Hierarchical Forecasting

Name of First Author and Name of Second Author

1 Introduction

Accurate forecasting of key macroeconomic variables such as Gross Domestic Product (GDP), inflation, industrial production, has been at the forefront of economic research over many decades. Early approaches involved univariate models or at best low dimensional multivariate systems. The era of big data has now led to the use of regularization and shrinkage methods such as dynamic factor models, Lasso, LARS, Bayesian VARs, in an effort to exploit the plethora of potentially useful predictors now available. These predictors commonly also include the components of the variables of interest. For instance, GDP is formed as an aggregate of consumption, government expenditure, investment and net exports with each of these components also formed as aggregates of other economic variables. While the macroeconomic forecasting literature regularly uses such sub-indices as predictors, it does so in ways that fail to exploit accounting identities that describe known deterministic relationships between macroeconomic variables.

In this chapter we take a different approach. Over the past decade there has been a growing literature on forecasting collections of time series that follow aggregation constraints, known as hierarchical times series. Initially the aim of this literature was to ensure that forecasts adhered to aggregation constraints thus ensuring aligned decision making. However in applications to tourism data, retail data, more examples and references the forecast reconciliation methods designed to deal with this problem have also been shown to improve forecast accuracy. Since both aligned decision making and forecast accuracy are key concerns for economic agents and policy makers we propose the application of state of the art forecast reconciliation methods to macroeconomic forecasting. Has anyone applied reconciliation to macro?.

Name of First Author
Name, Address of Institute, e-mail: name@email.address

Name of Second Author
Name, Address of Institute e-mail: name@email.address

The remainder of the paper is set out as follows. Section 2 introduces the concept of hierarchical time series, i.e. collections of time series with known linear constraints, with a particular emphasis on macroeconomic examples. Section 3 describes state-of-the-art forecast reconciliation techniques for point forecasts, while section 4 describes the more recent extension of these techniques to probabilistic forecasting. Section 5 describes the data used in our empirical case study, namely Australian GDP data that can be represented using two different hierarchies. Section 6 provides details on the setup of our empirical study including metrics used for the evaluation of both point and probabilistic forecasts. Section 7 presents results and Section 8 concludes providing future avenues for research that are of particular relevance to macroeconometrians.

2 Hierarchical time series

Fix this depending on Section 2 To simplify the introduction of some notation we use the simple two-level hierarchical structure shown in Figure 1. Denote as $y_{Tot,t}$ the value observed at time t for the most aggregate (Total) series corresponding to level 0 of the hierarchy. Below level 0, denote as $y_{i,t}$ the value of the series corresponding to node i, observed at time t. For example, $y_{A,t}$ denotes the tth observation of the series corresponding to node A at level 1, $y_{AB,t}$ denotes the tth observation of the series corresponding to node AB at level 2, and so on.

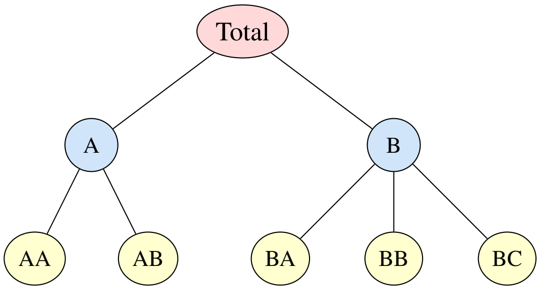

Fig. 1 A simple two-level hierarchical structure.

Let $\boldsymbol{y}_t = (y_{Tot,t}, y_{A,t}, y_{B,t}, y_{AA,t}, y_{AB,t}, y_{BA,t}, y_{BB,t}, y_{BC,t})'$, a vector containing observations across all series of the hierarchy at t. Similarly denote as $\boldsymbol{b}_t = (y_{AA,t}, y_{AB,t}, y_{BA,t}, y_{BB,t}, y_{BC,t})'$ a vector containing observations only for the bottom-level series. In general, $\boldsymbol{y}_t \in \mathbb{R}^n$ and $\boldsymbol{b}_t \in \mathbb{R}^m$ where n denotes the number of total series in the structure, m the number of series at the bottom level, and $n > m$ always. In the simple example of Figure 1, $n = 8$ and $m = 5$.

Aggregation constraints dictate that $y_{Tot} = y_{A,t} + y_{B,t} = y_{AA,t} + y_{AB,t} + y_{BA,t} + y_{BB,t} + y_{BC,t}$, $y_{A,t} = y_{AA,t} + y_{AB,t}$ and $y_B = y_{BA,t} + y_{BB,t} + y_{BC,t}$. Hence we can write

$$\boldsymbol{y}_t = \boldsymbol{S}\boldsymbol{b}_t, \tag{1}$$

where

$$\boldsymbol{S} = \begin{pmatrix} 1\ 1\ 1\ 1\ 1 \\ 1\ 1\ 0\ 0\ 0 \\ 0\ 0\ 1\ 1\ 1 \\ \boldsymbol{I}_5 \end{pmatrix}$$

an $n \times m$ matrix referred to as the *summing matrix* and $\boldsymbol{I}_m$ is an m-dimensional identity matrix. $\boldsymbol{S}$ reflects the linear aggregation constraints and in particular how the bottom-level series aggregate to levels above. Thus, columns of $\boldsymbol{S}$ span the linear subspace of $\mathbb{R}^n$ for which the aggregation constraints hold. We refer to this as the *coherent subspace* and denote it by $\mathfrak{s}$. Notice that pre-multiplying a vector in $\mathbb{R}^m$ by $\boldsymbol{S}$ will result in an n-dimensional vector that lies in $\mathfrak{s}$.

Property 1. A hierarchical time series has observations that are *coherent*, i.e., $\boldsymbol{y}_t \in \mathfrak{s}$ for all t. We use the term coherent to describe not just $\boldsymbol{y}_t$ but any vector in $\mathfrak{s}$.

Structures similar to the one portrayed in Figure 1 can be found in macroeconomics. For instance in Section 5 we consider two alternative hierarchical structures for the case of GDP and its components. However, while this motivating example involves aggregation constraints, the mathematical framework we use can be applied for any general linear constraints, examples of which are ubiquitous in macroeconomics. For instance, the trade balance is computed as exports minus imports, while the consumer price index is computed as a weighted average of sub-indices, which are in turn weighted averages of sub-sub-indices and so on. These structures can also be captured by an appropriately designed $\boldsymbol{S}$ matrix.

An important alternative aggregation structure also commonly found in macroeconomics, is one for which the most aggregate series is disaggregated by attributes of interest that are crossed, as distinct to nested which is the case for hierarchical time series. For example, industrial production may be disaggregated along the lines of geography or sector or both. We refer to this as a *grouped* structure. Figure 2 shows a simple example of such a structure. The Total series disaggregates into $y_{A,t}$ and $y_{B,t}$, but also into $y_{X,t}$ and $y_{Y,t}$, at level 1, and then into the bottom-level series, $\boldsymbol{b}_t = (y_{AX}, y_{AY}, y_{BX}, y_{BY})'$. Hence, in contrast to hierarchical, grouped time series do not naturally disaggregate in a unique manner.

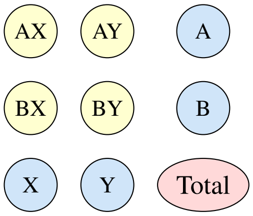

Fig. 2 A simple two-level grouped structure.

An important implementation of aggregation structures are *temporal hierarchies* introduced by **?**. In this case the aggregation structure spans the time dimension and dictates how higher frequency data (e.g., monthly) are aggregated to lower frequencies. There is a vast literature that studies the effects of temporal aggregation, going back to the seminal work of **????** and others such as, **????**. The main aim of this work is to find the single most optimum level of aggregation for modelling and forecasting time series. In this literature, the analyses, results (whether theoretical or empirical) and inferences, are extremely heterogeneous, making it very challenging to reach a consensus or some concrete conclusions. For example, **?** who study the effect of aggregation on several key macroeconomic variables state, “Quarterly data do not seem to suffer badly from temporal aggregation distortion, nor are they subject to the construction problems affecting monthly data. They therefore may be the optimal data for econometric analysis.” A similar conclusion is reached by

?. **?** consider forecasting French cash state deficit and provide empirical evidence of forecast accuracy gains from forecasting with the aggregate model rather than aggregating forecasts from the disaggregate model.

The overwhelming majority of the literature concentrates on a single level of temporal aggregation (there are some notable exceptions such as, **??**). **?** show that considering multiple levels of aggregation via temporal hierarchies and implementing forecast reconciliation approaches rather than single level approaches results in substantial gains in forecast accuracy across all levels of temporal aggregation. This is an example of the benefits of forecast reconciliation to which we now turn out attention to.

