Optimal forecast reconciliation with time series selection

Xiaoqian Wang, Rob J Hyndman, Shanika L Wickramasuriya

Optimal forecast reconciliation with time series selection

Xiaoqian Wang
Monash University, VIC 3800, Australia
Email: xiaoqian.wang@monash.edu
Corresponding author

Rob J Hyndman
Monash University, VIC 3800, Australia
Email: rob.hyndman@monash.edu

Shanika L Wickramasuriya
Monash University, VIC 3145, Australia
Email: shanika.wickramasuriya@monash.edu

2 October 2023

Optimal forecast reconciliation with time series selection

Abstract

Abstract

---

1 Introduction

Hierarchical time series and forecast reconciliation. Post-processing.

Single-level approaches, least squares-based reconciliation approaches, geometric intuition, other extensions with constraints. Forecast combinations.

However... Two issues. The choice of W can have significant effect on the quality of the reconciled forecasts. Some time series perform poorly. Forecast trimming.

In this paper, our focus will be on... selection. Other entries will be adjusted accordingly.

The remainder of the paper is structured as follows.

2 Preliminaries

2.1 Notation

We denote the set $\{1, \ldots, k\}$ by $[k]$ for any non-negative integer k. A *hierarchical time series* can be considered as an n-dimensional multivariate time series, $\{\boldsymbol{y}_t, t \in [T]\}$, that adheres to known linear constraints. Let $\boldsymbol{y}_t \in \mathbb{R}^n$ be a vector comprising observations of all time series in the hierarchy at time t, and $\boldsymbol{b}_t \in \mathbb{R}^{n_b}$ be a vector comprising observations of all bottom-level time series at time t. The full hierarchy at time t can be written as

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

where S is an $n \times n_b$ *summing matrix* that shows aggregation constraints present in the structure. We can write the summing matrix as $S = \begin{bmatrix} A \\ I_{n_b} \end{bmatrix}$, where A is an $n_a \times n_b$ *aggregation matrix* with $n = n_a + n_b$, and I_{n_b} is an n_b-dimensional identity matrix.
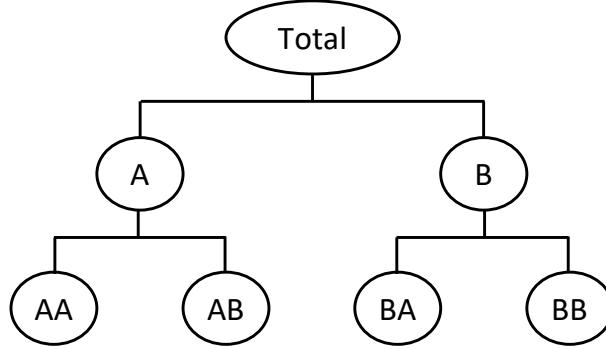

Figure 1: *An example of a two-level hierarchical time series.*

To clarify these notations, consider the example of the hierarchy in Figure 1. For this two-level hierarchy, $n = 7$, $n_b = 4$, $n_a = 3$, $\boldsymbol{y}_t = [y_{\text{Total},t}, y_{\text{A},t}, y_{\text{B},t}, y_{\text{AA},t}, y_{\text{AB},t}, y_{\text{BA},t}, y_{\text{BB},t}]'$, $\boldsymbol{b}_t = [y_{\text{AA},t}, y_{\text{AB},t}, y_{\text{BA},t}, y_{\text{BB},t}]'$, and

$$S = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ & I_4 & & \end{bmatrix}.$$
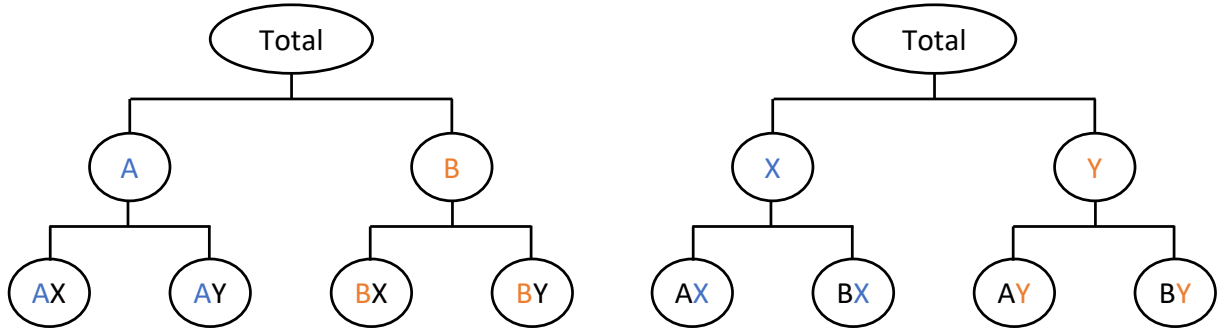

Figure 2: *An example of a two level grouped time series.*

When data structure does not naturally disaggregate in a unique hierarchical manner, we can combine these hierarchical structures to form a *grouped time series*. Thus, grouped time series can also be considered as hierarchical time series with more than one grouping structure. Figure 2 shows an example of a two level grouped time series with two alternative aggregation structures. For this example, $n = 9$, $n_b = 4$, $n_a = 5$, $\boldsymbol{y}_t = [y_{\text{Total},t}, y_{\text{A},t}, y_{\text{B},t}, y_{\text{X},t}, y_{\text{Y},t}, y_{\text{AX},t}, y_{\text{AY},t}, y_{\text{BX},t}, y_{\text{BY},t}]'$, $\boldsymbol{b}_t = [y_{\text{AX},t}, y_{\text{AY},t}, y_{\text{BX},t}, y_{\text{BY},t}]'$, and

$$
S = \begin{bmatrix}
1 & 1 & 1 & 1 \\
1 & 1 & 0 & 0 \\
0 & 0 & 1 & 1 \\
1 & 0 & 1 & 0 \\
0 & 1 & 0 & 1 \\
& I_4 &
\end{bmatrix}.
$$

2.2 Linear forecast reconciliation

Let $\hat{y}_{T+h|T} \in \mathbb{R}^n$ be a vector of h-step-ahead *base forecasts* for all time series in the hierarchy, given observations up to time T, and stacked in the same order as y_t. We can use any method to generate these forecasts, but In general they will not add up especially when we forecast each series independently.

When forecasting hierarchical time series, we expect the forecasts to be *coherent* (i.e., aggregation constraints are satisfied). Let $\tilde{y}_{T+h|T} \in \mathbb{R}^n$ denote a vector of h-step-ahead *reconciled forecasts* which are coherent by construction, ψ a mapping that reconciles base forecasts, $\hat{y}_{T+h|T}$. Then we have *forecast reconciliation* $\tilde{y}_{T+h|T} = \psi(\hat{y}_{T+h|T})$, which is essentially a post-processing method. In this paper, we focus on linear forecast reconciliation given by

$$
\tilde{y}_{T+h|T} = SG_h\hat{y}_{T+h|T},
$$

where

- G_h is an $n_b \times n$ weighting matrix that maps the base forecasts into the bottom level. In other words, it combines all base forecasts to form reconciled forecasts for bottom-level series.
- S is an $n \times n_b$ summing matrix that sums up bottom-level reconciled forecasts to produce coherent forecasts of all levels. It identifies the linear constraints involved in the hierarchy.

2.2.1   Minimum trace reconciliation

Let the h-step-ahead *base forecast errors* be defined as $\hat{e}_{T+h|T} = y_{T+h} - \hat{y}_{T+h|T}$, and the h-step-ahead *reconciled forecast errors* be defined as $\tilde{e}_{T+h|T} = y_{T+h} - \tilde{y}_{T+h|T}$. Wickramasuriya, Athanasopoulos & Hyndman (2019) formulated a linear reconciliation problem as minimizing the trace (MinT) of the h-step-ahead covariance matrix of the reconciled forecast errors, $\text{Var}(\tilde{e}_{T+h|T})$.

Under the assumption of unbiasedness, the unique solution of the minimization problem is given by

$$G_h = \left(S'W_h^{-1}S\right)^{-1}S'W_h^{-1}, \tag{1}$$

where W_h is the positive definite covariance matrix of the h-step-ahead base forecast errors, $\mathrm{Var}(\hat{e}_{T+h|T})$.

The trace minimization problem can be reformulated as a least squares problem with linear constraints given by

$$\min_{\tilde{y}_{T+h|T}} \quad \frac{1}{2}(\hat{y}_{T+h|T} - \tilde{y}_{T+h|T})'W_h^{-1}(\hat{y}_{T+h|T} - \tilde{y}_{T+h|T})$$
$$\text{s.t.} \quad \tilde{y}_{T+h|T} = S\tilde{b}_{T+h|T}, \tag{2}$$

where $\tilde{b}_{T+h|T} \in \mathbb{R}^{n_b}$ is the vector comprising h-step-ahead bottom-level reconciled forecasts, made at time T. Focusing on W_h, the intuitive behind the MinT reconciliation is that the larger the estimated variance of the base forecast errors, the larger the range of adjustments permitted for forecast reconciliation.

It's challenging to estimate W_h, especially for $h > 1$. Assuming that $W_h = k_h W_1$, $\forall h$, where $k_h > 0$, the MinT solution of G does not change with the forecast horizon, h. Hence, we will drop the subscript h for the ease of exposition. The most popularly used candidate estimators for W in the forecast reconciliation literature are listed as follows.

1. $W_{\mathrm{OLS}} = I$ is the *OLS estimator* proposed by Hyndman et al. (2011), assuming that the base forecast errors are uncorrelated and equivariant. In what follows, we denote this as **OLS**.
2. $W_{\mathrm{WLSs}} = \mathrm{diag}(S1)$ is the *WLS estimator applying structural scaling* proposed by Athanasopoulos et al. (2017). This estimator depends only on the aggregation structure of the hierarchy. It assumes that the variance of each bottom-level base forecast error is equivalent and uncorrelated between nodes. We denote this method as **WLSs**.
3. $W_{\mathrm{WLSv}} = \mathrm{diag}(\hat{W}_1)$ is the *WLS estimator applying variance scaling* proposed by Hyndman, Lee & Wang (2016), where $\hat{W}_1$ denotes the unbiased covariance estimator based on the in-sample one-step-ahead base forecast errors (i.e., residuals). In the results that follow, we denote this as **WLSv**.

4. $W_{\text{MinT}} = \hat{W}_1$ is referred to as the *MinT estimator* based on the sample covariance matrix proposed by Wickramasuriya, Athanasopoulos & Hyndman (2019). We denote this method as **MinT** in the results that follow.

5. $W_{\text{MinTs}} = \lambda \, \text{diag}(\hat{W}_1) + (1 - \lambda)\hat{W}_1$ is the *MinT shrinkage estimator* suggested by Wickramasuriya, Athanasopoulos & Hyndman (2019), in which off-diagonal elements of $\hat{W}_1$ are shrunk toward zero. We refer to this method as **MinTs**.

It's hard to say which estimator for W works better. Pritularga, Svetunkov & Kourentzes (2021) demonstrated that the performance of forecast reconciliation is affected by two sources of uncertainties, i.e., the base forecast uncertainty and the reconciliation weight uncertainty. Recall that the uncertainty in the MinT solution in Equation 1 is introduced by the uncertainty in the reconciliation weighting matrix as the summing matrix is fixed for a certain hierarchy. This indicates that OLS and WLSs estimators for W may lead to less volatile reconciliation performance compared to WLSv, MinT, and MinTs estimators. Panagiotelis et al. (2021) provided a geometric intuition for reconciliation and showed that, when considering the Euclidean distance loss function, OLS reconciliation yields results that are at least as favorable as the base forecasts, whereas MinT reconciliation performs poorly relative to the base forecasts. However, when considering the mean squared reconciled forecast error, Wickramasuriya (2021) indicated that MinT reconciliation is better than OLS reconciliation. Therefore, which estimator for W to use hinges on the specific hierarchical time series of interest, the targeted level or series, and the selected loss function.

2.2.2 Relaxation of the unbiasedness assumptions

Both Hyndman et al. (2011) and Wickramasuriya, Athanasopoulos & Hyndman (2019) impose two unbiasedness conditions, i.e., the base forecasts and the reconciled forecasts are unbiased. Ben Taieb & Koo (2019) proposed a reconciliation method relaxing the assumption of unbiasedness. Specifically, by expanding the training window forward by one observation until $T - h$, they formulated the reconciliation problem as a regularized empirical risk minimization (RERM) problem given by

$$\min_{G_h} \frac{1}{(T - T_1 - h + 1)n} \left\| Y_h^* - \hat{Y}_h^* G_h' S' \right\|_F^2 + \lambda \| \text{vec}(G_h) \|_1,$$

where T_1 denotes the minimum number of observations used for model training, $\|\cdot\|_F$ is the Frobenius norm, $Y_h^* = [y_{T_1+h}, \ldots, y_T]' \in \mathbb{R}^{(T-T_1-h+1)\times n}$, $\hat{Y}_h^* = [\hat{y}_{T_1+h|T_1}, \ldots, \hat{y}_{T|T-h}]' \in \mathbb{R}^{(T-T_1-h+1)\times n}$, and $\lambda \geq 0$ is a regularization parameter.

