Enhancing Forecast Reconciliation: A Study of Alternative Covariance Estimators

Vincent Su

Notation

- Scalar y_t
- Vector $\boldsymbol{y}_t$
- Matrix $\boldsymbol{S}$
- Covariance matrix of in-sample 1-step-ahead base forecast errors $\hat{\boldsymbol{W}}_1$
- Its shrinkage estimator with diagonal target $\hat{\boldsymbol{W}}_{1,D}^{shr}$
- Its NOVELIST estimator $\hat{\boldsymbol{W}}_{1,thr}^{shr}$

Abstract

This is pasted from the Project Description

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 3.1, 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, then city sales for each city within a country, and finally outlet 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.

This hierarchical structure is not unique to the Starbucks sales data; it can be found in many other domains, such as national tourism, electricity demand, or Gross Domestic Product (GDP). 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).

- Talk about the history and evolution of forecasting hierarchical time series, starting from the early heuristic methods to the modern statistical approaches.
    - Single level methods
    - Optimal combination methods (OLS, WLS)
    - MinT
    - Bayesian, Machine learning
    - Probabilistic methods

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 and disaggregate them down. Middle-out methods combine both approaches by forecasting a middle level 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. 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 3.1 – the disaggregation becomes more complex since the disaggregation paths are not unique.

- Discuss the interests in MinT and how it has become a standard method for forecast reconciliation.
- Discuss the MinT's reliance on a good estimate of the covariance matrix of base forecast errors and other gaps

  - Empirical evidence of MinT under perform
  - Comparison with other methods

- Explain why this paper focus on exploring alternative covariance estimators for MinT. And is there any paper talk about this.

  - Is better estimate of W_h really lead to better performance?

- Talk about the paper outline and what will be covered in the following sections.

- **Problem Statement:**

  The sample covariance matrix, although natural, suffers in high-dimensional settings. Especially when the number of series p is huge and larger than the time dimension T, the sample covariance matrix is non-positive definite (rank T if p>T).

  The shrinkage estimators come in to tackle this issue. The shrinkage estimator with diagonal target (often shrinking toward a diagonal matrix) is proven to produce a guaranteed PD matrix (Schäfer & Strimmer, 2005). However, as it shrinks the covariance matrix toward a diagonal one, it does not have flexibility and might neglect the prominent structure presented in the covariance matrix.

  An alternative approach is to perform shrinkage of the sample covariance towards its thresholded version, instead of a diagonal matrix. This is the NOVELIST (NOVEL Integration of the Sample and Thresholded covariance estimators) method proposed by Huang & Fryzlewicz (2019). They introduced thresholding functions applied only to off-diagonal elements, allowing for more flexibility in the estimation.

  … can include more estimators …

- **Research Aim:**

  This paper assesses the reconciled forecasting performance of MinT approach using various covariance estimators, with a focus on the NOVELIST estimator.

- **Paper Outline:**

  The paper is structured as follows:

  - A literature review of forecast reconciliation and covariance estimation.
  - A description of the methodology, including the NOVELIST estimator and its principal-component-adjusted variant.
  - An experimental design using both synthetic and real hierarchical time series.
  - Empirical results and discussion.
  - Conclusions and suggestions for future work.

2 Literature Review

2.1 Forecast Reconciliation in Hierarchical and Grouped Time Series

Forecast reconciliation converts 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. Early work focused on heuristic single-level strategies, including bottom-up, top-down, and middle-out (…), each of which exploits only part of the information in the hierarchy and can induce bias or high variance.

- Cite the single level approach
- Athanasopoulos et al., 2024

Hyndman et al. (2011) first showed that all single-level methods can be written as $\tilde{y} = SG\hat{y}$, where S is the summing matrix and G is a matrix that maps base forecasts $\hat{y}$ to into the bottom level. Treating reconciliation as a GLS regression problem, Hyndman et al. (2011) found that it yields a solution for G, but the required covariance matrix of reconcilation error is not identifiable in practice (Wickramasuriya et al., 2019).

