Enhancing Forecast Reconciliation: A Study of Alternative Covariance Estimators

Vincent Su Shanika Wickramasuriya (supv.)
George Athanasopoulos (supv.)

Abstract

A collection of time series connected via a set of linear constraints is known as hierarchical time series. Forecasting these series without respecting the hierarchical nature of the data can lead to incoherent forecasts across aggregation levels and lower accuracy. To mitigate this issue, various forecast reconciliation approaches have been proposed in the literature, where the individual forecasts are adjusted to satisfy the aggregation constraints. Among these, **MinT** (Minimum Trace) is widely used, however, it requires a good estimate of the covariance matrix of the base forecast errors. The current practice is to use the shrinkage estimator (often shrinking toward a diagonal matrix), but it lacks flexibility and might not fully utilise the prominent latent structure presented. In this project, we aim to assess the forecasting performance of MinT when different covariance estimators are used, namely NOVELIST (NOVEL Integration of the Sample and Thresholded Covariance), POET (Principal Orthogonal complEment Thresholding), and others.

1 Introduction

In time series forecasting, aggregation occurs in a variety of settings. While a formal definition of hierarchical time series can be found in Section 2.1, we can think of Starbucks sales data as an illustrative example. Starbucks operates in many countries, and each country has multiple cities where they have outlets. The sales data is *structured hierarchically*: the top level is the total sales across all countries, followed by national sales for each country, and then individual sales for each outlet in a city. As a result, there are over 40,000 individual outlet sales to forecast, plus additional series at higher levels of aggregation such as city and country. The hierarchy can be even more complex if we consider the sales of different kinds of drinks (e.g., coffees, teas, refreshers) at each aggregation level.

Forecasting data from such hierarchical structures also arises in many other decision-making contexts, from supply chains (Angam et al., 2025; Seaman & Bowman, 2022) and energy planning (Di Modica et al., 2021), to macroeconomics (El Gemayel et al., 2022; Li et al., 2019) and tourism analysis (Athanasopoulos et al., 2009). Stakeholders in these settings need forecasts at several aggregation levels to allocate resources and manage risk. The impact of methods for forecasting hierarchical time series has not been limited to academia, with industry also showing a strong interest. Many companies and organisations have adopted these methods in practice, including Amazon, the International Monetary Fund, IBM, SAP, and more (Athanasopoulos et al., 2024).

In practice, when forecasts are produced for all series (often called *base forecasts*), they typically violate the aggregation constraints observed in the data; such forecasts are *incoherent*. This can undermine downstream decisions that require internal consistency.

Traditionally, forecasting these hierarchical time series has been done using single-level methods, such as bottom-up, top-down, and middle-out approaches. Bottom-up methods involve generating forecasts for the bottom-level series and aggregating them to higher levels. Top-down methods start with forecasts for the only top-level series and disaggregate them down. Middle-out methods combine both approaches by forecasting middle-level series and then aggregating or disaggregating as needed. Despite their simplicity, these methods only anchor forecasts to a single level, implying a large loss of information on the hierarchy's inherent correlation structure. Additionally, the most disaggregated series often are very noisy or even intermittent, and the higher-level data might be smoother due to the aggregation. Furthermore, as we saw from the Starbucks example considering the sales of different kinds of drinks at each aggregation level – formally defined as grouped structure in Section 2.1 – the disaggregation becomes more complex since the disaggregation paths are not unique. Consequently, these single-level methods tend to give poor results across other levels of the hierarchy.

To overcome these issues, forecast reconciliation was introduced by Hyndman et al. (2011), and later developed by Erven & Cugliari (2015), Hyndman et al. (2016), Ben Taieb & Koo (2019), Wickramasuriya et al. (2019), and others to achieve coherency in point forecasts and enhance accuracy. Forecast reconciliation projects a collection of independent base forecasts into a set of coherent forecasts that respect the linear constraints defining a hierarchical or grouped time-series system.

Among the modern reconciliation strategies, the Min Trace (MinT) approach developed by Wickramasuriya et al. (2019) is widely used and perform significantly well under right conditions. Despite its properties, MinT relies on a good high-dimensional covariance estimator, which is positive-definite. However, not many researchers have explored this issue in the current literature, except Carrara et al. (2025), who introduced a new estimator for MinT–Double Shrinkage estimator. The current stage of this research piece is fairly early, leaving many potential gaps still not yet to explore. As a result, a study of alternative high-dimensional covariance estimators for MinT is timely and desirable.

The remainder of the paper is organised as follows. Section 2 provides the basic theoretical framework for hierarchical time series and forecasting, and Min Trace approach, introducing notations, terminologies, and motivations for alternative estimators. Section 3 walks through the main covariance estimators this paper explores, and argues their strengths and weaknesses. Section 4 covers the simulation design and currently explores the performance of NOVELIST on MinT. Section 5 shows a real-world application of MinT with NOVELIST, which produces results that did not occured in the simulation settings, suggesting further inspection and analysis.

2 Theoretical Framework

2.1 Hierarchical tructure

The *hierarchical structure* can be represented as a tree, as shown in Figure 1. The top level of the tree represents the total value of all series, while the lower levels represent the series at different levels of disaggregation. These hierarchical structures naturally form *aggregation constraints* (child levels must sum up to parents). When there are attributes of interest that are crossed, such as the Starbucks drinks sales at any aggregation level (brand-wise, national, or outlet) is also considered by kinds of drinks (e.g., coffees, refreshers), the structure is described as a grouped time series. As illustrated in Figure 2, the aggregation or disaggregation paths are not unique.

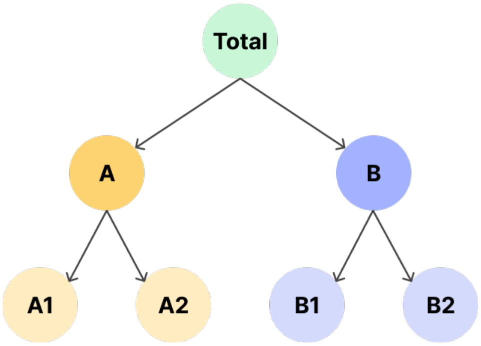


Figure 1: A 2-level hierarchical tree structure

For simplicity, we refer to both of these structures as hierarchical structure, we will distinguish between them if and when it is necessary. Taken together, a collection of time series organised in a hierarchy and subject to aggregation constraints is refered as *hierarchical time series.* All hierarchical time series can be represented using matrix algebra:

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

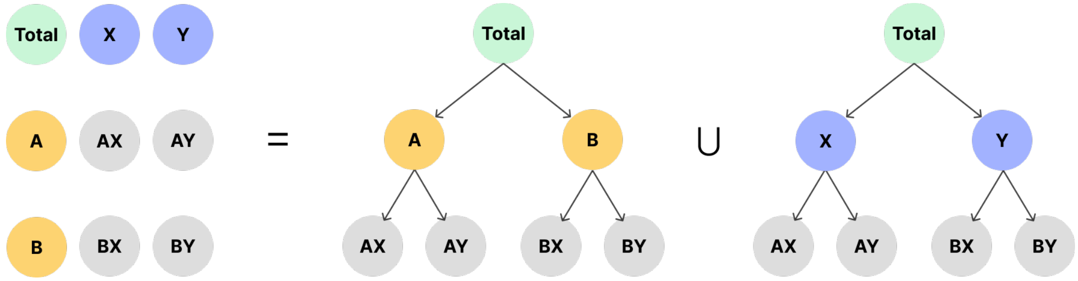


Figure 2: A 2-level grouped structure, which can be considered as the union of two hierarchical trees with common top and bottom level series

where $\boldsymbol{S}$ is a summing matrix of order $n \times n_b$ which aggregates the bottom-level series $\boldsymbol{b}_t \in \mathbb{R}^{n_b}$ to the series at aggregation levels above. The vector $\boldsymbol{y}_t \in \mathbb{R}^n$ contains all observations at time t. The summing matrix $\boldsymbol{S}$ for the tree structure in Figure 1 is:

$$\boldsymbol{S} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ & \multicolumn{2}{c}{\boldsymbol{I_4}} & \end{bmatrix} .$$

Assume we produce h-step-ahead base forecasts $\hat{\boldsymbol{b}}_{t+h|t}$ for the bottom-level series, obtained by any prediction methods, pre-multiplying them by $\boldsymbol{S}$ we get:

$$\tilde{\boldsymbol{y}}_{t+h|t} = \boldsymbol{S}\hat{\boldsymbol{b}}_{t+h|t} \, . \tag{1}$$

We refer to $\tilde{\boldsymbol{y}}_{t+h|t}$ as *coherent forecasts*, as they respect the aggregation constraints. We also refer to this way of obtaining coherent forecasts by summing the bottom-level forecasts as the bottom-up approach. However, generating forecasts this way is anchored only to prediction models at a single level, and will not be utilising the inherent information from other levels. This drawback applies to the top-down and middle-out approaches. For example, the bottom-level data can be very noisy or even intermittent, and the higher-level data might be smoother due to the aggregation.