3 Point forecasting

A requirement when forecasting hierarchial time series is that the forecasts adhere to the same aggregation constraints as the observed data, i.e., they are coherent.

Definition 1. A set of h-step ahead forecasts $\tilde{\boldsymbol{y}}_{T+h|T}$, stacked in the same order as $\boldsymbol{y}_t$ and generated using information up to and including time T, are said to be *coherent* if $\tilde{\boldsymbol{y}}_{T+h|T} \in \mathfrak{s}$.

Hence, coherent forecasts of lower level series aggregate up to their corresponding upper level series and vice versa.

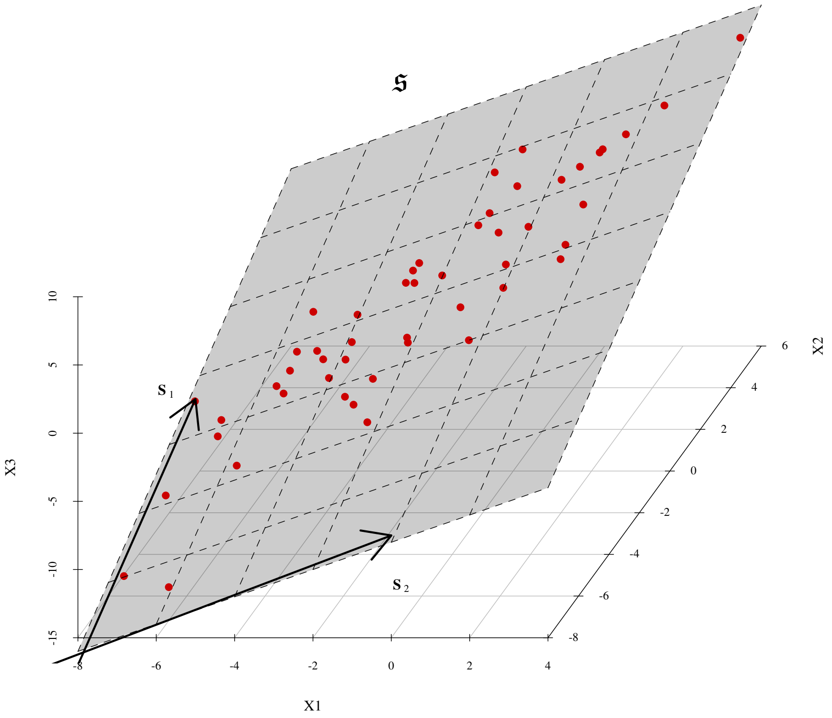

Fig. 3 Three dimensional schematic to represent the smallest possible hierarchy.

Fix the caption of figure 3.

Let us consider the smallest possible hierarchy with two bottom level series, A and B that add up to the top level Tot. Suppose $\check{\boldsymbol{y}}_{T+h|T}$ of this hierarchy is given by $\check{\boldsymbol{y}}_{T+h|T} = [\check{y}_{Tot,T+h|T}, \check{y}_{A,T+h|T}, \check{y}_{B,T+h|T}]$. Due to the aggregation structure we have

$\breve{y}_{Tot,T+h|T} = \breve{y}_{A,T+h|T} + \breve{y}_{B,T+h|T}$. This implies that, even though $\breve{\boldsymbol{y}}_{Tot,T+h|T} \in \mathbb{R}^3$, the points actually lie in $\mathfrak{s} \subset \mathbb{R}^3$, which is a two dimensional subspace within $\mathbb{R}^3$ space.

3.1 Single-level approaches

A common theme across all traditional approaches for forecasting hierarchical time series is that a single-level of aggregation is first selected and forecasts for that level are generated. These are then linearly combined to generate a set of coherent forecasts the rest of the structure.

3.1.1 Bottom-up

In the *bottom-up* approach, forecasts for the most disaggregate level are first generated. These are then aggregated to obtain forecasts for all other series of the hierarchy (Dunn et al. 1976). In general, this consists of first generating $\hat{\boldsymbol{b}}_{T+h|T} \in \mathbb{R}^m$, a set of h-step ahead forecasts for the bottom-level series. For the simple hierarchical structure of Figure 1, $\hat{\boldsymbol{b}}_{T+h|T} = (\hat{y}_{AA,T+h|T}, \hat{y}_{AB,T+h|T}, \hat{y}_{BA,T+h|T}, \hat{y}_{BB,T+h|T}, \hat{y}_{BC,T+h|T})$, where, $\hat{y}_{i,T+h|T}$ is the h-step ahead forecast of the series corresponding to node i. A set of coherent forecasts for the whole hierarchy is then given by,

$$\tilde{\boldsymbol{y}}^{BU}_{T+h|T} = \boldsymbol{S}\hat{\boldsymbol{b}}_{T+h|T}.$$

Generating bottom-up forecasts has the advantage of no information being lost due to aggregation. However, bottom-level data can potentially be highly volatile or very noisy and therefore challenging to forecast.

3.1.2 Top-down

In contrast *top-down* approaches involve first generating forecasts for the most aggregate level and then disaggregating these down the hierarchy. In general, coherent forecasts generated from top-down approaches are given by,

$$\tilde{\boldsymbol{y}}^{TD}_{T+h|T} = \boldsymbol{S}\boldsymbol{p}\hat{y}_{Tot,T+h|T},$$

where $\boldsymbol{p} = (p_1, ..., p_m)'$ is an m-dimensional vector consisting of a set of proportions which disaggregate the top-level forecast $\hat{y}_{Tot,T+h|T}$ to forecasts for the bottom-level series, hence $\boldsymbol{p}\hat{y}_{Tot,T+h|T} = \hat{\boldsymbol{b}}_{T+h|T}$. These are then aggregated up by the summing matrix $\boldsymbol{S}$.

Traditionally proportions have been calculated based on the observed historical data. Gross & Sohl (1990) present and evaluate twenty-one alternative approaches. The most convenient attribute of these approaches is their simplicity. Generating a set of coherent forecasts involves only modelling and generating forecasts for the most aggregate top-level series. In general, such top-down approaches seem to produce quite reliable forecasts for the aggregate levels and they are useful with low count data. However, a significant disadvantage is the loss of information due to aggregation. A limitation of such top-down approaches, is that characteristics of lower level series cannot be captured. To overcome this, **?** introduced a new top-down approach which disaggregates the top-level based on proportions of forecasts rather than the historical data and show evidence that this method outperforms the conventional top-down approaches. However, a limitation of all top-down is that they introduce bias to the forecasts even when the top-level forecast itself is unbiased. We discuss this in detail in Section 3.2 that follows.

3.1.3 Middle-out

A compromise between bottom-up and top-down approaches is the middle-out approach. It entails first forecasting the series of a selected middle-level. For series above the middle-level, coherent forecasts are generated using the bottom-up approach by aggregating the middle-level forecasts upwards. For series below the middle level, coherent forecasts are generated using a top-down approach by disaggregating the middle-level forecasts downwards. Since the middle-out approach involves generating top-down forecasts, it also introduces bias to the forecasts.

3.2 Point forecast reconciliation

All approaches discussed so far are limited to only using information from a single-level of aggregation. Furthermore, these ignore any correlations across levels of a hierarchy. An alternative framework that overcomes these limitations is one that involves forecast *reconciliation*. In a first step forecasts for all the series across all levels of the hierarchy are generated, ignoring any aggregation constraints. We refer to these as *base* forecasts and denote them by $\hat{\boldsymbol{y}}_{T+h|T}$. In general, base forecasts will not be coherent, although a notable exception is when a random walk is used to generate base forecasts. In this case, forecasts are simply equal to a previous realisation of the data and they inherit the property of coherence.

The second step is an ex-post adjustment that reconciles base forecasts so that they become coherent. In general, this is achieved by mapping the base forecasts $\hat{\boldsymbol{y}}_{T+h|T}$ onto the coherent subspace $\mathfrak{s}$, via a matrix $\boldsymbol{SG}$, resulting in a set of coherent forecasts $\tilde{\boldsymbol{y}}_{T+h|T}$. More specifically,

$$\tilde{\boldsymbol{y}}_{T+h|T} = \boldsymbol{SG}\hat{\boldsymbol{y}}_{T+h|T}, \tag{2}$$

where $\boldsymbol{G}$ is an $m \times n$ matrix that maps $\hat{\boldsymbol{y}}_{T+h|T}$ to $\mathbb{R}^m$, producing new forecasts for the bottom-level, which are in turn mapped to the coherent subspace by the summing matrix $\boldsymbol{S}$. We restrict our attention to projections on $\mathfrak{s}$ in which case $\boldsymbol{SGS} = \boldsymbol{S}$. This ensures that unbiasedness is preserved, i.e., for a set of unbiased base forecasts reconciled forecasts will also be unbiased.