When $\lambda = 0$, the problem reduces to an empirical risk minimization (ERM) problem without regularization. Assuming that the series in the hierarchy are jointly weakly stationary and $\hat{Y}_h^{*\prime}\hat{Y}_h^*$ is invertible, it has a closed-form solution given by

$$\hat{G}_h = B_h^{*\prime}\hat{Y}_h^* \left(\hat{Y}_h^{*\prime}\hat{Y}_h^*\right)^{-1},$$

where $B_h^* = [b_{T_1+h}, \ldots, b_T]' \in \mathbb{R}^{(T-T_1-h+1)\times n}$. If $\hat{Y}_h^{*\prime}\hat{Y}_h^*$ is not invertible, they suggested using a generalized inverse.

When $\lambda > 0$, imposing such a L_1 penalty on G_h will introduce sparsity and reduce estimation variance, albeit at the cost of introducing some bias. In addition, they also proposed another strategy that penalizes the matrix G_h towards the solution obtained by bottom-up method, i.e., $G_{\text{BU}} = [\mathbf{0}_{n_b \times n_a} \mid I_{n_b}]$.

Following the work, Wickramasuriya (2021) proposed an empirical MinT (**EMinT**) without the unbiasedness constraint by minimizing the trace of the covariance matrix of the reconciled forecast errors, $\text{Var}(\tilde{e}_{T+h|T})$. Assuming that the series are jointly weakly stationary, she derived the solution given by

$$\hat{G}_h = B_h'\hat{Y}_h \left(\hat{Y}_h'\hat{Y}_h\right)^{-1},$$

where $B_h = [b_h, \ldots, b_T]' \in \mathbb{R}^{(T-h+1)\times n}$, and $\hat{Y}_h = \left[\hat{y}_{h|0}, \ldots, \hat{y}_{T|T-h}\right]' \in \mathbb{R}^{(T-h+1)\times n}$. The difference between EMinT and ERM lies in the data sources used, as EMinT uses in-sample observations and base forecasts, while ERM relies on observations and base forecasts from a holdout validation set. We note that both ERM and EMinT consider an estimate of G that changes over the forecast horizon, which is why we keep the subscript h here.

In practice, a prevalent challenge in forecast reconciliation arises when the base forecasts of some time series within the hierarchical structure may perform poorly, especially for large hierarchies. This can be attributed to either the inherent complexity of forecasting these series or potential model misspecification. In such cases, the effectiveness of forecast reconciliation may diminish, as the role of the weighting matrix G is to assimilate *all* base forecasts and map them into bottom-level disaggregated forecasts which are subsequently summed by S. While the RERM method proposed by Ben Taieb & Koo (2019) introduces sparsity by shrinking some elements of G towards zero, it remains incapable of mitigating the adverse impact of underperforming base forecasts on the quality of the reconciled forecasts. Moreover, the method is time-consuming

because it uses expanding windows to recursively generate out-of-sample base forecasts, which are then used in the minimization problem.

We therefore propose two branches of innovative methods, constrained (out-of-sample-based) and unconstrained (in-sample-based) reconciliation with selection. These methods aim to identify and address the negative effect of some base forecasts of poor performance in a hierarchy on the overall performance of the reconciled forecasts. Additionally, through the incorporation of regularization in our objective function, our method has the potential to enhance reconciliation outcomes produced by using a "bad" choice of W, thus reducing the risk of choosing estimator of W. Moreover, our method generalizes to grouped hierarchies.

3 Forecast reconciliation with time series selection

In this section, we introduce our methods for keeping forecasts of an automatically selected set of series, identified as harmful to reconciliation, unused in forming reconciled forecasts, i.e., forecast reconciliation with series selection. Section 3.1 introduces constrained reconciliation methods with selection that formulate the problem based on out-of-sample base forecasts, while Section 3.2 presents an unconstrained reconciliation method with selection that formulates the problem based on in-sample observations and base forecasts.

3.1 Series selection with unbiasedness constraint

As S is fixed and $\hat{y}_{T+h|T}$ is given, the estimation of G carries the linear reconciliation performance, as shown in Equation 1. (Subscript h is dropped as we assume W and G do not change over the forecast horizon.) A natural way to keep forecasts of some series unused in reconciliation is through controlling the number of nonzero column entries in G. This leads to a generalization of the MinT optimization problem by applying an additional penalty to the objective function. More precisely, we consider the optimization problem given by

$$
\begin{aligned}
\min_{G} \quad & \frac{1}{2} \left(\hat{y}_{T+h|T} - SG\hat{y}_{T+h|T} \right)' W^{-1} \left(\hat{y}_{T+h|T} - SG\hat{y}_{T+h|T} \right) + \lambda \mathfrak{g}(G) \\
\text{s.t.} \quad & GS = I,
\end{aligned}
\tag{3}
$$

where $\mathfrak{g}(\cdot)$ is defined as an exterior penalty function designed to penalize the columns of G towards zero, with λ is the corresponding penalty coefficient. Thus, this can be considered as a grouped variable selection problem, with each group corresponding to a column of G. The constraint, $GS = I$, reflects the assumption that base forecasts and reconciled forecasts are

unbiased. When $\lambda = 0, \forall h$, the problem reduces to the MinT optimization problem in Equation 2 with a closed-form solution given by Equation 1.

Proposition 1. *Under the assumption of unbiasedness, the count of nonzero column entries of $\boldsymbol{G}$ (i.e., the number of time series selected for reconciliation), derived through solving Equation 3, is at least equal to the number of time series at the bottom level. In addition, we can restore the full hierarchical structure by aggregating/disaggregating the selected time series.*

Proof. According to the unbiasedness constraint $\boldsymbol{GS} = \boldsymbol{I}$, we have

$$\min\left(\operatorname{rank}(\boldsymbol{G}), \operatorname{rank}(\boldsymbol{S})\right) \geq \operatorname{rank}(\boldsymbol{I}_{n_b}) = n_b,$$

which indicates that the count of nonzero column entries of $\boldsymbol{G}$ is at least equal to n_b.

Let $\boldsymbol{X}_{.\mathbb{S}} \in \mathbb{R}^{r \times |\mathbb{S}|}$ denote the submatrix of the $r \times c$ matrix $\boldsymbol{X}$ with column indices forming a set $\mathbb{S}$ (and when $\mathbb{S} = \{j\}$, we simply use $\boldsymbol{X}_{.j}$). Here, $|\mathbb{S}|$ denotes the size of the set $\mathbb{S}$. Similarly, let $\boldsymbol{X}_{\mathbb{S}.} \in \mathbb{R}^{|\mathbb{S}| \times c}$ denote the submatrix of $\boldsymbol{X}$ whose rows are indexed by a set $\mathbb{S}$ (and when $\mathbb{S} = \{i\}$, we simply use $\boldsymbol{X}_{i.}$). Assuming that the set $\mathbb{S}$ consists of the indices of nonzero columns in the solution of Equation 3, the following equations hold:

$$\boldsymbol{GS} = \boldsymbol{G}_{.\mathbb{S}} \boldsymbol{S}_{\mathbb{S}.} \text{ and}$$

$$\min\left(\operatorname{rank}(\boldsymbol{G}_{.\mathbb{S}}), \operatorname{rank}(\boldsymbol{S}_{\mathbb{S}.})\right) \geq \operatorname{rank}(\boldsymbol{I}_{n_b}) = n_b.$$

Additionally, we have $\operatorname{rank}(\boldsymbol{S}_{\mathbb{S}.}) \leq n_b$ as $\boldsymbol{S}$ has n_b columns. Therefore, we can conclude that $\operatorname{rank}(\boldsymbol{S}_{\mathbb{S}.}) = n_b$, which implies that the hierarchical structure can be fully restored by aggregating/disaggregating the selected time series, $(\boldsymbol{y}_t)_{\mathbb{S}}$.

For example, consider the simple hierarchy shown in Figure 1, it is not possible for our constrained reconciliation methods with selection to simultaneously zero out columns of $\boldsymbol{G}$ associated with series AA and AB. However, it is possible to zero out columns related to series AA and BA simultaneously.

Proposition 2. *The optimization problem in Equation 3 can be reformulated as a least squares problem with regularization and linear equality constraint as follows:*

$$\min_{\text{vec}(G)} \quad \frac{1}{2} \left(\hat{\mathbf{y}}_{T+h|T} - \left(\hat{\mathbf{y}}'_{T+h|T} \otimes S \right) \text{vec}(G) \right)' W^{-1} \left(\hat{\mathbf{y}}_{T+h|T} - \left(\hat{\mathbf{y}}'_{T+h|T} \otimes S \right) \text{vec}(G) \right) + \lambda \mathfrak{g}\left(\text{vec}(G) \right)$$

$$\text{s.t.} \quad \left(S' \otimes I_{n_b} \right) \text{vec}(G) = \text{vec}(I_{nb}),$$

(4)

which is characterized as a high-dimensional problem in which the number of features, denoted as $p = n_b \times n$, is much larger than the number of observations, n.

Proof. Let $\text{vec}(A)$ denote the vectorization of a matrix A, which stacks the columns of A on top of one another. We have

$$\text{vec}\left(\hat{\mathbf{y}}_{T+h|T} \right) = \hat{\mathbf{y}}_{T+h|T},$$

$$\text{vec}\left(SG\hat{\mathbf{y}}_{T+h|T} \right) = \left(\hat{\mathbf{y}}'_{T+h|T} \otimes S \right) \text{vec}(G),$$

$$\text{vec}\left(GS \right) = \text{vec}\left(I_{nb}GS \right) = \left(S' \otimes I_{n_b} \right) \text{vec}(G).$$

Substituting the terms in Equation 3 with these expressions, the previous problem now takes the form of a regression problem with an additional regularization term and an equality constraint on the coefficients, as shown in Equation 4.

Moving forward, we present three classes of regularizations we use to establish forecast reconciliation with series selection, resulting in the consideration of three optimization problems: (i) group best-subset selection with ridge regularization, (ii) intuitive method with L_0 regularization, and (iii) group lasso method.

3.1.1 Group best-subset selection with ridge regularization

In high-dimensional regime with $p \gg n$, a common desiderata is to assume that the true regression coefficient (i.e., $\text{vec}(G)$ in our problem) is sparse. We propose to apply a combination of L_0 and L_2 regularization as the exterior penalty function to control the nonzero column entries in G:

$$\min_{\text{vec}(G)} \quad \frac{1}{2} \left(\hat{\mathbf{y}}_{T+h|T} - \left(\hat{\mathbf{y}}'_{T+h|T} \otimes S \right) \text{vec}(G) \right)' W^{-1} \left(\hat{\mathbf{y}}_{T+h|T} - \left(\hat{\mathbf{y}}'_{T+h|T} \otimes S \right) \text{vec}(G) \right)$$

$$+ \lambda_0 \sum_{j=1}^{n} \mathbf{1}\left(G_{\cdot j} \neq \mathbf{0} \right) + \lambda_2 \left\| \text{vec}(G) \right\|_2^2$$

(5)

$$\text{s.t.} \quad \left(S' \otimes I_{n_b} \right) \text{vec}(G) = \text{vec}(I_{nb}),$$

where $\mathbf{1}(\cdot)$ is the indicator function, $\lambda_0 \geq 0$ controls the number of nonzero columns of G selected, and $\lambda_2 \geq 0$ controls the strength of the ridge regularization. In a hierarchical time series context, the parameter of interest in Equation 5, $\text{vec}(G)$, has an inherent non-overlapping group structure, wherein each group corresponds to a single column of G, each with a size of n_b. Therefore, we refer to this reconciliation method as *group best-subset selection with ridge regularization*. In the results that follow, we label the method differently based on various estimators for W, referring to them as **OLS-subset**, **WLSs-subset**, **WLSv-subset**, **MinT-subset**, and **MinTs-subset**, respectively.

The inclusion of the ridge term in Equation 5 is motivated by earlier work on best-subset selection (e.g., Hazimeh & Mazumder 2020; Mazumder, Radchenko & Dedieu 2022), which suggests that additional ridge regularization can mitigate the poor predictive performance of best-subset selection method in the low signal-to-noise ratio (SNR) regimes.

We present a Big-M based mixed integer programming (MIP) formulation for problem in Equation 5 given by

$$
\min_{\text{vec}(G),z,\breve{e},g^+} \frac{1}{2}\breve{e}'W^{-1}\breve{e} + \lambda_0 \sum_{j=1}^{n} z_j + \lambda_2 g^{+\prime}g^+
$$

$$
\text{s.t.} \quad \left(S' \otimes I_{n_b}\right) \text{vec}(G) = \text{vec}\left(I_{n_b}\right)
$$

$$
\hat{y}_{T+h|T} - \left(\hat{y}'_{T+h|T} \otimes S\right)\text{vec}(G) = \breve{e}
$$

$$
\sum_{i=1}^{n_b} g^+_{i+(j-1)n_b} \leqslant \mathcal{M}z_j, \quad j \in [n]
$$

$$
g^+ \geqslant \text{vec}(G)
$$

$$
g^+ \geqslant -\text{vec}(G)
$$

$$
z_j \in \{0,1\}, \quad j \in [n],
$$

$$(6)$$

where $\mathcal{M}$ is a Big-M parameter (a-priori specified) that is sufficiently large such that some optimal solution, say g^{+*}, to Equation 6 satisfies $\max_{j \in [n]} \sum_{i=1}^{n_b} g^+_{i+(j-1)n_b} \leqslant \mathcal{M}$, the binary variable z_j controls whether all the regression coefficients, $\text{vec}(G)$, in group j are zero or not, i.e., $z_j = 0$ implies that $G_{\cdot j} = 0$, and $z_j = 1$ implies that $\sum_{i=1}^{n_b} g^+_{i+(j-1)n_b} \leqslant \mathcal{M}$. Such Big-M formulations are commonly used in MIP problems to model relations between discrete and continuous variables, and have been recently explored in regression with L_0 regularization, see Bertsimas, King & Mazumder (2016) for more dicussion. The problem is a mixed integer quadratic program (MIQP) that can be solved using commercial MIP solvers, e.g., Gurobi and CPLEX.