- (Talk more about how others transform it to OLS, WLS,..)
- Di Fonzo and Marini (2011)
- Athanasopoulos et al. (2009)

Wickramasuriya, Athanasopoulos & Hyndman (2019) reframed the problem by taking an optimisation approach rather than the regression. They formulated the problem as minimising the variances of all reconciled forecasts, which happens to be equivalent to minimising the trace of the covariance matrix (sum of the diagonal elements). This is known as the Minimum Trace (MinT) reconciliation method. The MinT solution is given by $G_h = (S'W_h^{-1}S)^{-1}S'W_h^{-1}$, and W_h is the covariance matrix of the h-step-ahead base forecast errors.

The MinT approach is an algebraical generalisation of the GLS, and the OLS and WLS methods are special cases of MinT when W_h is a diagonal or identity matrix, respectively. However, the MinT solution hinges on a reliable estimate of the h-step-ahead base forecast error covariance W_h. In high-dimensional setting, the usual sample covariance matrix is unstable, thus we need alternative covariance estimators.

- Structural Scaling, based only on the struc- ture of the hierarchy (Athanasopoulos et al., 2017)
- Shrinkage estimators (Schäfer & Strimmer, 2005; Ledoit & Wolf, 2004)

Empirical evaluations have demonstrated that MinT with an appropriate covariance estimate often outperforms earlier methods in both simulation and real data studies.

2.2 Covariance Estimation in High Dimensions

- Limitations of the sample covariance matrix.
- Estimators used by Wickramasuriya et al. (2019).
- Shrinkage estimators:
  - Diagonal shrinkage (e.g., Schäfer & Strimmer, Ledoit & Wolf).
  - NOVELIST estimator and its Cross-validation & PC-adjusted variant.
  - ...

2.3 Relevance to Forecast Reconciliation

- Discuss how covariance estimation affects MinT performance.
- Identify research gaps.

3 Theoretical Framework

3.1 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. When there are attributes of interest that are crossed, such as the Starbucks drinks sales at any aggregation level (brand-wise, national, city, 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.

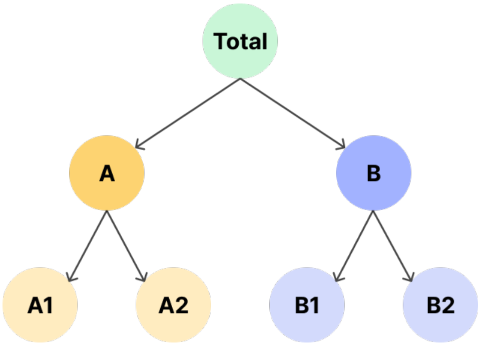

Figure 1: A 2-level hierarchical tree structure

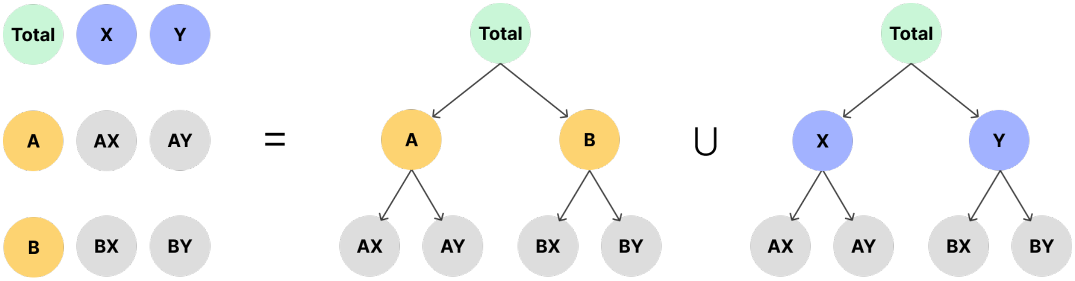

Figure 2: A 2-level grouped structure, which can be considered as the union of two hierarchical trees with common top and bottom level series

For simplicity, we refer to both of these structures as hierarchical time series, we will distinguish between them if and when it is necessary. All hierarchical structures can be represented using matrix algebra:

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

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:

$$\boldsymbol{S} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ \multicolumn{4}{c}{\boldsymbol{I}_4} \end{bmatrix}.$$

Assume we produce h-step-ahead base forecasts $\hat{\boldsymbol{b}}_{t+h|t}$ for the bottom-level series, obtained by any prediction methods. Then 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 structure. 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 $\boldsymbol{G}$.