Note that all single-level approaches discussed so far can also be represented by (2) using appropriately designed $\boldsymbol{G}$ matrices, however not all of these will be projections. For example for the bottom-up approach, $\boldsymbol{G} = \big(\mathbf{0}_{(m\times n-m)}\ \boldsymbol{I}_m\big)$ in which case $\boldsymbol{SGS} = \boldsymbol{S}$. For any top-down approach $\boldsymbol{G} = \big(\boldsymbol{p}\ \mathbf{0}_{(m\times n-1)}\big)$, for which case $\boldsymbol{SGS} \neq \boldsymbol{S}$.

3.2.1 Optimal MinT reconciliation

? build a unifying framework for much of the previous literature on forecast reconciliation. We present here a detailed outline of this approach and in turn relate it to previous significant contributions in forecast reconciliation.

Assume that $\hat{\boldsymbol{y}}_{T+h|T}$ is a set of unbiased base forecasts, i.e., $E_{1:T}(\hat{\boldsymbol{y}}_{T+h|T}) = E_{1:T}[\boldsymbol{y}_{T+h}|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_T]$, the true mean with the expectation taken over the observed sample up to time T. Let

$$\hat{\boldsymbol{e}}_{T+h|T} = \boldsymbol{y}_{T+h|T} - \hat{\boldsymbol{y}}_{T+h|T} \qquad (3)$$

denote a set of base forecast errors with $\mathrm{Var}(\hat{\boldsymbol{e}}_{T+h|T}) = \boldsymbol{W}_h$, and

$$\tilde{\boldsymbol{e}}_{T+h|T} = \boldsymbol{y}_{T+h|T} - \tilde{\boldsymbol{y}}_{T+h|T}$$

denote a set of coherent forecast errors. Lemma 1 in **?** shows that for any matrix $\boldsymbol{G}$ such that $\boldsymbol{SGS} = \boldsymbol{S}$, $\mathrm{Var}(\tilde{\boldsymbol{e}}_{T+h|T}) = \boldsymbol{SGW}_h\boldsymbol{S}'\boldsymbol{G}'$. Furthermore Theorem 1 shows that

$$\boldsymbol{G} = (\boldsymbol{S}'\boldsymbol{W}_h^{-1}\boldsymbol{S})^{-1}\boldsymbol{S}'\boldsymbol{W}_h^{-1} \qquad (4)$$

is the unique solution that minimises the $\mathrm{tr}[\boldsymbol{SGW}_h\boldsymbol{S}'\boldsymbol{G}']$ subject to $\boldsymbol{SGS} = \boldsymbol{S}$. Is there anything in the following sentence that has not already been said? MinT is optimal in the sense that given a set of unbiased base forecasts, it returns a set of best linear unbiased reconciled forecasts, using as $\boldsymbol{G}$ the unique solution that minimises the trace (hence MinT) of the variance of the forecast error of the reconciled forecasts. A significant advantage of the MinT reconciliation solution is that it is the first to incorporate the full correlation structure of the hierarchy via $\boldsymbol{W}_h$. However, estimating $\boldsymbol{W}_h$ is challenging, especially for $h > 1$. **?** present possible alternative estimators for $\boldsymbol{W}_h$ and show that these lead to different $\boldsymbol{G}$ matrices. We summarise these below.

- Set $\boldsymbol{W}_h = k_h\boldsymbol{I}_n$, for all h, where $k_h > 0$ is a proportionality constant. This simple assumption returns $\boldsymbol{G} = (\boldsymbol{S}'\boldsymbol{S})^{-1}\boldsymbol{S}'$ so that the base forecasts are orthogonally projected onto the coherent subspace $\mathfrak{s}$ minimising the Euclidean distance between $\hat{\boldsymbol{y}}_{T+h|T}$ and $\tilde{\boldsymbol{y}}_{T+h|T}$. **?** come to same solution, however from the perspective of the following regression model

$$\hat{\boldsymbol{y}}_{T+h|T} = \boldsymbol{S}\beta_{T+h|T} + \boldsymbol{\varepsilon}_{T+h|T}$$

where $\beta_{T+h|T} = E[\boldsymbol{b}_{T+h}|\boldsymbol{b}_1,.....,\boldsymbol{b}_T]$ is the unknown conditional mean of the bottom-level series and $\boldsymbol{\varepsilon}_{T+h|T}$ is the coherence or reconciliation error with mean zero and variance $\boldsymbol{V}$. The OLS solution leads to the same projection matrix $\boldsymbol{S}(\boldsymbol{S}'\boldsymbol{S})^{-1}\boldsymbol{S}'$, and due to this interpretation we continue to refer to this reconciliation method as OLS. A disadvantage of the OLS solution is that the homoscedastic diagonal entries do not account for the scale differences between the levels of the hierarchy due to aggregation. Furthermore, OLS does not account for the correlations across series. I have issues with the next sentence that need to be discussed. We should note that using the usual GLS estimator in this context is not possible as $\boldsymbol{V}$ is not identifiable as shown by **?** who provide the alternative solutions that follow.

- Set $\boldsymbol{W}_h = k_h \text{diag}(\hat{\boldsymbol{W}}_1)$ for all h, where $k_h > 0$ and

$$\hat{\boldsymbol{W}}_1 = \frac{1}{T}\sum_{T=1}^{T} \hat{\boldsymbol{e}}_t \hat{\boldsymbol{e}}_t'$$

  is the unbiased sample estimator of the in-sample one-step-ahead base forecast errors as defined in (3). Hence this estimator scales the base forecasts using the variance of the in-sample residuals and is therefore described and referred to as a weighted least squares (WLS) estimator applying variance scaling. A similar estimator was proposed by Hyndman et al. (2016).

  An alternative WLS estimator is proposed by **?** in the context of temporal hierarchies. Here $\boldsymbol{W}_h$ is proportional to $\text{diag}(\boldsymbol{S}\mathbf{1})$ where $\mathbf{1}$ being a unit column vector of dimension n. Here weights are proportional to the number of bottom level variables required to form an aggregate. For example in the hierarchy in Figure 1 the weights corresponding to the Total, series A and series B are proportional to 5, 2 and 3 respectively. This weighting scheme depends only on the aggregation structure and is referred to as structural scaling. Its advantage over OLS is that it assumes equivariant forecast errors only at the bottom-level of the structure and not across all levels. It is particularly useful in cases where forecast errors are not available; for example, in cases where the base forecasts are generated by judgemental forecasting.
- Set $\boldsymbol{W}_h = k_h \hat{\boldsymbol{W}}_1$, for all h, where $k_h > 0$, the unrestricted sample covariance estimator for $h = 1$. Although this is relatively simple to obtain and provides a good solution for small hierarchies, it does not provide reliable results as m grows compared to T. This is referred to this as the MinT(Sample) estimator.
- Set $\boldsymbol{W}_h = k_h \hat{\boldsymbol{W}}_1^D$, for all h, where $k_h > 0$, $\hat{\boldsymbol{W}}_1^D = \lambda_D \text{diag}(\hat{\boldsymbol{W}}_1) + (1-\lambda_D)\hat{\boldsymbol{W}}_1$ is a shrinkage estimator with diagonal target, and shrinkage intensity parameter

$$\hat{\lambda}_D = \frac{\sum_{i\neq j} \hat{Var}(\hat{r}_{ij})}{\sum_{i\neq j} \hat{r}_{ij}^2},$$

  where $\hat{r}_{ij}$ is the ijth element of $\hat{\boldsymbol{R}}_1$, the 1-step-ahead sample correlation matrix as proposed by Schäfer & Strimmer (2005). Why use r instead of w? Hence, off-diagonal elements of $\hat{\boldsymbol{W}}_1$ are shrunk towards zero while diagonal elements (variances) remain unchanged. This is referred to as the MinT(Shrink) estimator.

4 Hierarchical probabilistic forecasting

A limitation of point forecasts is that they provide no indication of uncertainty around the forecast. A richer description of forecast uncertainty can be obtained by providing a “probabilistic forecasts”, that is a full density for the target of interest. For a review of probabilistic forecasts, and methods for evaluating such forecasts known as *scoring rules* see (Gneiting & Katzfuss 2014). In recent years, the use of probabilistic forecasts and their evaluation via scoring rules has become pervasive in macroeconomic forecasting, for example need to find some references that use scoring rules for macro forecasting. Check Bayesian macro guys like Koop Korobilis, Josh Chan also Mike Smith’s work with Shaun Vahey.