Parameter tuning. $\lambda_0 \geq 0$ and $\lambda_2 \geq 0$ are tuning parameters. To avoid computationally-expensive cross-validation, we tune the parameters to minimize the sum of squared reconciled forecast errors on the truncated training set, comprising only the h observations closest to the forecast origin. Let $\lambda_0^1 = \frac{1}{2} \left(\hat{\boldsymbol{y}}_{T+h|T} - \tilde{\boldsymbol{y}}_{T+h|T}^{\text{bench}} \right)' \boldsymbol{W}^{-1} \left(\hat{\boldsymbol{y}}_{T+h|T} - \tilde{\boldsymbol{y}}_{T+h|T}^{\text{bench}} \right)$ that captures the scale of first term in the objective, where $\tilde{\boldsymbol{y}}_{T+h|T}^{\text{bench}}$ is a vector of reconciled forecasts obtained using Equation 1 with same estimator of $\boldsymbol{W}$, and define $\lambda_0^k = 0.0001\lambda_0^1$. For the parameter λ_0, we consider a grid of $k+1$ values, $\{\lambda_0^1, ..., \lambda_0^k, 0\}$, where $\lambda_0^j = \lambda_0^1 \left(\lambda_0^k / \lambda_0^1 \right)^{(j-1)/(k-1)}$ for $j \in [k]$. So $\lambda_0^1, ..., \lambda_0^k$ is a sequence decreasing on the log scale. We use a grid of six values for the parameter λ_2, $\{0, 10^{-2}, 10^{-1}, 10^0, 10^1, 10^2\}$. Therefore, we tune over a two-dimensional grid of $(k+1) \times 6$ values to find the optimal combination of λ_0 and λ_2.

Computation details. The MIQP problem in Equation 6 is NP-Hard and computationally intensive. Bertsimas, King & Mazumder (2016) showed that commercial MIP solvers are capable of tackling problem instances for p up to a thousand. To address larger instances, there has been impressive work on developing MIP-based approches for solving L_0-regularized regression problem, e.g., Bertsimas, King & Mazumder (2016), Hazimeh & Mazumder (2020), and Hazimeh, Mazumder & Saab (2022). However, it is challenging to extend their approaches to accommodate additional constraints within the optimization problem. Despite the potential sluggishness of handling large instances with commercial MIP solvers, in our experiments, we use Gurobi to solve our problem in Equation 6 by configuring parameters such as MIPGap = 0.001 and TimeLimit = 600 seconds for cases with $p > 1000$. This enables us to terminate the solver before reaching the global optimum and return a suboptimal solution instead. This strategy is motivated by our need to consider numerous parameter candidates, and the final solution will be validated against the training set, which helps prevent the utilization of a very poor estimate of $\boldsymbol{G}$.

3.1.2 Intuitive method with L_0 regularization

Instead of estimating the entire matrix $\boldsymbol{G}$ in Section 3.1.1, we leverage the MinT solution in Equation 1 to streamline the optimization problem under consideration. Specifically, we define $\bar{\boldsymbol{S}} = \boldsymbol{A}\boldsymbol{S}$, where $\boldsymbol{A} = \text{diag}(\boldsymbol{z})$ is an $n \times n$ diagonal matrix, and $\boldsymbol{z}$ is an n-dimensional vector with elements either equal to 0 or 1. Taking the MinT solution in Equation 1, we have $\bar{\boldsymbol{G}} = (\boldsymbol{S}'\boldsymbol{A}'\boldsymbol{W}^{-1}\boldsymbol{A}\boldsymbol{S})^{-1}\boldsymbol{S}'\boldsymbol{A}'\boldsymbol{W}^{-1}$. Given fixed $\boldsymbol{S}$ and estimation of $\boldsymbol{W}$, $\bar{\boldsymbol{G}}$ is entirely determined by $\boldsymbol{A}$. By this way, when the jth diagnal element of $\boldsymbol{A}$ equals zero, the jth column of $\bar{\boldsymbol{G}}$ becomes entirely composed of zeros. Therefore, the optimization problem can be reduced to an integer quadratic programming (IQP) problem in which all of the variables are restricted to be integers:

$$\min_{A} \quad \frac{1}{2}\left(\hat{y}_{T+h|T} - S\bar{G}\hat{y}_{T+h|T}\right)' W^{-1}\left(\hat{y}_{T+h|T} - S\bar{G}\hat{y}_{T+h|T}\right) + \lambda_0 \sum_{j=1}^{n} A_{jj}$$

$$\text{s.t.} \quad \bar{G} = (S'A'W^{-1}AS)^{-1}S'A'W^{-1}$$

$$\bar{G}S = I,$$

where $\lambda_0 \geq 0$ controls the number of nonzero diagonal elements in A, consequently affecting the number of nonzero columns (i.e., selected time series) in G. We refer to this reconciliation method as *intuitive method with L_0 regularization*. In the results that follow, we label the method differently based on various estimators for W, referring to them as **OLS-intuitive**, **WLSs-intuitive**, **WLSv-intuitive**, **MinT-intuitive**, and **MinTs-intuitive**, respectively.

We should note that implementing grouped variable selection with this optimization problem can be challenging because it imposes restrictions on the parameter of interest ($\bar{G}$) to ensure it adheres rigorously to the analytical solution of MinT while making the selection. Therefore, the resulting solution may not have zero columns.

To ensure the invertibility of $S'A'W^{-1}AS$ and make the problem compatible with Gurobi, we reformulate the problem as

$$\min_{A,\bar{G},C,\check{e},z} \quad \frac{1}{2}\check{e}'W^{-1}\check{e} + \lambda_0 \sum_{j=1}^{n} z_j$$

$$\text{s.t.} \quad \bar{G}S = I$$

$$\hat{y}_{T+h|T} - \left(\hat{y}'_{T+h|T} \otimes S\right)\text{vec}(\bar{G}) = \check{e} \tag{7}$$

$$\bar{G}AS = I$$

$$\bar{G} = CS'A'W^{-1}$$

$$z_j \in \{0,1\}, \quad j \in [n].$$

Parameter tuning. Similarly to the setup in Section 3.1.1, we select the tuning parameter, λ_0, by minimizing the sum of squared reconciled forecast errors on a truncated training set, comprising only the h observations occurred prior to the forecast origin. Let $\lambda_0^1 = \frac{1}{2}\left(\hat{y}_{T+h|T} - \tilde{y}_{T+h|T}^{\text{bench}}\right)' W^{-1}\left(\hat{y}_{T+h|T} - \tilde{y}_{T+h|T}^{\text{bench}}\right)$, and $\lambda_0^k = 0.0001\lambda_0^1$, the collection of candidate values for λ_0 we consider is $\{\lambda_0^1, ..., \lambda_0^k, 0\}$, where $\lambda_0^j = \lambda_0^1\left(\lambda_0^k/\lambda_0^1\right)^{(j-1)/(k-1)}$ for $j \in [k]$.

Computation details. Following a setup akin to that in Section 3.1.1, we employ Gurobi to solve Equation 7 by configuring parameters such as MIPGap = 0.001 and TimeLimit = 600 seconds for problems with $p > 1000$.

3.1.3 Group lasso method

Lasso is another popular method for selection and estimation of parameters in the context of linear regression. Yuan & Lin (2006) introduced the group lasso method that can be used when there is a grouped structure among the variables. Here, we consider *a group lasso problem under the unbiasedness assumption* given by

$$
\min_{G} \quad \frac{1}{2} \left(\hat{y}_{T+h|T} - \left(\hat{y}'_{T+h|T} \otimes S \right) \text{vec}(G) \right)' W^{-1} \left(\hat{y}_{T+h|T} - \left(\hat{y}'_{T+h|T} \otimes S \right) \text{vec}(G) \right)
$$
$$
+ \lambda \sum_{j=1}^{n} w_j \left\| G_{\cdot j} \right\|_2 \tag{8}
$$
$$
\text{s.t.} \quad \left(S' \otimes I_{n_b} \right) \text{vec}(G) = \text{vec}\left(I_{n_b} \right),
$$

where $\lambda \geq 0$ is a tuning parameter, $w_j \neq 0$ is the penalty weight assigned in $G_{\cdot j}$ to make model more flexible, and the second term in the objective is the penalty function that is intermediate between the L_1-penalty that is used in the lasso and the L_2-penalty that is used in ridge regression. In the results that follow, we label the method differently based on various estimators for W, referring to them as **OLS-lasso**, **WLSs-lasso**, **WLSv-lasso**, **MinT-lasso**, and **MinTs-lasso**, respectively.

Next, we present the second order cone programming (SOCP) formulation for the group lasso based estimators given by

$$
\min_{\text{vec}(G), \check{e}, g^+} \frac{1}{2} \check{e}' W_h^{-1} \check{e} + \lambda \sum_{j=1}^{n} w_j c_j
$$
$$
\text{s.t.} \quad \left(S' \otimes I_{n_b} \right) \text{vec}(G) = \text{vec}\left(I_{n_b} \right)
$$
$$
\hat{y}_{T+h|T} - \left(\hat{y}'_{T+h|T} \otimes S \right) \text{vec}(G) = \check{e} \tag{9}
$$
$$
c_j = \sqrt{\sum_{i=1}^{n_b} g_{i+(j-1)n_b}^2}, \quad j \in [n].
$$

Equation 9 includes additional auxiliary variables $c_j \in \mathbb{R}_{\geq 0}$, $j \in [n]$, and second order cone constraints, $c_j = \sqrt{\sum_{i=1}^{n_b} g_{i+(j-1)n_b}^2}$ for $j \in [n]$.

Compared to the previous two methods we proposed, the group lasso method is computationally friendlier. Nonetheless, Hazimeh, Mazumder & Radchenko (2023) demonstrated, both empirically and theoretically, that group L_0-regularized method exhibits advantages over its group lasso counterpart across a range of regimes. Group lasso can either be highly dense or possess non-zero coefficients that are overly shrunk. This issue becomes more pronounced when

the groups are correlated with each other as group lasso tends to retain all correlated groups instead of seeking a more concise model.

Penalty weights and parameter tuning. In the context of group lasso, the default choice for the penalty weight, w_j, is $\sqrt{p_j}$, where p_j is the size of each group (in our case, $p_j = n_b$). In our experiment, we allocate different penalty weights to each group by considering $w_j = 1/\left\|G_{\cdot j}^{\text{bench}}\right\|_2$, which allows us to account for variations in scale across different levels in the hierarchy.

We compute the group lasso over $k + 1$ values of the tuning parameter λ, and select the tuning parameter by optimizing the sum of squared reconciled forecast errors on a truncated training set, consisting only of h observations occurred prior to the forecast origin. The collection of candidate values for λ under consideration is $\{\lambda^1, ..., \lambda^k, 0\}$, where $\lambda^1 = \max_{j=1,...,n} \left\| -\left(\left(\hat{y}'_{T+h|T} \otimes S\right)_{\cdot j^*}\right)' W^{-1} \hat{y}_{T+h|T} \right\|_2 / w_j$, $\lambda^k = 0.0001\lambda^1$, and $\lambda^j = \lambda^1 \left(\lambda^k / \lambda^1\right)^{(j-1)/(k-1)}$ for $j \in [k]$.

Proposition 3. *Ignoring the unbiasedness constraint, we define λ^1 as the smallest λ value such that all predictors in the group lasso problem have zero coefficients. Then we have*

$$\lambda^1 = \max_{j=1,...,n} \left\| -\left(\left(\hat{y}'_{T+h|T} \otimes S\right)_{\cdot j^*}\right)' W^{-1} \hat{y}_{T+h|T} \right\|_2 / w_j,$$

where j^* denotes the column index of $\hat{y}'_{T+h|T} \otimes S$ that corresponds to the jth column of G.

Proof. Denote $\beta = \text{vec}(G)$, and the first term in the objective of Equation 8 as $L(\beta \mid D)$, where D is the working data $\{\hat{y}_{T+h|T}, \hat{y}'_{T+h|T} \otimes S\}$. Ignoring the unbiasedness constraint, we define λ^1 as the smallest λ value such that all predictors in the group lasso problem have zero coefficients, i.e., the solution at λ^1 is $\hat{\beta}^1 = 0$. (Note that there is no intercept in our problem.) Under the Karush-Kuhn-Tucker conditions, we have

$$\lambda^1 = \max_{j=1,...,n} \left\| \left[\nabla L\left(\hat{\beta}^1 \mid D\right)\right]^{(j)} \right\|_2 / w_j$$

$$= \max_{j=1,...,n} \left\| -\left(\left(\hat{y}'_{T+h|T} \otimes S\right)_{\cdot j^*}\right)' W^{-1} \hat{y}_{T+h|T} \right\|_2 / w_j.$$

Computation details. Due to the incorporation of the unbiasedness constraint, we can not directly use some open-source packages designed for group lasso. Consequently, we employ Gurobi to solve the SOCP problem in Equation 9, configuring it by setting OptimalityTol = 0.0001.