Another issue with expressing reconciled methods as in Equation 1 is that it restricts the reconciliation to only single-level approaches. Thus, Hyndman et al. (2011) suggested a generalised expression for all existing methods, which also provides a framework for new methods to be developed:

$$\tilde{\boldsymbol{y}}_{t+h|t} = \boldsymbol{S}\boldsymbol{G}\hat{\boldsymbol{y}}_{t+h|t} \; , \tag{2}$$

for a suitable $n_b \times n$ matrix $\boldsymbol{G}$. $\boldsymbol{G}$ maps the base forecasts of all levels $\hat{\boldsymbol{y}}_{t+h|t}$ down into the bottom level, which is then aggregated to the higher levels by $\boldsymbol{S}$. The choice of $\boldsymbol{G}$ determines the composition of reconciled forecasts $\tilde{\boldsymbol{y}}_{t+h|t}$, and modern reconciliation methods are developed to estimate $\boldsymbol{G}$.

2.2 The Minimum Trace (MinT) Reconciliation

Wickramasuriya et al. (2019) framed the problem as minimising the variances of all reconciled forecast errors $\text{Var}[y_{t+h} - \tilde{y}_{t+h|t}] = \boldsymbol{S}\boldsymbol{G}\boldsymbol{W}_h\boldsymbol{G}'\boldsymbol{S}'$, where $\boldsymbol{W}_h = \mathbb{E}(\hat{\boldsymbol{e}}_{t+h|t}\ \hat{\boldsymbol{e}}'_{t+h|t})$ is the positive definite covariance matrix of the h-step-ahead base forecast errors. They showed that this is equivalent to minimising the trace of the reconciled forecast error covariance matrix (sum of the diagonal elements - the variances). The Minimum Trace (MinT) solution is given by

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

Wickramasuriya et al. (2019) also showed that MinT is an algebraic generalisation of the GLS, and the OLS and WLS methods are special cases of MinT when $\boldsymbol{W}_h$ is an identity matrix I_{n_b} and a diagonal matrix $\text{diag}(\boldsymbol{W}_h)$, respectively. In this paper, we place our main focus on the MinT method.

The MinT solution hinges on a reliable, positive-definite estimate of $\boldsymbol{W}_h$, which is challenging to estimate in high-dimensional setting. The sample covariance matrix is unstable and non-positive-definite when the number of series n is huge and larger than the time dimension T. To tackle this issue, the original paper Wickramasuriya et al. (2019) adopted the diagonal-target shrinkage estimator from Schäfer & Strimmer (2005), given by

$$\hat{\boldsymbol{W}}_1^{shr} = \lambda_D \hat{\boldsymbol{W}}_{1,D} + (1 - \lambda_D)\hat{\boldsymbol{W}}_1\,,$$

where $\hat{\boldsymbol{W}}_{1,D}$ is a diagonal matrix comprising the diagonal entries $\text{diag}(\hat{\boldsymbol{W}}_1)$. We refer to any $\lambda \in [0,1]$ as the shrinkage intensity parameter, the subscript specifies which estimator it belongs to. This approach shrinks the covariance matrix $\hat{\boldsymbol{W}}_1$ towards its diagonal matrix, meaning the off-diagonal elements are shrunk towards zero while the diagonal ones remain unchanged.

Schäfer & Strimmer (2005) also proposed an estimate of the optimal shrinkage intensity parameter λ_D:

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

where $\hat{r}_{ij}$ is the ijth element of $\hat{\boldsymbol{R}}_1$, the 1-step-ahead sample correlation matrix (obtained from $\hat{\boldsymbol{W}}_1$). The optimal estimate is obtained by minimising $MSE(\hat{\boldsymbol{W}}_1) = Bias(\hat{\boldsymbol{W}}_1)^2 +$

$Var(\hat{\boldsymbol{W}}_1)$. More specifically, we trade the unbiasedness of the sample covariance matrix for a lower variance.

However, the hierarchical time series data often exhibit a prominent principal components structure, which is not fully taken advantage. Taking an example of the Australian domestic overnight trips data set (*Tourism Research Australia*, 2024), where the national trips are disaggregated into states and territories, and further into regions. We then fit ETS models to all series, using the algorithm from Fabletools R package (O'Hara-Wild et al., 2024), and compute the one-step-ahead in-sample base forecast error covariance matrix $\hat{\boldsymbol{W}}_1$. The twenty largest eigenvalues of the covariance matrix are plotted in Figure 3. We can see that the point of inflexion occurs at the component with 5th largest eigenvalue, indicating a prominent principal components structure.

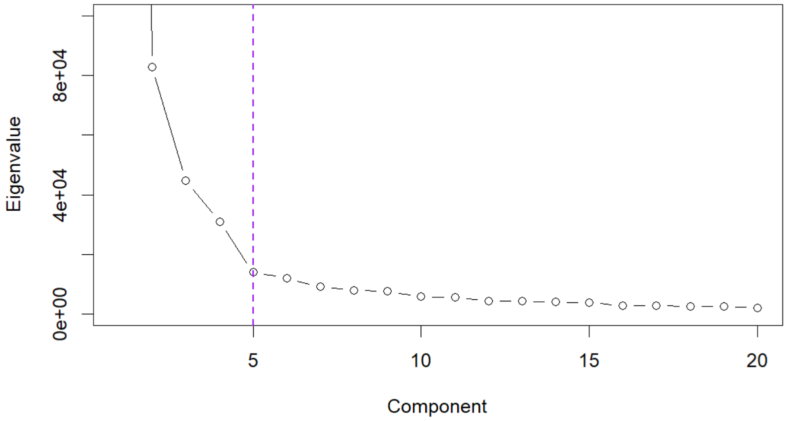


Figure 3: Twenty largest eigenvalues of one-step-ahead in-sample base forecast error covariance, Australian domestic overnight trips

Additionally, the shrinkage estimator shrinks all off-diagonal elements towards zeros with equal weights λ_D. We might prefer to better preserve strong signals, and largely reduce the effects of small, noisy correlations. In the next sections, we will explore several options that take these two issues into account.

3 Covariance Estimation Approaches

3.1 NOVELIST Estimator

The NOVELIST (NOVEL Integration of the Sample and Thresholded Covariance) estimator, proposed by Huang & Fryzlewicz (2019), introduces a way to control the target matrix's sparsity, retaining strong correlations while discarding weak, noisy effects. NOVELIST offers more flexibility than the shrinkage estimator, which is useful when we believe that only a few variables are truly correlated. However, it does not guarantee to be positive definite.

The method is based on the idea of soft-thresholding the sample covariance matrix, then performing shrinkage towards this thresholded version. This introduces an extra parameter, the threshold δ, which is used to control the amount of soft-thresholding. The NOVELIST estimator is given by:

$$\hat{\boldsymbol{W}}_1^N = \lambda_\delta \hat{\boldsymbol{W}}_{1,\delta} + (1 - \lambda_\delta)\hat{\boldsymbol{W}}_1, \tag{3}$$

where $\hat{\boldsymbol{W}}_{1,\delta}$ is the thresholded version of $\hat{\boldsymbol{W}}_1$. By convenient setting, we can rewrite it in terms of correlation:

$$\hat{\boldsymbol{R}}_1^N = \lambda_\delta \hat{\boldsymbol{R}}_{1,\delta} + (1 - \lambda_\delta)\hat{\boldsymbol{R}}_1, \tag{4}$$

In this setting, $\hat{\boldsymbol{R}}_{1,\delta}$ is the thresholded correlation matrix, where each element is regularised by:

$$\hat{r}_{1,ij}^{\delta} = \operatorname{sign}(\hat{r}_{1,ij}) \max(|\hat{r}_{1,ij}| - \delta,\ 0), \tag{5}$$

where $\delta \in [0, 1]$ is the threshold parameter. For a given threshold δ, Huang & Fryzlewicz (2019) derived an analytical expression for the optimal shrinkage intensity parameter $\lambda(\delta)$ using Ledoit-Wolf's lemma (Ledoit & Wolf, 2003), following similar logic to Schäfer & Strimmer (2005). It can be computed as:

$$\hat{\lambda}(\delta) = \frac{\sum_{i \neq j} \widehat{Var}(\hat{r}_{1,ij})\ \mathbf{1}(|\hat{r}_{1,ij}| \leq \delta)}{\sum_{i \neq j} (\hat{r}_{1,ij} - \hat{r}_{1,ij}^{\delta})^2}, \tag{6}$$

where $\mathbf{1}(.)$ is the indicator function.

On the other hand, the optimal threshold δ^* does not have a closed-form solution, and is typically obtained by executing a rolling-window cross-validation procedure. The idea is to find the threshold $\hat{\delta^*}$, with the corresponding $\hat{\lambda}^*$ and $\hat{\boldsymbol{R}}_1^N(\hat{\delta}^*, \hat{\lambda}^*)$, that minimises the average out-of-sample 1-step-ahead reconciled forecast mean squared error over all windows. The formal algorithm is given in the Section 6.2 Appendix. Although it is not required to fit forecasting models multiple times, the cross-validation procedure is still computationally expensive as it computes the NOVELIST estimator and perform reconciliation for each threshold value.

We also tested out minimising 1- to h-step-ahead MSE in the cross-validation procedure. Surprisingly, it returns almost the same best threshold parameter is in the 1-step-ahead case above. Note that when $\delta \in \left[\max_{i \neq j} |\hat{r}_{1,ij}|,\ 1\right]$, the NOVELIST estimator collapses to the shrinkage estimator, and when $\delta = 0$, it becomes the sample covariance matrix.