The literature on hierarchical probabilistic forecasting is still an emerging area of interest. To the best of our knowledge the first attempt to even define coherence in the setting of probabilistic forecasting is provided by Taieb et al. (2017) who define a coherent forecast in terms of a convolution. An equivalent definition, provided by Gamakumara et al. (2018) defines a coherent probabilistic forecast as a probability measure on the coherent subspace $\mathfrak{s}$. Gamakumara et al. (2018) also generalise the concept of forecast reconciliation to the probabilistic setting.

Definition 2. Let $\mathscr{A}$ be a subset[1] of $\mathfrak{s}$ and let $\mathscr{B}$ be all points in $\mathbb{R}^n$ that are mapped onto $\mathscr{A}$ after premultiplication by $\boldsymbol{SG}$. Letting $\hat{\nu}$ be a ‘base’ probabilistic forecast for the full hierarchy, the coherent measure $\tilde{\nu}$ reconciles $\hat{\nu}$ if $\tilde{\nu}(\mathscr{A}) = \hat{\nu}(\mathscr{B})$ for all $\mathscr{A}$.

In practice this definition leads to two approaches. For some parametric distributions, for instance the multivariate normal, a reconciled probabilistic forecast can be derived analytically. However, in macroeconomic forecasting, non-standard distributions such as bimodal distributions are often required to take different policy regimes into account worth checking if any (marginal) predictives are bimodal before we include this statement. In such cases a non-parametric approach based on bootstrapping in-sample errors proposed Gamakumara et al. (2018) can be used. These scenarios are now covered in detail.

4.1 Probabilistic forecast reconciliation in the Gaussian framework

In the case where the base forecasts are probabilistic forecasts characterised by elliptical distributions Gamakumara et al. (2018) show that reconciled probabilistic forecasts will also be elliptical. This is particularly straightforward for the Gaussian distribution which is completely characterised by two moments. Letting the base probabilistic forecast be $\mathscr{N}(\hat{\boldsymbol{y}}_{T+h|T}, \hat{\boldsymbol{\Sigma}}_{T+h|T})$, then the reconciled probabilistic forecast will be $\mathscr{N}(\tilde{\boldsymbol{y}}_{T+h|T}, \tilde{\boldsymbol{\Sigma}}_{T+h|T})$, where,

[1] Strictly speaking $\mathscr{A}$ is a Borel set

$$\tilde{\boldsymbol{y}}_{T+h|T} = \boldsymbol{S}\boldsymbol{G}\hat{\boldsymbol{y}}_{T+h|T}, \tag{5}$$

and

$$\tilde{\boldsymbol{\Sigma}}_{T+h|T} = \boldsymbol{S}\boldsymbol{G}\hat{\boldsymbol{\Sigma}}_{T+h|T}\boldsymbol{G}'\boldsymbol{S}'. \tag{6}$$

There are several options for obtaining the base probabilistic forecast and in particular the variance covariance matrix $\hat{\boldsymbol{\Sigma}}$. One option is to fit multivariate models either level by level or for the hierarchy as a whole leading respectively to a $\hat{\boldsymbol{\Sigma}}$ that is block diagonal or dense. Another alternative is to fit univariate models for each individual series in which case $\hat{\boldsymbol{\Sigma}}$ is a diagonal matrix. Due to the large number of series under investigation here we consider the latter option. However we emphasise that correlation will enter the probabilistic forecast after reconciliation. The reconciled probabilistic forecast will ultimately depend on the choice of $\boldsymbol{G}$; the same choices of $\boldsymbol{G}$ matrices used in section 3 can be used.

Need to check with Puwasala that base forecasts have diagonal sigma hat I have taken the variance covariance matrix of the base forecast errors as the $\hat{\boldsymbol{\Sigma}}$. Thus $\hat{\boldsymbol{\Sigma}}$ is same as $\hat{\boldsymbol{W}}_1$, and I have taken the shrinkage estimator for $\hat{\boldsymbol{W}}_1$

4.2 *Probabilistic forecast reconciliation in the non-parametric framework*

In many applications, including macroeconomic forecasting, it may not reasonable to assume Gaussian predictive distributions. Therefore, non-parametric approaches has been widely used for probabilistic forecasts in different disciplines. For example, ensemble forecasting in weather applications (Gneiting & Raftery (2005), Gneiting & Katzfuss (2014), Gneiting et al. (2008)), bootstrap based approaches (Manzan & Zerom (2008), Vilar & Vilar (2013)). Check/replace these references with references that show heavy tails/skewness in macro applications.

Due to these concerns, we employ a reconciliation method proposed by Gamakumara et al. (2018) that does not make parametric assumptions about the predictive distribution. An important result that this method exploits is that applying methods for point forecast reconciliation to the draws from incoherent base predictive distribution results in a sample from the reconciled predictive distribution. This process, is summarised

1. Fit univariate models to each series in the hierarchy over a training set from $t = 1$ to $t = T$
2. For each series compute h-step ahead point forecasts, for all h up to H. Collect these into a $n \times H$ matrix $\hat{\boldsymbol{Y}} := (\hat{\boldsymbol{y}}_{T+1|T}, \ldots, \hat{\boldsymbol{y}}_{T+H|T})$, where $\hat{\boldsymbol{y}}_{T+h|T}$ is a $n \times 1$ vector of h-step point forecasts for all series in the hierarchy.
3. Compute one-step ahead in-sample forecasting errors. Collect these into an $n \times T$ matrix $\hat{\boldsymbol{E}} = (\hat{\boldsymbol{e}}_1, \hat{\boldsymbol{e}}_2, \ldots\ldots, \hat{\boldsymbol{e}}_T)$, where the $n \times 1$ vector $\hat{\boldsymbol{e}}_t = \boldsymbol{y}_t - \hat{\boldsymbol{y}}_{t|t-1}$. Here, $\hat{\boldsymbol{y}}_{t|t-1}$ is a vector of forecasts made for time t using information up to and including $t-1$. These are called in-sample forecasts since they depend on past values but

information from the entire training sample is used to estimate parameters that forecasts are based on.

4. Block bootstrap from $\hat{\boldsymbol{E}}$, that is choose H consecutive columns of $\hat{\boldsymbol{E}}$ at random, repeating this process B times. Denote the $n \times H$ matrix obtained at iteration b as $\hat{\boldsymbol{E}}^b$ for b=1,...,B.
5. For all b, compute $\hat{\boldsymbol{\Upsilon}}^b := \hat{\boldsymbol{Y}} + \hat{\boldsymbol{E}}^b$. Each row of $\hat{\boldsymbol{\Upsilon}}^b$ is a sample path of h forecasts for a single series. Each column of $\hat{\boldsymbol{\Upsilon}}^b$ is a realisation from the joint predictive distribution at a particular horizon.
6. For each $b = 1, \dots, B$ select the h^{th} column of $\hat{\boldsymbol{\Upsilon}}^b$ and stack these to form a $n \times B$ matrix $\hat{\boldsymbol{\Upsilon}}_{T+h|T}$
7. For a given $\boldsymbol{G}$ matrix and for each $h = 1, \dots, H$ compute $\tilde{\boldsymbol{\Upsilon}}_{T+h|T} = \boldsymbol{S}\boldsymbol{G}\hat{\boldsymbol{\Upsilon}}_{T+h|T}$. Each column of $\tilde{\boldsymbol{\Upsilon}}_{T+h|T}$ is a realisation from the joint h-step ahead reconciled predictive distribution.

Check with Puwasala that this is exactly what she has done. these steps are correct. Notation may need work to bring in line with previous sections.

5 Empirical Study: Australian GDP

In our empirical data we consider Gross Domestic Product (GDP) of Australia with quarterly data spanning the period 1984:Q4-2018:Q3. The Australian Bureau of Statistics (ABS) measures GDP using three main approaches namely, Production, Income and Expenditure. The final GDP figure is obtained as an average of these three figures. Each of these measures are aggregates of economic variables which are also targets of interests for the macroeconomic forecaster. This suggests a hierarchical approach to forecasting could be used to improve forecasts of all series in the hierarchy including headline GDP.

We concentrate on the Income and Expenditure approaches as nominal data are available only for these two. We focus only on nominal data due to the fact that real data are constructed via a chain price index approach with different price deflators used for each series. As a result, real GDP data are not coherent - the aggregate series is not a linear combination of the disaggregate series. For similar reasons we do not use seasonally adjusted data; the process of seasonal adjustment results in data that are not coherent. Finally, although there is a small statistical discrepancy between each series and the headline GDP figure, we simply treat this statistical discrepancy, which is also published by the ABS, as a time series in its own right. For further details on the data we refer the reader to Australian Bureau of Statistics (2018).

