat{\boldsymbol{G}} = \boldsymbol{B}' \hat{\boldsymbol{Y}} (\hat{\boldsymbol{Y}}' \hat{\boldsymbol{Y}})^{-1}$$

where $\boldsymbol{B} = [\boldsymbol{b}_{T_1+h}, \ldots, \boldsymbol{b}_T]'$. When $\hat{\boldsymbol{Y}}' \hat{\boldsymbol{Y}}$ is non-invertible, the solution is not unique, and a generalised inverse can be used.

Inspired by this development, Wickramasuriya (2021) proposed minimizing the trace of the forecast error covariance matrix without an unbiased constraint to create an unconstrained version of MinT, which she called "MinT-U". She also derived an estimate of the resulting $\boldsymbol{G}$ matrix in the case where the series are jointly weakly stationary, dubbing the resulting method "EMinT-U" (empirical MinT unconstrained).

3.6. Adding optimisation constraints

Any approach to reconciliation based on optimisation uses a form of constrained optimisation since reconciled forecasts must lie on the coherent subspace. However, at times, additional constraints may be implemented. The first is the case where reconciled forecasts must be non-negative. In general, even if base forecasts are constrained to be positive (which can be achieved by modelling on the log scale and back-transforming), there is no guarantee that the usual reconciliation approaches, such as OLS and MinT, will maintain the non-negativity of forecasts. The usual optimisation problem can be augmented with non-negativity constraints on the reconciled forecasts to address this issue. Such optimisation problems can be solved using quadratic programming, with Wickramasuriya et al. (2020) providing an early example for forecast reconciliation and Di Fonzo and Girolimetto (2023b) a more recent example.

Kourentzes and Athanasopoulos (2021) also consider the case of non-negative reconciled forecasts. However, instead of a constrained optimisation approach, they propose a heuristic to adjust the reconciled predictions to be non-negative iteratively. Although this does not guarantee optimal solutions, their proposed algorithm has the interesting feature that it distributes adjustments of forecasts across the hierarchy, which can be useful in various situations, such as applying judgemental adjustments on specific nodes of the hierarchy.

Di Fonzo and Girolimetto (2023b) also discuss an effective nonnegative heuristic called "set-negative-to-zero", whereby the negative reconciled forecasts at the bottom level are set to zero, and the remaining forecasts computed via aggregation.

Another constraint of interest is where some particular base forecasts remain unchanged. For instance, Hollyman et al. (2021) consider the case of reconciliation where the top-level base forecast is retained. This differs from truly top-down approaches in that it can be done while preserving the unbiasedness of base forecasts. To briefly illustrate the main idea, for a three-variable hierarchy where $y_{Tot,t} = y_{A,t} + y_{B,t}$, either setting

$$\begin{pmatrix} \tilde{y}_{Tot,t} \\ \tilde{y}_{A,t} \\ \tilde{y}_{B,t} \end{pmatrix} = \begin{pmatrix} \hat{y}_{Tot,t} \\ \hat{y}_{A,t} \\ \hat{y}_{Tot,t} - \hat{y}_{A,t} \end{pmatrix} \text{ or } \begin{pmatrix} \tilde{y}_{Tot,t} \\ \tilde{y}_{A,t} \\ \tilde{y}_{B,t} \end{pmatrix} = \begin{pmatrix} \hat{y}_{Tot,t} \\ \hat{y}_{Tot,t} - \hat{y}_{B,t} \\ \hat{y}_{B,t} \end{pmatrix}$$

will lead to coherent forecasts that preserve unbiasedness. Any average between these two solutions will have the same properties. Hollyman et al. (2021) generalise this idea to more complex hierarchies, and the properties of their methods are investigated by Di Fonzo and Girolimetto (2023). Zhang et al. (2023) further generalise this idea to a setting where reconciliation can be carried out while keeping a subset of base forecasts unchanged and not just the top level. Conditions on how a set of such "immutable" series can be selected are also provided by Zhang et al. (2023).

3.7. Machine learning and regularisation

Machine Learning (ML), including Artificial Intelligence (AI), methods have been used to provide various modifications of the optimal combination approach of Hyndman et al. (2011). Most contributions attempt to replace the linear regression formulation with a less restrictive method to obtain combinations of forecasts from the various hierarchical levels. Coherence is achieved via a bottom-up approach or by embedding coherence in the ML training. A minority of contributions focus on other aspects of hierarchical forecasting, such as selecting the best reconciliation method.

Qiao and Huang (2018) focus on earnings forecasting and rely on the problem structure of hierarchical forecasting to address this, recognising the forecast combination at its core. However, instead of combining forecasts across hierarchical levels, they combine forecasts across alternative hierarchical mappings and then proceed to achieve coherence using a bottom-up approach. The different mappings are the product of how one can account for the different components that contribute to the net earnings. This raises the question of which hierarchy is best to use and how to search across the different hierarchies efficiently. They resolve the construction of the hierarchy using a genetic algorithm to avoid the computationally infeasible greedy search across mappings of the hierarchy. Although they do not discuss this in their work, this approach could be used to relax the conventionally rigid hierarchies and identify re-mappings that can potentially improve the final result. LSTM networks generate forecasts for each time series and each different hierarchy mapping. The forecasts are then combined across these mappings to give a final prediction, with encompassed forecasts being rejected from the combination.

Spiliotis et al. (2021) rely on random forests and gradient boosting machines, specifically XGBoost, to facilitate the combination of forecasts implied by hierarchical forecasting. They show superior performance to the linear approach. However, it is unclear whether the gains are due to the nonlinear capabilities of the ML methods or due to the differential combination over various forecast horizons that are considered and are typically omitted by the linear counterparts. Furthermore, it should be noted that the objective of training the ML methods is obtaining the minimum forecast combination errors rather than minimum reconciliation errors. Coherent forecasts across the complete hierarchy are obtained via bottom-up aggregation. Burba and Chen (2021) propose an alternative use of ML to achieve coherent forecasts. They recast the reconciliation step as an encoder–decoder setup, where a trainable encoder processes base forecasts to produce the reconciled bottom-level forecasts. These are then decoded using the summing matrix, as with a standard bottom-up setup. The encoder is implemented using a shallow feed-forward neural network. They find that this approach demonstrates increasing gains for deeper hierarchies.

Gleason (2020) attempts to overcome the lack of focus on coherence by adjusting the objective function. Using neural network forecasts, he includes a regularisation term that penalises incoherences in the generated forecasts. This follows from Mishchenko et al. (2019), who proposed a similar regularisation term to obtain reconciled forecasts directly from the base forecasts. The disadvantage of these regularisation approaches is that they result in soft constraints that do not guarantee coherence. Gleason (2020) provides two alternatives for the regularisation term and shows that the resulting forecasts can outperform standard MinT reconciliation. However, when the regularisation is used in conjunction with MinT, this results in both coherent and the most accurate forecasts. Han et al. (2021) propose a similar approach, where a regularisation term is added in the loss function, again based on coherence constraints. They also consider a regularised loss for producing coherent quantile forecasts. The authors demonstrate the use of the proposed regularised loss on various linear and ML models and empirically show the negative effect of regularisation on coherence.

Shiratori et al. (2020) introduces a regularisation term, based on the coherence constraints, in the objective function to push bottom-level forecasts to fit both on their target series and their aggregate counterparts. They forecast the bottom-level series, but as the regularisation cannot ensure coherence, these are used in a bottom-up setting to produce coherent forecasts for the rest of the hierarchy. The authors demonstrate the efficacy of this in the context of neural networks. They find that these outperform conventionally trained networks whose forecasts are reconciled with bottom-up or MinT.

Paria et al. (2021) propose a regularised neural network with sequence-to-sequence architecture. Focusing on the hierarchical part of the contribution, a regularisation term is added to incorporate the coherence constraints. As with the previous work, this does not guarantee coherence yet forces the final forecasts to be approximately coherent. The regularisation is embedded in the loss function of the network, achieving an integrated approach. In contrast to the previous work, the network outputs forecasts for all the levels.

The contribution by Anderer and Li (2022) can belong loosely in the regularisation approaches. Focusing on the M5 competition dataset, the authors use different methods to produce separate forecasts for the hierarchy's top and bottom levels (NBEATS and LightGBM, respectively). To achieve coherent forecasts, they modify the loss function of the bottom-level method, where positive errors can be multiplied by a factor. This factor is identified by minimising the incoherence error between the summed bottom-level forecasts and the top-level forecast. Note that this factor is calibrated by keeping the top-level forecasts fixed. Therefore, any coherence is obtained by modifying the bottom-level forecasts.

Rangapuram et al. (2021) propose another way to achieve an integrated forecast-reconciliation mechanism.

First, a global neural network is used to forecast the time series in the hierarchy, which can produce sample forecasts with the assumption of a forecast distribution. The sample forecasts are subsequently reconciled, obtaining distributions of coherent probabilistic forecasts. The calculation of the loss for the training of the model makes use of the reconciled forecasts, allowing an end-to-end approach that parametrises the model to achieve both accurate and coherent forecasts. Note that the methodology allows for various assumptions on the forecast distribution and the relaxing of these. The resulting forecasts guarantee coherence, in contrast to the previous integrated approaches. The substantive difference with conventional hierarchical approaches that post-process base forecasts is that the global network can model richer interconnections between the time series in the hierarchy to generate the forecasts. Furthermore, the authors compare the results of the proposed hierarchical model against a global learner without coherent forecasts and demonstrate, on average, better performance, suggesting that the integrated methodology offers gains beyond any achieved by global learning.

Wang, Zhou, et al. (2022) contribute with a similar formulation. The important differences are that they use an autoregressive transformer to produce the base forecasts and that their approach does not require assuming particular predictive distributions. Instead, they rely on empirical estimation (conditional normalising flow) for the distributions. Furthermore, their approach focuses on obtaining bottom-level forecasts, which are internally aggregated in a bottom-up fashion to the complete hierarchy. Similar to Rangapuram et al. (2021), the errors of the final forecasts are used during training, realising an end-to-end hierarchical forecasting method. The authors demonstrate gains in performance over a non-hierarchical version of their method, as well as over various benchmarks.

Focusing on temporal hierarchies, Theodosiou and Kourentzes (2021b) provide an end-to-end neural network-based method. Similarly to Burba and Chen (2021), they explore a series of encoder–decoders to achieve reconciliation, considering fixed and trained decoder weights, using the complete or only the bottom-level of the hierarchy. These can be used instead of conventional hierarchical methods, demonstrating good performance in a global learning setting. To achieve an end-to-end integration, they pass the temporal hierarchy data through a convolutional layer to an LSTM. The convolutional layer compresses the abstracts of the hierarchical time series, and the LSTM models the dynamics over time. An end-to-end methodology is obtained by appending an encoder–decoder to the outputs of the LSTM. The authors demonstrate the gains due to the various components and investigate the time series needed to perform well in a global training setting. By modifying the training loss function, Theodosiou and Kourentzes (2021a) extend the method to provide quantile forecasts.

Abolghasemi, Tarr, et al. (2022) use ML to address the challenge of selecting the best method to perform the hierarchical reconciliation. To achieve this, they rely on a meta-learning classifier, which is trained to identify, given the time series features of a dataset, the best reconciliation method to employ. The approach is quite flexible in terms of features and classifiers, as well as the forecasting models and reconciliation methods.

Abolghasemi et al. (2019) use ML to improve the performance of non-combination hierarchical approaches. They consider various ML methods to obtain better prorations to decompose upper-level forecasts to lower levels. Feng and Zhang (2020) provide a standard implementation of hierarchical methods using base forecasts from ML models and find the hierarchically reconciled forecasts to be the most accurate. Punia et al. (2020) leverage LSTM networks to produce forecasts for the cross-temporal case. Forecasts are generated for the various temporal aggregation levels, reconciled using temporal hierarchies and then cross-sectional. Sprangers et al. (2021) propose a method for probabilistic gradient boosting machines and benchmark its performance on the hierarchical M5 dataset, demonstrating good performance against other non-hierarchical ML methods. Although the proposed method can provide probabilistic predictions for all time series of the hierarchy, coherence is not established. Mancuso et al. (2021) use a deep neural network to produce reconciled forecasts directly. The neural network captures the structure of the hierarchy. It links the relationship between time series features extracted at any level of the hierarchy and explanatory variables into an end-to-end neural network. Sagheer et al. (2021) propose a deep, long, short-term memory approach developed for the hierarchical time-series setting.

3.8. Bayesian versions

Papers tackling hierarchical forecasting from a Bayesian angle focus on different aspects of the problem but have much in common due to certain advantages of Bayesian inference. These are the suitability of Bayesian inference for models with latent states such as state space models, the natural way uncertainty is propagated via Bayes' rule leading to probabilistic forecasts, and the use of priors to incorporate judgement into forecasts.

Park and Nassar (2014) propose a top-down approach, similar to Athanasopoulos et al. (2009), that forecasts the bottom-level series as proportions of the top-level forecast rather than forecasting them directly. To this end, a state space model is proposed where latent states are mapped to proportions via the softmax function. A variational approximation factorised into states and remaining parameters is employed with Evidence Lower Bound (ELBO) optimised via the EM algorithm. For the data considered, the method improves upon the top-down method of Athanasopoulos et al. (2009) and gives more accurate bottom-level forecasts than bottom-up and OLS. However, as a top-down method, there is no scope for top-level forecasts to be improved using bottom-level series. For the data considered in the paper, both bottom-up and OLS generate more accurate forecasts for the top-level series.

Roque et al. (2021) also employ Bayesian modelling for forecasting hierarchical data in a way that differs from the two-step reconciliation approach. In particular, they decompose time series into a trend modelled by a piecewise linear component and a stationary component modelled as a sum of Gaussian Processes (GPs). Rather than modelling individual GPs for each series in the hierarchy, these are fit group-wise, with groups determined according to the hierarchical structure. The method outperforms MinT in an empirical study for the Australian prison population data but not for the Australian tourism data.

Another strain of the literature brings a Bayesian approach to the regression model interpretation of forecast reconciliation. Novak et al. (2017) recognise that the posterior of $\boldsymbol{\beta}_h$ can act as a probabilistic forecast for the bottom-level series. Using Markov chain Monte Carlo to obtain a sample from this posterior and then aggregating gives a probabilistic forecast for the entire hierarchy. Eckert et al. (2021) also obtain a posterior on $\boldsymbol{\beta}_h$. Still, their focus is on augmenting the reconciliation regression equation with a vector of intercepts that allows base forecasts to be biased and evolve according to a state space representation. Both Novak et al. (2017) and Eckert et al. (2021) suggest that in a Bayesian setting, judgement can be incorporated via the prior, in the latter case via an explicit empirical example where prior information about a structural break in data classification can be exploited. Also, while both papers recognise the potential of Bayesian inference to obtain probabilistic forecasts, neither paper makes this the focus of empirical evaluation. Novak et al. (2017) minimise loss functions over the posterior sample and then use this as a point forecast, while Eckert et al. (2021) use maximum a posteriori (MAP) estimates as point forecasts.