3.2 PC-adjusted Estimator

This method takes the latent factors directly into its construction, and is appealing when there are common drivers in the time series within the hierarchy, as we saw in the Australian tourism example. It starts by decomposing the correlation matrix $\hat{\boldsymbol{R}}_1$ into a prominent principle components part (low-rank) and a orthogonal complement part $\hat{\boldsymbol{C}}_1^K$ (the correlation matrix after removing the first K principal components). Then we can apply either shrinkage or NOVELIST estimator to $\hat{\boldsymbol{R}}_{1,K}$:

$$\hat{\boldsymbol{R}}_1^{g,K} = \sum_{k=1}^{K} \hat{\gamma}_k \hat{\boldsymbol{\xi}}_k \hat{\boldsymbol{\xi}}_k' + g(\hat{\boldsymbol{C}}_1^K)$$

where $g(.)$ is either the shrinkage or NOVELIST estimator, $\hat{\gamma}_k$ and $\hat{\boldsymbol{\xi}}_k$ are the k-th largest eigenvalue and the corresponding eigenvector of the sample covariance matrix, respectively. Similar to the NOVELIST estimator, $\hat{\boldsymbol{R}}_1^{*,K}$ is not guaranteed to be positive definite.

3.3 Scaled Variance

Apply shrinkage or NOVELIST to the correlation matrix of 1-step-ahead base forecast errors, then scale it back to covariance matrix of h-step-ahead base forecast errors using the h-step-ahead standard deviations:

$$\hat{\boldsymbol{W}}_h^{sv} = \boldsymbol{D}_h^{1/2} g(\hat{\boldsymbol{R}}_1) \boldsymbol{D}_h^{1/2},$$

where $\boldsymbol{D}_h = \text{diag}(\hat{\sigma}_{1,h}, \hat{\sigma}_{2,h}, \dots, \hat{\sigma}_{n_b,h})$, and $\hat{\sigma}_{i,h}$ is the standard deviation of the i-th series' h-step-ahead base forecast errors. The asterisk $*$ indicates either shrinkage or NOVELIST estimator.

3.4 Constructing from h-step-ahead residuals

Construct the sample covariance matrix directly from h-step-ahead base forecast errors, then apply either shrinkage or NOVELIST estimator:

$$\hat{\boldsymbol{W}}_h^g = g(\hat{\boldsymbol{W}}_h)$$

where $\hat{\boldsymbol{R}}_h$ is the sample correlation matrix of h-step-ahead base forecast errors, and $\circ$ denotes the Hadamard product (element-wise product). The asterisk $*$ indicates either shrinkage or NOVELIST estimator.

4 Simulation Design

The general design of data generating process for bottom-level series is a stationary VAR(1) process, with the following structure:

$$\boldsymbol{b}_t = \boldsymbol{A}\boldsymbol{b}_{t-1} + \boldsymbol{\epsilon}_t,$$

where $\boldsymbol{A}$ is a $n_b \times n_b$ block diagonal matrix of autoregressive coefficients $\boldsymbol{A} = diag(\boldsymbol{A}_1, \ldots, \boldsymbol{A}_m)$, with each $\boldsymbol{A}_i$ being a $n_{b,i} \times n_{b,i}$ matrix. The block diagonal structure ensures that the time series are grouped into m groups, with each group having its own autoregressive coefficients. This aim to simulate the interdependencies between the time series within each group, where reconciliation will be expected to better performed than the usual base forecasts.

The model is added with a Gaussian innovation process $\boldsymbol{\epsilon}_t$, with covariance matrix $\boldsymbol{\Sigma}$. The covariance matrix $\boldsymbol{\Sigma}$ is generated specifically using the Algorithm 1 in Hardin et al. (2013):

1. A compound symmetric correlation matrix is used for each block of size $n_{b,i}$ in $\boldsymbol{A}_i$, where the entries ρ_i for each block i are sampled from a uniform distribution between 0 and 1. They are baseline correlations within group.
2. A constant correlation, which is smaller than $\min\{\rho_1, \rho_2, \ldots, \rho_m\}$, is imposed on the entries between different blocks. It serves as baseline correlations between group.
3. The entry-wise random noise is added on top of the entire correlation matrix.
4. The covariance matrix $\boldsymbol{\Sigma}$ is then constructed by uniform sampling of standard deviations, in a range of $[\sqrt{2}, \sqrt{6}]$, for all n_b series.

We will randomly flip the signs of the covariance elements, which will create a more realistic structure in the innovation process. This can be done by pre- and post-multiplying $\boldsymbol{\Sigma}$ by a random diagonal matrix $\boldsymbol{V}$ with diagonal entries sampled from $\{-1, 1\}$, yielding $\boldsymbol{\Sigma}^* = \boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{V}$.

For all hierarchies in our experiments, we simulate two panel lengths, $T = 54$ and $T = 304$, reserving the final four observations as an out-of-sample test set. In each Monte Carlo replication ($M = 500$), we fit univariate ARIMA models (base models) to the training observations using an automatic AICc minimization algorithm from the fabletools package (O'Hara-Wild et al., 2024), generating incoherent 1–4-step base forecasts. We then reconcile these forecasts under three covariance estimators: the raw sample covariance (mint_sample), the shrinkage estimator (mint_shr), and the NOVELIST estimator (mint_n).

All data generation, covariance estimation, and reconciliation routines were implemented in the ReconCov R package, a package developed solely for this research, and is available under an open-source license on GitHub (Su, 2025).

4.1 Exploring Effects of Hierarchy's Size

In our first set of experiments, we examine how MinT combined with the NOVELIST estimator scales as the hierarchy expands. We generate synthetic data from the same VAR(1) framework described earlier, but vary the number of bottom-level series, n_b, across three different structures: the smallest with two groups of two ($n_b = 4$), an intermediate case with six groups of six ($n_b = 36$), and a much larger configuration with two clusters of fifty ($n_b = 100$). In the 4-series hierarchy, each pair collapses into a single level-1 series, which then sums to the top. In the 36-series case, each block of six forms a level-1 aggregate, and those six aggregates form the national total.

The 100-series design employs a deliberately intricate aggregation path to stress-test reconciliation methods. We first sum the one hundred bottom series into ten intermediate series by grouping them in contiguous blocks of ten. These ten series are then organised into three level-2 aggregates—four, three, and four series, respectively—before finally summing to a single top node. This asymmetric hierarchy creates overlapping correlation patterns: some level-2 series share bottom-level groups, while others draw from both, emulating practical scenarios such as regional sales aggregations that span multiple product categories or overlapping territories.

To save space, we illustrate only the six-by-six's VAR(1) and correlation configuration in Figure 4; the two-by-two and the one-hundred-series structures appear in Appendix Section 6.1.

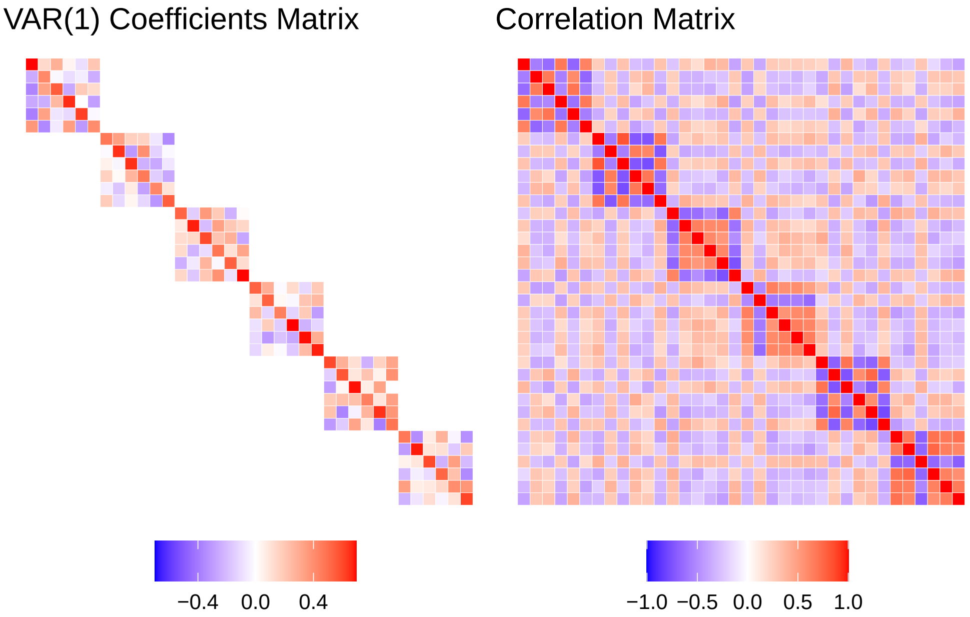


Figure 4: Heatmaps of the VAR(1) coefficient matrix and correlation matrix for the 2x50 structure.