3.2 Series selection method without unbiasedness constraint

In this section, we relax the unbiasedness constraint, $GS = I$, and introduce a reconciliation method with selection that relies on in-sample observations and fitted values. Let $Y \in \mathbb{R}^{T \times n}$ denote a matrix comprising observations from all time series on the training set in the structure, and $\hat{Y} \in \mathbb{R}^{T \times n}$ denote a matrix of in-sample one-step-ahead forecasts (i.e., fitted values) for all time series, where T is the length of the training set. The proposed *empirical group lasso* method considers the optimization problem given by

$$\min_{G} \quad \frac{1}{2T} \left\| Y - \hat{Y} G' S' \right\|_F^2 + \lambda \sum_{j=1}^{n} w_j \left\| G_{\cdot j} \right\|_2 ,$$

where $\|\cdot\|_F$ is the Frobenius norm, and $\lambda \geq 0$ is a tuning parameter, $w_j \neq 0$ is the penalty weight assigned in $G_{\cdot j}$ to make a more flexible model. Following the work by Ben Taieb & Koo (2019), using the fact that $\|X\|_F^2 = \|\text{vec}(X)\|_2^2$ and the useful formulation that $\text{vec}(ABC) = (C' \otimes A) \text{vec}(B)$, we rewrite the problem as

$$\min_{\text{vec}(G)} \quad \frac{1}{2N} \left\| \text{vec}(Y) - (S \otimes \hat{Y}) \text{vec}(G') \right\|_2^2 + \lambda \sum_{j=1}^{n} w_j \left\| G_{\cdot j} \right\|_2 ,$$

which becomes a standard group lasso problem, with $\text{vec}(Y)$ serving as the dependent variable and $S \otimes \hat{Y}$ as the covariate matrix. We denote this as **Elasso** in the results that follow.

Upon relaxing the unbiasedness constraint, the number of non-zero column entries in the solution for G may be less than the number of time series at the bottom level. This differs from the series selection methods with an unbiasedness constraint that we introduced in Section 3.1. In an extreme scenario, the solution may take the form of a top-down $G_{TD} = [p \mid O_{n_b \times (n-1)}]$, where only the column corresponding to the top level (most aggregated level) retains non-zero values, and $p = (p_1, p_2, \dots, p_{n_b})$ is a proportionality vector obtained based on in-sample reconcilied forecast errors such that $\sum_{i=1}^{n_b} p_i = 1$.

We also explored the empirical version of group best-subset selection with ridge regularization and intuitive method with L_0 regularization in which we do not impose the unbiasedness constraint. It is worth mentioning that Hazimeh, Mazumder & Radchenko (2023) presented a new algorithmic framework for formulating the group L_0 problem with ridge regularization and provided the **L0Group** Python package[1] for implementation. However, our experiments showed that this algorithm can not terminate within five hours for typical instances with $p \sim 10^4$.

---

[1]The L0Group Python package is available on github at https://github.com/hazimehh/L0Group.

Therefore, in this paper, we only present the empirical group lasso method for series selection without unbiasedness constraint.

Penalty weights and parameter tuning. Similarly to the setup in Section 3.1.3, we assign different penalty weights to each group by setting $w_j = 1/\left\|G^{\mathrm{OLS}}_{\cdot j}\right\|_2$, where G^{OLS} is the solution obtained by the OLS estimator of W. We select the tuning parameter, λ, by minimizing the sum of squared reconciled forecast errors on a truncated training set, comprising only the h observations closest to the forecast origin. Specifically, we form the set of candidate values for λ as $\{\lambda^1, ..., \lambda^k, 0\}$, where $\lambda^1 = \max_{j=1,...,n}\left\|-\frac{1}{N}\left((S \otimes \hat{Y})_{\cdot j*}\right)' \mathrm{vec}(Y)\right\|_2 / w_j$, $\lambda^k = 0.0001\lambda^1$, and $\lambda^j = \lambda^1 \left(\lambda^k/\lambda^1\right)^{(j-1)/(k-1)}$ for $j \in [k]$. Following the same derivation as in the proof of **Proposition 3**, λ^1 is the smallest λ value such that all predictors in the empirical group lasso problem have zero coefficients, i.e., $G = O$.

Computation details. While there are open-source packages available for solving a group lasso problem, they are still relatively slow when applied to large instance for practical usage. For example, given a specific value for the tuning parameter, λ, our experiments observed that, using the **gglasso** R package, we can not obtain a solution within five hours for typical instances with $p \sim 10^4$. Instead, we use Gurobi to solve the problem based on the SOCP formulation for the empirical group lasso. The formulation aligns with Equation 9 but omits the unbiasedness constraint.

4 Monte Carlo simulations

To evaluate the performance of various reconciliation methods with time series selection outlined in Section 3, we carry out two simulations with different designs. In both simulations, we consider a hierarchy comprising two levels of aggregation, as shown in Figure 1. Specifically, the structure has four time series at the bottom level, and seven time series in total, i.e., $n_b = 4$, and $n = 7$. In addition, the bottom-level series are first generated and then summed appropriately to obtain aggregated time series at higher levels.

In particular, Section 4.1 delves into a setup where the bottom-level series are generated using a structural time series model, but model misspecification exists for some series within the hierarchical structure. In Section 4.2, we explore the impact of correlation between series on the performance of reconciled forecasts.

4.1 Setup 1: Exploring the effect of model misspecification

In this simulation design, we follow a simulation setup similar to Wickramasuriya, Athanasopoulos & Hyndman (2019), assuming that the bottom-level time series are generated using the basic structural time series model

$$b_t = \mu_t + \gamma_t + \eta_t,$$

where μ_t, γ_t, and η_t are trend, seasonality, and error components, respectively. The trend and seasonality components are defined by

$$
\begin{aligned}
\mu_t &= \mu_{t-1} + v_t + \varrho_t, & \varrho_t &\sim \mathcal{N}\left(\mathbf{0}, \sigma_\varrho^2 I_4\right), \\
v_t &= v_{t-1} + \zeta_t, & \zeta_t &\sim \mathcal{N}\left(\mathbf{0}, \sigma_\zeta^2 I_4\right), \\
\gamma_t &= -\sum_{i=1}^{s-1} \gamma_{t-i} + \omega_t, & \omega_t &\sim \mathcal{N}\left(\mathbf{0}, \sigma_\omega^2 I_4\right),
\end{aligned}
$$

where ϱ_t, ζ_t, and ω_t are error terms independent of each other and over time. The error term η_t is generated independently from an $\text{ARIMA}(p, 0, q)$ process, where p and q take values of 0 or 1 with equal probability. The coefficients for the AR and MA components in the ARIMA process are sampled randomly from a uniform distribution within the range $[0.5, 0.7]$, and the contemporaneous error covariance matrix is given by

$$
\begin{bmatrix}
5 & 3 & 2 & 1 \\
3 & 4 & 2 & 1 \\
2 & 2 & 5 & 3 \\
1 & 1 & 3 & 4
\end{bmatrix},
$$

which enables correlations among time series in a hierarchical structure.

We set $s = 4$ for quarterly data with error variances $\sigma_\varrho^2 = 2$, $\sigma_\zeta^2 = 0.007$, and $\sigma_\omega^2 = 7$, respectively. The initial values for μ_0, v_0, γ_0, γ_1, and γ_2 are generated independently from a multivariate normal distribution with zero mean and identity covariance matrix. For each series at the bottom level, we generate a total of $T + h = 180$ observations, with the last $h = 16$ observations serving as the test set. Recall that the bottom-level series are aggregated to obtain the data for the aggregated levels. This process is repeated 500 times.

We use ETS models to generate base forecasts for all time series in the hierarchy, using the default settings as implemented in the **forecast** R package (Hyndman et al. 2023). To introduce model misspecification into our experiment, we deliberately undermine the quality of in-sample and out-of-sample forecasts (i.e., fitted values and base forecasts) for some specific time series. Specifically, we investigate three scenarios characterized by artificial model misspecification, where a 1.5 multiplier is applied to in-sample and out-of-sample forecasts for one series, i.e., series AA at the bottom level, series A at the middle level, and series Total at the top level, resulting in Scenario I, Scenario II, and Scenario III, respectively.

The results for Scenario I, II, and III are presented in Table 1, Table 2, and Table 3, respectively. Each table reports the average root mean squared error (RMSE) for each level as well as the whole structure (denoted as *Average*). The *Base* row shows the average RMSE of the base forecasts, while entries below this row reporting the percentage decrease (negative) or increase (positive) in the average RMSE of the reconciled forecasts compared to the base forecasts.

Table 1: *Out-of-sample forecast results for the simulated data in Scenario I (model misspecification introduced in series AA). "Base" shows the average RMSE of the base forecasts. Entries below this row indicate the percentage decrease (negative) or increase (positive) in the average RMSE of the reconciled forecasts compared to the base forecasts. The entries with the lowest values in each column are highlighted in blue. In each panel, methods that outperform the benchmark method are marked in bold.*

Method	Top h=1	1-4	1-8	1-16	Middle h=1	1-4	1-8	1-16	Bottom h=1	1-4	1-8	1-16	Average h=1	1-4	1-8	1-16
Base	9.6	10.7	12.6	15.6	6.3	7.3	8.6	10.8	6.4	7.5	8.3	9.8	6.8	7.9	9.0	10.9
BU	57.8	68.5	53.7	38.9	58.2	61.8	48.1	34.4	0.0	0.0	0.0	0.0	27.0	29.6	23.8	17.7
OLS	0.6	2.2	1.8	1.4	7.1	6.4	4.6	3.1	-7.6	-8.6	-8.2	-7.3	-2.1	-2.5	-2.7	-2.6
OLS-subset	0.6	**1.8**	**1.5**	**1.3**	7.2	**5.2**	**3.8**	**2.6**	**-8.3**	**-12.9**	**-11.6**	**-9.9**	**-2.4**	**-5.2**	**-4.8**	**-4.1**
OLS-intuitive	0.8	2.6	2.1	1.8	7.5	**6.1**	**4.4**	**3.0**	**-9.0**	**-12.8**	**-11.6**	**-9.9**	**-2.7**	**-4.8**	**-4.5**	**-3.8**
OLS-lasso	0.6	2.2	1.8	1.6	7.4	6.7	4.8	3.2	-7.6	-8.5	-8.1	-7.2	-2.0	-2.4	-2.6	-2.5
WLSs	7.3	10.6	8.1	5.9	15.6	16.0	11.8	8.0	-6.9	-7.8	-7.4	-6.4	1.9	2.0	1.0	0.2
WLSs-subset	**5.0**	**5.7**	**4.6**	**3.6**	**12.3**	**10.0**	**7.5**	**5.2**	**-7.6**	**-10.5**	**-9.6**	**-8.2**	**0.2**	**-2.0**	**-2.1**	**-2.0**
WLSs-intuitive	**7.1**	**9.2**	**7.1**	**5.2**	16.5	15.5	11.5	7.9	**-6.8**	**-9.2**	**-8.4**	**-7.3**	2.1	**0.9**	**0.1**	**-0.4**
WLSs-lasso	7.3	**10.3**	**8.0**	5.9	15.7	16.1	11.8	8.1	**-7.0**	-7.8	-7.3	-6.4	1.9	2.0	1.0	0.2
WLSv	1.0	2.9	2.3	1.9	4.5	4.3	3.2	2.1	-25.8	-26.4	-22.7	-18.3	-12.4	-12.6	-10.7	-8.4
WLSv-subset	-1.0	0.3	0.4	0.5	0.6	0.6	0.5	0.3	-32.3	-32.2	-27.3	-21.7	-17.3	-17.3	-14.2	-10.9
WLSv-intuitive	-0.5	0.2	0.3	0.5	0.9	0.7	0.5	0.3	-32.3	-32.3	-27.4	-21.7	-17.1	-17.3	-14.2	-10.9
WLSv-lasso	0.4	1.5	1.5	1.4	3.0	2.5	2.0	1.3	-28.5	-29.2	-24.9	-19.9	-14.4	-14.9	-12.3	-9.5
MinT	-0.4	0.7	0.9	0.6	0.7	0.7	0.6	0.3	-32.9	-33.4	-28.3	-22.5	-17.5	-17.8	-14.6	-11.3
MinT-subset	**-0.6**	0.7	**0.8**	0.7	**0.6**	0.8	0.6	0.3	**-33.0**	-33.1	-28.0	-22.3	**-17.6**	-17.6	-14.5	-11.2
MinT-intuitive	-0.4	0.7	0.9	0.6	0.7	0.7	0.6	0.3	-32.9	-33.4	-28.3	-22.5	-17.5	-17.8	-14.6	-11.3
MinT-lasso	**-0.7**	**0.3**	**0.6**	0.4	0.3	0.4	0.4	0.1	**-33.2**	**-33.7**	**-28.5**	**-22.6**	**-17.8**	**-18.1**	**-14.8**	**-11.4**
MinTs	-0.9	0.6	0.7	0.5	0.6	0.6	0.5	0.2	-32.9	-33.5	-28.3	-22.5	-17.6	-17.9	-14.6	-11.3
MinTs-subset	-0.7	0.9	1.1	1.0	0.7	0.8	0.7	0.4	**-33.0**	-33.1	-27.9	-22.2	-17.6	-17.5	-14.3	-11.0
MinTs-intuitive	-0.9	0.6	0.7	0.5	0.6	0.6	0.5	0.2	-32.9	-33.5	-28.3	-22.5	-17.6	-17.9	-14.6	-11.3
MinTs-lasso	-0.9	**0.4**	**0.6**	0.5	0.6	0.4	0.4	0.1	**-33.2**	-33.6	-28.4	**-22.6**	-17.7	-18.0	-14.8	**-11.4**
EMinT	2.2	2.9	2.5	1.7	2.5	2.9	2.3	1.3	-31.9	-32.3	-27.5	-22.0	-15.9	-16.2	-13.4	-10.5
Elasso	**1.5**	**2.8**	**2.4**	1.7	**2.1**	**2.8**	2.3	1.3	**-32.1**	-32.2	-27.4	-21.9	**-16.3**	-16.2	-13.3	-10.5