Probabilistic forecast reconciliation is the motivation and focus of the Bayesian algorithm proposed by Corani et al. (2021). In particular, a prior is placed on the bottom-level series with the mean set to point forecasts obtained in the first step of forecast reconciliation and variance given by the variance–covariance matrix of one-step ahead errors. This prior is updated using the top-level forecasts obtained in the first stage of forecast reconciliation via Bayes' rule. The method generalises MinT because the posterior mean is equivalent to the usual MinT approach. The necessary updates via Bayes' rule parallel the Kalman filter since the reconciliation problem is recast as a linear Gaussian model. The empirical results are evaluated using scoring rules for probabilistic forecasts, including the CRPS and energy score. This approach has also recently been extended to the challenging case of forecast reconciliation for discrete data by Corani et al. (2022) and Zambon et al. (2022).

Bayesian methods are likely to continue to play an important role in the development of the forecast reconciliation literature. Some promising avenues will be incorporating information from the hierarchical structure via the prior and using Bayesian methods to obtain non-Gaussian probabilistic forecasts. The challenges are likely to be computational, as the scalability of MCMC methods to large hierarchies may be difficult. The development of fast alternatives, such as variational inference, represents a promising way forward.

3.9. In-built coherence

So far, all the approaches discussed have involved two steps — first, compute the base forecasts $\hat{\boldsymbol{y}}_h$, and then reconcile them to produce $\tilde{\boldsymbol{y}}_h$. The computationally slow part is producing the base forecasts because they usually involve fitting models to each series individually, with the estimation requiring non-linear optimisation. However, as shown by Ashouri et al. (2022), if the base forecasts $\hat{\boldsymbol{y}}$ are produced using a linear regression model, the base forecasts and reconciliation can be combined, giving coherent forecasts directly in a single closed-form equation. Further, the computation is extremely fast, provided sparse matrix algebra is used.

Another approach which aims to produce coherent forecasts directly is due to Pennings and van Dalen (2017), who propose the state space model

$$\begin{aligned} \boldsymbol{y}_t &= \boldsymbol{S}\boldsymbol{\mu}_t + \boldsymbol{\varepsilon}_t, && \boldsymbol{\varepsilon}_t \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_\varepsilon), \\ \boldsymbol{\mu}_t &= \boldsymbol{\mu}_{t-1} + \boldsymbol{\eta}_t, && \boldsymbol{\eta}_t \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_\eta). \end{aligned} \tag{3}$$

Variations are also considered, including covariates in the measurement Eq. (3). Coherent forecasts arise naturally using the Kalman filter, as discussed by Simon (2010). However, the covariance matrices are difficult to estimate with anything other than small hierarchies.

A related state space approach was proposed by Villegas and Pedregal (2018), who show that their formulation subsumes bottom-up, top-down, and some forms of forecast reconciliation and combination forecasting. In-built constraints to form coherent forecasts have also been considered in an ARIMA modelling context (Cholette, 1982; De Alba, 1993; Guerrero, 1989) and for exponential smoothing forecasts (Rosas & Guerrero, 1994).

4. Temporal and cross-temporal reconciliation

4.1. Early temporal aggregation papers

Studying the effects of temporal aggregation on forecasts goes back to the seminal works of Amemiya and Wu (1972), Tiao (1972), and Brewer (1973). Wei (1979), Lütkepohl (1984b), Lütkepohl (1986), Stram and Wei (1986), Hotta and Cardoso Neto (1993) and Rossana and Seater (1995) study the effect of temporal aggregation on seasonal and non-seasonal ARIMA processes, respectively, with aligned theoretical results. These generally show that aggregation to the annual frequency simplifies the dynamics of ARIMA processes generated at monthly or quarterly frequencies. They state that "quarterly data may be the best compromise among frequency of observation, measurement error, and temporal aggregation distortion". Such observations are not unusual. A similar conclusion is reached by Nijman and Palm (1990). Silvestrini and Veredas (2008) provide a detailed literature review, focusing on the implications of the model structure and identifiability for univariate ARIMA and multivariate GARCH processes. Pino et al. (1987) study the temporal and contemporaneous aggregation of scale and vector ARIMA processes and provide general results and forecast comparisons.

In parallel to the investigations in the econometric literature up to the early 2000s, temporal aggregation was becoming popular in high-frequency time series forecasting, albeit implicitly. That period's relatively limited computational resources forced researchers to develop clever ways to handle the long time series appearing in applications such as daily electricity load forecasting. The dominant approach had become to split the daily time series into seven weekly series, each corresponding to a stock measurement at a specific week of the day (Hippert et al., 2001).

From an accuracy standpoint, in the last decade, there has been a revival of temporal aggregation in the forecasting literature. Luna and Ballini (2011) study 111 weekly time series related to cash withdrawals. They examine forecasts generated by two top-down approaches and find considerable improvements compared to forecasts generated by the daily models directly. Motivated by intermittent demand forecasting, Nikolopoulos et al. (2011) proposed the ADIDA method, where a time series is temporally aggregated to a less intermittent level, forecasted, and subsequently disaggregated. The authors find promising accuracy gains, although only heuristic guidance exists for the temporal aggregation level. Spithourakis et al. (2011) demonstrate the benefits of using ADIDA in fast-moving consumer goods, while Spithourakis et al. (2014) attempt to develop the theoretical background for the method, investigating the aggregation and disaggregation mechanism, and trying to identify how to select well-performing temporal aggregation levels. Rostami-Tabar et al. (2013) and Rostami-Tabar et al. (2014) derive the optimal temporal aggregation level for AR(1), MA(1), and ARMA(1,1), when simple exponential smoothing is used to produce the forecasts.

Notably, the discussion so far has been on non-overlapping temporal aggregation. This form of aggregation acts as a moving average filter (Kourentzes et al., 2014) while substantially reducing the available sample size. Boylan and Babai (2016) investigate the effect of overlapping temporal aggregation, where a moving window is used to aggregate the time series, which is moved iteratively over the original series, including and dropping one observation at a time. They provide the conditions for which overlapping temporal aggregation outperforms its non-overlapping counterpart for independently and identically distributed demand processes. Earlier work on overlapping aggregation by Hotta et al. (1992) explore the effects on ARIMA, and by Mohammadipour and Boylan (2012) on INARMA models. Similarly, Petropoulos et al. (2016) motivated by intermittent demand problems, investigate empirically the usefulness of aggregating over unequal time periods, finding cases that can be beneficial. Babai et al. (2022) provide an extensive review of the aggregation literature.

The use of temporal aggregation in demand forecasting is natural, as we are typically interested in the demand of the lead time period, which lends itself to temporal aggregation and remains an active research area (for example, Rostami-Tabar et al., 2019; Saoud et al., 2022).

4.2. Temporal reconciliation

Motivated by the resurgence in the interest in temporal aggregation, following the work by Nikolopoulos et al. (2011), Kourentzes et al. (2014) propose the Multiple Aggregation Prediction Algorithm (MAPA), where a time series is modelled independently at multiple temporal aggregation levels with a state-space model, such as exponential smoothing, and then these models are combined by state. Temporal aggregation filters high-frequency components in the time series, making low-frequency ones more prominent. MAPA takes advantage of this to provide forecasts that perform well both in the short- and long-term forecast horizons. A by-product of the algorithm is that the resulting forecasts are coherent at the various temporal aggregation levels, which is a benefit the authors stress from a decision-making perspective. In follow-up work Kourentzes and Petropoulos (2016) investigate the impact of temporal aggregation on promotional indicator variables. They demonstrate the usefulness of the method in the presence of promotional information. Similar benefits are seen in the case of intermittent demand when relying on multiple temporal aggregation levels instead of a single level (Petropoulos & Kourentzes, 2015), as well as in inventory management (Barrow & Kourentzes, 2016; Petropoulos et al., 2019), again alluding to the connection between temporal aggregation and supply chain management. Kourentzes and Petropoulos (2016) compare optimally identified single temporal aggregation levels to using multiple levels. They find the latter preferable even when the underlying data-generating process is known. Using multiple aggregation levels results in combinations of forecasts, therefore reducing the modelling risk.

Athanasopoulos et al. (2017) combine the idea of using the multiple temporal aggregation levels of MAPA with hierarchical reconciliation, introducing the notion of temporal hierarchies. With temporal hierarchies, multiple temporal aggregation levels of a time series (typically up to the annual level) are constructed from the original series and modelled independently. Subsequently, the forecasts are reconciled using the approaches outlined previously.

Let the original time series y_t, with $t = 1, \dots, T$, be observed at a sampling frequency $1/m$ (e.g., $m = 12$ for monthly data). The aggregation levels $\{k_1, \dots, k_p\}$ are the p factors of m in ascending order, where $k_1 = 1$ and $k_p = m$. For each factor k of m, the non-overlapping temporally aggregated time series is constructed as:

$$x_j^{[k]} = \sum_{t=t^*+(j-1)k}^{t^*+jk-1} y_t,$$

where $j = 1, \dots, \lfloor T/k \rfloor$ and $t^* = T - \lfloor T/m \rfloor m + 1$, ensuring that all aggregation levels have complete aggregation windows. Note that $x_j^{[1]} = y_t$. The complete hierarchy progresses at the observation index of the most aggregate level, which we define as τ (this corresponds to j at that level). For each aggregation level, we stack the observations in column vectors

$$\boldsymbol{x}_\tau^{[k]} = \begin{bmatrix} x_{m_k(\tau-1)+1}^{[k]} \\ x_{m_k(\tau-1)+2}^{[k]} \\ \vdots \\ x_{m_k\tau}^{[k]} \end{bmatrix},$$

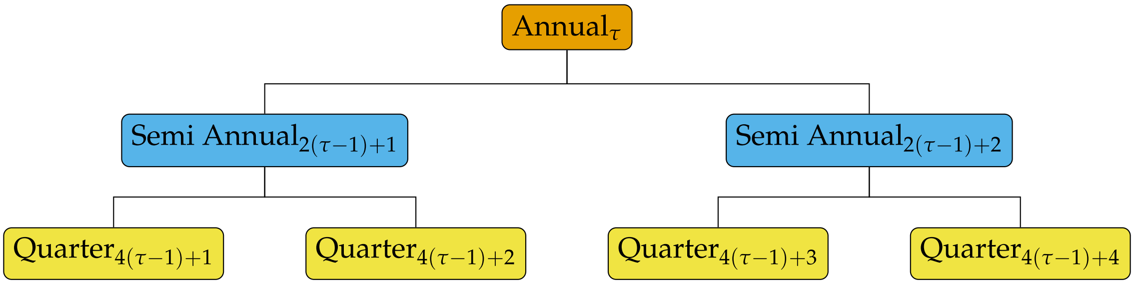

Fig. 5. A temporal hierarchy for a quarterly time series at year τ, with $m = 4$.

where $m_k = m/k$, $\tau = 1, \ldots, N$, and $N = T/m$. Collecting these in one column vector, we obtain

$$\boldsymbol{x}_{\tau} = \begin{bmatrix} x_{\tau}^{[m]} \\ \boldsymbol{x}_{\tau}^{[m-1]} \\ \vdots \\ \boldsymbol{x}_{\tau}^{[1]} \end{bmatrix}.$$

The structural representation becomes $\boldsymbol{x}_{\tau} = \boldsymbol{S}\boldsymbol{x}_{\tau}^{[1]}$ with $\boldsymbol{S} = \begin{bmatrix} \boldsymbol{A} \\ \boldsymbol{I} \end{bmatrix}$ and

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{1}'_m \\ \boldsymbol{I}_{m/k_{p-1}} \otimes \boldsymbol{1}'_{k_{p-1}} \\ \vdots \\ \boldsymbol{I}_{m/k_2} \otimes \boldsymbol{1}'_{k_2} \end{bmatrix}.$$

We have used $\boldsymbol{x}_{\tau}$ in the notation to clearly distinguish a temporal hierarchy from a cross-sectional one that uses $\boldsymbol{y}_t$. An example for a quarterly time series ($m = 4$) is provided in Fig. 5. If multiple seasonalities are not integer multiples of each other, the resulting additional temporal aggregations can simply be stacked in $\boldsymbol{x}_{\tau}$, and $\boldsymbol{A}$ can be extended accordingly.

An important difference between a temporal hierarchy and the data structure in its predecessor MAPA is that the use of the p factors of m simplifies the hierarchical structure and allows the direct implementation of hierarchical reconciliation, with the major advantage being that now there are no model restrictions. At each level, different forecasting models/methods can be used. Further, there is added flexibility in combining the forecasts between levels through $\boldsymbol{G}$. In contrast to cross-sectional hierarchies, temporal hierarchies can be constructed for any time series, requiring no additional data. However, this comes at the cost of estimation inefficiencies. Since there are only $N = T/m$ observations of the complete hierarchy, this significantly affects any estimation required in the approximation of $\boldsymbol{W}_h$, which motivated Athanasopoulos et al. (2017) to propose the aforementioned structural scaling that requires no estimation. Nystrup et al. (2020) propose and demonstrate the benefits of more advanced approximations that take advantage of the autocorrelation in the forecast errors when there is sufficient data and demonstrate their merits in a short-term electricity load forecasting application. Nystrup et al. (2021) go one step further and propose an estimator based on an eigendecomposition of the temporal correlation matrix, which can perform well even with relatively limited data compared with the dimension of the temporal hierarchy.

Athanasopoulos et al. (2017) provide simulation and empirical evidence of the accuracy benefits of forecasting with temporal hierarchies. These benefits increase with added modelling uncertainty and at more temporally aggregate levels, echoing the arguments and evidence in Kourentzes et al. (2014) and Kourentzes and Petropoulos (2016). The excellent performance of temporal hierarchies has been evidenced in various follow-up studies (for example, Jeon et al., 2019; Kourentzes et al., 2021; Nystrup et al., 2021, 2020; Yang, Quan, Disfani, & Rodríguez-Gallegos, 2017).

4.3. Cross-temporal reconciliation

Observe that cross-sectional hierarchies are described at time t for time series $\boldsymbol{y}_t$, while temporal hierarchies at time τ for $\boldsymbol{x}_{\tau}$ contain temporally aggregate views of $\boldsymbol{y}_t$. Given that hierarchical methods are motivated to support forecasting at various hierarchy levels, cross-sectional and temporal hierarchies may have limitations. For example, consider the case of forecasting for a grocery retailer. Let $y_t \in \boldsymbol{y}_t$ describe the sales of a particular ice cream product. Cross-sectionally, this may be grouped with other similar products or with product sales within a geographic demarcation, and so on. The further we aggregate, the less relevant a forecast becomes for the specific period t and at the granularity of y_t. Although we may be interested in the daily sales of a specific ice cream product, it is unlikely that we are interested in the daily sales at the top level of the hierarchy describing the total company sales. Similarly, from the temporal point of view, we are unlikely to be interested in the sales of y_t in time increments τ, for example, the sales of a particular ice cream product in several years. A more aggregate view across products and time units is typically more relevant for decision-makers, with many nodes in a hierarchy having the role of statistical devices that improve the quality of the overall coherent forecast rather than being directly connected with some supported decision (Athanasopoulos & Kourentzes, 2023).