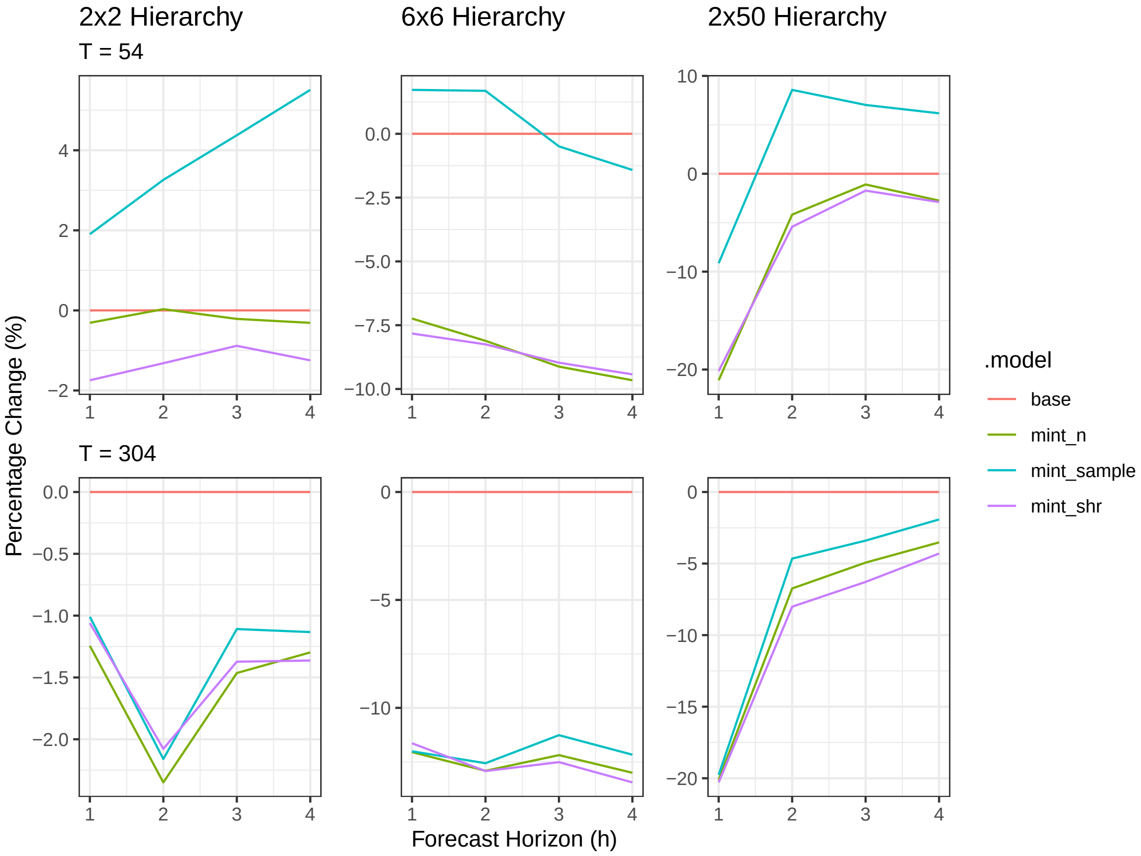


Figure 5: Relative improvement of the MSE of reconciled forecasts over the base forecasts for the 2x2, 6x6, and 2x50 hierarchical structures, for 1 to 4 steps ahead forecasts, with 2 time series lengths (T = 54 and T = 304).

Figure Figure 5 illustrates that as hierarchy size and complexity increase, the relative improvements in MSE over the base forecasts from reconciliation amplify, across all 1- to 4-step horizons. While extending the training window (from 50 to 300 in-sample observations) uniformly improves the accuracy of base ARIMA forecasts, it does not alter the relative improvements of mint_shr or mint_n. This is evidence that both high-dimensional estimators robustly handle the "large $p \gg T$" regime. By contrast, mint_sample achieves significant accuracy when more data make the sample covariance more reliable.

Despite their theoretical differences, mint_shr and mint_n exhibit nearly identical performance in these canonical settings. Yet NOVELIST's thresholding ability to minimise effects of weak correlations and retain strong ones suggests it may excel in contexts where correlation matrix has small, noisy entries. This motivates our subsequent exploration of sparse covariance structures, in which many off-diagonals are truly zero.

4.2 Exploring the sparsity of the DGP covariance matrix

In our second simulation study, we design a data-generating process that contrasts "dense" and "sparse" correlation regimes among bottom-level series, reflecting settings one might encounter in practice. We consider the same hierarchical structure of two large groups of 50 bottom series as above. Specifically, the two groups would have strong within-group dependencies throughout and either modest between-group correlations (the dense scenario) or complete independence (the sparse scenario). These correlation matrices are depicted in Figure Figure 6. Both scenarios share the same VAR(1) coefficient structure as in our previous simulations; only the innovation covariance changes. Such a setup mirrors real-world contexts where, for example, sales within a product line may exhibit strong co-movements, while those in a separate line operate nearly independently.

Figure Figure 7 presents out-of-sample mean squared error improvements over the base forecasts for each reconciliation strategy under both dense and sparse settings, with two panel lengths–54 and 304 observations–reserving the last four points for testing. In both short and long samples, MinT using either the shrinkage estimator (mint_shr) or the NOVELIST estimator (mint_n) delivers pronounced gains over incoherent ARIMA forecasts, particularly at the one-step horizon where cross-series correlations most directly inform the forecast adjustments. Although the two MinT variants perform almost indistinguishably overall, mint_n edges out mint_shr in the immediate horizon, whereas mint_shr slightly outperforms for longer horizons. By contrast, MinT with the raw sample covariance (mint_sample) suffers in small-sample settings; as expected, its performance improves dramatically with 304 data points, since the sample covariance becomes more reliable with larger n. This highlights the practical necessity of regularized estimators in high-dimensional, low-sample contexts, a situation common in real applications where histories are short relative to the number of series.

Additional designs (varying block sizes, aggregation paths, correlation configurations) also failed to separate NOVELIST from Shrinkage. Their nearly identical performance under these

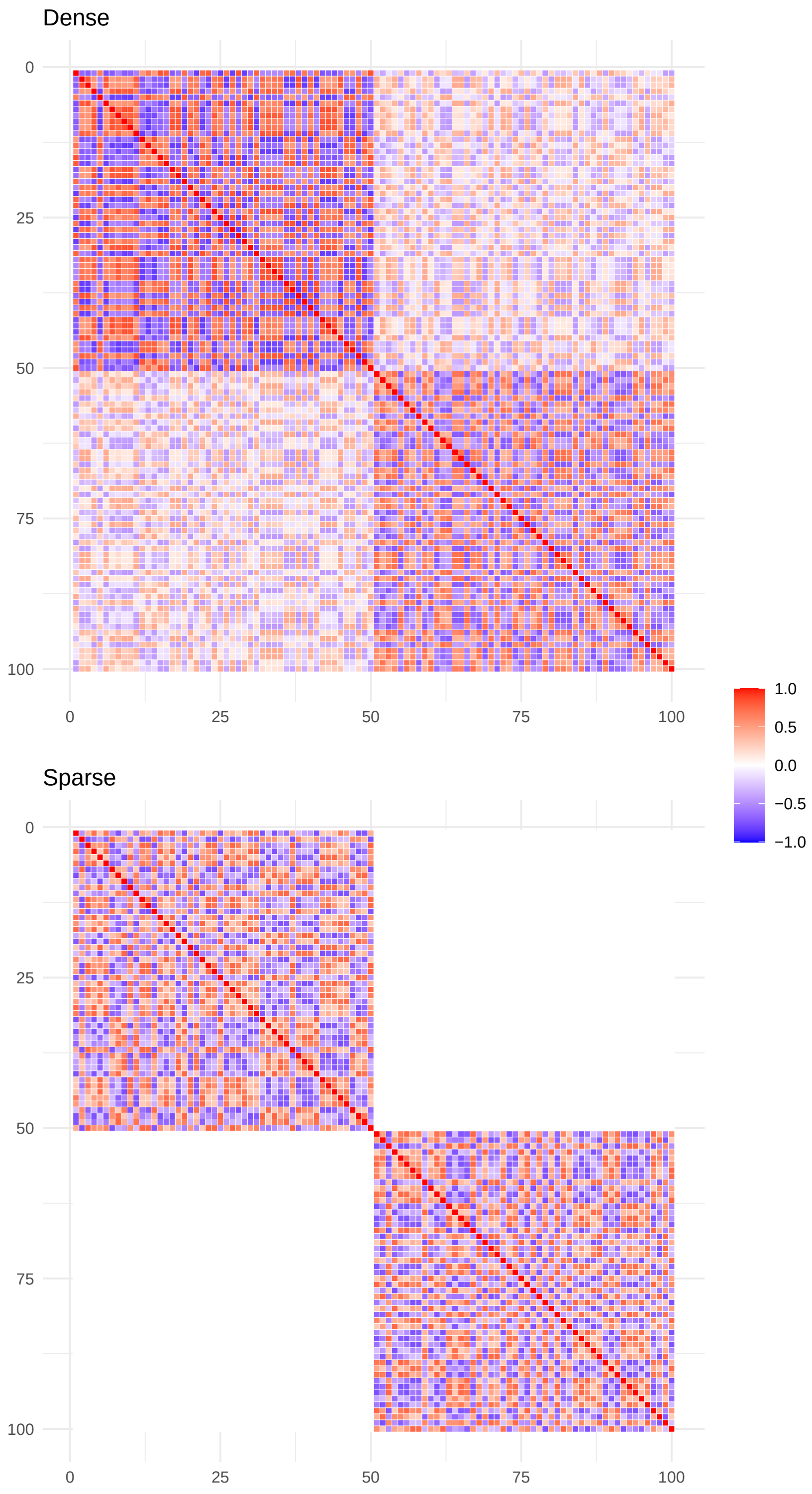


Figure 6: Heatmaps of the dense and sparse correlation matrix of the data generating process.