Table 2: *Out-of-sample forecast results for the simulated data in Scenario II (model misspecification introduced in series A). "Base" shows the average RMSE of the base forecasts. Entries below this row indicate the percentage decrease (negative) or increase (positive) in the average RMSE of the reconciled forecasts compared to the base forecasts. The entries with the lowest values in each column are highlighted in blue. In each panel, methods that outperform the benchmark method are marked in bold.*

Method	Top				Middle				Bottom				Average			
	h=1	1-4	1-8	1-16	h=1	1-4	1-8	1-16	h=1	1-4	1-8	1-16	h=1	1-4	1-8	1-16
Base	9.6	10.7	12.6	15.6	12.1	14.4	15.3	17.0	4.2	4.9	5.9	7.5	7.2	8.5	9.6	11.4
BU	**-1.0**	0.4	0.6	0.7	**-47.7**	**-49.6**	**-43.6**	**-36.2**	**0.0**	**0.0**	**0.0**	**0.0**	**-23.0**	**-24.0**	**-19.8**	**-15.3**
OLS	8.5	13.9	10.4	7.6	-28.2	-29.4	-26.7	-23.1	22.9	23.9	17.0	11.3	-4.2	-3.8	-4.2	-4.1
OLS-subset	**-0.5**	**0.5**	**0.6**	**0.7**	**-46.3**	**-49.0**	**-43.2**	**-35.9**	**2.2**	**1.0**	**0.7**	**0.5**	**-21.5**	**-23.4**	**-19.4**	**-15.0**
OLS-intuitive	**-0.5**	**0.5**	**0.6**	**0.6**	**-46.5**	**-49.0**	**-43.2**	**-36.0**	**2.2**	**1.2**	**0.7**	**0.5**	**-21.6**	**-23.4**	**-19.4**	**-15.0**
OLS-lasso	**-0.2**	**1.5**	**1.4**	**1.3**	**-46.9**	**-48.9**	**-43.1**	**-35.8**	**0.9**	**0.8**	**0.5**	**0.3**	**-22.1**	**-23.3**	**-19.3**	**-14.9**
WLSs	12.1	18.6	14.0	10.2	-34.4	-35.1	-31.7	-26.9	15.6	17.0	12.0	8.0	-9.0	-8.0	-7.6	-6.5
WLSs-subset	**-0.1**	**1.2**	**1.1**	**1.1**	**-46.7**	**-48.8**	**-43.1**	**-35.8**	**1.5**	**1.1**	**0.8**	**0.6**	**-21.8**	**-23.2**	**-19.2**	**-14.8**
WLSs-intuitive	**0.0**	**1.2**	**1.0**	**0.9**	**-46.5**	**-48.8**	**-43.1**	**-35.9**	**1.7**	**1.3**	**0.9**	**0.6**	**-21.6**	**-23.1**	**-19.2**	**-14.9**
WLSs-lasso	**-0.1**	**1.5**	**1.5**	**1.3**	**-46.7**	**-48.9**	**-43.1**	**-35.8**	**0.9**	**0.8**	**0.5**	**0.3**	**-22.0**	**-23.2**	**-19.3**	**-14.9**
WLSv	-0.8	2.3	1.8	1.6	-46.3	-47.9	-42.3	-35.2	1.6	1.9	1.2	0.8	-21.7	-22.2	-18.6	-14.4
WLSv-subset	-0.7	**1.3**	**1.4**	**1.4**	**-46.9**	**-48.7**	**-42.9**	**-35.6**	**1.0**	**1.0**	**0.8**	**0.6**	**-22.2**	**-23.1**	**-19.1**	**-14.7**
WLSv-intuitive	-0.4	**1.5**	**1.4**	**1.2**	**-46.9**	**-48.6**	**-42.8**	**-35.6**	**0.9**	**1.2**	**0.9**	**0.7**	**-22.2**	**-23.0**	**-19.0**	**-14.7**
WLSv-lasso	-0.6	**1.3**	**1.3**	**1.3**	**-47.2**	**-48.9**	**-43.0**	**-35.7**	**0.6**	**0.8**	**0.5**	**0.4**	**-22.4**	**-23.3**	**-19.2**	**-14.8**
MinT	0.2	0.5	0.6	0.5	-47.5	-49.4	-43.5	-36.1	1.1	0.5	0.3	0.1	-22.3	-23.7	-19.6	**-15.3**
MinT-subset	**-0.1**	0.8	0.9	0.9	-46.9	-49.1	-43.3	-36.0	1.7	0.9	0.5	0.3	-21.9	-23.4	-19.4	-15.1
MinT-intuitive	0.2	0.5	0.6	0.5	-47.5	-49.4	-43.5	-36.1	1.1	0.5	0.3	0.1	-22.3	-23.7	-19.6	**-15.3**
MinT-lasso	**-0.3**	**0.3**	0.6	0.5	**-47.6**	-49.4	-43.5	-36.1	**0.8**	**0.3**	**0.2**	0.1	**-22.5**	**-23.9**	**-19.7**	**-15.3**
MinTs	-0.3	0.3	**0.4**	**0.4**	-47.6	-49.5	**-43.6**	**-36.2**	0.7	0.2	0.1	**0.0**	-22.6	-23.9	**-19.8**	**-15.3**
MinTs-subset	**-0.8**	0.5	0.8	0.8	-47.2	-49.2	-43.4	-36.0	1.0	0.7	0.4	0.3	-22.3	-23.6	-19.5	-15.1
MinTs-intuitive	-0.3	0.3	**0.4**	**0.4**	-47.6	-49.5	**-43.6**	**-36.2**	0.7	0.2	0.1	**0.0**	-22.6	-23.9	**-19.8**	**-15.3**
MinTs-lasso	**-0.9**	**0.2**	0.5	0.5	**-47.7**	-49.5	**-43.6**	**-36.2**	**0.5**	0.2	0.1	0.1	**-22.8**	**-24.0**	**-19.8**	**-15.3**
EMinT	2.2	2.9	2.5	1.7	-46.2	-48.1	-42.4	-35.3	3.6	2.9	2.0	1.1	-20.5	-21.9	-18.2	-14.3
Elasso	**1.4**	**2.7**	**2.4**	**1.6**	**-46.4**	**-48.2**	**-42.4**	**-35.4**	**3.1**	3.2	2.1	1.2	**-20.9**	-21.9	-18.2	-14.3

Focusing on the results of the benchmark reconciliation methods, we find that the bottom-up (BU) approach performs the best in both Scenario II and III but ranks as the worst overall in Scenario I. This is not surprising, as bottom-level base forecasts are deteriorated in Scenario I, while higher-level base forecasts are deteriorated in Scenario II and III. Moreover, the WLSv, MinT, and MinTs approaches perform especially well, benefiting from their ability to consider the in-sample covariance of base forecast errors, allowing for larger range of adjustments for reconciliation. EMinT also provides accurate reconciled forecasts in our setup, where the in-sample forecasts for specific series are intentionally undermined, a situation that can be detected by the in-sample information based EMinT method. However, OLS and WLSs perform much worse than other benchmark methods in this simulation setup.

In all three scenarios, our proposed methods consistently produce either improved or comparable reconciled forecasts compared to the benchmark methods. The improvements are particularly pronounced when using OLS and WLSs estimators of W, which do not take into account the in-sample covariance of base forecast errors. One advantage of using the forecast

Table 3: *Out-of-sample forecast results for the simulated data in Scenario III (model misspecification introduced in series Total). "Base" shows the average RMSE of the base forecasts. Entries below this row indicate the percentage decrease (negative) or increase (positive) in the average RMSE of the reconciled forecasts compared to the base forecasts. The entries with the lowest values in each column are highlighted in blue. In each panel, methods that outperform the benchmark method are marked in bold.*

Method	Top h=1	1-4	1-8	1-16	Middle h=1	1-4	1-8	1-16	Bottom h=1	1-4	1-8	1-16	Average h=1	1-4	1-8	1-16
Base	25.0	30.3	30.9	32.3	6.3	7.3	8.6	10.8	4.2	4.9	5.9	7.5	7.8	9.2	10.3	12.0
BU	-62.0	-64.4	-59.0	-51.5	-0.3	0.0	0.1	0.0	0.0	0.0	0.0	0.0	-28.5	-30.2	-25.3	-19.8
OLS	-34.8	-35.5	-33.5	-30.1	45.3	50.6	37.7	25.1	27.7	29.9	21.2	13.7	3.1	3.8	1.6	-0.2
OLS-subset	-35.3	-41.9	-39.2	-35.0	43.9	39.5	29.5	19.6	27.1	23.6	16.8	10.9	2.4	-3.5	-4.2	-4.5
OLS-intuitive	-41.2	-49.2	-45.5	-40.0	35.1	26.8	20.3	13.7	21.9	15.9	11.5	7.6	-4.0	-12.2	-10.9	-9.1
OLS-lasso	-61.8	-63.6	-58.1	-50.9	0.4	1.3	1.3	0.7	0.3	0.8	0.6	0.4	-28.2	-29.3	-24.5	-19.2
WLSs	-50.9	-52.4	-48.7	-43.3	17.6	20.0	14.5	9.3	9.6	11.3	7.7	4.9	-16.3	-16.7	-14.9	-12.5
WLSs-subset	-61.8	-63.6	-58.1	-50.7	0.3	1.4	1.4	0.9	0.3	0.9	0.7	0.6	-28.2	-29.3	-24.4	-19.0
WLSs-intuitive	-61.8	-63.8	-58.3	-50.9	0.0	1.0	1.0	0.7	0.3	0.7	0.6	0.5	-28.3	-29.5	-24.6	-19.2
WLSs-lasso	-61.7	-63.5	-58.0	-50.7	0.5	1.5	1.4	0.9	0.3	0.9	0.7	0.5	-28.1	-29.2	-24.4	-19.1
WLSv	-61.1	-63.4	-58.1	-50.8	1.0	1.7	1.3	0.8	0.7	1.0	0.6	0.4	-27.6	-29.1	-24.5	-19.2
WLSv-subset	-61.9	-63.6	-58.2	-50.9	0.2	1.3	1.2	0.8	0.1	0.8	0.6	0.5	-28.3	-29.3	-24.5	-19.2
WLSv-intuitive	-61.8	-63.8	-58.3	-51.0	0.0	1.1	1.1	0.6	0.1	0.6	0.5	0.4	-28.4	-29.5	-24.7	-19.3
WLSv-lasso	-61.8	-63.9	-58.4	-51.1	0.2	0.9	0.9	0.5	0.1	0.5	0.4	0.3	-28.3	-29.6	-24.8	-19.4
MinT	-62.1	-64.3	-58.9	-51.6	-0.2	0.6	0.5	0.2	0.8	0.5	0.3	0.1	-28.3	-29.9	-25.1	-19.8
MinT-subset	-61.8	-63.7	-58.2	-50.9	0.4	1.2	1.3	0.8	0.8	1.0	0.7	0.5	-28.0	-29.3	-24.5	-19.2
MinT-intuitive	-62.1	-64.3	-58.9	-51.6	-0.2	0.6	0.5	0.2	0.8	0.5	0.3	0.1	-28.3	-29.9	-25.1	-19.8
MinT-lasso	-62.1	-64.4	-58.9	-51.5	-0.3	0.3	0.4	0.1	0.6	0.3	0.1	0.1	-28.4	-30.1	-25.2	-19.8
MinTs	-62.2	-64.4	-59.0	-51.6	-0.3	0.3	0.4	0.1	0.4	0.3	0.1	0.0	-28.5	-30.1	-25.2	-19.8
MinTs-subset	-62.0	-63.8	-58.4	-51.1	0.4	1.1	1.2	0.7	0.5	0.9	0.7	0.5	-28.2	-29.5	-24.6	-19.3
MinTs-intuitive	-62.2	-64.4	-59.0	-51.6	-0.3	0.3	0.4	0.1	0.4	0.3	0.1	0.0	-28.5	-30.1	-25.2	-19.8
MinTs-lasso	-62.2	-64.4	-58.9	-51.5	-0.2	0.3	0.4	0.1	0.2	0.2	0.1	0.0	-28.5	-30.1	-25.2	-19.8
EMinT	-60.7	-63.5	-58.2	-51.0	2.5	2.9	2.3	1.3	3.6	2.9	2.0	1.1	-26.2	-28.3	-23.8	-18.9
Elasso	-60.9	-63.6	-58.2	-51.1	2.3	2.8	2.3	1.3	3.1	3.1	2.1	1.2	-26.5	-28.3	-23.8	-18.9

reconciliation methods with selection proposed in this paper is that it reduces the difference introduced by using different estimates of W, thereby mitigating the risk of estimator selection. In some cases, such as Scenarios II and III, we can align the forecasting accuracy achieved with different methods to match the best results we can obtain through time series selection in forecast reconciliation.