Motivated by this, Kourentzes and Athanasopoulos (2019) proposed the notion of cross-temporal hierarchies, where the hierarchy spans across both cross-sections and time, more accurately mapping the various forecasts required by decision-makers and stakeholders. They show that sequential reconciliation across the cross-sectional and temporal dimensions, irrespective of order, does not always result in coherent forecasts, and address this by estimating all cross-sectional $\boldsymbol{G}_k$, across the k temporal aggregation levels, which are then averaged in a common cross-temporal $\bar{\boldsymbol{G}}$. They obtain holistically coherent forecasts, which are also the most accurate. In their experiments, the temporal reconciliation provided the biggest accuracy gain.

Motivated by the sequential algorithm of Kourentzes and Athanasopoulos (2019), Di Fonzo and Girolimetto (2023a) propose an iterative version whereby the forecasts are alternately reconciled in temporal and cross-sectional dimensions in a cyclic fashion and find that it produces more accurate forecasts.

Several contributions recognise potential accuracy benefits in reconciling across both cross-sectional and temporal dimensions. Spiliotis et al. (2020), Yagli et al. (2019), and Punia et al. (2020), apply sequentially temporal and cross-sectional approaches, with Yagli et al. (2019) experimenting with the order of reconciliation as well. They all identify accuracy benefits but do not establish holistic coherence. This sequential approach is discussed and improved by Di Fonzo and Girolimetto (2023b).

Rather than separately reconciling the cross-sectional and temporal dimensions, Di Fonzo and Girolimetto (2023a) proposed a single reconciliation step using the full cross-temporal hierarchy. Following our notation, from the cross-sectional $\boldsymbol{y}_t$ at the most temporally disaggregate level, let $y_{i,t}$ denote its ith element, $i = 1, \dots, n$. For each i, we construct all the temporally aggregated variants, giving a vector of length p:

$$\boldsymbol{x}_{i,\tau} = \begin{bmatrix} x_{i,\tau}^{[m]} \\ \vdots \\ \boldsymbol{x}_{i,\tau}^{[1]} \end{bmatrix}.$$

These can then be stacked into a long vector:

$$\boldsymbol{x}_{\tau} = \begin{bmatrix} \boldsymbol{x}_{1,\tau} \\ \vdots \\ \boldsymbol{x}_{n,\tau} \end{bmatrix}.$$

With $\boldsymbol{S}_{cs}$ and $\boldsymbol{S}_{te}$ denoting the structural matrices for the cross-sectional and temporal reconciliations respectively, the cross-temporal structural matrix is $\boldsymbol{S}_{ct} = \boldsymbol{S}_{cs} \otimes \boldsymbol{S}_{te}$, so that

$$\boldsymbol{x}_{\tau} = \boldsymbol{S}_{ct}\boldsymbol{b}_{\tau}^{[1]},$$

where the bottom-level series

$$\boldsymbol{b}_{\tau}^{[1]} = \begin{bmatrix} \boldsymbol{b}_{1,\tau}^{[1]} \\ \vdots \\ \boldsymbol{b}_{n_b,\tau}^{[1]} \end{bmatrix}.$$

Di Fonzo and Girolimetto (2023a) develop optimal cross-temporal reconciliation and evaluate it against the heuristic approach of Kourentzes and Athanasopoulos (2019) and variants. They report a relatively larger contribution to accuracy from the temporal side and find that the optimal approaches tend to be outperformed by the heuristic approaches. We note that in their experiments, all the approximations used for the cross-temporal $\boldsymbol{W}_h$ required some estimation, which, given the size of the cross-temporal matrices, may explain the findings (Pritularga et al., 2021). Building on the results of Di Fonzo and Girolimetto (2023a), Girolimetto et al. (2023) consider cross-temporally reconciled probabilistic forecasts.

Cross-temporally reconciled forecasts offer relatively limited accuracy gains compared to one-way reconciled forecasts (primarily temporally reconciled), with their major benefit being the qualitative difference of being coherent across both dimensions. This is impactful within a decision-making context and can be seen as a tool to sidestep organisational information silos and achieve aligned plans across different functions in organisations (Kourentzes & Athanasopoulos, 2019). Kourentzes (2022) argues that cross-temporally coherent forecasts offer a pathway towards so-called "one-number" forecasts, enabling the integration of independent forecasts built for different functions and decisions that are typically based on different information, with different horizons, and purposes. If these forecasts remain disconnected, they can lead to misaligned decisions and organisational friction. In the cross-temporal case, some hierarchy nodes are by-products of the structure rather than directly connected with some decision (Athanasopoulos & Kourentzes, 2023). This raises questions about how to best evaluate the quality of these forecasts, given that the relevant metrics for different decisions may vary in an organisational context.

5. Probabilistic hierarchical forecasting and reconciliation

The period that saw growth in the development of methods for hierarchical forecasting and reconciliation coincided with an increasing awareness of the importance of probabilistic forecasting. Therefore, it is unsurprising that several papers have attempted to tackle the problem of probabilistic hierarchical forecasting in recent years. As for point forecasting, methods for obtaining probabilistic hierarchical forecasts can be split into bottom-up, top-down and reconciliation approaches. However, some algorithms combine elements of more than one approach. Bayesian approaches to probabilistic forecast reconciliation are discussed in Section 3.8.

Bottom-up methods of probabilistic hierarchical forecasting were introduced by Ben Taieb, Taylor, et al. (2017) and subsequently expanded on in Ben Taieb et al. (2021). The algorithm is initialised by generating a Monte Carlo sample from each bottom-level variable's predictive distribution, which is independent by construction. These samples are first ranked and then permuted series-wise to induce dependence. The permutations are designed to ensure that the samples from bottom-level predictive distributions have the same empirical copula as 'in-sample' forecast errors (residuals), taking care to exploit the hierarchical structure to avoid dealing with very high-dimensional copulas. The rationale for matching the empirical copulas of the samples and residuals is to bring dependence information into the reconciliation procedure since this is known to work well in point forecasting. The samples are then aggregated to yield a sample from

$$\begin{array}{ccccccccccc} A\downarrow & B\downarrow & & A\circlearrowright & B\circlearrowright & & A & & B & & T \\ \begin{bmatrix}1.4\\2.3\\1.7\\2.1\end{bmatrix} & \begin{bmatrix}3.6\\5.3\\2.2\\6.4\end{bmatrix} & \overset{\text{rank}}{\Rightarrow} & \begin{bmatrix}1.4\\1.7\\2.1\\2.3\end{bmatrix} & \begin{bmatrix}2.2\\3.6\\5.3\\6.4\end{bmatrix} & \overset{\text{permute}}{\Rightarrow} & \begin{bmatrix}1.4\\1.7\\2.1\\2.3\end{bmatrix} & + & \begin{bmatrix}3.6\\2.2\\6.4\\5.3\end{bmatrix} & \overset{\text{aggregate}}{\Rightarrow} & \begin{bmatrix}5.0\\3.9\\8.5\\7.6\end{bmatrix} \end{array}$$

Fig. 6. A toy example describing bottom-up approaches for probabilistic forecasting for a simple 3-variable hierarchy with $T = A + B$. A sample of size $K = 4$ has been drawn for two bottom-level series A and B. These are then ranked from smallest to largest and permuted so that their empirical copula matches that of the residuals. The residuals are not shown here, but in this example, the smallest residual in A coincides with the second smallest residual in B, the second smallest residual in A coincides with the smallest residual in B and so on. Finally, A and B are aggregated to give a sample from the predicted distribution of the total series, T.

the predictive distribution of all top-level series. This is summarised with a simple example in Fig. 6.

While the bottom-up algorithm does not use a top-level forecast at all, both Ben Taieb, Taylor, et al. (2017) and Ben Taieb et al. (2021) propose an extension that incorporates top-level information by adjusting the mean of each series to be equal to a reconciled point forecast; for example, in Ben Taieb et al. (2021), MinT is used. A shortcoming of the bottom-up approach and its extension is that the sample drawn from the predictive distribution and the sample of training data must be of equal size, making it ill-suited to problems with a small amount of training data. This is overcome by Panamtash and Zhou (2018) and Zhao et al. (2019), who estimate predictive quantiles directly via quantile regression rather than using Monte Carlo.

Panamtash and Zhou (2018) also propose a top-down method for producing probabilistically coherent forecasts whereby quantile forecasts are first produced for all series. Proportions for top-down disaggregation are then found by taking the ratio of a forecast of a child node to the ratio of the forecast of the parent node. These are then applied to the original top-level forecasts. To the best of our knowledge, the only other top-down method for coherent forecasting has been proposed by Das et al. (2023), who model future proportions based on past proportions using a combination of an LSTM and multi-head self-attention architecture. A sample from the predictive distribution of the top level is generated, and each observation from this sample is disaggregated according to the forecast proportions.

Similar to the point forecasting case, progress has been made in extending the two-step reconciliation approach whereby probabilistic forecasts are produced from all series and then reconciled to be coherent in a second step. In the temporal reconciliation framework, Jeon et al. (2019) propose drawing a sample from the predictive distribution of each series (both top- and bottom-level) and then stacking these into a matrix. The matrix can then be pre-multiplied by a projection matrix $\boldsymbol{SG}$ to obtain a sample from the coherent multivariate predictive distribution. One algorithm proposed by Jeon et al. (2019), which they refer to as the 'ranked sample', orders the observations drawn from each predictive distribution before pre-multiplying by $\boldsymbol{SG}$. This approach is described in Fig. 7 and corresponds to reconciling quantiles, an idea with antecedents in Shang and Hyndman (2017) who reconcile prediction intervals. Quantiles are only preserved under linear combinations when the data are perfectly dependent (Kolassa, 2023). However, in cases where dependence between series is high, the method of Jeon et al. (2019) performs well.

Panagiotelis et al. (2023) make a number of contributions to probabilistic forecast reconciliation by providing formal definitions for coherence and reconciliation that justify Monte Carlo approaches as well as finding reconciled probabilistic forecasts for elliptical distributions (including the Gaussian). The Panagiotelis et al. (2023) framework allows any set of base forecasts, either univariate or multivariate (the latter until then had not been considered in the hierarchical forecasting literature), to be reconciled using any reconciliation method. Reconciliation weights are trained by optimising with respect to a multivariate scoring rule.

Rangapuram et al. (2021) also optimise with respect to scoring rules for probabilistic forecasts, but rather than train reconciliation weights, they assume least squares reconciliation, with the novelty coming from training forecasting models and reconciling in an end-to-end fashion rather than via the usual two-step approach.

Building on the work of Panagiotelis et al. (2023) and Di Fonzo and Girolimetto (2023a), Girolimetto et al. (2023) consider cross-temporally reconciled probabilistic forecasts. They propose the parametric Gaussian and non-parametric bootstrap approaches to draw samples from an incoherent cross-temporal distribution. They explore and study various strategies to overcome estimation issues related to the high dimensionality of the cross-temporal setting.

Despite very recent progress in forecast reconciliation, a number of open research questions remain. In the case of point forecasting, the optimality of certain reconciliation techniques, such as MinT, is now well understood. Similar results have not been derived for reconciliation methods in the probabilistic setting with the exception of Wickramasuriya (2023), who derive the optimality of MinT in the Gaussian case. We expect a better understanding of the theoretical and empirical properties of different probabilistic forecast reconciliation techniques in coming years, including which methods produce well-calibrated forecasts and whether their prediction intervals achieve correct coverage rates.

6. Significant applications

6.1. Tourism

The new concept of forecast reconciliation was first implemented on tourism data in Athanasopoulos et al.

$$\begin{array}{ccc} T\downarrow & A\downarrow & B\downarrow \\ \begin{bmatrix}6.2\\8.3\\3.6\\3.4\end{bmatrix} & \begin{bmatrix}1.4\\2.3\\1.7\\2.1\end{bmatrix} & \begin{bmatrix}3.6\\5.3\\2.2\\6.4\end{bmatrix} \end{array} \overset{\text{rank}}{\Longrightarrow} \begin{array}{ccc} T & A & B \\ \begin{bmatrix}3.4\\3.6\\6.2\\8.3\end{bmatrix} & \begin{bmatrix}1.4\\1.7\\2.1\\2.3\end{bmatrix} & \begin{bmatrix}2.2\\3.6\\5.3\\6.4\end{bmatrix} \end{array}$$

$$\overset{\text{stack}}{\Downarrow}$$

$$\begin{array}{c} T \\ A \\ B \end{array}\begin{bmatrix}3.4 & 3.6 & 6.2 & 8.3\\ 1.4 & 1.7 & 2.1 & 2.3\\ 2.2 & 3.6 & 5.3 & 6.4\end{bmatrix} \overset{\text{reconcile}}{\Longrightarrow} \boldsymbol{SG}\begin{bmatrix}3.4 & 3.6 & 6.2 & 8.3\\ 1.4 & 1.7 & 2.1 & 2.3\\ 2.2 & 3.6 & 5.3 & 6.4\end{bmatrix}$$

Fig. 7. A toy example describing the ranked sample approach to probabilistic forecast reconciliation for a three-variable hierarchy where $T = A + B$. A sample of size $K = 4$ has been drawn for the top-level series T and two bottom-level series A and B. These are then ranked from smallest to largest, then stacked, and reconciliation is applied (final reconciled forecasts depend on the choice of **G** and are not shown here).

(2009). Tourism flows comprise aggregation structures across various dimensions. The most obvious of these are geographic hierarchies. At an international level, we have inbound travel from multiple countries or regions within source countries to multiple destination countries, airports, or regions within the destination countries. The same applies to outbound travel, while similar geographic divisions are natural for domestic travel. Further, policymakers and business planners are also interested in various characteristics of tourists. For example, holidaymakers' expenditure patterns differ significantly from business travellers or visiting friends and relatives. Hence, grouped structures where geographic hierarchies are crossed with various attributes of interest also naturally arise. An attribute commonly observed for tourism flows is the purpose of travel, which typically comprises holidays, visiting friends and relatives, business, and others.

Athanasopoulos et al. (2009) focus on two aggregation structures based on quarterly tourism flows. A grouped structure where geographic divisions (Australia, States, Capital city versus other) are crossed with the purpose of travel (also considered in Hyndman et al., 2011); and a pure geographic hierarchy where States, Zones and Regions disaggregate Australian domestic tourism flows. Coherent forecasts were generated from traditional bottom-up and top-down approaches based on historical proportions and a new top-down approach based on forecast proportions and reconciliation. The paper found that the proposed top-down and reconciliation approaches improve forecast accuracy compared to the traditional approaches and provide detailed forecasts for Australian domestic tourism flows, identifying some key features at both the aggregate and disaggregate levels crucial for informing policymakers. Variations of these Australian tourism data have since become ubiquitous for benchmarking forecast reconciliation methods. For example, Abolghasemi, Tarr, et al. (2022) propose what they refer to as conditional hierarchical forecasting, an approach based on machine learning classification methods that use time series features to select the reconciliation method for a hierarchy and evaluate the performance of the method based on the pure geographic hierarchy. Gleason (2020) introduces an embedding reconciliation term that penalises deviation from an aggregation structure and uses the grouped structure to evaluate the forecasting performance of the proposed method, claiming improvements over MinT.

An updated and richer monthly Australian tourism data set was introduced in Wickramasuriya et al. (2019). The geographic hierarchy comprises 111 series. In particular, the total tourism flow is disaggregated into seven states, 27 zones and 76 regions. The hierarchical structure is crossed with the four purposes of travel, resulting in a total of 555 time series, of which 525 are unique (Di Fonzo & Girolimetto, 2023, online appendix). This data set, or close variations of it, is considered in several studies, including Kourentzes and Athanasopoulos (2019), Hollyman et al. (2021), Spiliotis et al. (2021) and Di Fonzo and Girolimetto (2023).