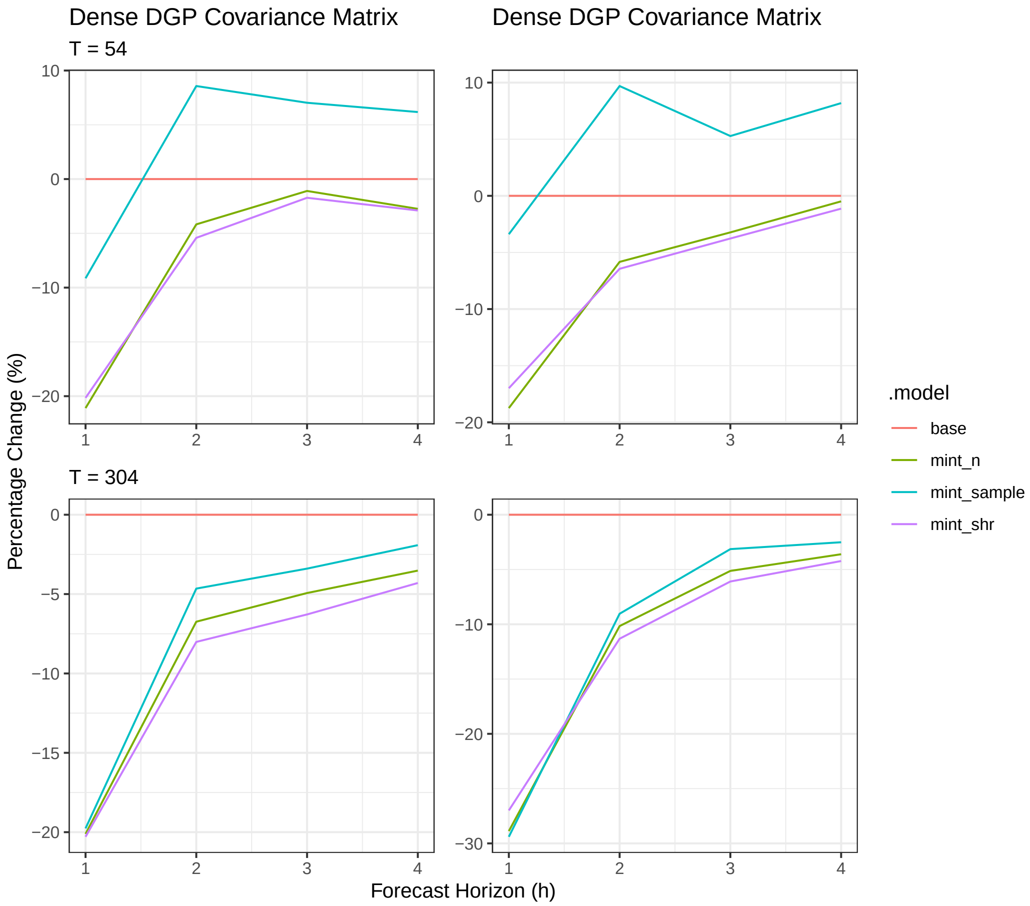


Figure 7: Relative improvement of the MSE of reconciled forecasts over the base forecasts for the 2x50 hierarchical structure with dense and sparse DGP's correlation matrix, for 1 to 4 steps ahead forecasts, with 2 time series lengths (T = 54 and T = 304).

synthetic scenarios suggests that our current simulation may not unveil the full advantages of the thresholding estimators. This finding motivates our turn to empirical data, where latent structural features and regime shifts, which we will discuss in the next section, may reveal performance differences.

5 Forecasting Australian Domestics Tourism

Domestic tourism flows in Australia exhibit a natural hierarchical and grouped structure, driven both by geography and by purpose of travel. At the top of this hierarchy lies the national total, which splits into the seven states and territories. Each state is further subdivided into tourism zones, which in turn break down into 77 regions. A complete illustration of this geographic hierarchy appears in Appendix Section 6.3. Intersecting this geographic hierarchy is a second dimension–travel motive–partitioning tourism flows into four categories: holiday, business, visiting friends and relatives, and other. Altogether, this yields a grouped system of 560 series, from the most disaggregated regional-purpose cells up to the full national aggregate. Table 1 depicts this structure.

Table 1: Hierarchical and grouped structure of Australian domestic tourism flows

Geographical division	Number of series per geographical division	Number of series per purpose	Total number of series
Australia	1	4	5
States	7	28	35
Zones	27	108	135
Regions	77	308	385
Total	112	448	560

We quantify tourism demand via “visitor nights”, the total number of nights spent by Australians away from home. The data is collected via the National Visitor Survey, managed by Tourism Research Australia, using computer assisted telephone interviews from nearly 120,000 Australian residents aged 15 years and over (*Tourism Research Australia*, 2024).

The data are monthly time series spanning from January 1998 to December 2016, resulting in 228 observations per series, producing a canonical “$n \ll p$” setting which is ideal for evaluating reconciliation approaches that rely on high-dimensional covariance estimation. The extreme dimensionality over sample size mirrors many contemporary business problems, for instance, Starbucks drink sales. Tourism demand is also economically vital yet highly volatile, with geographical and purpose-specific patterns create a realistic stress-test for reconciliation algorithms.

Wickramasuriya et al. (2019) also argued that modelling spatial autocorrelations directly from the start would be challenging as in this case of a large collection of time series. Post-processing reconciliation approaches have the advantage to implicitly model this spatial autocorrelation structure, especially true for MinT.

To assess forecasting performance, we adopt a rolling-window cross-validation scheme. Beginning with the first 120 monthly observations (January 1998-December 2005) as the initial training set, we obtain the best-fitted ARIMA model for each of the 560 series via the automatic algorithm by minimising AICc from Hyndman & Khandakar (2008). The 1- to 12-step-ahead base forecasts are then generated by these ARIMA models, and then reconciled using multiple approaches. We then roll the training window forward by one month and refit all models, rebuild reconciliations, and produce another batch of 1- to 12-step-ahead forecasts, repeating until the training set reaches December 2015. In total, this results in 133 out-of-sample windows.

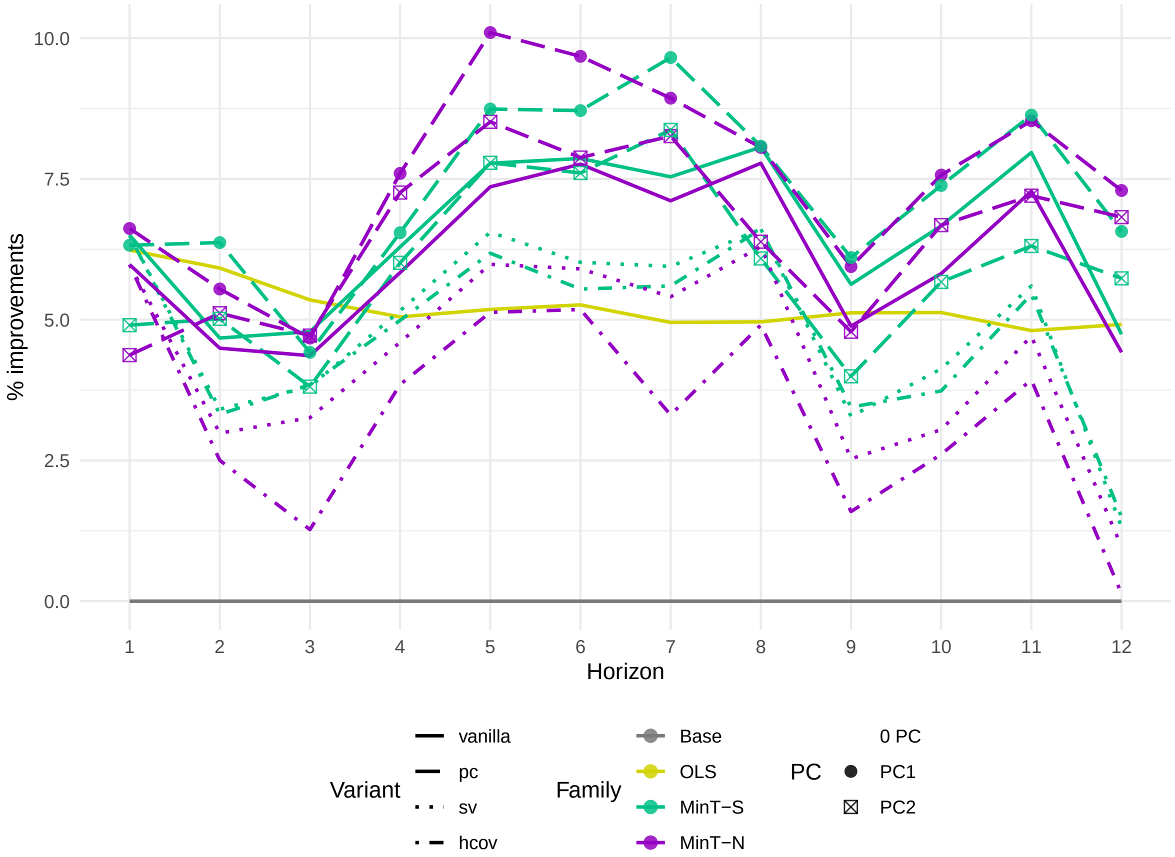


Figure 8: Percentage relative improvement in the mean squared error (MSE) of different reconciled forecasts over the base forecasts for the Australian domestic tourism data, for 1 to 12 steps ahead forecasts. The positive entries indicate an decrease in MSE.