In addition, we report the proportion of time series being selected from the application of our proposed methods in 500 simulation instances, as shown in Table 4, Table 5, and Table 6 for each respective scenario. Clearly, our proposed methods select fewer time series from the hierarchy for forecast reconciliation, and generally improve forecast accuracy over the benchmark methods. Furthermore, we observe that the subset method tends to return fewer time series compared to the intuitive and lasso methods, which aligns with our expectations. As discussed in Section 3, the intuitive and lasso methods tends to produce dense estimates.

Table 4: *Proportion of time series being selected after using the proposed reconciliation methods with selection in Scenario I. The last column displays a stacked barplot for each method, based on the total number of selected series data from 500 simulation instances, with a darker sub-bar indicating a larger number.*

	Top	A	B	AA	AB	BA	BB	Summary
OLS-subset	0.52	0.79	0.57	0.79	1	0.91	0.85	
OLS-intuitive	0.80	0.90	0.81	0.80	1	0.85	0.86	
OLS-lasso	0.90	1.00	0.68	1.00	1	1.00	1.00	
WLSs-subset	0.85	0.91	0.86	0.90	1	0.97	0.97	
WLSs-intuitive	0.92	0.95	0.67	0.92	1	0.92	0.95	
WLSs-lasso	0.72	1.00	0.72	1.00	1	1.00	1.00	
WLSv-subset	0.50	0.62	0.42	0.19	1	0.81	0.87	
WLSv-intuitive	0.59	0.55	0.49	0.17	1	0.76	0.86	
WLSv-lasso	0.40	1.00	0.41	0.77	1	1.00	1.00	
MinT-subset	0.66	0.90	0.61	0.72	1	0.91	0.93	
MinT-intuitive	1.00	1.00	1.00	1.00	1	1.00	1.00	
MinT-lasso	0.80	0.96	0.84	0.72	1	0.98	0.97	
MinTs-subset	0.57	0.88	0.52	0.67	1	0.89	0.92	
MinTs-intuitive	1.00	1.00	1.00	1.00	1	1.00	1.00	
MinTs-lasso	0.68	1.00	0.66	0.74	1	1.00	1.00	
Elasso	0.82	0.63	0.69	1.00	1	1.00	1.00	

4.2 Setup 2: Exploring the effect of correlation

We now consider to simulate a hierarchical structure with correlated series. A simular simulation to Wickramasuriya (2021) is implemented in this section. Using the same hierarchical sturcture as shown in Figure 1, we assume the data generating process for the time series at the bottom level follows a stationary first-order vector autoregressive model, i.e., VAR(1), given by

$$b_t = c + \begin{bmatrix} A_1 & 0 \\ 0 & A_2 \end{bmatrix} b_{t-1} + \varepsilon_t,$$

where c is a constant vector with all entries set to 1, A_1 and A_2 are 2×2 matrices with eigenvalues $z_{1,2} = 0.6[\cos(\pi/3) \pm i \sin(\pi/3)]$ and $z_{3,4} = 0.9[\cos(\pi/6) \pm i \sin(\pi/6)]$, respectively, and $\varepsilon_t \sim \mathcal{N}(0, \Sigma)$, where

Table 5: *Proportion of time series being selected after using the proposed reconciliation methods with selection in Scenario II. The last column displays a stacked barplot for each method, based on the total number of selected series data from 500 simulation instances, with a darker sub-bar indicating a larger number.*

	Top	A	B	AA	AB	BA	BB	Summary
OLS-subset	0.55	0.04	0.41	0.74	0.78	0.79	0.83	
OLS-intuitive	0.61	0.04	0.52	0.75	0.69	0.69	0.83	
OLS-lasso	0.04	0.35	0.02	1.00	1.00	1.00	1.00	
WLSs-subset	0.45	0.06	0.36	0.81	0.84	0.81	0.87	
WLSs-intuitive	0.61	0.06	0.48	0.75	0.71	0.73	0.84	
WLSs-lasso	0.02	0.33	0.02	1.00	1.00	1.00	1.00	
WLSv-subset	0.54	0.29	0.46	0.91	0.94	0.86	0.89	
WLSv-intuitive	0.59	0.32	0.53	0.82	0.86	0.77	0.86	
WLSv-lasso	0.27	0.42	0.26	1.00	1.00	1.00	1.00	
MinT-subset	0.69	0.64	0.66	0.95	0.96	0.90	0.90	
MinT-intuitive	1.00	1.00	1.00	1.00	1.00	1.00	1.00	
MinT-lasso	0.82	0.74	0.83	1.00	0.99	0.97	0.97	
MinTs-subset	0.62	0.63	0.58	0.95	0.96	0.90	0.86	
MinTs-intuitive	1.00	1.00	1.00	1.00	1.00	1.00	1.00	
MinTs-lasso	0.68	0.75	0.68	1.00	1.00	1.00	1.00	
Elasso	0.78	0.95	0.68	1.00	1.00	1.00	1.00	

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_1 & 0 \\ 0 & \boldsymbol{\Sigma}_2 \end{bmatrix}, \text{ and } \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \begin{bmatrix} 2 & \sqrt{6}\rho \\ \sqrt{6}\rho & 3 \end{bmatrix},$$

and $\rho \in \{0, \pm 0.2, \pm 0.4, \pm 0.6, \pm 0.8\}$ controls the error correlation in the simulated hierarchy.

For each time series at the bottom level, we generate a total of 101 observations, with the last one observation serving as the test set, i.e., $T = 100$ and $h = 1$. Once again, the data at the higher levels are obtained by aggreating the bottom-level series. The process is repeated 500 times for each candidate correlation, ρ.

For each series in the hierarchy, base forecasts are generated from ARMA models based on a training data comprising 100 observations. Specifically, we identify the best ARMA model with the minimum AICc (corrected Akaike information criterion) value for each series by using the automated algorithm implemented in the **forecast** R package. Additionally, when fitting ARMA models for series AA and BA, we introduce bias by omitting the constant term, which is

Table 6: *Proportion of time series being selected after using the proposed reconciliation methods with selection in Scenario III. The last column displays a stacked barplot for each method, based on the total number of selected series data from 500 simulation instances, with a darker sub-bar indicating a larger number.*

	Top	A	B	AA	AB	BA	BB	Summary
OLS-subset	0.75	0.45	0.44	0.82	0.79	0.83	0.80	
OLS-intuitive	0.47	0.70	0.69	0.86	0.92	0.90	0.89	
OLS-lasso	0.38	0.01	0.01	1.00	1.00	1.00	1.00	
WLSs-subset	0.08	0.42	0.41	0.87	0.85	0.84	0.89	
WLSs-intuitive	0.06	0.55	0.50	0.66	0.87	0.69	0.88	
WLSs-lasso	0.35	0.03	0.03	1.00	1.00	1.00	1.00	
WLSv-subset	0.31	0.67	0.65	0.88	0.90	0.91	0.90	
WLSv-intuitive	0.34	0.63	0.60	0.80	0.89	0.84	0.87	
WLSv-lasso	0.45	0.35	0.36	1.00	1.00	1.00	1.00	
MinT-subset	0.69	0.78	0.80	0.91	0.91	0.91	0.91	
MinT-intuitive	1.00	1.00	1.00	1.00	1.00	1.00	1.00	
MinT-lasso	0.75	0.89	0.86	0.97	0.97	0.97	0.97	
MinTs-subset	0.67	0.74	0.76	0.90	0.89	0.88	0.91	
MinTs-intuitive	1.00	1.00	1.00	1.00	1.00	1.00	1.00	
MinTs-lasso	0.77	0.72	0.73	1.00	1.00	1.00	1.00	
Elasso	0.95	0.64	0.64	1.00	1.00	1.00	1.00	

a common scenario in practice where forecasting bottom-level series is challenging and therefore prone to model misspecification.

5 Applications

In this section we perform two empirical applications to investigate the performance of our proposed methods and compare them with state-of-the-art reconciliation approaches. Section 5.1 considers Australian domestic tourism flows with a natural geographical hierarchy, while Section 5.2 focuses on a grouped hierarchy built using the Australian labour force survey data released by the Australian Bureau of Statistics.

5.1 Forecasting Australian domestic tourism

We consider Australian domestic tourism flows, measured as the number of overnight trips Australians spend away from home, and create a hierarchical structure using geographic divisions. The data are sourced from the National Visitor Survey and collected through computer-assisted telephone interviews involving approximately 120,000 Australian residents aged 15 years and older. The hierarchical structure starts with the national total tourism flow as the top-level aggregation, then disaggregates it into seven states and territories (referred to as *State* level hereafter), further divides them into 27 zones, and finally, into 76 regions, thus forming a natural geographical hierarchy.

Therefore, the hierarchy under consideration involves 76 monthly time series at the bottom level and 111 monthly series in total, i.e., $n_b = 76$ and $n = 111$. Each series in the hierarchy spans the period from January 1998 to December 2017, with a total of 240 observations.

Figure 3 shows the aggregate tourism flows for Australia as well as individual states, revealing pronounced seasonal patterns across the national total and states, albeit with varying seasonal patterns among the series. Notably, there was a significant growth starting from around 2010 for the national total flow and some states such as NSW (New South Wales), VIC (Victoria), QLD (Queensland), and WA (Western Australia). While flows are relatively flat for SA (South Australia), TAS (Tasmania), and NT (Northern Territory). Moreover, the time plot displays that there was a large decrease in tourism flows for WA occurred in 2016.

Our objective is to forecast tourism flows for each series in the geographical hierarchy while ensuring the coherence of forecasts across all levels. We split the data into a training set comprising 228 observations and test set with 12 observations, i.e., $T = 228$ and $h = 12$. To generate base forecasts, we select an optimal ETS model for each series using the automatic algorithm implemented in the **forecast** package for R. The base forecasts are then reconciled using our proposed methods and some state-of-the-art reconciliation methods.

Table 7 reports the RMSE values for base forecasts generated by ETS models, along with the percentage relative improvements in average RMSE obtained by a particular reconciliation method relative to the base forecasts. We also show the reconciliation errors across each of the 111 series across the four levels in the hierarchy in Figure 4. The results show that, in the tourism application, the OLS method stands out as a strong performer when comparing to other benchmark methods such as WLSv and MinTs, which account for the in-sample covariance of base forecast errors. This highlights the effectiveness of the OLS method despite its simplicity.
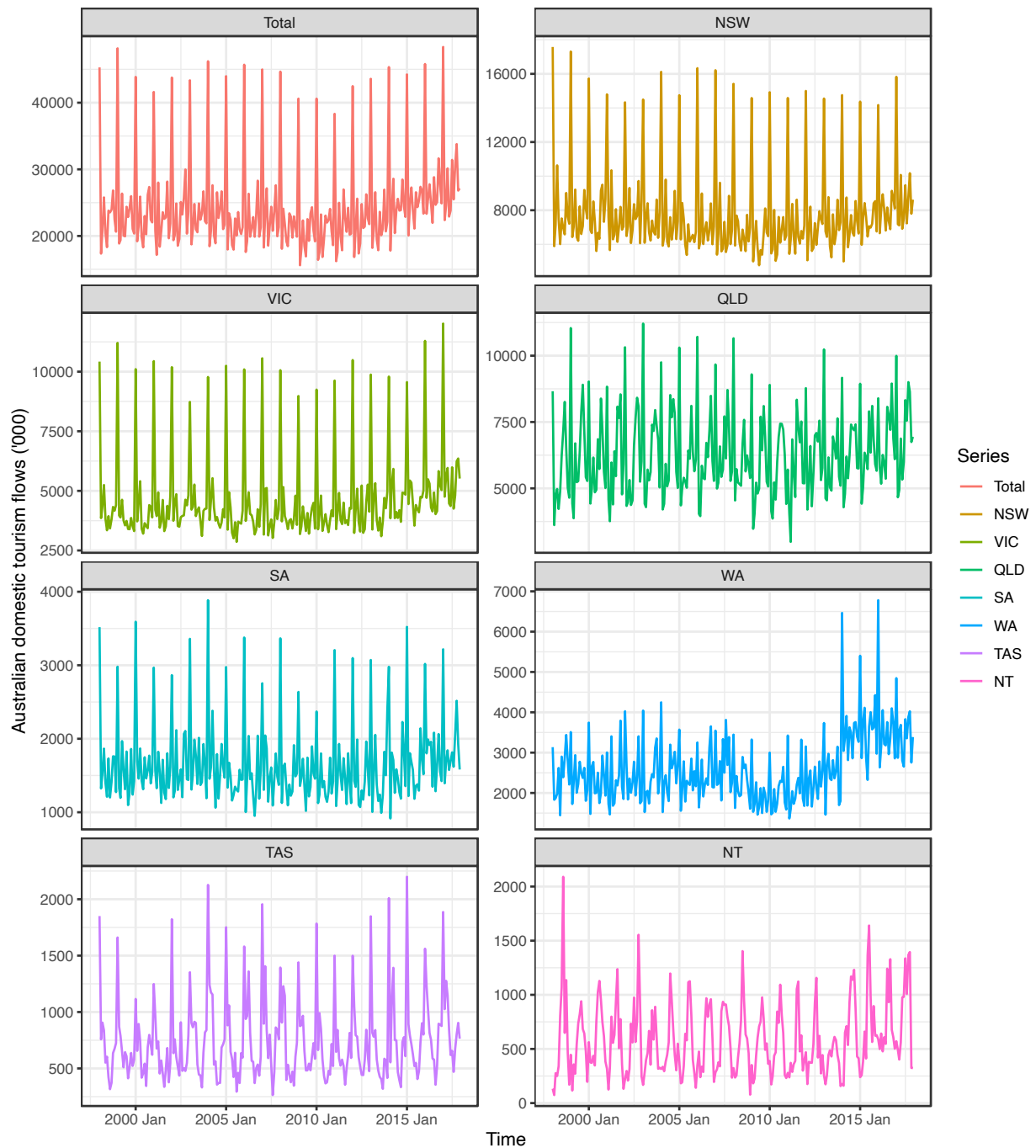
Figure 3: *Domestic tourism flows from January 1998 to December 2017 for the whole of Australia as well as the states.*