Income approach

Using the income approach, GDP is calculated by aggregating all income flows. In particular, GDP at purchaser's price is the sum of all factor incomes and taxes, minus subsidies on production and imports. are these descriptions correct? Can we please double-check. (Australian Bureau of Statistics 2015):

$$\begin{aligned}
GDP = {} & \textit{Compensation of employees} + \textit{Gross operating surplus} \\
& + \textit{Gross mixed income} + \textit{Taxes on production and imports} \\
& - \textit{Subsidies on production and imports} + \textit{Statistical discrepancy (I)}
\end{aligned}$$

Figure 4 shows the full hierarchical structure capturing all components aggregated to form GDP using the income approach. The hierarchy has two levels of aggregation below the top-level, with a total of $n = 16$ series across the whole structure and $m = 10$ series at the bottom-level.

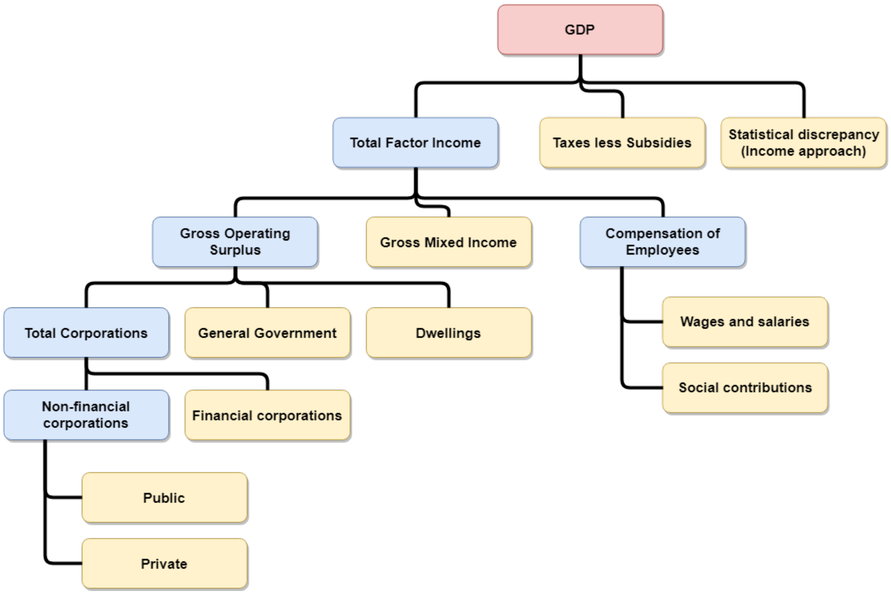

Fig. 4 Hierarchical structure of the income approach for GDP. The pink cell contains GDP the most aggregate series. The blue cells contain intermediate-level series and the yellow cells to the most disaggregate bottom-level series.

Expenditure approach

In the expenditure approach, GDP is calculated as the aggregation of final consumption expenditure, gross fixed capital formation (GFCF), changes in inventories of finished goods, work-in-progress and raw materials and the value of exports less imports of the goods and services (Australian Bureau of Statistics 2015). Underline equation is,

$$\begin{aligned} GDP &= Final\ consumption\ expenditure + Gross\ fixed\ capital\ formation \\ &\quad + Changes\ in\ inventories + Trade\ balance + Statistical\ discrepancy\ (E) \end{aligned}$$

Figures 5, 6 and 7 show the full hierarchical structure capturing all components aggregated to form GDP using the expenditure approach. The hierarchy has three levels of aggregation below the top-level, with a total of $n = 80$ series across the whole structure and $m = 53$ series at the bottom-level.

THIS NEEDS TO BE ADDED: description of each series in these hierarchies along with the series ID assigned by the ABS is given in the tables **??**, **??**, **??** and **??** in the supplementary materials.

Figure 8 displays time series from the income and expenditure approaches. Top panel shows the most aggregate GDP series. The bottom panel shows series from below levels for income hierarchy in left panel and expenditure hierarchy in right panel. The plots show the diverse features of the time series with some displaying positive and others negative trending behaviour, some showing no trends but possibly a cycle, and some having a strong seasonal component. These highlight the need to account and model all information and diverse signals from each series in the hierarchy which can only be achieved through a forecast reconciliation approach.

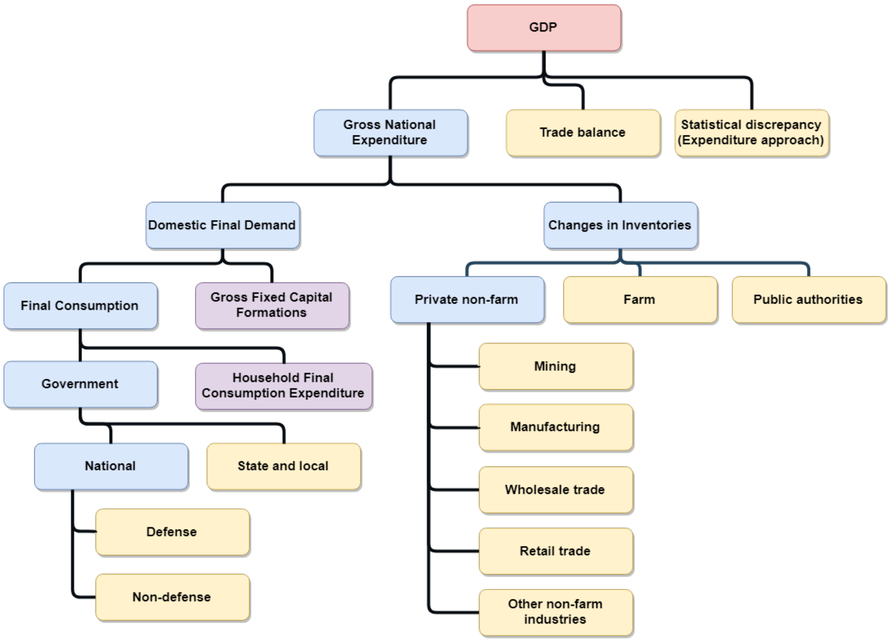

Fig. 5 Hierarchical structure of the expenditure approach for GDP. The pink cell contains GDP, the most aggregate series. The blue and purple cells contain intermediate-level series with the series in the purple cells further disaggregated in Figures 6 and 7. The yellow cells contain the most disaggregate bottom-level series.

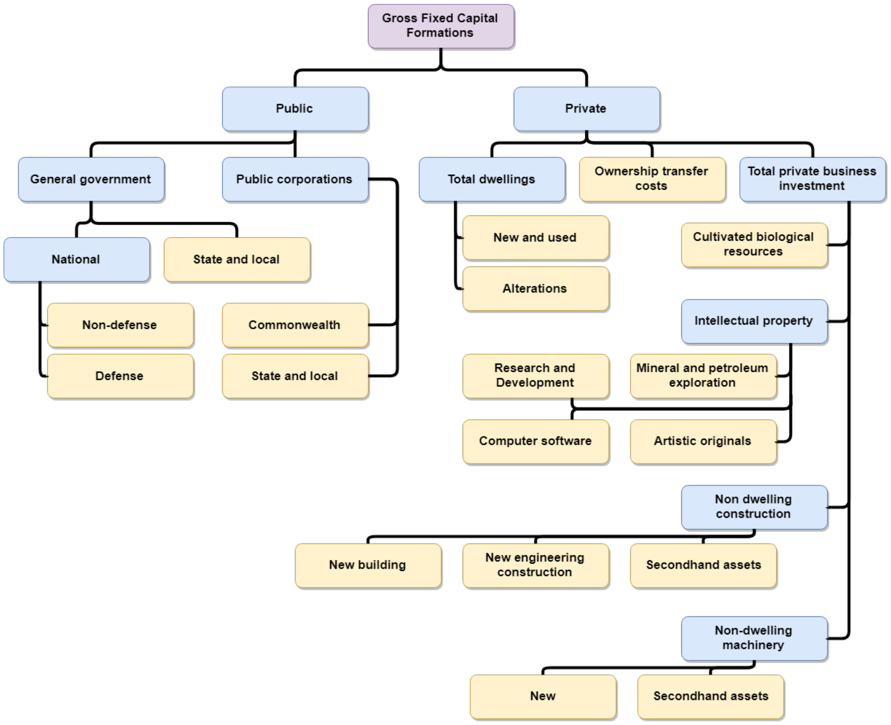

Fig. 6 Hierarchical structure for Gross Fixed Capital Formations under the expenditure approach for GDP, continued from Figure 5. Blue cells contain intermediate-level series and the yellow cells to the most disaggregate bottom-level series.