Figure 8 and Figure 9 show that reconciliation is beneficial for point forecasts: across hori-

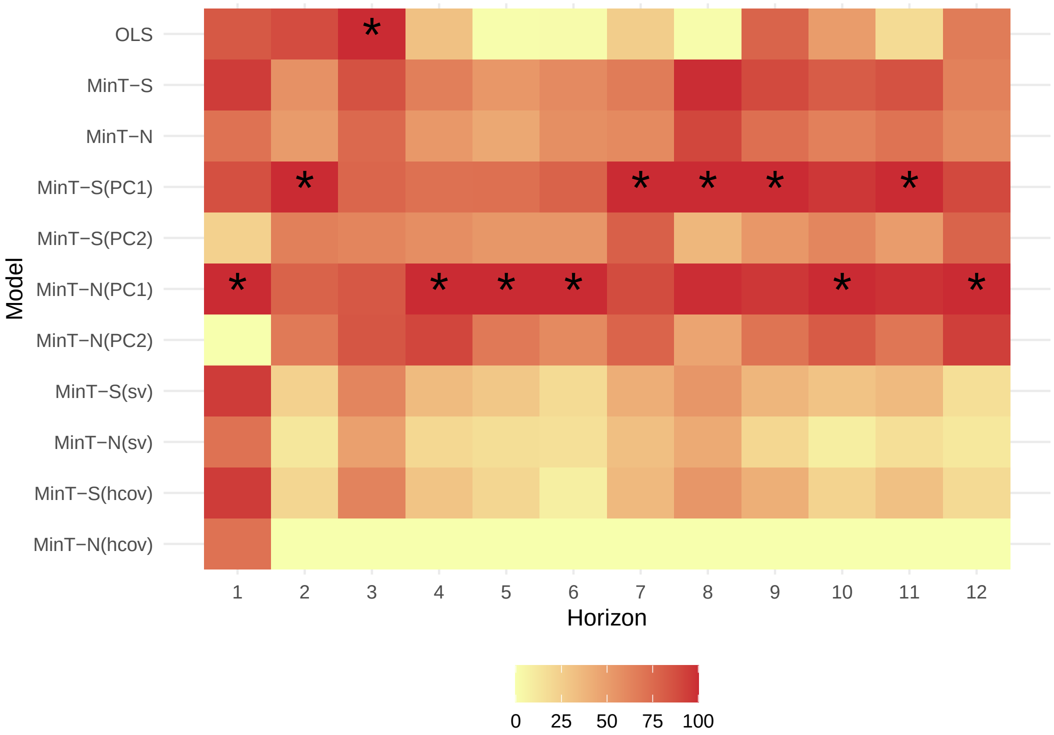


Figure 9: Heatmap of relative improvement in the mean squared error (MSE) of different reconciled forecasts over the base forecasts for the Australian domestic tourism data, for 1 to 12 steps ahead forecasts. The values are scaled to the range of 0 to 100 for better visualisation, with darker colors indicating greater improvement and best performance is noted by a star.

zons $h = 1, 2, \ldots, 12$, all reconciled methods improve MSE over incoherent base ARIMA on average. Among vanilla MinT variants, the shrinkage estimator (*MinT-S*) marginally outperforms NOVELIST (*MinT-N*) at most horizons, though the gap is small. Adjusting for a single dominant factor via PC decomposition tightens performance further: both *MinT-S(PC1)* and *MinT-N(PC1)* beat their unadjusted counterparts, with *MinT-N(PC1)* narrowly ahead. Adding more than one PC brings no additional benefit, likely due to injecting estimation noise from weaker components. Variants that modify the multi-step covariance, either via scaled-variance or direct h-step residual covariances, underperform standard MinT, suggesting that extra estimation at horizon $h > 1$ is not rewarded in this setting.

Turning to probabilistic forecasts (1-step-ahead forecasts), Figure 10 shows that *MinT-N* consistently outperforms *MinT-S* across CRPS and both Winkler scores, and PC1-adjustment again yields the best improvements. Results are aligned across probabilistic scoring rules, but the 95% Winkler reveals a notable pattern, as all MinT variants lose to the base while *OLS* performs best.

This points to tail-calibration sensitivity at high nominal coverage. The reconciliation that helps central accuracy can degrade extreme-tail interval performance when uncertainty is misallocated across the hierarchy.

In the multivariate evaluation, the Energy score places *OLS* close to the PC1-adjusted MinT methods, indicating that simple coherent projection might be able to capture a substantial share of cross-series dependence benefits even without sophisticated covariance regularisation.

The summary radar graph in Figure 11 consolidates these findings: among the selected MinT models by probabilistic criteria, *MinT-N(PC1)* clearly leads across CRPS, W80, W95, and Energy, extending the improvements seen in the single-metric panels.

Taken together, the evidence supports the use of reconciliation for both point and probabilistic forecasts in this high-dimensional setting. For point forecasts, MinT with shrinkage is a solid default; for probabilistic forecasts, NOVELIST performs more reliably. PC adjustment with a single dominant factor consistently enhances performance and should be considered when a dominant latent factor is evident. More complex adjustments, such as multiple PCs or horizon-specific covariances, add estimation noise without clear benefits and can be omitted for parsimony.

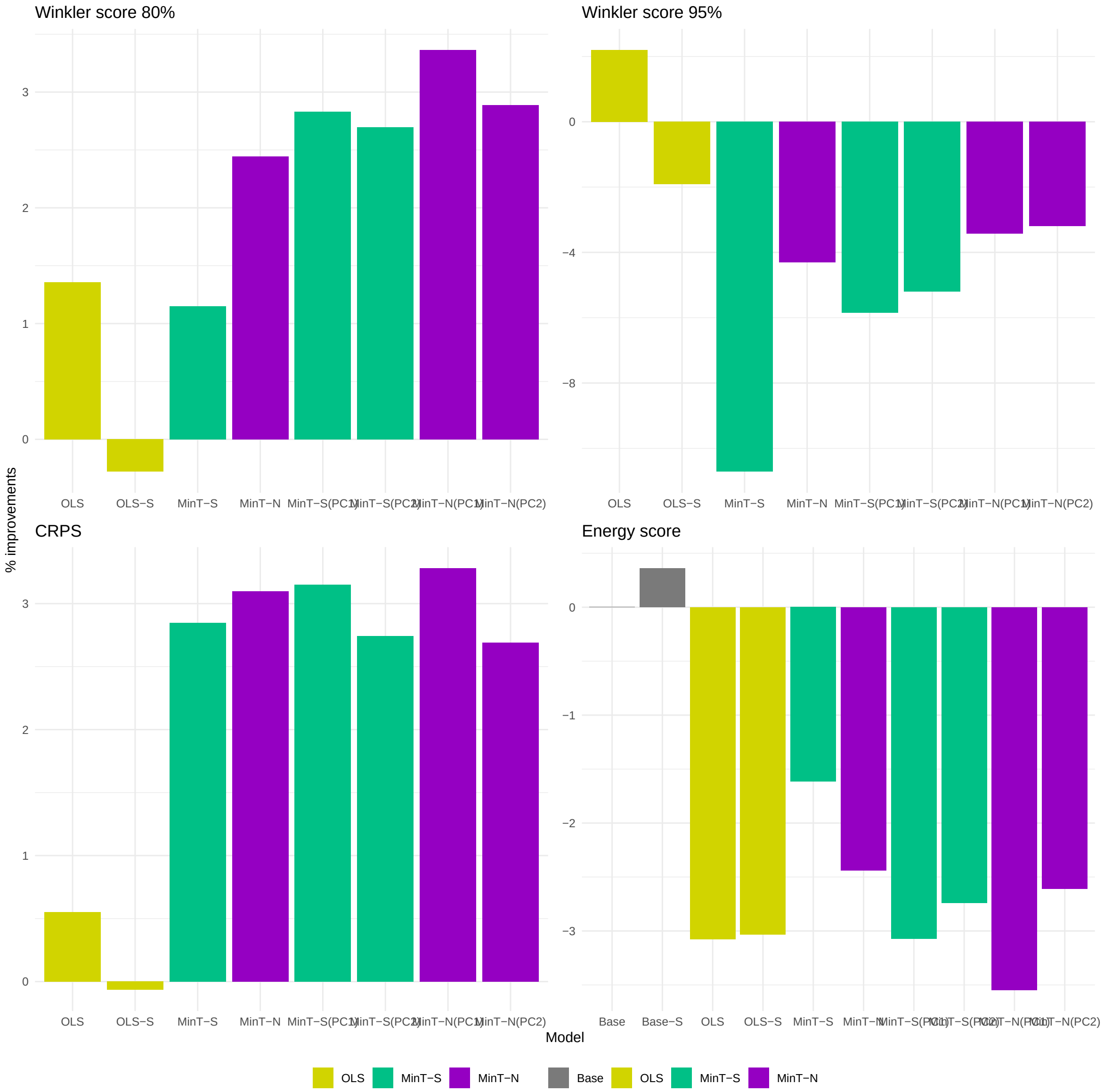


Figure 10: Percentage relative improvement in the Winkler score at 80% and 95% nominal coverage, CRPS, and Energy score of multiple reconciled forecasts over the base forecasts for the Australian domestic tourism data, for 1-step-ahead forecasts. The positive entries indicate an decrease in the probabilistic scores.