Karmy and Maldonado (2019) study ten different data sets, including three related to tourism. The first one is a variation of the quarterly data set introduced in Athanasopoulos et al. (2009), using only the hierarchical structure based on the geographic divisions (also used in Karmy et al., 2023). They also consider a weekly dataset of flight passengers between Melbourne and Sydney for an airline, disaggregated by flight ticket class type (first class, business, and economy), and monthly departures from Australia disaggregated into permanent, long-term, and short-term departures, and then between residents and visitors for the cases of long- term and short-term departures. The paper presents forecasting methods for hierarchical time series based on Support Vector Machines. These are compared to traditional single-level approaches with the conclusion that a major limitation of the proposed methods is the lack of data in a time series context.

Athanasopoulos et al. (2023) and Kourentzes et al. (2021) focus on tourism flows amid the COVID-19 pandemic. Athanasopoulos et al. (2023) model both international inbound and domestic flows for the case of Australia, while Kourentzes et al. (2021) analyse international arrivals for the case of South Africa. The papers argue for using forecast reconciliation to generate robust forecasts for tourism flows during the pre-COVID period.

6.2. Macroeconomics

Since macroeconomics studies aggregate economic phenomena, it is unsurprising that this field has provided fertile ground for hierarchical data. For example, Gross Domestic Product (GDP) is constructed as an aggregate of individual components. The expenditure method constructs GDP as an aggregate of expenditure characterised by consumption, investment, imports and exports. In contrast, the income approach aggregates variables such as the gross operating surplus of firms with employee compensation. Thus, there are two hierarchies involved with the same aggregated series. The structural formulation of the reconciliation problem does not allow for this scenario, but the more general constraint formulation (Section 2.2) does.

As noted in Section 3.1, reconciliation of estimates (distinct from forecast reconciliation) has a long history in macroeconomics. For forecasting problems, forecasts of the disaggregate series may be of direct interest. Otherwise, the objective may be to improve accuracy by leveraging forecasts of the disaggregate series via reconciliation. Athanasopoulos et al. (2020), Bisaglia et al. (2020), and Di Fonzo and Girolimetto (2022c) find evidence in favour of forecast reconciliation in both of these settings for Australian GDP. In particular, for both point and probabilistic forecasts, improvements in forecast accuracy over base and bottom-up methods can be achieved using forecast reconciliation. The MinT method works best overall, while when attention is restricted to only the top-level series, weighted least squares are more accurate. Di Fonzo and Girolimetto (2023a) use the constraint representation with a cross-temporal framework and show that it leads to better forecasts on this data set.

Another important macroeconomic variable that admits a hierarchical structure is the consumer price index (CPI). The CPI is a weighted price index of a basket of goods. The novel aspect of reconciliation in this setting is that rather than a summing matrix consisting only of ones and zeros, the $\boldsymbol{S}$ matrix includes weights that can change over time. In an early application of reconciliation methodology, Capistrán et al. (2010) find that OLS reconciliation can improve upon a bottom-up approach for some but not all components of Mexican CPI. For overall CPI, OLS improves upon a bottom-up approach. However, the difference between the forecast accuracy of these methods is not found to be statistically significant. Weiss (2018) considers UK CPI and finds that reconciliation using OLS performs better than traditional approaches for 1-month-ahead forecasts but that middle-out approaches work better for longer horizons. Weiss (2018) also considers inflation volatility; in this case, reconciliation methods do not outperform bottom-up.

While hierarchical structures are common in macroeconomics, certain details concerning the construction of these datasets suggest new directions in reconciliation methodology. One example, Koop et al. (2023) consider a geographic hierarchy of productivity, with a model that includes growth rates in national and regional output. However, since chain volume measures are used to construct output with different price deflators for each region, regional growth rates only add up to the national growth rate as an approximation. Since the usual aggregation constraint only holds approximately, Koop et al. (2023) recommend shrinking towards the aggregation constraint via a Bayesian approach rather than imposing a hard constraint. Considering unemployment data from multiple labour force surveys in Brazil, Lila et al. (2022) introduce robust estimation in the reconciliation stage with the primary aim to address measurement issues occurring in the original time series. These methods will likely generalise to other hierarchical forecasting problems in macroeconomics and other disciplines.

6.3. Energy

Energy applications are widespread in the forecast reconciliation literature due to the natural geographic hierarchies that arise in energy distribution. For example, both the GEFCom2012 (Hong et al., 2014) and GEFcom2017 (Hong et al., 2019) energy forecasting competitions included hierarchical electricity load data, although none of the participants took advantage of the hierarchical structure to improve their forecasts.

One of the earliest uses of forecast reconciliation applied to energy data was van Erven and Cugliari (2015), discussed earlier, who applied their methods to electricity demand data from Électricité de France, disaggregated into 17 tariff groups. Other early uses of reconciliation methods for forecasting short-term (24 h) electricity demand include Almeida et al. (2016) who applied OLS reconciliation to a hierarchy disaggregated by grid supply point and voltage level, da Silva et al. (2019) who use Bayesian estimation, and Meira et al. (2023) who propose methods robust to the influence of outliers. Further examples of applications to load forecasting include Feng and Zhang (2020), who compared point forecasts obtained from bottom-up, OLS and MinT on hourly load data from 13 buildings in Texas, USA and Nespoli et al. (2020) who applied reconciliation methods to electricity load data, comparing several of the probabilistic forecasting methods discussed in Section 5 when applied to a small set of 24 power meters located in Rolle, Switzerland.

Three interesting examples of forecast reconciliation of electricity load are due to Ben Taieb and his co-authors. Ben Taieb, Yu, et al. (2017) propose a regularised version of MinT reconciliation with penalties analogous to those used in LASSO and elastic-net regressions, giving sparse adjustments to the base forecasts. They apply the method to electricity consumption measured by 5701 smart meters with a rich geographic hierarchy. Ben Taieb, Taylor, et al. (2017) and Ben Taieb et al. (2021) each developed new probabilistic reconciliation methods (discussed in Section 5), and applied them to the same dataset.

Zhao et al. (2019) proposed a computationally simpler variation of the method of Ben Taieb, Taylor, et al. (2017) and applied it to two public data sets: the ISO New England data from Hong et al. (2019) and some Irish smart meter data aggregated to groups of 100 customers. Roach (2019) also studies the ISO New England data from the Hong et al. (2019) and proposes a method based on generating reconciled quantile forecasts using a gradient-boosted model, which is shown to outperform the benchmark.

Brégère and Huard (2022) also tackle load forecasting but introduce an innovation by including an "aggregation algorithm" before reconciling the forecasts. This aggregation algorithm involves computing revised base forecasts that are linear combinations of the base forecasts of all series. The revised base forecasts are then reconciled using OLS reconciliation. This approach means any cross-sectional relationships between series are modelled in the aggregation step rather than the reconciliation step. The authors applied this approach to the same dataset used by Ben Taieb, Taylor, et al. (2017).

Forecast reconciliation has also been used in solar generation forecasts. Here, power generated by distributed photovoltaics (PV) is naturally disaggregated in a geographic hierarchy such as transmission zones, distribution nodes, PV plants, subsystems and inverters. Yang, Quan, Disfani, and Liu (2017) explored the application of MinT reconciliation and some of its special cases to hourly generation data from 318 power plants in California, USA. In contrast, Yagli et al. (2020) applies probabilistic (Gaussian) MinT reconciliation to the same data set. The data set is used again in Yang, Quan, Disfani, and Rodríguez-Gallegos (2017), applying temporal reconciliation methods, and in Yagli et al. (2019), where both cross-sectional and temporal reconciliation are considered. They apply the cross-sectional and temporal reconciliations sequentially rather than simultaneously. Di Fonzo and Girolimetto (2022a) and Di Fonzo and Girolimetto (2023a) critique this two-step approach and argue for simultaneous cross-temporal reconciliation but show that the two approaches can be equivalent if the covariance matrices used in both steps are constant across levels and time granularities. They further show how the forecasts can be constrained to be non-negative using the simple approach of setting any negative bottom-level reconciled forecasts to zero and then aggregating the results. Another application to solar power is Panamtash and Zhou (2018), who applied the probabilistic reconciliation methods they developed (discussed in Section 5) to 5-minute solar power data for about 6000 simulated PV plants in Florida, USA.

Forecast reconciliation has been extensively applied to wind power forecasting; see Jeon et al. (2019), Bai and Pinson (2019) and Hansen et al. (2023). Also of particular note is Gilbert et al. (2018), who produce probabilistic wind power forecasts for individual turbines and the entire wind farm. An innovative feature is the use of a weighted aggregation based on an elastic net penalised regression.

Other applications of forecast reconciliation in the general field of energy include Buzna et al. (2021) who consider probabilistic forecasting of loads at electric vehicle charging stations, and Bergsteinsson et al. (2021) who consider heat load forecasting. The latter propose an adaptive reconciliation method for temporal hierarchies by allowing time-varying weights.

6.4. Mortality rates

Applications of forecast reconciliation in mortality started with Shang and Hyndman (2017), who forecast Japanese mortality rates disaggregated by age, sex, and a geographic hierarchy of 47 prefectures within eight regions. The base forecasts were obtained using a functional data method. Because mortality rates do not sum directly, they proposed an aggregation matrix $\boldsymbol{A}$ comprising population ratios. For example, for the mortality rate of 50-year-old females within a region, the non-zero values of the corresponding row of the $\boldsymbol{A}$ matrix contain the 50-year-old female population of each prefecture divided by the total 50-year-old female population of the region. Thus, the aggregation matrix $\boldsymbol{A}$ is time-varying, and the values for the future time periods were forecast using univariate time series models. WLS reconciliation was used. Shang and Haberman (2017) provide an application for annuity pricing using an identical approach to the same data set. Li and Hyndman (2021) used a similar formulation but applied MinT reconciliation to US age-sex-specific mortality rates and explored future mortality inequality.

Li et al. (2019) consider forecasting mortality due to different causes of death and show that forecast accuracy is improved by reconciliation. Although the data form a simple two-level hierarchy with bottom-level series and their aggregate, the authors combine individual causes of death into middle-level series. This is done via hierarchical clustering of the data. Augmenting the hierarchy with middle-level series in this fashion further improves forecast accuracy.

Li and Tang (2019) use a forecast reconciliation approach to forecast a longevity divergence index (LDI) used to compute the Swiss Re Kortis bond. The most disaggregated series in the hierarchy are age-specific mortality improvement rates in the US and UK. In contrast, the most aggregated series are LDI values, expressed as linear combinations of these disaggregated series. They use a MinT approach to generate probabilistic forecasts, following Jeon et al. (2019).

6.5. Retail demand & supply chain

Forecasting for demand planning has been an attractive application for hierarchical forecasting, and therefore, it has attracted attention in the literature (Dannecker et al., 2013). Rostami-Tabar et al. (2015) look at the conditions where top-down or bottom-up is favourable, assuming that the disaggregate series are ARIMA(0,1,1) processes, using the reconciliation approach as a benchmark. The latter is found to be more accurate overall, with the bottom-up method outperforming it in some cases, particularly for the bottom-level time series. Yang et al. (2016) using the publicly available Dominick's Finer Food dataset, find that the shrinkage estimator for $\boldsymbol{W}_1$ performed better than classic hierarchical methods (bottom-up and top-down). Oliveira and Ramos (2019) find the same conclusion on data from a Portuguese supermarket. Mircetic et al. (2022) look at sales of a major European brewery and find that hierarchical reconciliation performs better than base forecasts. They also propose combining the bottom-level forecasts of different hierarchical forecasting methods and constructing them for the rest of the hierarchy using a bottom-up approach. They argue that this method has the advantage that it eliminates the need to select a hierarchical approach and find small gains over the reconciliation method. Karmy and Maldonado (2019) explore the performance of hierarchical forecasting on sales in the travel retail industry. They do not consider the reconciliation approach and find bottom-up to be best.

Villegas and Pedregal (2018) propose encapsulating cross-sectional hierarchical reconciliation in a state-space formulation with the forecasting model. Using simulations and an empirical investigation on a Spanish grocery retailer, they find that the standard reconciliation performs well for short horizons (1–3 days), while for longer horizons (4–7 days), the state-space-based reconciliation is best and is also the overall most accurate method.

Other applications include the contributions of Abolghasemi, Hyndman, et al. (2022), Abolghasemi, Tarr, et al. (2022), and Spiliotis et al. (2021) who investigate the application of hierarchical forecasting on data from a food manufacturer in Australia. Finally, the M5 forecasting competition used data from Walmart, a major US retailer, with a grouped time series structure, providing a test bed for reconciliation methods (Makridakis et al., 2022). We provide further details of these works in Section 3.7.

6.6. Intermittent demand

Hierarchical forecasting has seen some application in intermittent demand modelling. However, in reviewing these papers, it is useful to consider that the literature has progressed substantially over the last years in terms of how to evaluate forecasts of intermittent demand time series (Athanasopoulos & Kourentzes, 2023; Kolassa, 2016; Kourentzes, 2014), recognising that classic error metrics, especially those based on absolute errors, are often inappropriate. Direct evaluation of decision metrics, such as inventory or predictive distribution, is preferable.

One of the earliest works using hierarchical forecasting for intermittent demand is by Moon et al. (2012), who looked at top-down hierarchical forecasts against combination and base forecasts for predicting spare parts for the South Korean Navy. In the reported inventory cost, the top-down approach offers benefits in some settings but is outperformed by combination methods. It should be noted that the researchers constructed the hierarchy used in this work, and its eventual structure may have been significant for the findings.

Petropoulos and Kourentzes (2015) provide a translation of the temporal MAPA algorithm (Kourentzes et al., 2014) for intermittent demand. It is used to forecast spare parts for the UK Royal Air Force. They find it outperforms various benchmarks, including ADIDA that relies on a single temporal aggregation level (Nikolopoulos et al., 2011), and various combinations of forecasts. However, the empirical evaluation lacks decision or predictive distribution-related metrics.

Li and Lim (2018) provide a top-down-like hierarchical forecasting method for predicting intermittent demand in fashion retailing. Their approach produces separate daily forecasts of the aggregate demand across multiple items and a forecast of the inter-demand interval and demand size for each individual item. The latter forecasts are used to prorate the total forecast into the individual items. Although the proposed algorithm performs well against benchmarks, the empirical evaluation lacks other cross-sectional reconciliation benchmarks and strong performance metrics.

Kourentzes and Athanasopoulos (2021) investigate using temporal hierarchies for intermittent demand forecasting for aerospace spare parts. Their motivation is that a method that predicts well the intermittent pattern should be able to demonstrate the various patterns (such as seasonality and trend) that may appear when the data are explored at lower sampling frequency. They demonstrate that using temporal hierarchies allows capturing these patterns at higher temporal aggregation levels. Combining this information with intermittent demand forecasts of the original time series results in reconciled forecasts that dominate the base forecasts on various metrics and horizons. They obtain prediction intervals using the empirical distribution of the reconciled forecasts and provide a heuristic to ensure non-negative forecasts. The use of temporal hierarchies in an intermittent demand setting has more recently been considered by Sanguri et al. (2023) who focus on the estimation of forecasting downward trends in a scenario of obsolescence.