Overall, all subset methods with different estimators of *G*, i.e., OLS-subset, WLSs-subset, WLSv-subset, and MinTs-subset, give the same level of accuracy as their respective benchmark methods for one-step ahead forecasts, while showing gains for longer forecast horizons. The intuitive and lasso methods produce results identical to the corresponding benchmark methods, which is not surprising as ETS models typically do not yield extremely poor forecasts, making them challenging to be selected out using methods that tend to return dense estimates. When we relax the unbiasedness constraint, EMinT is the worst performing method across all levels. This is

attributed to its assumption that the series in the hierarchy are jointly weakly stationary, which is evidently not the case in the tourism application. The Elasso method presents significant improvement compared to the EMinT method, and it also outperforms other methods across almost all levels except for the bottom level.

Table 7: *Out-of-sample forecast results for Australian domestic tourism data. "Base" shows the average RMSE of the base forecasts, while entries below this row indicate the percentage decrease (negative) or increase (positive) in the average RMSE of the reconciled forecasts compared to the base forecasts. The entries with the lowest values in each column are highlighted in blue. In each panel, methods that outperform the benchmark method are marked in bold.*

| Method	Top				State				Zone				Region				Average			
	h=1	1-4	1-8	1-12	h=1	1-4	1-8	1-12	h=1	1-4	1-8	1-12	h=1	1-4	1-8	1-12	h=1	1-4	1-8	1-12
Base	1158.2	716.6	1279.5	1907.6	452.7	323.3	349.9	424.8	165.5	163.6	160.7	179.7	100.8	89.4	88.2	94.1	148.3	127.9	133.1	152.1
BU	89.1	132.8	53.4	42.0	-4.6	10.3	17.0	19.7	1.1	-2.4	0.4	1.0	0.0	0.0	0.0	0.0	5.7	7.6	7.6	8.5
OLS	-4.7	-0.4	0.5	1.4	-3.0	-3.9	-1.6	-1.5	-2.1	-4.2	-5.6	-7.5	1.0	-0.4	-1.9	-3.2	-1.0	-2.1	-2.7	-3.6
OLS-subset	-4.7	8.0	**-1.4**	**-14.1**	-3.0	5.5	0.3	**-7.9**	-2.1	-1.5	-3.7	**-8.7**	1.0	1.7	-0.1	-2.3	-1.0	1.7	-1.2	**-6.5**
OLS-intuitive	-4.7	-0.4	0.5	1.4	-3.0	-3.9	-1.6	-1.5	-2.1	-4.2	-5.6	-7.5	1.0	-0.4	-1.9	-3.2	-1.0	-2.1	-2.7	-3.6
OLS-lasso	-4.7	-0.4	0.5	1.4	-3.0	-3.9	-1.6	-1.5	-2.1	-4.2	-5.6	-7.5	1.0	-0.4	-1.9	-3.2	-1.0	-2.1	-2.7	-3.6
WLSs	25.1	55.2	20.8	19.1	-15.8	-5.0	3.5	6.2	-5.9	-5.4	-4.7	-5.0	-0.2	-0.8	-1.6	-2.2	-3.0	-0.1	0.3	0.9
WLSs-subset	25.1	**18.7**	**0.8**	**-7.8**	-15.8	-2.7	**-2.1**	**-6.2**	-5.9	-4.1	**-4.8**	**-8.5**	-0.2	0.3	-1.0	**-2.5**	-3.0	**-0.6**	**-2.1**	**-5.5**
WLSs-intuitive	25.1	55.2	20.8	19.1	-15.8	-5.0	3.5	6.2	-5.9	-5.4	-4.7	-5.0	-0.2	-0.8	-1.6	-2.2	-3.0	-0.1	0.3	0.9
WLSs-lasso	25.1	55.2	20.8	19.1	-15.8	-5.0	3.5	6.2	-5.9	-5.4	-4.7	-5.0	-0.2	-0.8	-1.6	-2.2	-3.0	-0.1	0.3	0.9
WLSv	38.2	76.2	29.6	25.6	-17.4	-3.1	7.0	9.9	-5.0	-4.3	-3.1	-3.2	-4.2	-1.6	-1.8	-2.1	-3.9	1.3	2.0	2.8
WLSv-subset	38.2	**34.5**	**10.7**	**8.5**	-17.4	**-8.8**	**-0.8**	1.4	-5.0	**-5.5**	**-5.3**	**-6.7**	-4.1	**-2.0**	**-2.6**	**-3.4**	-3.9	**-2.3**	**-2.0**	**-2.2**
WLSv-intuitive	38.2	76.2	29.6	25.6	-17.4	-3.1	7.0	9.9	-5.0	-4.3	-3.1	-3.2	-4.2	-1.6	-1.8	-2.1	-3.9	1.3	2.0	2.8
WLSv-lasso	38.2	76.2	29.6	25.6	-17.4	-3.1	7.0	9.9	-5.0	-4.3	-3.1	-3.2	-4.2	-1.6	-1.8	-2.1	-3.9	1.3	2.0	2.8
MinTs	20.6	53.6	21.6	19.0	-22.2	-7.2	3.5	6.3	-12.1	-6.6	-5.1	-5.3	-5.3	-2.6	-2.8	-3.1	-8.6	-1.8	-0.3	0.4
MinTs-subset	20.6	**20.0**	**6.4**	**5.6**	-22.2	**-11.3**	**-2.5**	**-0.1**	-12.1	**-7.5**	**-6.4**	**-7.8**	-5.3	**-2.9**	**-3.2**	**-3.9**	-8.6	**-4.5**	**-3.2**	**-3.3**
MinTs-intuitive	20.6	53.6	21.6	19.0	-22.2	-7.2	3.5	6.3	-12.1	-6.6	-5.1	-5.3	-5.3	-2.6	-2.8	-3.1	-8.6	-1.8	-0.3	0.4
MinTs-lasso	20.6	53.6	21.6	19.0	-22.2	-7.2	3.5	6.3	-12.1	-6.6	-5.1	-5.3	-5.3	-2.6	-2.8	-3.1	-8.6	-1.8	-0.3	0.4
EMinT	116.5	97.8	-15.8	-13.7	149.4	114.5	63.5	47.5	108.4	68.4	60.6	54.2	122.1	103.1	90.2	78.2	123.2	93.9	67.9	55.5
Elasso	-84.5	-50.4	-16.3	-16.4	-18.3	0.6	-9.0	-11.4	-7.8	-8.8	-7.5	-10.4	2.9	1.6	4.1	0.3	-10.2	-4.4	-3.2	-6.7
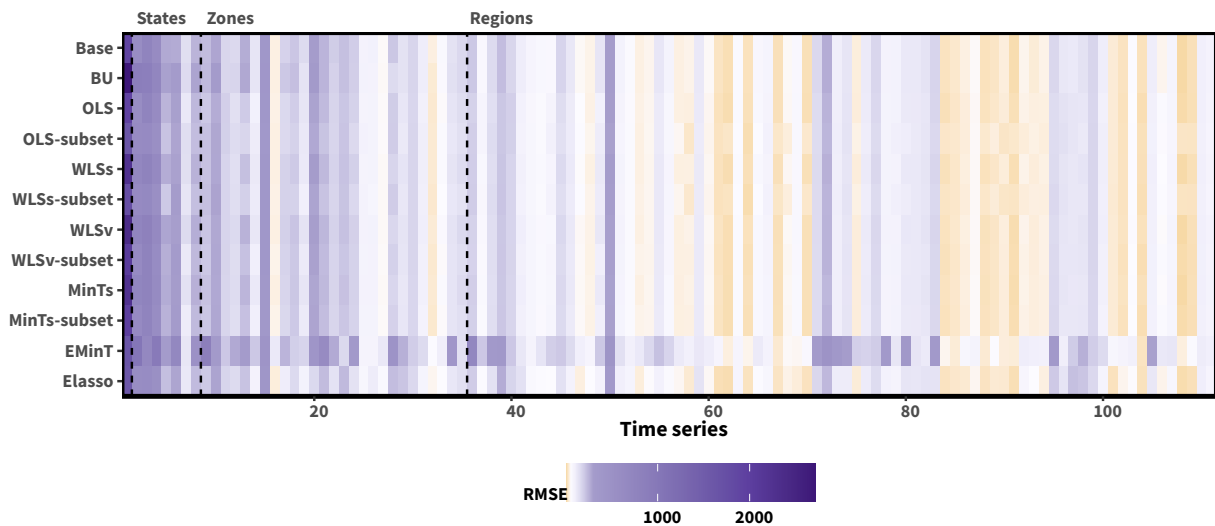
Figure 4: *Average out-of-sample forecasting performance, measured in terms of RMSE (from 1- to 12-step-ahead), for each series across different reconciliation methods. Time series are arranged along the horizontal axis.*

Table 8 presents a summary of the number of series selected using different proposed methods for each level as well as the optimal tuning parameter values identified. Here we only give the results of the subset and Elasso methods since they are useful in the tourism application. Note that the variation in the scale of the optimal parameters for different methods comes from the

difference in the scales of objective. We observe that the OLS-subset and WLSs-subset methods exclude some series at the Satate and Zone levels for forecast reconciliation. In contrast, the WLSv and MinTs methods retain all series, which is reasonable because they take into account the in-sample covariance, making themselves allow for larger adjustments made to series with large in-sample forecast error variances in forecast reconciliation. Nonetheless, the WLSv and MinTs methods can still enhance the quality of reconciled forecasts due to the inclusion of shrinkage through additional ridge regularization. It is surprising that Elasso performs exceptionally well despite using only 13 series for reconciliation.

Table 8: *Number of time series selected using different proposed methods and the optimal parameter values identified in the tourism application.*

	Number of time series retained					Optimal parameters		
	Top	State	Zone	Region	Total	λ	λ_0	λ_2
None	1	7	27	76	111	-	-	-
OLS-subset	1	2	13	76	92	-	27.98	10.00
WLSs-subset	1	1	15	76	93	-	18.73	10.00
WLSv-subset	1	7	27	76	111	-	0.03	0.01
MinTs-subset	1	7	27	76	111	-	0.05	0.01
Elasso	1	4	0	8	13	71759.21	-	-

5.2 Forecasting Australian labour force

This section evaluates the performance of the proposed methods using a grouped hierarchy built using the Australian labour force dataset. The dataset from the Labour Force Survey are released by the Australian Bureau of Statistics, consisting of monthly data on the number of unemployed persons in Australia for the period from January 2010 to July 2023[2].

There are a few missing values in the Australian labour force data. To deal with the missing observations, we use a random walk to give linear interpolation between points.

We construct a grouped hierarchy by disaggregating the number of unemployed persons over two independent attributes, duration of job search (referred to as *Duration* level), and State and Territory (referred to as *STT* level). The two attributes are crossed, but none are nested within the others. At the bottom level, the data are disaggregated by both attributes. We refer to the bottom level as *Duration* × *STT* level. Specifically, there are six different groups of job search duration, under 1 month, 1-3 months, 3-6 months, 6-12 months, 1-2 years, and 2 years and over.

[2]The Labour Force Survey data is publicly available at https://www.abs.gov.au/statistics/labour/employment-and-unemployment/labour-force-australia-detailed/aug-2023.

Additionally, the number of unemployed persons in Australia can be disaggregated by eight states and territories, NSW, VIC, QLD, SA, WA, TAS, NT, ACT (Australian Capital Territory). So the final grouped hierarchy consists of the top series, 6 series at the Duration level, 8 series at the STT level, 48 series at the Duration $\times$ STT level, giving 63 time series in total, each of length 163 observations.