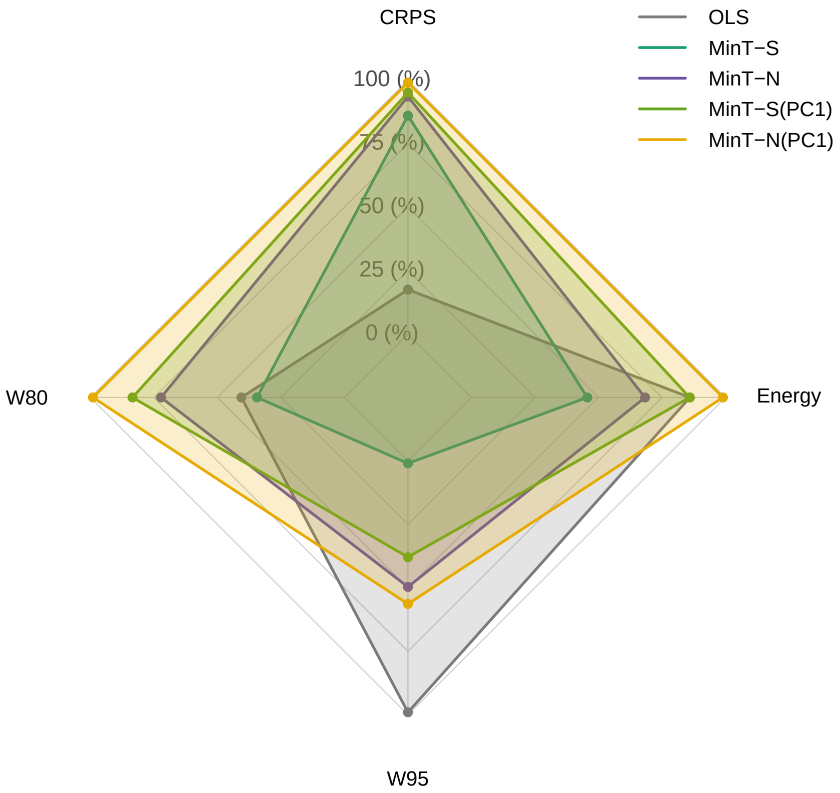


Figure 11: Radar plot of relative improvements in probabilistic scores (Winkler 80, Winkler 95, CRPS, Energy) over the base forecasts. The scores are scaled to a range of 0 to 100, with larger values indicating better performance. Only the top 5 models are shown.

6 Appendix

6.1 Simulation Settings: Supplementary Figures

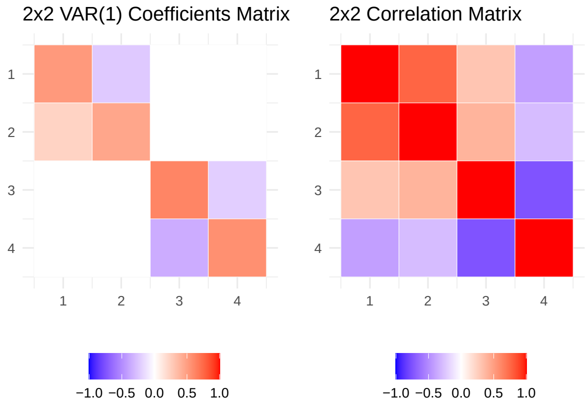
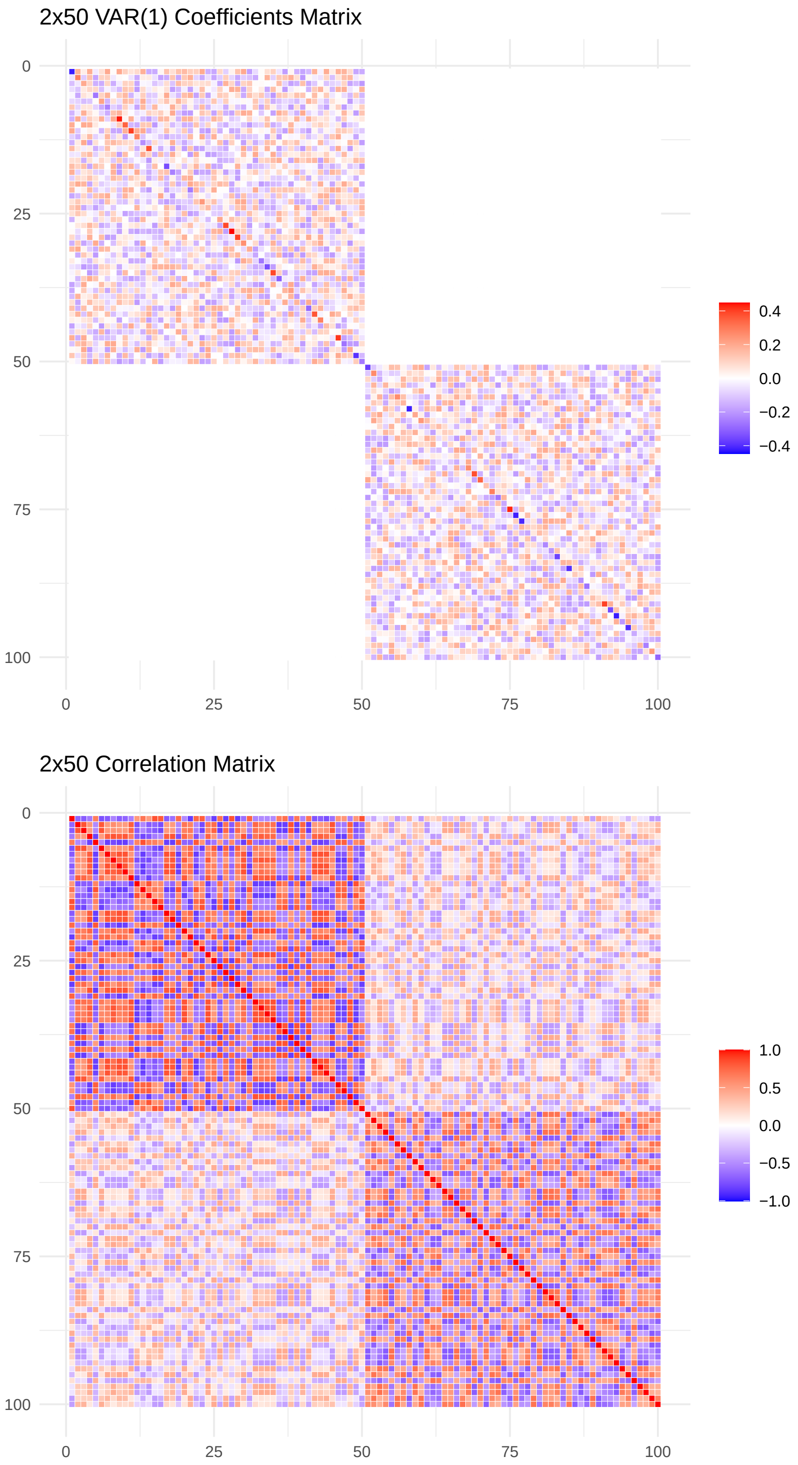


Figure 12: Heatmaps of the VAR(1) coefficient matrix and correlation matrix for the 2x2 structure.



2x50 VAR(1) Coefficients Matrix

2x50 Correlation Matrix

Figure 13: Heatmaps of the VAR(1) coefficient matrix and correlation matrix for the 6×6 structure.

6.2 Algorithm: NOVELIST cross-validation for optimal threshold δ^*

The cross-validation algorithm for NOVELIST is available in the ReconCov package (Su, 2025).