6.7. Healthcare, accidents & emergencies

Athanasopoulos et al. (2017) applied temporal hierarchies to predict weekly admissions for Accidents & Emergency wards in UK hospitals. The volume of patients relates to different decisions in the operations of the wards, from staff scheduling to procuring consumables, training and hiring of staff, etc. They showed that temporally coherent forecasts dominated base forecasts in all cases.

Pritularga et al. (2021) looked at weekly Accidents & Emergency cases for a specific hospital and were interested in producing cross-sectionally coherent forecasts across various patient demarcations. They compared a variety of approximations for $\boldsymbol{W}_h$, controlling for the sample size, and found that hierarchical forecasts were universally better than the base forecasts. They also showed that the complexity of the approximation is important, with simpler ones performing best at smaller sample sizes and more complex ones gaining a substantial advantage when there were sufficiently long time series.

Weiss (2018) investigated improving the staffing for a large UK teaching hospital. Hierarchical forecasts were found to provide more accurate forecasts than the benchmark used by the hospital. Further, when the forecasts were used in a staffing model, they resulted in cheaper operations and less under-staffing.

Gibson et al. (2021) forecast weighted influenza-like illness (wILI), with weights corresponding to the population size of different U.S. regions. Enforcing probabilistic coherence increases forecast skill for most models when tested over multiple flu seasons. This provides further evidence of the benefits of using forecast reconciliation with geographical hierarchies.

7. Open-source software implementations

The first available open-source implementation of forecast reconciliation methods was the `hts` package for R (Hyndman et al., 2010), which helped popularise cross-sectional point forecast reconciliation methods in business and industry. The `hts` package has continued to be developed, and its latest version (Hyndman et al., 2021) includes implementations of Wickramasuriya et al. (2019) and Wickramasuriya et al. (2020). The game-theoretic approach of van Erven and Cugliari (2015) is implemented in the `gtop` package for R (Cugliari & van Erven, 2015). Temporal point forecast reconciliation is provided in the `thief` package for R (Hyndman & Kourentzes, 2018). Cross-sectional, temporal, and cross-temporal point forecast reconciliation with optional non-negativity constraints is provided by the `FoReco` package for R (Girolimetto & Di Fonzo, 2022). The score optimisation approach of Panagiotelis et al. (2023) is implemented in the `ProbReco` package (Panagiotelis, 2020). Probabilistic cross-sectional forecast reconciliation is also included in the `fabletools` package for R (O'Hara-Wild et al., 2023), with a simple user interface for specifying complicated hierarchical and grouping structures. The `Bayesrecon` package (Azzimonti et al., 2023) implements reconciliation methods based on conditioning such as those proposed by Corani et al. (2021).

There are also several Python implementations of the methods, including the `pyhts` package of Zhang et al. (2022), which is a Python translation of Hyndman et al. (2021), and Darts (Herzen et al., 2022, 2023) which provides similar functionality. The `hierarchicalforecast` package (Olivares, Garza, et al., 2022; Olivares et al., 2023) provides a more comprehensive suite of functions covering both point and probabilistic forecast reconciliation including the methods of Ben Taieb, Taylor, et al. (2017), Ben Taieb and Koo (2019), Wickramasuriya et al. (2019), and Panagiotelis et al. (2023). The method of Rangapuram et al. (2021) is implemented in `GluonTS` (Alexandrov et al., 2023).

Many datasets studied in the hierarchical forecasting literature are publicly available. We provide references and discuss some specific examples that have been used extensively across several studies. Monthly tourism data analysed in Wickramasuriya et al. (2019) (and many follow-up studies) are available at https://robjhyndman.com/publications/mint. A quarterly version used in Hyndman and Athanasopoulos (2021) is available in the `tsibble` package (Wang, Cook, & Hyndman, 2020; Wang, Cook, et al., 2022). Quarterly GDP data, originally studied in Athanasopoulos et al. (2020) and most recently in Girolimetto et al. (2023), are available from the latter paper's public GitHub repository.[4] Quarterly prison data introduced in Hyndman and Athanasopoulos (2021) are available on the textbook website.[5] Daily electricity data introduced in Panagiotelis et al. (2023) can be downloaded from the paper's public GitHub repository.[6] Store level Walmart retail data used in the M5 competition (Makridakis et al., 2022) can be found on the Kaggle website.[7]

8. Conclusion

Research into hierarchical time series and forecast reconciliation has seen great success and impact, particularly over the last decade. We wish to speculate on what the next decade holds by identifying some key open questions. We anticipate growth in the following areas.

The sheer size of some hierarchical forecasting problems can lead to computational difficulties. It is common to have n_b, the number of the most disaggregated series, well over 1 million. The number of aggregated series can be even larger with many grouping factors. This leads to very large matrices that need to be inverted, even when using the more efficient constraint matrix approach of Section 2.2. Sparse matrix algebra has helped considerably with this problem, and when coupled with the Lanczos algorithm (as implemented by Paige & Saunders, 1982), very large problems can be handled (see Hyndman et al., 2011). For a more modern treatment of sparse algebra, see also Fong and Saunders (2011) and Davis et al. (2016). However, further improvements may be possible using dimension reduction methods along the lines of Wang, Chen, et al. (2020).

A related problem arises with the MinT approach, where the covariance matrix of the base forecasts needs to be computed. Here, shrinkage methods have been used to good effect (Wickramasuriya et al., 2019), but these don't scale to very large matrices. Alternative sparse or low-rank approaches (Lam, 2020) would be a welcome addition to the literature.

The breadth of data to which reconciliation methods need to be applied necessitates extending methods to non-standard domains. This includes non-negative data (for which there is existing work by Wickramasuriya et al., 2020, amongst others), discrete data (for which Corani et al. (2021), Olivares, Meetei, et al. (2022) and Zambon et al. (2022) give early attempts to address the issues), and finally mixtures of discrete and continuous data. The latter could be potentially useful for zero-inflated data (which arise in intermittent sales data) or where there are hierarchies where some bottom-level series are best modelled as discrete, while the top-level series are best modelled as continuous. The development of algorithms to handle these cases, as well as understanding the theoretical properties of reconciliation in such settings, represents a bold research agenda.

While recent progress has been made in probabilistic forecasting, as discussed in Section 5, there are a number of open questions on the properties of probabilistic forecast reconciliation. For example, what are the coverage properties of prediction intervals derived from reconciled forecasts, how do these depend on the coverage properties of the base forecasts themselves and does reconciliation even improve coverage relative to base forecasts?

4 https://github.com/daniGiro/ctprob
5 https://OTexts.com/fpp3/extrafiles/prison_population.csv
6 https://github.com/PuwasalaG/Probabilistic-Forecast-Reconciliation/tree/master/EnergyApplication/Data
7 https://kaggle.com/competitions/m5-forecasting-accuracy

These questions are particularly vexing since reconciled probabilistic forecasts are defined on a domain that is a linear subspace.

An often-stated advantage of coherent forecasts is that they have the potential to lead to aligned decisions. However, the attempts to quantify this effect have been limited. Where gains in forecast accuracy due to forecast reconciliation have been established, forecast evaluation is often based on general-purpose metrics such as RMSE and MAE as well as scaled versions thereof. These metrics do not explicitly penalise incoherence in forecasts, particularly when, as is often the case, the forecasts of different variables are evaluated individually. The disconnect between metrics of forecast evaluation and the operational considerations of hierarchical forecasting may explain why hierarchical methods have not been popular in forecasting competitions such as the M5, even where the data follow a hierarchical structure. Therefore, we anticipate the development of new forecast evaluation metrics that account for the multivariate and hierarchical nature of the data. Further, we concur with the view of Athanasopoulos and Kourentzes (2023) that forecast evaluation must be integrated with the decisions made by agents at different levels of the hierarchies.

This issue opens up additional questions related to the game theoretic aspects of hierarchical forecasting. Most empirical work shows that even where reconciliation improves forecast accuracy, these improvements do not occur across all hierarchy levels. In some applications, improvements may be seen in forecasts of bottom-level series after reconciliation, while base forecasts at the top level still outperform the top-level reconciled forecast. In other applications, the reverse may be true. This has a number of interesting implications in an organisational setting where those making forecasts at different levels can be treated as separate agents. Can reconciliation methods be found that lead to Pareto improvements across the hierarchy, meaning all agents gain from reconciliation? If not, in a cooperative setting, can forecast reconciliation and the decisions agents make based on these forecasts be aligned to improve the overall welfare of the organisation? Also, in a competitive setting, can compensation mechanisms be developed to encourage agents at different levels of the hierarchy, each making forecasts based on their own information sets to share base forecasts for reconciliation? These open questions should stimulate research for years to come.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank Tommaso Di Fonzo, Xiaoqian Wang and Daniele Girolimetto for providing helpful comments on an earlier draft of this paper. Rob J Hyndman was funded by the Australian Government through the Australian Research Council Industrial Transformation Training Centre in Optimisation Technologies, Integrated Methodologies, and Applications (OPTIMA), Project ID IC200100009.

References

Abolghasemi, M., Hyndman, R. J., Spiliotis, E., & Bergmeir, C. (2022). Model selection in reconciling hierarchical time series. *Machine Learning*, *111*(2), 739–789. http://dx.doi.org/10.1007/s10994-021-06126-z.

Abolghasemi, M., Hyndman, R. J., Tarr, G., & Bergmeir, C. (2019). Machine learning applications in time series hierarchical forecasting. URL https://arxiv.org/abs/1912.00370.

Abolghasemi, M., Tarr, G., & Bergmeir, C. (2022). Machine learning applications in hierarchical time series forecasting: Investigating the impact of promotions. *International Journal of Forecasting 40*(2), 597–615.

Ahmed, R. A. (2009). *Forecasting hierarchical time series* (Ph.D. thesis), Monash University.

Alexandrov, A., Benidis, K., Bohlke-Schneider, M., Flunkert, V., Gasthaus, J., Januschowski, T., Maddix, D. C., Rangapuram, S., Salinas, D., Schulz, J., Stella, L., Türkmen, A. C., & Wang, Y. (2023). GluonTS: Probabilistic time series modeling in Python. Python package v0.11.8 URL https://ts.gluon.ai/.

Almeida, V., Ribeiro, R., & Gama, J. (2016). Hierarchical time series forecast in electrical grids. In K. J. Kim, & N. Joukov (Eds.), *Information science and applications (ICISA)* (pp. 995–1005). Singapore: Springer, http://dx.doi.org/10.1007/978-981-10-0557-2.

Amemiya, T., & Wu, R. Y. (1972). The effect of aggregation on prediction in the autoregressive model. *Journal of the American Statistical Association*, *67*(339), 628–632. http://dx.doi.org/10.1080/01621459.1972.10481264.

Anderer, M., & Li, F. (2022). Hierarchical forecasting with a top-down alignment of independent level forecasts. *International Journal of Forecasting*, *38*(4), 1405–1414. http://dx.doi.org/10.1016/j.ijforecast.2021.12.015.

Ashouri, M., Hyndman, R. J., & Shmueli, G. (2022). Fast forecast reconciliation using linear models. *Journal of Computational & Graphical Statistics*, *31*(1), 263–282. http://dx.doi.org/10.1080/10618600.2021.1939038.

Athanasopoulos, G., Ahmed, R. A., & Hyndman, R. J. (2009). Hierarchical forecasts for Australian domestic tourism. *International Journal of Forecasting*, *25*(1), 146–166. http://dx.doi.org/10.1016/j.ijforecast.2008.07.004.

Athanasopoulos, G., Gamakumara, P., Panagiotelis, A., Hyndman, R. J., & Affan, M. (2020). Hierarchical forecasting. In P. Fuleky (Ed.), *Macroeconomic forecasting in the era of big data* (pp. 689–719). Springer, http://dx.doi.org/10.1007/978-3-030-31150-6_21.

Athanasopoulos, G., Hyndman, R. J., Kourentzes, N., & O'Hara-Wild, M. (2023). Probabilistic forecasts using expert judgement: the road to recovery from COVID-19. *Journal of Travel Research*, *62*(1), 233–258. http://dx.doi.org/10.1177/00472875211059240.

Athanasopoulos, G., Hyndman, R. J., Kourentzes, N., & Petropoulos, F. (2017). Forecasting with temporal hierarchies. *European Journal of Operational Research*, *262*(1), 60–74. http://dx.doi.org/10.1016/j.ejor.2017.02.046.

Athanasopoulos, G., & Kourentzes, N. (2023). On the evaluation of hierarchical forecasts. *International Journal of Forecasting*, *39*(4), 1502–1511. http://dx.doi.org/10.1016/j.ijforecast.2022.08.003.

Azzimonti, D., Rubattu, N., Zambon, L., & Corani, G. (2023). bayesRecon: Probabilistic reconciliation via conditioning. R package version 0.1.2 URL https://CRAN.R-project.org/package=bayesRecon.

Babai, M. Z., Boylan, J. E., & Rostami-Tabar, B. (2022). Demand forecasting in supply chains: a review of aggregation and hierarchical approaches. *International Journal of Production Research*, *60*(1), 324–348. http://dx.doi.org/10.1080/00207543.2021.2005268.

Bai, L., & Pinson, P. (2019). Distributed reconciliation in Day-Ahead wind power forecasting. *Energies*, *12*(6), 1–19. http://dx.doi.org/10.3390/en12061112, URL https://www.mdpi.com/1996-1073/12/6/1112.

Barrow, D. K., & Kourentzes, N. (2016). Distributions of forecasting errors of forecast combinations: implications for inventory management. *International Journal of Production Economics*, *177*, 24–33. http://dx.doi.org/10.1016/j.ijpe.2016.03.017.

Ben Taieb, S., & Koo, B. (2019). Regularized regression for hierarchical forecasting without unbiasedness conditions. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 1337–1347). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3292500.3330976.

Ben Taieb, S., Taylor, J. W., & Hyndman, R. J. (2017). Coherent probabilistic forecasts for hierarchical time series. In D. Precup, & Y. W. Teh (Eds.), *Proceedings of machine learning research*: *vol. 70*, *Proceedings of the 34th international conference on machine learning* (pp. 3348–3357). PMLR, URL https://proceedings.mlr.press/v70/taieb17a.html.

Ben Taieb, S., Taylor, J. W., & Hyndman, R. J. (2021). Hierarchical probabilistic forecasting of electricity demand with smart meter data. *Journal of the American Statistical Association*, *116*(533), 27–43. http://dx.doi.org/10.1080/01621459.2020.1736081.

Ben Taieb, S., Yu, J., Barreto, M. N., & Rajagopal, R. (2017). Regularization in hierarchical time series forecasting with application to electricity smart meter data. In *31st AAAI conference on artificial intelligence, vol. 31, no. 1* (pp. 4474–4480). URL https://ojs.aaai.org/index.php/AAAI/article/view/11167.

Bergsteinsson, H. G., Møller, J. K., Nystrup, P., Pálsson, Ó. P., Guericke, D., & Madsen, H. (2021). Heat load forecasting using adaptive temporal hierarchies. *Applied Energy*, *292*, Article 116872. http://dx.doi.org/10.1016/j.apenergy.2021.116872.