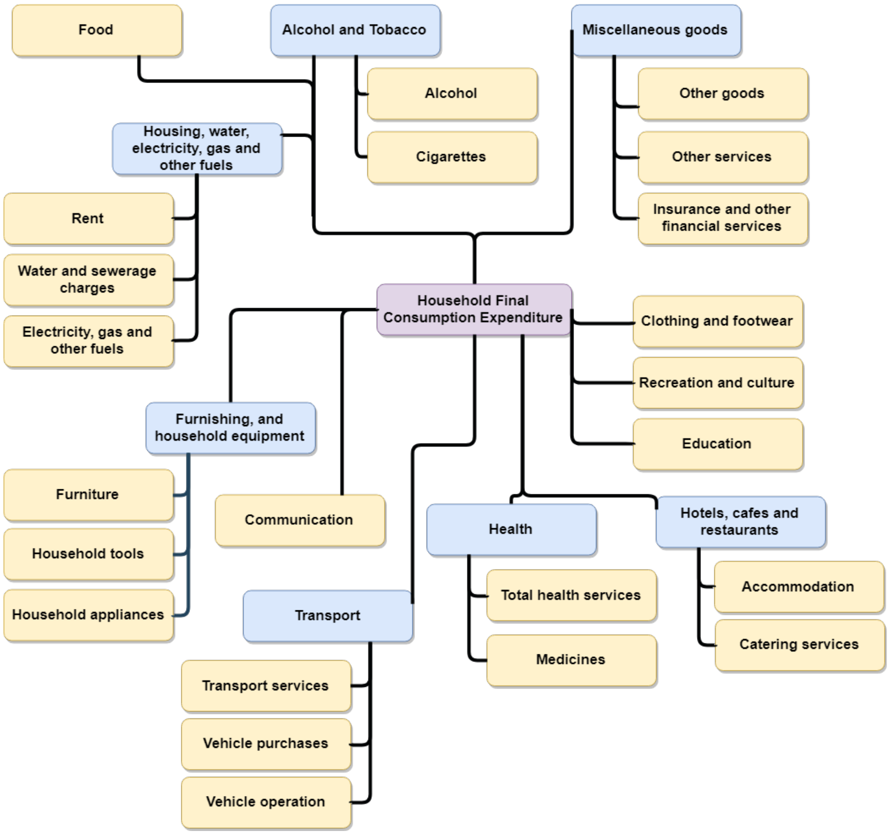

Fig. 7 Hierarchical structure for Household Final Consumption Expenditure under expenditure approach for GDP, continued from Figure 5. Blue cells contain intermediate-level series and the yellow cells to the most disaggregate bottom-level series..

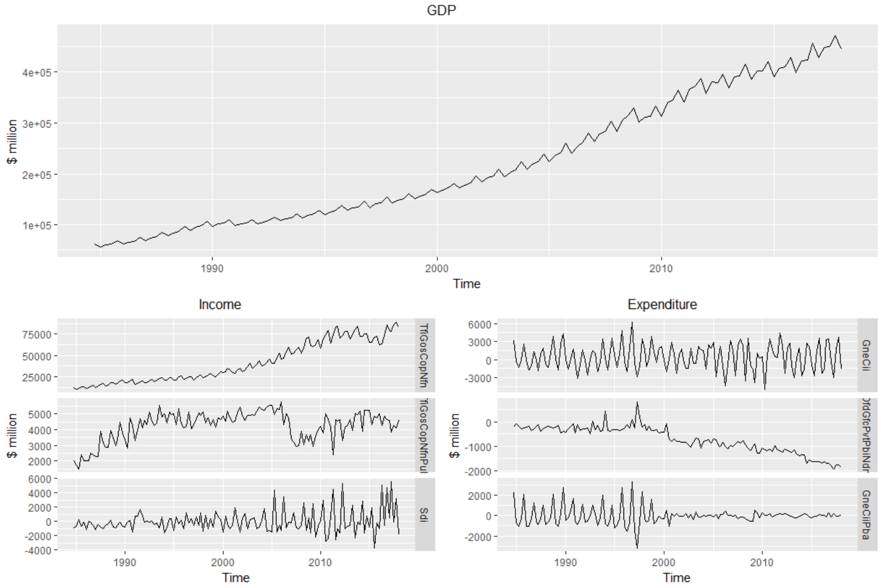

Fig. 8 Time plots for series from different levels of income and expenditure hierarchies.

6 Methodology

We now demonstrate the potential for reconciliation methods to improve forecast accuracy for Australian GDP data. We consider forecasts from $h = 1$ quarter ahead forecasts up to $h = 4$ quarter ahead using an *expanding* window. First, the training sample is set to Q4 of 1984 to Q3 of 1994 and forecasts are produced for Q4 of 1994 to Q3 of 1995. Then the training window is expanded one period ahead, i.e. from Q4 of 1984 to Q4 of 2017 with forecasts produces for the last observation at Q1 of 2018. All up this leads to 94 1-step-ahead, 93 2-step-ahead, 92 3-step-ahead and 91 4-step-ahead forecasts.

Need to get these dates off Puwasala. Also last section may need to be changed

6.1 Models

The first task in forecasting reconciliation is to obtain base forecasts for all series in the hierarchy. In the case of the income approach this necessitates forecasting $n = 16$ separate time series while in the case of the expenditure approach forecasts for $n = 80$ separate time series must be obtained. Given the diversity in these time series discussed in Section 5, our focus was on a methodology that was fast but flexible. We consider simple univariate ARIMA models, where model order is selected via a combination of unit root testing and AIC using an algorithm developed by XXX and implement in the auto.arima function in XXX. Cite this to Rob's satisfaction. A similar approach was also undertaken using the ETS framework to produce base forecasts. This had minimal impact on our conclusions with respect to forecast reconciliation methods, and in most cases ARIMA forecasts outperformed ETS forecasts. Consequently, results for ETS models are excluded but are available from the authors upon request again do we put this in an appendix? Could referring to Puwasala's thesis be an option. There will be some editing of the book before it comes out. Just a suggestion.. We note that a number of more complicated approaches could have been used to obtain base forecasts including multivariate models such as vector autoregression and models and methods that handle a large number of predictors such as factor models or least angle regresssion. However, **?** show that univariate ARIMA models are highly competitive for forecasting Australian GDP even compared to these methods, and in any case our primary motivation is to demonstrate the potential of forecast reconciliation.

The forecast reconciliation approaches that we consider are bottom up, OLS, WLS with variance scaling and the MinT (shrink) approach. The MinT (sample) approach was also used but due to the size of the hierarchy forecasts reconciled via this approach were less stable. Finally, all forecasts both base and reconciled are compared to a naïve benchmark. Since the data are not deseasonalised, the naïve benchmark is a seasonal random walk, i.e. the forecast for GDP (or one of its components) is the realised GDP in the same quarter of the previous year. The naïve forecast is by construction coherent and therefore does not need to be reconciled.

6.2 Evaluation

For evaluating point forecasts we consider two metrics, the Mean Squared Error (MSE) and the Mean Absolute Scaled Error (MASE). The absolute scaled error is defined as

$$q_{T+h} = \sum \frac{|\breve{e}_{T+h|T}|}{(T-4)^{-1}\sum_{t=5}^{T}|y_t - y_{t-4}|},$$

where $\breve{e}_{t+h}$ is the difference between any forecast and the realisation[2] and 4 is used due to the quarterly nature of the data we consider. An advantage of using MASE is that it is a scale independent measure. This is particularly relevant for hierarchical time series, since aggregate series by their very nature are on a larger scale than disaggregate series. As such scale dependent metrics may unfairly favour methods that perform well for the aggregate series but poorly for disaggregate series. For more details on different point forecast accuracy measures refer to (Chapter 3 of **?**).