Algorithm 1 Cross-validation procedure

```
Input: Observations and fitted values $\boldsymbol{y}_t, \hat{\boldsymbol{y}}_t \in \mathbb{R}^n$ for $t = 1, \dots, T$, set of threshold candidates $\Delta$, window size $v$.
$\hat{\boldsymbol{e}}_t = \boldsymbol{y}_t - \hat{\boldsymbol{y}}_t$ for $t = 1, \dots, T$
for $i = v : T - 1$ do
    $j = i - v + 1$
    $\hat{\boldsymbol{W}}_j = \frac{1}{v}\sum_{t=j}^{i} \hat{\boldsymbol{e}}_t \hat{\boldsymbol{e}}_t'$
    $\hat{\boldsymbol{D}}_j = \mathrm{diag}(\hat{\boldsymbol{W}}_j)$
    $\hat{\boldsymbol{R}}_j = \hat{\boldsymbol{D}}_j^{-1/2} \hat{\boldsymbol{W}}_j \hat{\boldsymbol{D}}_j^{-1/2}$
    for $\delta \in \Delta$ do
        Compute thresholded correlation $\hat{\boldsymbol{R}}_{j,\delta}$ using Equation 5
        Compute $\hat{\lambda}_{j,\delta}$ using Equation 6
        Compute $\hat{\boldsymbol{R}}_{j,\delta}^{N}$ using Equation 4
        $\hat{\boldsymbol{W}}_{j,\delta}^{N} = \hat{\boldsymbol{D}}_j^{1/2} \hat{\boldsymbol{R}}_{j,\delta}^{N} \hat{\boldsymbol{D}}_j^{1/2}$
        $\boldsymbol{G} = (\boldsymbol{S}' \hat{\boldsymbol{W}}_{j,\delta}^{N^{-1}} \boldsymbol{S})^{-1} \boldsymbol{S}' \hat{\boldsymbol{W}}_{j,\delta}^{N^{-1}}$
        Reconciled forecasts $\tilde{\boldsymbol{y}}_{i+1|\delta} = \boldsymbol{S}\boldsymbol{G}\hat{\boldsymbol{y}}_{i+1}$
        $\tilde{\boldsymbol{e}}_{i+1|\delta} = \boldsymbol{y}_{i+1} - \tilde{\boldsymbol{y}}_{i+1|\delta}$
    end for
end for
$\mathrm{MSE}_\delta = \frac{1}{T-v}\sum_{i=v}^{T-1} (\tilde{\boldsymbol{e}}_{i+1|\delta})^2$ for each $\delta \in \Delta$
$\hat{\delta}^* = \arg\min_{\delta \in \Delta} \mathrm{MSE}_\delta$
Compute $\hat{\lambda}^*$ on all training data using $\hat{\delta}^*$
Compute $\hat{\boldsymbol{R}}_1^*$ using $\hat{\delta}^*$ and $\hat{\lambda}^*$ on all training data, using Equation 3
Output: Estimate of optimal $\hat{\delta}^*$
```

6.3 Appendix: Australian Domestic Tourism Geographical Hierarchy

Table 2: Geographical divisions of Australia.

Series	Name	Label	Series	Name	Label
1	Australia	Total	57	Bundaberg	CAA
2	NSW	A	58	Capricorn	CAB
3	NT	B	59	Fraser Coast	CAC
4	QLD	C	60	Gladstone	CAD
5	SA	D	61	Mackay	CAE

6	TAS	E	62	Southern Queensland Country	CAF
7	VIC	F	63	Outback Queensland	CBA
8	WA	G	64	Brisbane	CCA
9	ACT	AA	65	Gold Coast	CCB
10	Metro NSW	AB	66	Sunshine Coast	CCC
11	Nth Coast NSW	AC	67	Townsville	CDA
12	Nth NSW	AD	68	Tropical North Queensland	CDB
13	Sth Coast NSW	AE	69	Whitsundays	CDC
14	Sth NSW	AF	70	Clare Valley	DAA
15	Central NT	BA	71	Flinders Ranges and Outback	DAB
16	Nth Coast NT	BB	72	Murray River, Lakes and Coorong	DAC
17	Central Coast QLD	CA	73	Riverland	DAD
18	Inland QLD	CB	74	Adelaide	DBA
19	Metro QLD	CC	75	Adelaide Hills	DBB
20	Nth Coast QLD	CD	76	Barossa	DBC
21	Inland SA	DA	77	Fleurieu Peninsula	DCA
22	Metro SA	DB	78	Kangaroo Island	DCB
23	Sth Coast SA	DC	79	Limestone Coast	DCC
24	West Coast SA	DD	80	Eyre Peninsula	DDA
25	Nth East TAS	EA	81	Yorke Peninsula	DDB
26	Nth West TAS	EB	82	East Coast	EAA
27	Sth TAS	EC	83	Launceston and the North	EAB
28	East Coast VIC	FA	84	North West	EBA
29	Metro VIC	FB	85	West Coast	EBB
30	Nth East VIC	FC	86	Hobart and the South	ECA
31	Nth West VIC	FD	87	Gippsland	FAA
32	West Coast VIC	FE	88	Lakes	FAB
33	Nth WA	GA	89	Phillip Island	FAC
34	Sth WA	GB	90	Geelong and the Bellarine	FBA
35	West Coast WA	GC	91	Melbourne	FBB
36	Canberra	AAA	92	Peninsula	FBC
37	Central Coast	ABA	93	Central Murray	FCA
38	Sydney	ABB	94	Goulburn	FCB
39	Hunter	ACA	95	High Country	FCC
40	North Coast NSW	ACB	96	Melbourne East	FCD
41	Blue Mountains	ADA	97	Murray East	FCE
42	Central NSW	ADB	98	Upper Yarra	FCF
43	New England North West	ADC	99	Ballarat	FDA
44	Outback NSW	ADD	100	Bendigo Loddon	FDB
45	South Coast	AEA	101	Central Highlands	FDC
46	Capital Country	AFA	102	Macedon	FDD
47	Riverina	AFB	103	Mallee	FDE
48	Snowy Mountains	AFC	104	Spa Country	FDF
49	The Murray	AFD	105	Western Grampians	FDG
50	Alice Springs	BAA	106	Wimmera	FDH
51	Barkly	BAB	107	Great Ocean Road	FEA
52	Lasseter	BAC	108	Australia's North West	GAA
53	MacDonnell	BAD	109	Australia's Golden Outback	GBA
54	Darwin	BBA	110	Australia's Coral Coast	GCA
55	Katherine Daly	BBB	111	Australia's South West	GCB
56	Litchfield Kakadu Arnhem	BBC	112	Destination Perth	GCC

References

Angam, B., Beretta, A., De Poorter, E., Duvinage, M., & Peralta, D. (2025). Forecast reconciliation for vaccine supply chain optimization. In *Communications in computer and information science* (pp. 101–118). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-74650-5/_6

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

Athanasopoulos, G., Hyndman, R. J., Kourentzes, N., & Panagiotelis, A. (2024). *Forecast reconciliation: A review*. *40*(2), 430–456. https://www.sciencedirect.com/science/article/pii/S0169207023001097

Ben Taieb, S., & Koo, B. (2019). Regularized regression for hierarchical forecasting without unbiasedness conditions. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. https://doi.org/10.1145/3292500.3330976

Carrara, C., Zambon, L., Azzimonti, D., & Corani, G. (2025). A novel shrinkage estimator of the covariance matrix for hierarchical time series. In *Italian statistical society series on advances in statistics* (pp. 140–145). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-96736-8/_24

Di Modica, C., Pinson, P., & Ben Taieb, S. (2021). Online forecast reconciliation in wind power prediction. *Electric Power Systems Research*, *190*(106637), 106637. https://doi.org/10.1016/j.epsr.2020.106637

El Gemayel, J., Lafarguette, R., Itd, K. M., et al. (2022). *United arab emirates: Technical assistance reportliquidity management and forecasting*.

Erven, T. van, & Cugliari, J. (2015). Game-theoretically optimal reconciliation of contemporaneous hierarchical time series forecasts. In *Modeling and stochastic learning for forecasting in high dimensions* (pp. 297–317). Springer International Publishing. https://doi.org/10.1007/978-3-319-18732-7/_15

Hardin, J., Garcia, S. R., & Golan, D. (2013). A method for generating realistic correlation matrices. *The Annals of Applied Statistics*, *7*(3), 1733–1762. https://www.jstor.org/stable/23566492

Huang, N., & Fryzlewicz, P. (2019). NOVELIST estimator of large correlation and covariance matrices and their inverses. *Test (Madrid, Spain)*, *28*(3), 694–727. https://doi.org/10.1007/s11749-018-0592-4

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

Hyndman, R. J., & Khandakar, Y. (2008). Automatic time series forecasting: The forecast package for R. *Journal of Statistical Software*, *27*, 1–22. https://doi.org/10.18637/JSS.V027.I03

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

https://doi.org/10.1016/j.csda.2015.11.007

Ledoit, O., & Wolf, M. (2003). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance*, *10*(5), 603–621. https://doi.org/10.1016/s0927-5398(03)00007-0

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

O'Hara-Wild, M., Hyndman, R. J., & Wang, E. (2024). *Fabletools R package* (Version v0.5.0). https://fabletools.tidyverts.org/

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), Article32. https://doi.org/10.2202/1544-6115.1175

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

Su, V. (2025). *ReconCov R package* (Version beta). https://github.com/lordtahdus/ReconCov

Tourism research australia. (2024). https://www.tra.gov.au/.

Wickramasuriya, S. L., Athanasopoulos, G., & Hyndman, R. J. (2019). Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization. *Journal of the American Statistical Association*, *114*(526), 804–819. https://doi.org/10.1080/01621459.2018.1448825