Bikker, R., Daalmans, J., & Mushkudiani, N. (2013). Benchmarking large accounting frameworks: a generalized multivariate model. *Economic Systems Research*, *25*(4), 390–408.

Bisaglia, L., Di Fonzo, T., & Girolimetto, D. (2020). Fully reconciled GDP forecasts from income and expenditure sides. In A. Pollice, N. Salvati, & F. Schirripa Spagnolo (Eds.), *Book of short papers SIS 2020* (pp. 951–956). Pearson.

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

Brégère, M., & Huard, M. (2022). Online hierarchical forecasting for power consumption data. *International Journal of Forecasting*, *38*, 339–351. http://dx.doi.org/10.1016/j.ijforecast.2021.05.011.

Brewer, K. (1973). Some consequences of temporal aggregation and systematic sampling for ARMAand ARMAX models. *Journal of Econometrics*, *1*(2), 133–154. http://dx.doi.org/10.1016/0304-4076(73)90015-8.

Burba, D., & Chen, T. (2021). A trainable reconciliation method for hierarchical time-series. URL https://arxiv.org/abs/2101.01329.

Buzna, L., De Falco, P., Ferruzzi, G., Khormali, S., Proto, D., Refa, N., Straka, M., & van der Poele, G. (2021). An ensemble methodology for hierarchical probabilistic electric vehicle load forecasting at regular charging stations. *Applied Energy*, *283*, Article 116337. http://dx.doi.org/10.1016/j.apenergy.2020.116337.

Byron, R. P. (1978). The estimation of large social account matrices. *Journal of the Royal Statistical Society, Series A*, *141*(3), 359–367. http://dx.doi.org/10.2307/2344807, URL http://www.jstor.org/stable/2344807.

Byron, R. P. (1979). Corrigenda: The estimation of large social account matrices. *Journal of the Royal Statistical Society. Series A*, *142*(3), 405. http://dx.doi.org/10.2307/2982515, URL https://www.jstor.org/stable/10.2307/2982515?origin=crossref.

Capistrán, C., Constandse, C., & Ramos-Francia, M. (2010). Multi-horizon inflation forecasts using disaggregated data. *Economic Modelling*, *27*(3), 666–677. http://dx.doi.org/10.1016/j.econmod.2010.01.006.

Chase, C. W. (2013). Using big data to enhance demand-driven forecasting and planning. *Journal of Business Forecasting*, *32*(2), 27–32.

Cholette, P. A. (1982). Prior information and ARIMA forecasting. *Journal of Forecasting*, *1*(4), 375–383.

Chow, G. C., & Lin, A.-l. (1971). Best linear unbiased interpolation, distribution, and extrapolation of time series by related series. *The Review of Economics and Statistics*, *53*(4), 372–375.

Claeskens, G., Magnus, J. R., Vasnev, A. L., & Wang, W. (2016). The forecast combination puzzle: A simple theoretical explanation. *International Journal of Forecasting*, *32*(3), 754–762.

Collins, D. W. (1976). Predicting earnings with sub-entity data: Some further evidence. *Journal of Accounting Research*, *14*(1), 163. http://dx.doi.org/10.2307/2490463.

Corani, G., Azzimonti, D., Augusto, J. P. S. C., & Zaffalon, M. (2021). Probabilistic reconciliation of hierarchical forecast via Bayes' rule. In F. Hutter, K. Kersting, J. Lijffijt, & I. Valera (Eds.), *Machine learning and knowledge discovery in databases* (pp. 211–226). Cham: Springer International Publishing, http://dx.doi.org/10.1007/978-3-030-67664-3_13.

Corani, G., Rubattu, N., Azzimonti, D., & Antonucci, A. (2022). Probabilistic reconciliation of count time series. URL https://arxiv.org/abs/2207.09322.

Corona, F., Guerrero, V. M., & López-Peréz, J. (2021). Optimal reconciliation of seasonally adjusted disaggregates taking into account the difference between direct and indirect adjustment of the aggregate. *Journal of Official Statistics*, *37*(1), 31–51.

Cugliari, J., & van Erven, T. (2015). gtop: Game-theoretically OPtimal (GTOP) reconciliation method. R package v0.2.0 URL https://CRAN.R-project.org/package=gtop.

da Silva, F. L. C., Cyrino Oliveira, F. L., & Souza, R. C. (2019). A bottom-up Bayesian extension for long term electricity consumption forecasting. *Energy*, *167*, 198–210. http://dx.doi.org/10.1016/j.energy.2018.10.201.

Dagum, E. B., & Cholette, P. A. (2006). *Lecture notes in statistics*: *vol. 186*, *Benchmarking, temporal distribution, and reconciliation methods for time series*. Springer New York, http://dx.doi.org/10.1007/0-387-35439-5, URL https://link.springer.com/book/10.1007/0-387-35439-5.

Dangerfield, B. J., & Morris, J. S. (1992). Top-down or bottom-up: aggregate versus disaggregate extrapolations. *International Journal of Forecasting*, *8*(2), 233–241. http://dx.doi.org/10.1016/0169-2070(92)90121-O.

Dannecker, L., Lorenz, R., Rösch, P., Lehner, W., & Hackenbroich, G. (2013). Efficient forecasting for hierarchical time series. In *CIKM '13 proceedings of the 22nd ACM international conference on information & knowledge management* (pp. 2399–2404). http://dx.doi.org/10.1145/2505515.2505622.

Das, A., Kong, W., Paria, B., & Sen, R. (2023). Dirichlet proportions model for hierarchically coherent probabilistic forecasting. URL https://arxiv.org/abs/2204.10414.

Davis, T. A., Rajamanickam, S., & Sid-Lakhdar, W. M. (2016). A survey of direct methods for sparse linear systems. *Acta Numerica*, *25*, 383–566.

De Alba, E. (1993). Constrained forecasting in autoregressive time series models: A Bayesian analysis. *International Journal of Forecasting*, *9*(1), 95–108.

Di Fonzo, T. (1990). The estimation of *M* disaggregate time series when contemporaneous and temporal aggregates are known. *The Review of Economics and Statistics*, *72*(1), 178–182. http://dx.doi.org/10.2307/2109758, URL http://www.jstor.org/stable/2109758.

Di Fonzo, T., & Girolimetto, D. (2022a). Enhancements in cross-temporal forecast reconciliation, with an application to solar irradiance forecasts. URL https://arxiv.org/abs/2209.07146.

Di Fonzo, T., & Girolimetto, D. (2022c). Fully reconciled probabilistic GDP forecasts from income and expenditure sides. In A. Balzanella, M. Bini, C. Cavicchia, & R. Verde (Eds.), *Book of short papers SIS 2022* (pp. 1376–1381). Pearson.

Di Fonzo, T., & Girolimetto, D. (2023). Forecast combination-based forecast reconciliation: Insights and extensions. *International Journal of Forecasting*, *40*(2), 490–514. http://dx.doi.org/10.1016/j.ijforecast.2022.07.001.

Di Fonzo, T., & Girolimetto, D. (2023a). Cross-temporal forecast reconciliation: Optimal combination method and heuristic alternatives. *International Journal of Forecasting*, *39*(1), 39–57. http://dx.doi.org/10.1016/j.ijforecast.2021.08.004.

Di Fonzo, T., & Girolimetto, D. (2023b). Spatio-temporal reconciliation of solar forecasts. *Solar Energy*, *251*, 13–29.

Di Fonzo, T., & Marini, M. (2011). Simultaneous and two-step reconciliation of systems of time series: methodological and practical issues. *Journal of the Royal Statistical Society. Series C. Applied Statistics*, *60*(2), 143–164.

Dunn, D. M., Williams, W. H., & Dechaine, T. L. (1976). Aggregate versus subaggregate models in local area forecasting. *Journal of the American Statistical Association*, *71*(353), 68–71. http://dx.doi.org/10.1080/01621459.1976.10481478.

Eckert, F., Hyndman, R. J., & Panagiotelis, A. (2021). Forecasting Swiss exports using Bayesian forecast reconciliation. *European Journal of Operational Research*, *291*(2), 693–710. http://dx.doi.org/10.1016/j.ejor.2020.09.046.

Edwards, J. B., & Orcutt, G. H. (1969). Should aggregation prior to estimation be the rule? *The Review of Economics and Statistics*, *51*(4), 409–420.

Espasa, A. (1994). Comments on 'Time-series analysis, forecasting and econometric modelling: The structural econometric modelling, time-series analysis (SEMTSA) approach' by A. Zellner. *Journal of Forecasting*, *13*(2), 215–233. http://dx.doi.org/10.1002/for.3980130213.

Feng, C., & Zhang, J. (2020). Assessment of aggregation strategies for machine-learning based short-term load forecasting. *Electric Power Systems Research*, *184*, Article 106304. http://dx.doi.org/10.1016/j.epsr.2020.106304.

Fliedner, G. (1999). An investigation of aggregate variable time series forecast strategies with specific subaggregate time series statistical correlation. *Computers & Operations Research*, *26*, 1133–1149.

Fliedner, G. (2001). Hierarchical forecasting: issues and use guidelines. *Industrial Management and Data Systems*, *101*(1), 5–12. http://dx.doi.org/10.1108/02635570110365952.

Fliedner, E. B., & Mabert, V. A. (1992). Constrained forecasting: some implementation guidelines. *Decision Sciences*, *23*, 1143–1161.

Fong, D. C. L., & Saunders, M. (2011). LSMR: An iterative algorithm for sparse least-squares problems. *SIAM Journal on Scientific Computing*, *33*(5), 2950–2971.

Gibson, G. C., Moran, K. R., Reich, N. G., & Osthus, D. (2021). Improving probabilistic infectious disease forecasting through coherence. *PLoS Computational Biology*, *January 2021*, http://dx.doi.org/10.1371/journal.pcbi.1007623.

Gilbert, C., Browell, J., & McMillan, D. (2018). A hierarchical approach to probabilistic wind power forecasting. In *2018 IEEE international conference on probabilistic methods applied to power systems* (pp. 1–6). ieeexplore.ieee.org, http://dx.doi.org/10.1109/PMAPS.2018.8440571.

Girolimetto, D., Athanasopoulos, G., Di Fonzo, T., & Hyndman, R. J. (2023). Cross-temporal probabilistic forecast reconciliation. *International Journal of Forecasting*, 1–28 http://dx.doi.org/10.1016/j.ijforecast.2023.10.003, (in press).

Girolimetto, D., & Di Fonzo, T. (2022). FoReco: Point forecast reconciliation. R package v0.2.5 URL https://CRAN.R-project.org/package=FoReco.

Girolimetto, D., & Di Fonzo, T. (2023). Point and probabilistic forecast reconciliation for general linearly constrained multiple time series. *forthcoming*,

Gleason, J. L. (2020). Forecasting hierarchical time series with a regularized embedding space. In *MileTS '20: 6th KDD workshop on mining and learning from time series 2020* (pp. 883–894). URL https://kdd-milets.github.io/milets2020/papers/MiLeTS2020_paper_13.pdf.

Gross, C. W., & Sohl, J. E. (1990). Disaggregation methods to expedite product line forecasting. *Journal of Forecasting*, *9*(3), 233–254. http://dx.doi.org/10.1002/for.3980090304.

Grunfeld, Y., & Griliches, Z. (1960). Is aggregation necessarily bad? *The Review of Economics and Statistics*, *42*(1), 1–13.

Guerrero, V. M. (1989). Optimal conditional ARIMA forecasts. *Journal of forecasting*, *8*(3), 215–229.

Han, X., Dasgupta, S., & Ghosh, J. (2021). Simultaneously reconciled quantile forecasting of hierarchically related time series. In *Proceedings of the 24th international conference on artificial intelligence and statistics, vol. 130*. URL https://arxiv.org/abs/2102.12612.

Hansen, M. E., Peter, N., Møller, J. K., & Henrik, M. (2023). Reconciliation of wind power forecasts in spatial hierarchies. *Wind Energy*.

Herzen, J., Lässig, F., Piazzetta, S. G., Neuer, T., Tafti, L., Raille, G., Pottelbergh, T. V., Pasieka, M., Skrodzki, A., Huguenin, N., Dumonal, M., Kościsz, J., Bader, D., Gusset, F., Benheddi, M., Williamson, C., Kosinski, M., Petrik, M., & Grosch, G. (2022). Darts: User-friendly modern machine learning for time series. *Journal of Machine Learning Research*, *23*(124), 1–6, URL http://jmlr.org/papers/v23/21-1177.html.

Herzen, J., Lässig, F., Piazzetta, S. G., Neuer, T., Tafti, L., Raille, G., Pottelbergh, T. V., Pasieka, M., Skrodzki, A., Huguenin, N., Dumonal, M., Kościsz, J., Bader, D., Gusset, F., Benheddi, M., Williamson, C., Kosinski, M., Petrik, M., & Grosch, G. (2023). Darts: Time series made easy in Python. Python package v0.23.1 URL https://unit8co.github.io/darts/.

Hippert, H. S., Pedreira, C. E., & Souza, R. C. (2001). Neural networks for short-term load forecasting: A review and evaluation. *IEEE Transactions on Power Systems*, *16*(1), 44–55.

Hollyman, R., Petropoulos, F., & Tipping, M. E. (2021). Understanding forecast reconciliation. *European Journal of Operational Research*, *294*(1), 149–160. http://dx.doi.org/10.1016/j.ejor.2021.01.017.

Hong, T., Pinson, P., & Fan, S. (2014). Global energy forecasting competition 2012. *International Journal of Forecasting*, *30*(2), 357–363. http://dx.doi.org/10.1016/j.ijforecast.2013.07.001, URL http://www.sciencedirect.com/science/article/pii/S0169207013000745.

Hong, T., Xie, J., & Black, J. (2019). Global energy forecasting competition 2017: Hierarchical probabilistic load forecasting. *International Journal of Forecasting*, *35*(4), 1389–1399. http://dx.doi.org/10.1016/j.ijforecast.2019.02.006, URL http://www.sciencedirect.com/science/article/pii/S016920701930024X.

Hotta, L. K., & Cardoso Neto, J. (1993). The effect of aggregation on prediction in autoregressive integrated moving-average models. *Journal of Time Series Analysis*, *14*(3), 261–269.

Hotta, L. K., Morettin, P. A., & 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.

Hubrich, K. (2005). Forecasting euro area inflation: Does aggregating forecasts by HICP component improve forecast accuracy? *International Journal of Forecasting*, *21*(1), 119–136.

Huddlestone, S. H., Porter, J. H., & Brown, D. E. (2015). Improving forecasts for noisy geographic time series. *Journal of Business Research*, *68*(8), 1810–1818.

Hyndman, R. J. (2022). Notation for forecast reconciliation. URL https://robjhyndman.com/hyndsight/reconciliation-notation.html. (Accessed 17 October 2023).

Hyndman, R. J., Ahmed, R. A., Athanasopoulos, G., & Shang, H. L. (2011). Optimal combination forecasts for hierarchical time series. *Computational Statistics & Data Analysis*, *55*(9), 2579–2589. http://dx.doi.org/10.1016/j.csda.2011.03.006.