Forecast accuracy of probabilistic forecasts can be evaluated using scoring rules (Gneiting & Katzfuss 2014). Let $\breve{F}$ be a probabilistic forecast and let $\breve{\boldsymbol{y}} \sim \breve{F}$ where breve is used to denote that either base forecast or reconciled forecast can be evaluated. The accuracy of multivariate probabilistic forecasts will be measured by the energy score given by

$$eS(\breve{F}_{T+h|T}, \boldsymbol{y}_{T+h}) = E_{\breve{F}}\|\breve{\boldsymbol{y}}_{T+h} - \boldsymbol{y}_{T+h}\|^{\alpha} - \frac{1}{2}E_{\breve{F}}\|\breve{\boldsymbol{y}}_{T+h} - \breve{\boldsymbol{y}}^{*}_{T+h}\|^{\alpha},$$

where $\boldsymbol{y}_{T+h}$ is the realisation at time $T+h$, $\alpha \in (0,2]$. What did we use for alpha?we use alpha = 1. The expectations can be evaluated numerically as long as a sample from $\breve{F}$ is available which is the case for all methods we employ. An advantage of using energy score is that in the univariate case it simplifies to the commonly used cumulative rank probability score (CRPS) given by

$$\text{CRPS}(\breve{F}_i, y_{i,T+h}) = E_{\breve{F}_i}|\breve{y}_{i,T+h} - y_{i,T+h}| - \frac{1}{2}E_{\breve{F}_i}|\breve{y}_{i,T+h} - \breve{y}^{*}_{i,T+h}|,$$

where the subscript i is used to denote that CRPS measures forecast accuracy for a single variable in the hierarchy.

As an alternative to the energy score, log score and variogram scores were also considered. The log score was disregarded since Gamakumara et al. (2018) prove that the log score is improper with respect to the class of incoherent probabilistic forecasts when the true DGP is coherent. The variogram score gave similar results to the energy score; variogram score results are omitted for brevity but are available from the authors upon request. or we put them in an appendix

[2] Breve is used instead of a hat or tilde to denote that this can be the error either a base or reconciled forecast

7 Results

7.1 Base forecasts

Due to the different features in each time series a variety of ARIMA models were selected to be used as base models. Some generalisations about models. Figure 9 gives some indication of the performance of these base forecasting models relative to the naïve forecast over different horizons. Panels on the left refer to results for the Income hierarchy with panels on the right referring to the expenditure hierarchy. The top panels summarise results over all series in the hierarchy How exactly do we do this. Do we compute MSE for each series then add? Do we compute skill score for each series and average these?. we compute MSE for each series then average over series The clear result is that base forecasts are more accurate than naïve forecasts, however as the forecasting horizon increases, the difference becomes smaller. This is to be expected since the naïve model here is a seasonal random walk, for horizons $h < 4$ the forecast from an ARIMA model is based on more recent information.

Similar results are obtained when MASE is used as the metric for evaluating forecast quality. However one disadvantage of the base forecasts relative to the naïve forecast is that base forecasts are no longer coherent. As such we now turn out attention to investigating whether reconciliation methods can lead to further improvements in forecast accuracy relative to base forecasts.

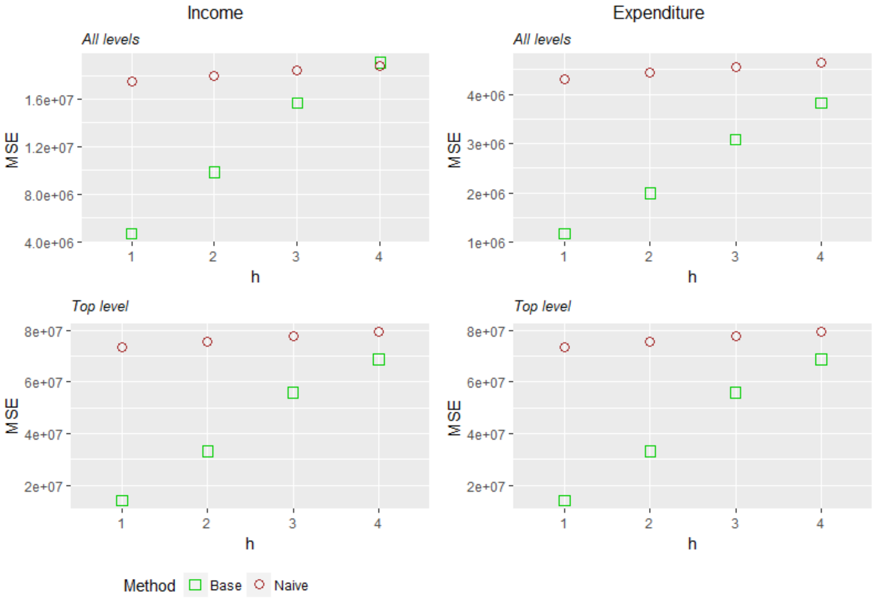

Fig. 9 Mean squared errors for naive and base forecasts. Top panels refer to results summarised over all series and bottom panels refer to the top-level GDP series. Left panels refer to the income hierarchy and right panels to the expenditure hierarchy

7.2 *Point Forecast Reconciliation*

We now turn our attention to an evaluation of point forecasts obtained using different reconciliation methods. All results in subsequent figures are presented as the percentage changes in a forecasting metric relative to the base forecast, a measure known in the forecasting literature as *skill scores*. Skill scores are computed such that positive values represent an improvement in forecasting performance over the base forecast and negative values represent a deterioration in forecast quality. The top row of Figure 10 shows skill scores based on the MSE and MASE. These are aggregated over all series for both the income hierarchy and expenditure hierarchy and over different forecast horizons Puwasala to confirm whether we are looking at a skill score based on average MSE or an average of skill scores We are looking at a skill score based on average MSE. We conclude that reconciliation methods generally improve forecast accuracy relative to base forecasts regardless of the hierarchy used, the forecasting horizon, the forecasting metric used to evaluate forecasts or the specific reconciliation method employed. We do however note that while all reconciliation methods improve forecast performance, MinT (shrink) is the best forecasting method in most cases.

To further investigate the differences between reconciliation methods we break down these results by different levels of each hierarchy. The second row of Figure

10 shows the forecasting performance a single series, namely GDP which represents the top level of both hierarchies. The third row shows results for all series excluding those on the bottom level, while the final row shows results for the bottom level series only. Here, we see two general features, the first is that OLS reconciliation performs poorly on the bottom level series, the second is that bottom up performs relatively poorly on aggregate series. These two features are particularly exacerbated for the larger expenditure hierarchy. These results are consistent with other findings in the forecast reconciliation literature see for instance XXX. Some papers from forecast reconciliation literature to be cited here

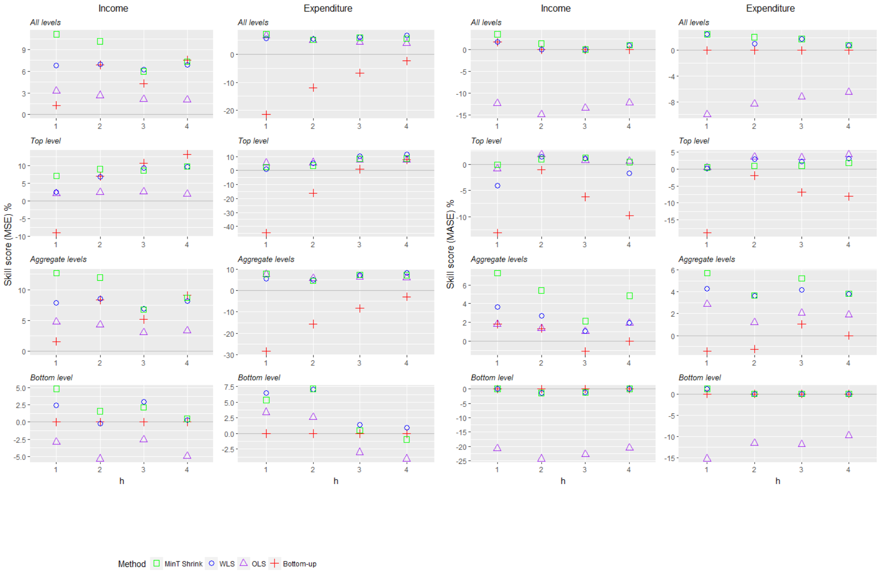

Fig. 10 Skill score point forecasts from different reconciliation methods (with reference to base forecasts). Left two panels refer to skill score using MSE for income and expenditure hierarchies. Similarly right two panels refer to skill score using MASE. First row refers to results summarised over all series, second row to top-level GDP series, third row to aggregate levels and last row to the bottom level.

Tas thinks two 4x2 panels would be better

7.3 *Probabilistic Forecast Reconciliation*

We now turn our attention towards results for probabilistic forecasts. Figure 11 reports the energy score which as a multivariate score summarises forecast accuracy over the entire hierarchy. Once again all results are presented as skill scores relative to base forecasts. The top panels refer to assuming Gaussian probabilistic forecasts as described in Section 4.1 while the bottom panels refer to the method described in Section 4.2. The left panels correspond to the income hierarchy, the right panels to the expenditure hierarchy. For the income hierarchy, all methods improve upon base forecasts at all horizons. In nearly all cases the best performing reconciliation method is MinT (shrink), a notable result since the optimal properties for MinT have thus far only been established theoretically in the point forecasting case. For the larger expenditure hierarchy results are little more mixed. While bottom up tends to perform poorly, all other reconciliation methods improve upon base forecasts (with the single exception of MinT (shrink) in the Gaussian framework four quarters ahead). Interestingly, OLS performs best under the assumption of Gaussianity - this may indicate that OLS is a more robust method under model misspecification but further investigation is required.