The top panel in Figure 5 shows the total number of unemployed persons in Australia from January 2010 to July 2023, representing the top-level series in the grouped hierarchical structure. The monthly series shows strong seasonality within each year, marked by large peaks occuring every February. In addition, lower peaks occur in July, August, September, and October. There was a substantial increase in the number of unemployed persons during April, May, June, July of 2020, followed by a subsequent downward trend. The bottom-left panel displays the breakdown of unemployed individuals by state and territory, while the bottom-right panel presents the breakdown by the duration of job search. The plots display diverse and rich dynamics both within and between different levels of hierarchy. For example, there was noticeable growth observed during 2020 for some states such as NSW, VIC, and QLD, whereas other states did not experience such such significant growth. Additionally, there is a resemblance in the seasonal patterns between NSW and QLD, while the seasonal pattern in VIC appears relatively heterogeneous. When comparing the series at the STT level and Duration level, we notice that the seasonal patterns in the Duration level series is more consistent and potentially easier to forecast.

For each series in the grouped hierarchy, we use the data spanning from January 2010 to July 2022 as the training set to select an optimal ETS model with the automatic algorithm implemented in the **forecast** package for R. Using these fitted ETS models, base forecasts are generated for $h = 1$ to 12-steps ahead. Subsequently, these base forecasts are used to perform forecast reconciliation, using the proposed methods and other state-of-the-art reconciliation approaches.
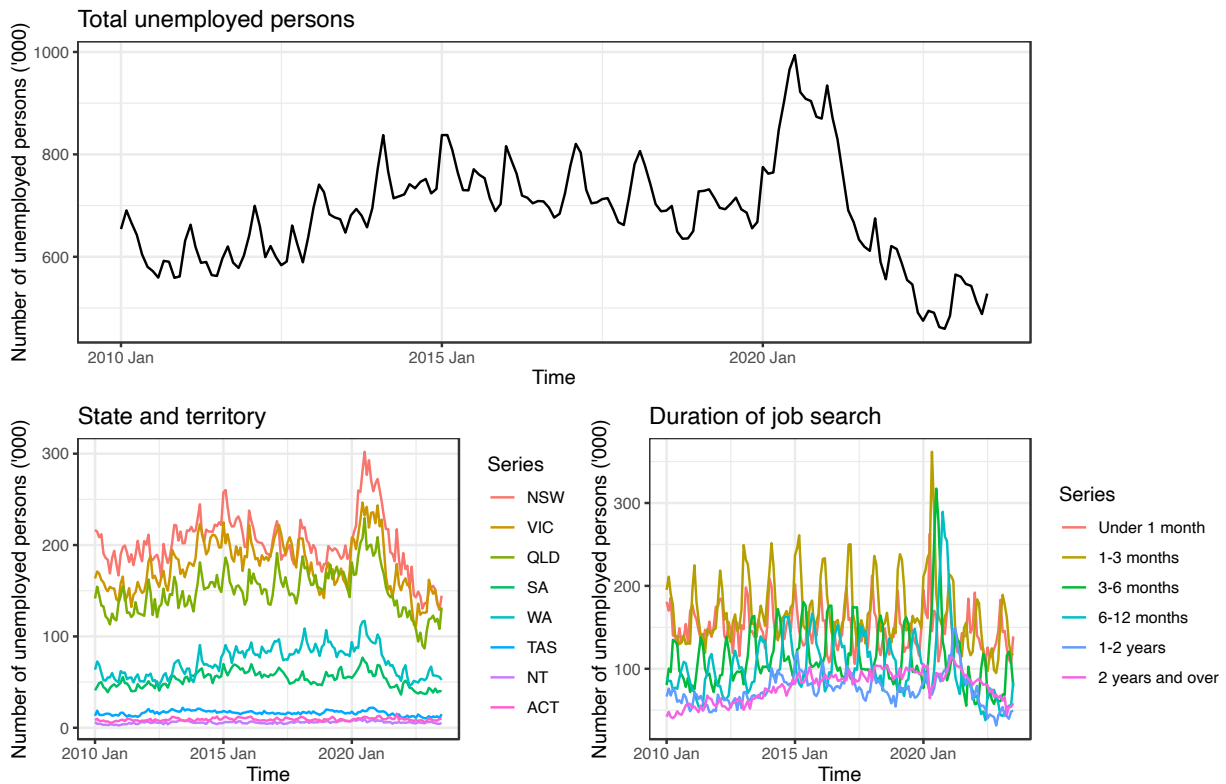
Results analysis

6 Conclusion

Figure 5: *Australia unemployed persons, disaggregated by state and territory, and by duration of job search.*

Table 9: *Out-of-sample forecast results for Australian labour force data. "Base" shows the average RMSE of the base forecasts, while entries below this row indicate the percentage decrease (negative) or increase (positive) in the average RMSE of the reconciled forecasts compared to the base forecasts. The entries with the lowest values in each column are highlighted in blue. In each panel, methods that outperform the benchmark method are marked in bold.*

Method	Top h=1	1-4	1-8	1-12	Duration h=1	1-4	1-8	1-12	STT h=1	1-4	1-8	1-12	Duration x STT h=1	1-4	1-8	1-12	Average h=1	1-4	1-8	1-12
Base	18.5	13.6	18.3	28.3	11.8	12.7	13.9	16.9	6.7	6.0	6.0	6.3	2.3	2.6	2.7	2.9	4.1	4.1	4.4	5.1
BU	**-81.5**	33.4	-19.9	-45.0	**-30.7**	-9.2	-7.9	-10.1	-12.9	-10.4	-13.4	-13.5	0.0	0.0	0.0	0.0	**-17.1**	-2.8	-5.9	-9.3
OLS	-16.2	-14.2	-13.4	-10.4	2.5	-2.6	-2.7	-0.6	-1.8	-0.9	-1.9	0.3	6.7	5.1	5.1	4.9	2.1	0.7	0.4	1.1
OLS-subset	**-17.0**	-2.1	**-31.2**	**-38.4**	**2.0**	-1.7	**-5.0**	**-2.7**	**-2.7**	**-4.2**	**-8.6**	**-7.3**	6.7	5.2	**3.3**	**3.7**	**1.7**	1.1	**-3.5**	**-3.8**
OLS-intuitive	**-79.6**	**-23.9**	**-31.9**	**-32.1**	**-13.0**	0.4	-0.8	0.3	**-8.9**	4.9	7.0	13.2	**6.3**	12.3	12.4	11.6	**-8.5**	5.6	4.7	4.4
OLS-lasso	-16.2	-14.2	-13.4	-10.4	2.5	-2.6	-2.7	-0.6	-1.8	-0.9	-1.9	0.3	6.7	5.1	5.1	4.9	2.1	0.7	0.4	1.1
WLSs	-60.6	-29.4	-44.0	-38.6	-12.0	-7.6	-6.9	-5.9	-6.5	-8.1	-9.4	-8.0	3.2	1.7	1.7	1.6	-7.7	-4.4	-5.8	-5.9
WLSs-subset	**-61.6**	-22.4	**-47.3**	**-50.4**	-12.0	**-8.0**	**-10.5**	**-7.8**	**-6.6**	**-10.7**	**-14.3**	**-12.8**	3.2	5.6	4.1	5.9	**-7.8**	-2.8	**-6.7**	**-6.4**
WLSs-intuitive	-60.6	-29.4	-44.0	-38.6	-12.0	-7.6	-6.9	-5.9	-6.5	-8.1	-9.4	-8.0	3.2	1.7	1.7	1.6	-7.7	-4.4	-5.8	-5.9
WLSs-lasso	-60.6	-29.4	-44.0	-38.6	-12.0	-7.6	-6.9	-5.9	-6.5	-8.1	-9.4	-8.0	3.2	1.7	1.7	1.6	-7.7	-4.4	-5.8	-5.9
WLSv	-60.6	-29.1	-41.4	-36.6	-14.5	-8.7	-5.6	-4.8	-3.3	-6.8	-8.0	-7.0	5.5	2.6	2.6	3.1	-6.7	-4.0	-4.5	-4.6
WLSv-subset	-51.6	**-32.7**	-36.6	-29.6	**-18.3**	-9.8	**-10.5**	**-10.9**	-1.1	-4.3	**-8.1**	**-7.3**	**2.5**	**2.1**	**2.3**	**1.8**	**-7.9**	**-4.3**	**-5.8**	**-6.5**
WLSv-intuitive	-60.6	-29.1	-41.4	-36.6	-14.5	-8.7	-5.6	-4.8	-3.3	-6.8	-8.0	-7.0	5.5	2.6	2.6	3.1	-6.7	-4.0	-4.5	-4.6
WLSv-lasso	-60.6	-29.1	-41.4	-36.6	-14.5	-8.7	-5.6	-4.8	-3.3	-6.8	-8.0	-7.0	5.5	2.6	2.6	3.1	-6.7	-4.0	-4.5	-4.6
MinTs	-27.0	-21.1	-22.9	-21.8	-9.1	-7.9	-6.7	-4.6	-3.6	-9.0	-10.5	-7.6	7.7	3.7	3.0	3.4	-1.9	-3.3	-3.9	-3.1
MinTs-subset	**-41.4**	-9.3	-17.8	**-45.1**	**-12.2**	-5.0	**-8.2**	**-6.8**	**-6.1**	-9.8	-7.1	**-9.3**	5.3	4.7	3.8	3.6	**-5.3**	-1.5	-3.0	**-6.1**
MinTs-intuitive	-27.0	-21.1	-22.9	-21.8	-9.1	-7.9	-6.7	-4.6	-3.6	-9.0	-10.5	-7.6	7.7	3.7	3.0	3.4	-1.9	-3.3	-3.9	-3.1
MinTs-lasso	-27.0	-21.1	-22.9	-21.8	-9.1	-7.9	-6.7	-4.6	-3.6	-9.0	-10.5	-7.6	7.7	3.7	3.0	3.4	-1.9	-3.3	-3.9	-3.1
EMinT	-60.4	-14.0	1.4	-29.9	-6.0	12.0	10.7	-6.7	16.7	-0.9	-12.4	**-21.0**	23.3	17.2	16.7	10.1	7.7	10.8	9.0	-3.7
Elasso	-4.2	-3.3	**-22.3**	-8.0	**-19.7**	**-9.9**	**-19.9**	**-25.3**	**-24.6**	**-24.3**	**-22.6**	-14.6	**-10.8**	**-3.8**	**-0.2**	**-4.9**	**-15.7**	**-9.3**	**-11.4**	**-13.2**

References

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

Table 10: *Number of time series selected using different proposed methods and the optimal parameter values identified in the labour application.*

	Number of time series retained				Optimal parameters		
	Top	Duration	STT	Duration x STT	Total	λ	λ_0	λ_2
None	1	6	8	48	63	-	-	-
OLS-subset	0	5	1	48	54	-	4.16	1.00
WLSs-subset	0	5	1	46	52	-	0.38	0.10
WLSv-subset	1	5	7	48	61	-	0.51	1.00
MinTs-subset	0	1	1	47	49	-	0.03	0.01
Elasso	1	5	2	3	11	213.59	-	-

Ben Taieb, S & B Koo (2019). Regularized Regression for Hierarchical Forecasting Without Unbiasedness Conditions. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. KDD '19. Anchorage, AK, USA: Association for Computing Machinery, pp.1337–1347.

Bertsimas, D, A King & R Mazumder (2016). Best subset selection via a modern optimization lens. en. *The Annals of Statistics* **44**(2), 813–852.

Hazimeh, H & R Mazumder (2020). Fast Best Subset Selection: Coordinate Descent and Local Combinatorial Optimization Algorithms. *Operations Research* **68**(5), 1517–1537.

Hazimeh, H, R Mazumder & P Radchenko (2023). Grouped variable selection with discrete optimization: Computational and statistical perspectives. en. *The Annals of Statistics* **51**(1), 1–32.

Hazimeh, H, R Mazumder & A Saab (2022). Sparse regression at scale: branch-and-bound rooted in first-order optimization. *Mathematical Programming* **196**(1), 347–388.

Hyndman, RJ, RA Ahmed, G Athanasopoulos & HL Shang (2011). Optimal combination forecasts for hierarchical time series. *Computational Statistics & Data Analysis* **55**(9), 2579–2589.

Hyndman, RJ, G Athanasopoulos, C Bergmeir, G Caceres, L Chhay, M O'Hara-Wild, F Petropoulos, S Razbash, E Wang & F Yasmeen (2023). *forecast: Forecasting functions for time series and linear models*. R package version 8.21.1. https://pkg.robjhyndman.com/forecast/.

Hyndman, RJ, AJ Lee & E Wang (2016). Fast computation of reconciled forecasts for hierarchical and grouped time series. *Computational Statistics & Data Analysis* **97**, 16–32.

Mazumder, R, P Radchenko & A Dedieu (2022). Subset Selection with Shrinkage: Sparse Linear Modeling When the SNR Is Low. *Operations Research*.

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

Pritularga, KF, I Svetunkov & N Kourentzes (2021). Stochastic coherency in forecast reconciliation. *International Journal of Production Economics* **240**, 108221.

Wickramasuriya, SL (2021). Properties of point forecast reconciliation approaches. arXiv: 2103. 11129 [stat.ME].

Wickramasuriya, SL, G Athanasopoulos & RJ Hyndman (2019). Optimal Forecast Reconciliation for Hierarchical and Grouped Time Series Through Trace Minimization. *Journal of the American Statistical Association* **114**(526), 804–819.

Yuan, M & Y Lin (2006). Model selection and estimation in regression with grouped variables. en. *Journal of the Royal Statistical Society. Series B, Statistical Methodology* **68**(1), 49–67.