Hyndman, R. J., Ahmed, R. A., & Shang, H. L. (2010). hts: Hierarchical time series. R package v1. URL https://cran.r-project.org/src/contrib/Archive/hts/hts_1.0.tar.gz.

Hyndman, R. J., & Athanasopoulos, G. (2014). Optimally reconciling forecasts in a hierarchy. *Foresight: International Journal of Applied Forecasting*, *35*, 42–48, URL https://robjhyndman.com/papers/Foresight-hts-final.pdf.

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and Practice* (2nd ed.). Melbourne, Australia: OTexts, URL https://otexts.com/fpp2/.

Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and practice* (3rd ed.). Melbourne, Australia: OTexts, URL https://otexts.com/fpp3/.

Hyndman, R. J., & Kourentzes, N. (2018). thief: Temporal HIErarchical forecasting. URL https://CRAN.R-project.org/package=thief.

Hyndman, R. J., Lee, A. J., & Wang, E. (2016). Fast computation of reconciled forecasts for hierarchical and grouped time series. *Computational Statistics & Data Analysis*, *97*, 16–32. http://dx.doi.org/10.1016/j.csda.2015.11.007.

Hyndman, R. J., Lee, A., Wang, E., & Wickramasuriya, S. (2021). hts: Hierarchical and grouped time series. R package v6.0.2 URL https://CRAN.R-project.org/package=hts.

Ilmakunnas, P. (1990). Aggregation vs disaggregation in forecasting construction activities. In T. Barker, & M. H. Pesaran (Eds.), *Disaggregation in Econometric Modelling* (pp. 73–86). Routledge.

Jeon, J., Panagiotelis, A., & Petropoulos, F. (2019). Probabilistic forecast reconciliation with applications to wind power and electric load. *European Journal of Operational Research*, *279*(2), 364–379. http://dx.doi.org/10.1016/j.ejor.2019.05.020.

Kahn, K. B. (1998). Revisiting top-down versus bottom-up forecasting. *The Journal of Business Forecasting Methods & Systems*, *17*(2), 14–19.

Karmy, J. P., López, J., & Maldonado, S. (2023). Pooling information across levels in hierarchical time series forecasting via kernel methods. *Expert Systems with Applications*, *213*(PA), Article 118830. http://dx.doi.org/10.1016/j.eswa.2022.118830.

Karmy, J. P., & Maldonado, S. (2019). Hierarchical time series forecasting via support vector regression in the European travel retail industry. *Expert Systems with Applications*, *137*, 59–73. http://dx.doi.org/10.1016/j.eswa.2019.06.060.

Kinney, W. R., Jr. (1971). Predicting earnings: entity versus subentity data. *Journal of Accounting Research*, *9*(1), 127–136.

Kolassa, S. (2016). Evaluating predictive count data distributions in retail sales forecasting. *International Journal of Forecasting*, *32*(3), 788–803.

Kolassa, S. (2023). Do we want coherent hierarchical forecasts, or minimal MAPEs or MAEs? (we won't get both!) *International Journal of Forecasting*, *39*(4), 1512–1517.

Koop, G., Mcintyre, S., Mitchell, J., & Poon, A. (2023). Using stochastic hierarchical aggregation constraints to nowcast regional economic aggregates. *International Journal of Forecasting* (in press) URL http://escoe-website.s3.amazonaws.com/wp-content/uploads/2022/03/03155343/ESCoE-DP-2022-04.pdf.

Kourentzes, N. (2014). On intermittent demand model optimisation and selection. *International Journal of Production Economics*, *156*, 180–190. http://dx.doi.org/10.1016/j.ijpe.2014.06.007.

Kourentzes, N. (2022). Toward a one-number forecast: cross-temporal hierarchies. *Foresight: The International Journal of Applied Forecasting*, (67), 32–38, URL https://ideas.repec.org/a/for/ijafaa/y2022i67p32-38.html.

Kourentzes, N., & Athanasopoulos, G. (2019). Cross-temporal coherent forecasts for Australian tourism. *Annals of Tourism Research*, *75*, 393–409. http://dx.doi.org/10.1016/j.annals.2019.02.001.

Kourentzes, N., & Athanasopoulos, G. (2021). Elucidate structure in intermittent demand series. *European Journal of Operational Research*, *288*(1), 141–152. http://dx.doi.org/10.1016/j.ejor.2020.05.046.

Kourentzes, N., & Petropoulos, F. (2016). Forecasting with multivariate temporal aggregation: The case of promotional modelling. *International Journal of Production Economics*, *181*, 145–153. http://dx.doi.org/10.1016/j.ijpe.2015.09.011.

Kourentzes, N., Petropoulos, F., & Trapero, J. R. (2014). Improving forecasting by estimating time series structural components across multiple frequencies. *International Journal of Forecasting*, *30*(2), 291–302. http://dx.doi.org/10.1016/j.ijforecast.2013.09.006.

Kourentzes, N., Rostami-Tabar, B., & Barrow, D. K. (2017). Demand forecasting by temporal aggregation: Using optimal or multiple aggregation levels? *Journal of Business Research*, *78*, 1–9. http://dx.doi.org/10.1016/j.jbusres.2017.04.016.

Kourentzes, N., Saayman, A., Jean-Pierre, P., Provenzano, D., Sahli, M., Seetaram, N., & Volo, S. (2021). Visitor arrivals forecasts amid COVID-19: A perspective from the Africa team. *Annals of Tourism Research*, *88*, Article 103197. http://dx.doi.org/10.1016/j.annals.2021.103197.

Kremer, M., Siemsen, E., & Thomas, D. J. (2016). The sum and its parts: judgmental hierarchical forecasting. *Management Science*, *62*(9), 2745–2764. http://dx.doi.org/10.1287/mnsc.2015.2259.

Lam, C. (2020). High-dimensional covariance matrix estimation. *Wiley Interdisciplinary Reviews: Computational Statistics*, *12*(2), Article e1485.

Lapide, L. (1998). A simple view of top-down vs bottom-up forecasting. *Journal of Business Forecasting Methods and Systems*, *17*, 28–31.

Li, H., & Hyndman, R. J. (2021). Assessing mortality inequality in the U.S.: What can be said about the future? *Insurance: Mathematics & Economics*, *99*, 152–162. http://dx.doi.org/10.1016/j.insmatheco.2021.03.014.

Li, H., Li, H., Lu, Y., & Panagiotelis, A. (2019). A forecast reconciliation approach to cause-of-death mortality modeling. *Insurance: Mathematics & Economics*, *86*, 122–133. http://dx.doi.org/10.1016/j.insmatheco.2019.02.011.

Li, C., & Lim, A. (2018). A greedy aggregation–decomposition method for intermittent demand forecasting in fashion retailing. *European Journal of Operational Research*, *269*(3), 860–869. http://dx.doi.org/10.1016/j.ejor.2018.02.029.

Li, H., & Tang, Q. (2019). Analyzing mortality bond indexes via hierarchical forecast reconciliation. *Astin Bulletin*, *49*(3), 823–846. http://dx.doi.org/10.1017/asb.2019.19.

Lila, M. F., Meira, E., & Cyrino Oliveira, F. L. (2022). Forecasting unemployment in Brazil: A robust reconciliation approach using hierarchical data. *Socio-Economic Planning Sciences*, *82*(PB), Article 101298. http://dx.doi.org/10.1016/j.seps.2022.101298.

Luna, I., & Ballini, R. (2011). Top-down strategies based on adaptive fuzzy rule-based systems for daily time series forecasting. *International Journal of Forecasting*, *27*(3), 708–724. http://dx.doi.org/10.1016/j.ijforecast.2010.09.006.

Lütkepohl, H. (1984a). Forecasting contemporaneously aggregated vector. *Journal of Business & Economic Statistics*, *2*, 201–214. http://dx.doi.org/10.2307/1391703.

Lütkepohl, H. (1984b). Linear transformations of vector ARMA processes. *Journal of Econometrics*, *26*, 283–293.

Lütkepohl, H. (1986). Forecasting temporally aggregated vector ARMA processes. *Journal of Forecasting*, *5*, 85–95, URL http://onlinelibrary.wiley.com/doi/10.1002/for.3980050202/abstract.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2022). The M5 accuracy competition: Results, findings and conclusions, vol. 38, no. 4. (pp. 1346–1364). URL https://www.researchgate.net/publication/344487258.

Mancuso, P., Piccialli, V., & Sudoso, A. M. (2021). A machine learning approach for forecasting hierarchical time series. *Expert Systems with Applications*, *182*, Article 115102. http://dx.doi.org/10.1016/j.eswa.2021.115102.

Meira, E., Lila, M. F., & Oliveira, F. L. C. (2023). A novel reconciliation approach for hierarchical electricity consumption forecasting based on resistant regression. *Energy*, *269*, Article 126794.

Mircetic, D., Rostami-Tabar, B., Nikolicic, S., & Maslaric, M. (2022). Forecasting hierarchical time series in supply chains: an empirical investigation. *International Journal of Production Research*, *60*(8), 2514–2533. http://dx.doi.org/10.1080/00207543.2021.1896817.

Mishchenko, K., Montgomery, M., & Vaggi, F. (2019). A self-supervised approach to hierarchical forecasting with applications to groupwise synthetic controls. https://arxiv.org/abs/1906.10586.

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

Moon, S., Hicks, C., & Simpson, A. (2012). The development of a hierarchical forecasting method for predicting spare parts demand in the South Korean Navy—A case study. *International Journal of Production Economics*, *140*(2), 794–802.

Nespoli, L., Medici, V., Lopatichki, K., & Sossan, F. (2020). Hierarchical demand forecasting benchmark for the distribution grid. *Electric Power Systems Research*, *189*, Article 106755. http://dx.doi.org/10.1016/j.epsr.2020.106755.

Nijman, T. E., & Palm, F. C. (1990). Predictive accuracy gain from disaggregate sampling in ARIMA models. *Journal of Business & Economic Statistics*, *8*(4), 405–415. http://dx.doi.org/10.1080/07350015.1990.10509811.

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

Novak, J., McGarvie, S., & Garcia, B. E. (2017). A Bayesian model for forecasting hierarchically structured time series. URL https://arxiv.org/abs/1711.04738.

Nystrup, P., Lindström, E., Møller, J. K., & Madsen, H. (2021). Dimensionality reduction in forecasting with temporal hierarchies. *International Journal of Forecasting*, *37*(3), 1127–1146. http://dx.doi.org/10.1016/j.ijforecast.2020.12.003.

Nystrup, P., Lindström, E., Pinson, P., & Madsen, H. (2020). Temporal hierarchies with autocorrelation for load forecasting. *European Journal of Operational Research*, *280*(3), 876–888. http://dx.doi.org/10.1016/j.ejor.2019.07.061.

O'Hara-Wild, M., Hyndman, R. J., & Wang, E. (2023). fabletools: Core tools for packages in the 'fable' framework. R package v0.3.2 URL https://CRAN.R-project.org/package=fabletools.

Olivares, K. G., Garza, F., Luo, D., Challú, C., Mergenthaler, M., Taieb, S. B., Wickramasuriya, S. L., & Dubrawski, A. (2022). Hierarchical-Forecast: Probabilistic hierarchical forecasting with statistical and econometric methods. Python package v0.2.1 URL https://nixtla.github.io/hierarchicalforecast.

Olivares, K. G., Garza, F., Luo, D., Challú, C., Mergenthaler, M., Taieb, S. B., Wickramasuriya, S. L., & Dubrawski, A. (2023). HierarchicalForecast: A reference framework for hierarchical forecasting in Python. URL https://arxiv.org/abs/2207.03517.

Olivares, K. G., Meetei, O. N., Ma, R., Reddy, R., Cao, M., & Dicker, L. (2022). Probabilistic hierarchical forecasting with deep Poisson mixtures. URL https://arxiv.org/abs/2110.13179v6.

Oliveira, J. M., & Ramos, P. (2019). Assessing the performance of hierarchical forecasting methods on the retail sector. *Entropy*, *21*(4), 436. http://dx.doi.org/10.3390/e21040436.

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

Paige, C. C., & Saunders, M. A. (1982). Algorithm 583: LSQR: Sparse linear equations and least squares problems. *ACM Transactions on Mathematical Software*, *8*(2), 195–209.

Panagiotelis, A. (2020). ProbReco: Score optimal probabilistic forecast reconciliation. R package version 0.1.0.1 URL https://CRAN.R-project.org/package=ProbReco.

Panagiotelis, A., Athanasopoulos, G., Gamakumara, P., & Hyndman, R. J. (2021). Forecast reconciliation: A geometric view with new insights on bias correction. *International Journal of Forecasting*, *37*(1), 343–359. http://dx.doi.org/10.1016/j.ijforecast.2020.06.004.

Panagiotelis, A., Gamakumara, P., Athanasopoulos, G., & Hyndman, R. J. (2023). Probabilistic forecast reconciliation: properties, evaluation and score optimisation. *European Journal of Operational Research*, *306*(2), 693–706. http://dx.doi.org/10.1016/j.ejor.2022.07.040.

Panamtash, H., & Zhou, Q. (2018). Coherent probabilistic solar power forecasting. In *2018 IEEE international conference on probabilistic methods applied to power systems* (pp. 1–6). http://dx.doi.org/10.1109/PMAPS.2018.8440483.

Paria, B., Sen, R., Ahmed, A., & Das, A. (2021). Hierarchically regularized deep forecasting. URL https://arxiv.org/abs/2106.07630.

Park, M., & Nassar, M. (2014). Variational Bayesian inference for forecasting hierarchical time series. In *International conference on machine learning, workshop on divergence methods for probabilistic inference*. URL https://privacy-preserving-machine-learning.github.io/ICMLworkshop_PARK_NASSAR.pdf.

Pennings, C. L., & van Dalen, J. (2017). Integrated hierarchical forecasting. *European Journal of Operational Research*, *263*(2), 412–418. http://dx.doi.org/10.1016/j.ejor.2017.04.047.

Petropoulos, F., & Kourentzes, N. (2015). Forecast combinations for intermittent demand. *Journal of the Operational Research Society*, *66*(6), 914–924. http://dx.doi.org/10.1057/jors.2014.62.

Petropoulos, F., Kourentzes, N., & Nikolopoulos, K. (2016). Another look at estimators for intermittent demand. *International Journal of Production Economics*, *181*, 154–161. http://dx.doi.org/10.1016/j.ijpe.2016.04.017.

Petropoulos, F., Wang, X., & Disney, S. M. (2019). The inventory performance of forecasting methods: Evidence from the M3 competition data. *International Journal of Forecasting*, *35*(1), 251–265.

Pino, F. A., Morettin, P. A., & Mentz, R. P. (1987). Modelling and forecasting linear combinations of time series. *Revue Internationale de Statistique*, (International Statistical Review) 295–313.

Pritularga, K. F., Svetunkov, I., & Kourentzes, N. (2021). Stochastic coherency in forecast reconciliation. *International Journal of Production Economics*, *240*, Article 108221. http://dx.doi.org/10.1016/j.ijpe.2021.108221.

Punia, S., Singh, S. P., & Madaan, J. K. (2020). A cross-temporal hierarchical framework and deep learning for supply chain forecasting. *Computers & Industrial Engineering*, *149*, Article 106796. http://dx.doi.org/10.1016/j.cie.2020.106796.