Finally, Figure 12 displays the skill scores based on the cumulative ranked probability score for a single series, namely top-level GDP. The cause of the poor performance of bottom up reconciliation as a failure to accurately forecast aggregate series is apparent here.

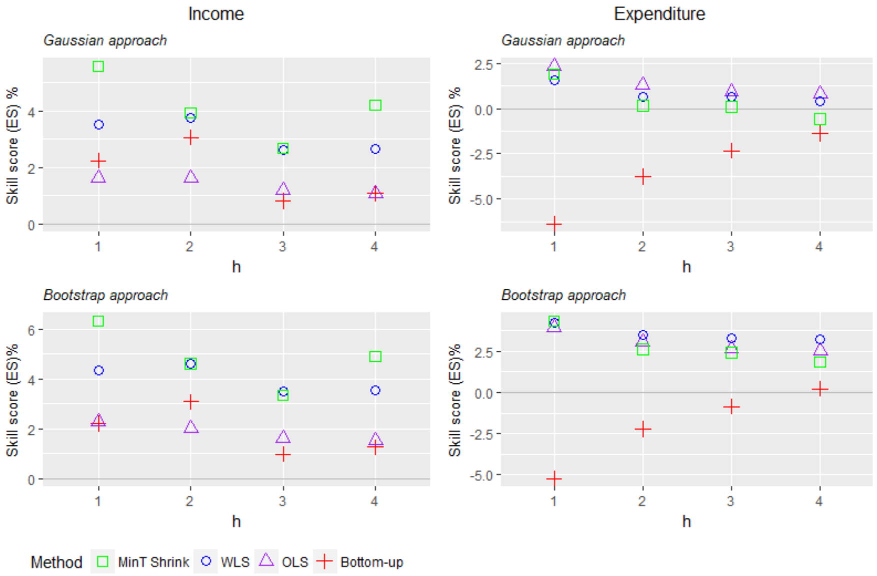

Fig. 11 Skill score for multivariate probabilistic forecasts from different reconciliation methods (with reference to base forecasts) using energy score. Top panels refer to the results for Gaussian approach and bottom panels to the non-parametric bootstrap approach. Left panels refer to the income hierarchy and right panels to the expenditure hierarchy.

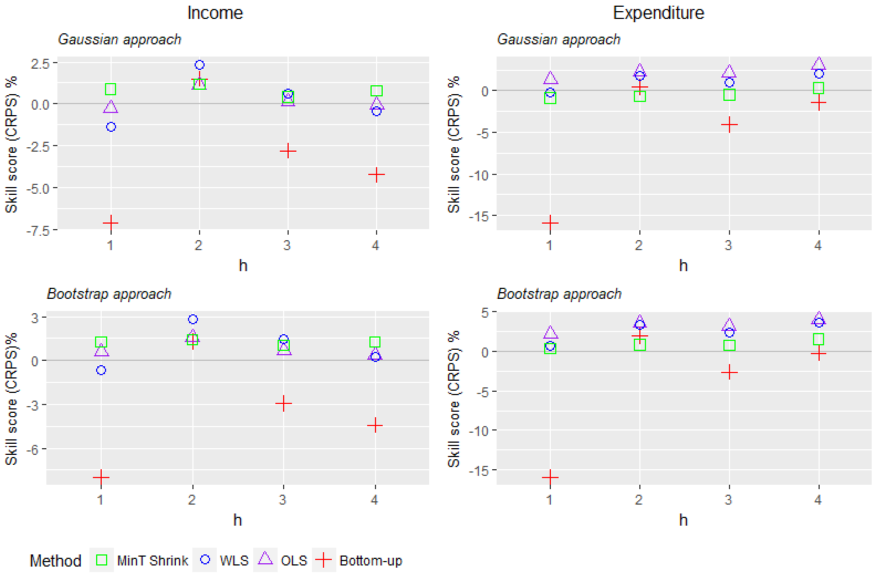

Fig. 12 Skill score for probabilistic forecasts of top-level GDP from different reconciliation methods (with reference to base forecasts) using CRPS. Top panels refer to the results for Gaussian approach and bottom panels refer to the non-parametric bootstrap approach. Left panel refers to the income hierarchy and right panel to the expenditure hierarchy.

Possible discuss skill scores in more detail especially w.r.t how they are combined

8 Conclusions

In the macroeconomic setting, we have demonstrated the potential for forecast reconciliation methods to not only provide coherent forecasts, but to also improve overall forecast accuracy. This result holds for both point forecasts and probabilistic forecasts, for two different hierarchies and over different forecasting horizons. Even where the objective is to only forecast a single series, for instance top-level GDP, the application of forecast reconciliation methods improves forecast accuracy.

By comparing results from different forecast reconciliation techniques we make a number of conclusions. Despite its simplicity, bottom up can perform poorly at more aggregated levels of the hierarchy, a result found elsewhere in the literature. Meanwhile, when forecast accuracy at the bottom level is evaluated, OLS tends to break down in some instances. Overall, the WLS and MinT (shrink) methods, and particularly the latter tend to yield the highest improvements in forecast accuracy.

There are a number of open avenues for research in the literature on forecast reconciliation, some of which are particularly relevant to macroeconomic applications. First there is scope to consider more complex aggregation structures, for instance in addition to the hierarchies we have already considered, data on GDP and GDP components disaggregated along geographical lines are also available. This leads to a grouped hierarchy. Also, given the substantial literature on the optimal frequency at which to analyse macroeconomic data, a study on forecasting GDP or other variables as a temporal hierarchy may be of interest. In this chapter we have only shown that reconciliation methods can be used to improve forecast accuracy when univariate ARIMA models are used to produce base forecasts. It will be interesting to evaluate whether such results hold when multivariate approaches, e.g. a Bayesian VAR or dynamic factor model, is used as a base forecasting model, or whether the gains from forecast reconciliation would be more modest. Finally, a current limitation of the forecast reconciliation literature is that it only applies to collections of time series that adhere to linear constraints. In macroeconomics there are many examples of data that adhere to non-linear constraints, for instance real GDP is a complicated but deterministic function of GDP components and price deflators. The extension of forecast reconciliation methods to non-linear constraints potentially holds great promise for continued improvement in macroeconomic forecasting.

References

Australian Bureau of Statistics (2015), Australian System of National Accounts: Concepts, Sources and Methods, Technical report.
URL: *http://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/5216.02015?OpenDocument*

Australian Bureau of Statistics (2018), Australian National Accounts : National Income, Expenditure and Product, Technical report.
URL: *http://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/5206.0Sep 2018?OpenDocument*

Dunn, D. M., Williams, W. H. & Dechaine, T. L. (1976), 'Aggregate Versus Subaggregate Models in Local Area Forecasting', *Journal of American Statistical Association* **71**(353), 68–71.

Gamakumara, P., Panagiotelis, A., Athanasopoulos, G. & Hyndman, R. J. (2018), Probabilisitic Forecasts in Hierarchical Time Series.

Gneiting, T. & Katzfuss, M. (2014), 'Probabilistic Forecasting', *Annual Review of Statistics and Its Application* **1**, 125–151.

Gneiting, T. & Raftery, A. E. (2005), 'Weather forecasting with ensemble methods', *Science* **310.5746**, 248–249.

Gneiting, T., Stanberry, L. I., Grimit, E. P., Held, L. & Johnson, N. A. (2008), 'Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds', *Test* **17**(2), 211–235.

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

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

Manzan, S. & Zerom, D. (2008), 'A bootstrap-based non-parametric forecast density', *International Journal of Forecasting* **24**(3), 535–550.

Schäfer, J. & Strimmer, K. (2005), 'A Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Functional Genomics', *Statistical Applications in Genetics and Molecular Biology* **4**(1).
URL: *https://www.degruyter.com/view/j/sagmb.2005.4.issue-1/sagmb.2005.4.1.1175/sagmb.2005.4.1.1175.xml*

Taieb, S. B., Taylor, J. W. & Hyndman, R. J. (2017), 'Hierarchical Probabilistic Forecasting of Electricity Demand with Smart Meter Data', pp. 1–30.

Vilar, J. A. & Vilar, J. A. (2013), 'Time series clustering based on nonparametric multidimensional forecast densities', *Electronic Journal of Statistics* **7**(1), 1019–1046.