Qiao, M., & Huang, K.-W. (2018). Hierarchical accounting variables forecasting by deep learning methods. In *ICIS 2018 proceedings 7*. International Conference on Information Systems, URL https://aisel.aisnet.org/icis2018/crypto/Presentations/7.

Quenneville, B., & Fortier, S. (2012). Restoring accounting constraints in time series: methods and software for a statistical agency. In W. R. Bell, S. H. Holan, & T. S. McElroy (Eds.), *Economic Time Series: Modeling and Seasonality* (pp. 231–253). Taylor & Francis, URL http://books.google.com.au/books?id=6-2HZ2STqGEC.

Rangapuram, S. S., Werner, L. D., Benidis, K., Mercado, P., Gasthaus, J., & Januschowski, T. (2021). End-to-end learning of coherent probabilistic forecasts for hierarchical time series. In *Proceedings of the 38th International Conference on Machine Learning* (pp. 8832–8843). URL http://proceedings.mlr.press/v139/rangapuram21a.html.

Rehman, H. U., Wan, G., & Rafique, R. (2023). A hybrid approach with step-size aggregation to forecasting hierarchical time series. *Journal of forecasting*, *42*(1), 176–192. http://dx.doi.org/10.1002/for.2895.

Roach, C. (2019). Reconciled boosted models for GEFCom2017 hierarchical probabilistic load forecasting. *International Journal of Forecasting*, *35*(4), 1439–1450. http://dx.doi.org/10.1016/j.ijforecast.2018.09.009.

Romagnoli, J. A., & Sanchez, M. C. (2000). *Data processing and reconciliation for chemical process operations*. Academic Press, URL https://play.google.com/store/books/details?id=dzHoeR9OENcC.

Roque, L., Torgo, L., & Soares, C. (2021). Automatic hierarchical time-series forecasting using Gaussian processes. *Engineering Proceedings*, *5*(1), 49. http://dx.doi.org/10.3390/engproc2021005049.

Rosas, A. L., & Guerrero, V. M. (1994). Restricted forecasts using exponential smoothing techniques. *International Journal of Forecasting*, *10*(4), 515–527.

Rossana, R., & Seater, J. (1995). Temporal aggregation and economic times series. *Journal of Business & Economic Statistics*, *13*(4), 441–451. http://dx.doi.org/10.2307/1392389.

Rossi, N. (1982). A note on the estimation of disaggregate time series when the aggregate is known. *The Review of Economics and Statistics*, *64*(4), 695–696.

Rostami-Tabar, B., Babai, M. Z., Ali, M., & Boylan, J. E. (2019). The impact of temporal aggregation on supply chains with ARMA (1, 1) demand processes. *European Journal of Operational Research*, *273*(3), 920–932.

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

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

Sagheer, A., Hamdoun, H., & Youness, H. (2021). Deep LSTM-based transfer learning approach for coherent forecasts in hierarchical time series. *Sensors*, *21*(13), 4379.

Sanguri, K., Patra, S., & Punia, S. (2023). Forecast reconciliation in the temporal hierarchy: Special case of intermittent demand with obsolescence. *Expert Systems with Applications*, *218*, Article 119566.

Saoud, P., Kourentzes, N., & Boylan, J. E. (2022). Approximations for the lead time variance: A forecasting and inventory evaluation. *Omega*, *110*, 102614. http://dx.doi.org/10.1016/j.omega.2022.102614.

Sbrana, G., & 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. http://dx.doi.org/10.1016/j.ijpe.2013.06.022.

Schwarzkopf, A. B., Tersine, R. J., & Morris, J. S. (1988). Top-down versus bottom-up forecasting strategies. *International Journal of Production Research*, *26*(11), 1833–1843. http://dx.doi.org/10.1080/00207548808947995.

Seaman, B., & Bowman, J. (2022). Applicability of the M5 to forecasting at Walmart. *International Journal of Forecasting*, *38*(4), 1468–1472. http://dx.doi.org/10.1016/j.ijforecast.2021.06.002.

Shang, H. L., & Haberman, S. (2017). Grouped multivariate and functional time series forecasting: An application to annuity pricing. *Insurance: Mathematics & Economics*, *75*, 166–179. http://dx.doi.org/10.1016/j.insmatheco.2017.05.007.

Shang, H. L., & Hyndman, R. J. (2017). Grouped functional time series forecasting: an application to age-specific mortality rates. *Journal of Computational and Graphical Statistics*, *26*(2), 330–343. http://dx.doi.org/10.1080/10618600.2016.1237877.

Shiratori, T., Kobayashi, K., & Takano, Y. (2020). Prediction of hierarchical time series using structured regularization and its application to artificial neural networks. *PLOS ONE*, *15*(11), Article e0242099. http://dx.doi.org/10.1371/journal.pone.0242099.

Shlifer, E., & Wolff, R. W. (1979). Aggregation and proration in forecasting. *Management Science*, *25*(6), 594–603, URL http://www.jstor.org/stable/2630330.

Silvestrini, A., & Veredas, D. (2008). Temporal aggregation of univariate and multivariate time series models: A survey. *Journal of Economic Surveys*, *22*(3), 458–497. http://dx.doi.org/10.1111/j.1467-6419.2007.00538.x.

Simon, D. (2006). *Optimal state estimation: Kalman, H_∞, and nonlinear approaches*. Hoboken, NJ: John Wiley and Sons.

Simon, D. (2010). Kalman filtering with state constraints: a survey of linear and nonlinear algorithms. *IET Control Theory & Applications*, *4*(8), 1303–1318.

Simon, D., & Chia, T. L. (2002). Kalman filtering with state equality constraints. *IEEE transactions on Aerospace and Electronic Systems*, *38*(1), 128–136.

Smith, R. J., Weale, M. R., & Satchell, S. E. (1998). Measurement error with accounting constraints: Point and interval estimation for latent data with an application to UK gross domestic product. *Review of Economic Studies*, *65*(1), 109–134.

Spavound, S., & Kourentzes, N. (2022). Making forecasts more trustworthy. *Foresight: The International Journal of Applied Forecasting*, (66), 21–25, URL https://forecasters.org/making-forecasts-more-trustworthy/.

Spiliotis, E., Abolghasemi, M., Hyndman, R. J., Petropoulos, F., & Assimakopoulos, V. (2021). Hierarchical forecast reconciliation with machine learning. *Applied Soft Computing*, *112*, Article 107756. http://dx.doi.org/10.1016/j.asoc.2021.107756.

Spiliotis, E., Petropoulos, F., Kourentzes, N., & Assimakopoulos, V. (2020). Cross-temporal aggregation: Improving the forecast accuracy of hierarchical electricity consumption. *Applied Energy*, *261*, Article 114339. http://dx.doi.org/10.1016/j.apenergy.2019.114339.

Spithourakis, G. P., Petropoulos, F., Babai, M. Z., Nikolopoulos, K., & Assimakopoulos, V. (2011). Improving the performance of popular supply chain forecasting techniques. *Supply Chain Forum: An International Journal*, *12*(4), 16–25.

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

Sprangers, O., Schelter, S., & de Rijke, M. (2021). Probabilistic gradient boosting machines for large-scale probabilistic regression. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery and data mining* (pp. 1510–1520). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3447548.3467278.

Stone, R. (1961). *Input-output and national accounts*. Paris: O.E.E.C..

Stone, R., Champernowne, D. G., & Meade, J. E. (1942). The precision of national income estimates. *Review of Economic Studies*, *9*(2), 111–125. http://dx.doi.org/10.2307/2967664.

Stram, D. O., & Wei, W. W. S. (1986). Temporal aggregation in the Arima process. *Journal of Time Series Analysis*, *7*(4), 279–292. http://dx.doi.org/10.1111/j.1467-9892.1986.tb00495.x.

Strijbosch, L. W. G., Heuts, R. M. J., & Moors, J. J. A. (2008). Hierarchical estimation as a basis for hierarchical forecasting. *IMA Journal of Management Mathematics*, *19*(2), 193–205. http://dx.doi.org/10.1093/imaman/dpm032.

Theodosiou, F., & Kourentzes, N. (2021a). Deep learning temporal hierarchies for interval forecasts. URL https://www.researchgate.net/publication/358425154_Deep_Learning_Temporal_Hierarchies_for_Interval_Forecasts.

Theodosiou, F., & Kourentzes, N. (2021b). Forecasting with deep temporal hierarchies. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3918315.

Tiao, G. C. (1972). Asymptotic behaviour of temporal aggregates of time series. *Biometrika*, *59*(3), 525–531. http://dx.doi.org/10.1093/biomet/59.3.525.

van Erven, T., & Cugliari, J. (2015). Game-theoretically optimal reconciliation of contemporaneous hierarchical time series forecasts. In A. Antoniadis, J.-M. Poggi, & X. Brossat (Eds.), *Modeling and stochastic learning for forecasting in high dimension* (pp. 297–317). Cham: Springer International Publishing, URL https://hal.inria.fr/hal-00920559.

Villegas, M. A., & Pedregal, D. J. (2018). Supply chain decision support systems based on a novel hierarchical forecasting approach. *Decision Support Systems*, *114*, 29–36.

Wan, S. K., Wang, S. H., & Woo, C. K. (2013). Aggregate vs. disaggregate forecast: Case of Hong Kong. *Annals of Tourism Research*, *42*, 434–438. http://dx.doi.org/10.1016/j.annals.2013.03.002.

Wang, X., Chen, B., Sheng, J., Zheng, H., Dan, T., & Wu, X. (2020). An improved Lanczos algorithm for principal component analysis. In *Proceedings of 2020 the 6th international conference on computing and data engineering* (pp. 70–74).

Wang, E., Cook, D., & Hyndman, R. J. (2020). A new tidy data structure to support exploration and modeling of temporal data. *Journal of Computational and Graphical Statistics*, *29*(3), 466–478. http://dx.doi.org/10.1080/10618600.2019.1695624.

Wang, E., Cook, D., Hyndman, R. J., & O'Hara-Wild, M. (2022). tsibble: Tidy temporal data frames and tools. R package version 1.1.3 URL https://CRAN.R-project.org/package=tsibble.

Wang, S., Zhou, F., Sun, Y., Ma, L., Zhang, J., Zheng, Y., Lei, L., & Hu, Y. (2022). End-to-end modeling hierarchical time series using autoregressive transformer and conditional normalizing flow based reconciliation. https://arxiv.org/abs/2212.13706.

Wanke, P., & Saliby, E. (2007). Top-down or bottom-up forecasting? *Pesquisa Operacional*, *27*, 591–605.

Weale, M. (1992). Estimation of data measured with error and subject to linear restrictions. *Journal of Applied Econometrics*, *7*(2), 167–174.

Wei, W. W. S. (1979). Some consequences of temporal aggregation in seasonal time series models. In A. Zellner (Ed.), *Seasonal analysis of economic time series* (pp. 433–448). NBER, URL https://www.nber.org/system/files/chapters/c4333/c4333.pdf.

Weiss, C. (2018). *Essays in hierarchical time series forecasting and forecast combination* (Ph.D. thesis), University of Cambridge, http://dx.doi.org/10.17863/CAM.21895.

Wickramasuriya, S. L. (2021). Properties of point forecast reconciliation approaches. URL https://arxiv.org/abs/2103.11129.

Wickramasuriya, S. L. (2023). Probabilistic forecast reconciliation under the Gaussian framework. *Journal of Business & Economic Statistics* http://dx.doi.org/10.1080/07350015.2023.2181176, (in press).

Wickramasuriya, S. L., Athanasopoulos, G., & 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. http://dx.doi.org/10.1080/01621459.2018.1448825.

Wickramasuriya, S. L., Turlach, B. A., & Hyndman, R. J. (2020). Optimal non-negative forecast reconciliation. *Statistics and Computing*, *30*(5), 1167–1182. http://dx.doi.org/10.1007/s11222-020-09930-0.

Widiarta, H., Viswanathan, S., & 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. http://dx.doi.org/10.1093/imaman/dpm039.

Williams, B. D., & Waller, M. A. (2011). Top-down versus bottom-up demand forecasts: The value of shared point-of-sale data in the retail supply chain. *Journal of Business Logistics*, *32*(1), 17–26. http://dx.doi.org/10.1111/j.2158-1592.2011.01002.x.

Yagli, G. M., Yang, D., & Srinivasan, D. (2019). Reconciling solar forecasts: Sequential reconciliation. *Solar Energy*, *179*, 391–397. http://dx.doi.org/10.1016/j.solener.2018.12.075.

Yagli, G. M., Yang, D., & Srinivasan, D. (2020). Reconciling solar forecasts: Probabilistic forecasting with homoscedastic Gaussian errors on a geographical hierarchy. *Solar Energy*, *210*, 59–67. http://dx.doi.org/10.1016/j.solener.2020.06.005.

Yang, D., Goh, G. S., Jiang, S., & Zhang, A. N. (2016). Forecast UPC-level FMCG demand, part III: Grouped reconciliation. In *Proceedings - 2016 IEEE international conference on big data* (pp. 3813–3819). Santa Clara, CA, USA: IEEE, http://dx.doi.org/10.1109/BigData.2016.7841053.

Yang, D., Quan, H., Disfani, V. R., & Liu, L. (2017). Reconciling solar forecasts: Geographical hierarchy. *Solar Energy*, *146*, 276–286. http://dx.doi.org/10.1016/j.solener.2017.02.010.

Yang, D., Quan, H., Disfani, V. R., & Rodríguez-Gallegos, C. D. (2017). Reconciling solar forecasts: Temporal hierarchy. *Solar Energy*, *158*, 332–346. http://dx.doi.org/10.1016/j.solener.2017.09.055.

Zambon, L., Azzimonti, D., & Corani, G. (2022). Efficient probabilistic reconciliation of forecasts for real-valued and count time series. URL https://arxiv.org/abs/2210.02286.

Zellner, A., & Tobias, J. (2000). A note on aggregation, disaggregation and forecasting performance. *Journal of Forecasting*, *19*(5), 457–469. http://dx.doi.org/10.1002/1099-131X(200009)19:5.

Zhang, B., Kang, Y., & Li, F. (2022). A python package for hierarchical forecasting. Python package v0.2.0 URL https://angelpone.github.io/pyhts/.

Zhang, B., Kang, Y., Panagiotelis, A., & Li, F. (2023). Optimal reconciliation with immutable forecasts. *European Journal of Operational Research*, *308*(2), 650–660. http://dx.doi.org/10.1016/j.ejor.2022.11.035.

Zhao, T., Wang, J., & Zhang, Y. (2019). Day-ahead hierarchical probabilistic load forecasting with linear quantile regression and empirical copulas. *IEEE Access*, *7*, 80969–80979. http://dx.doi.org/10.1109/ACCESS.2019.2922744.

Zotteri, G., & 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. http://dx.doi.org/10.1016/j.ijpe.2006.12.030.

Zotteri, G., Kalchschmidt, M., & Caniato, F. (2005). The impact of aggregation level on forecasting performance. *International Journal of Production Economics*, *93–94*(8), 479–491. http://dx.doi.org/10.1016/j.ijpe.2004.06.044.