3.2 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 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 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.

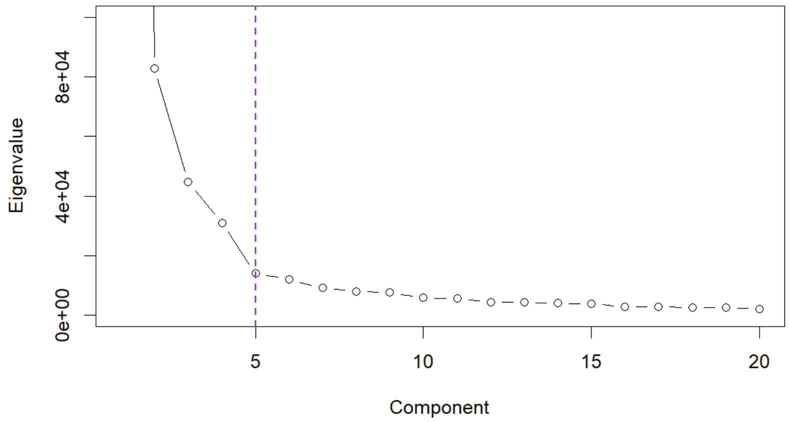

Figure 3: Twenty largest eigenvalues of one-step-ahead in-sample base forecast error covariance, Australian domestic overnight trips

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.

4 Covariance Estimation Approaches

4.1 The NOVELIST Estimator

$$g(\hat{W}_1) = \hat{W}^{shr}_{1,thr} = \lambda_\delta \hat{W}_{1,\delta} + (1 - \lambda_\delta)\hat{W}_1$$

is the NOVELIST shrinkage estimator, proposed by Huang & Fryzlewicz (2019). By convenient setting, we rewrite as sample correlation:

$$\hat{R}^{shr}_{1,thr} = \lambda_\delta \hat{R}_{1,\delta} + (1 - \lambda_\delta)\hat{R}_1,$$

$\hat{R}_{1,\delta}$ is a thresholded correlation matrix, in which thresholding is applied only to each off-diagonal element. This approach will shrink the sample correlation matrix $\hat{R}_1$ towards its thresholded version. There are various choices for the thresholding function, in this work, we use the soft-thresholding operator, defined as:

$$\hat{r}^{\delta}_{1,ij} = \text{sign}(\hat{r}_{1,ij})\,(|\hat{r}_{1,ij}| - \delta)_+.$$

After calculated the NOVELIST correlation matrix, we can re-obtain the covariance matrix $\hat{W}^{shr}_{1,thr} = \hat{D}_1^{1/2} \hat{R}^{shr}_{1,thr}\, \hat{D}_1^{1/2}$, where $\hat{D}_1 = diag(\hat{W}_1)$ is the diagonal matrix and the elements are given by the variances of the 1-step-ahead base forecast errors.

For a given threshold δ, the optimal shrinkage intensity parameter $\lambda(\delta)$ can be estimated as:

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

The formula derived using Ledoit-Wolf's lemma (Ledoit and Wolf, 2003) by Huang & Fryzlewicz. For the threshold parameter δ, we use a cross-validation to select the optimal value

- ALGORITHM WALKTHROUGH

Principal-Component Adjustment

When a factor structure is present, the procedure is:

4.2 Alternative Covariance Estimators

Brief overview of other high-dimensional estimators used in the literature for benchmarking.

Others:

- Principal Orthogonal complEment Thresholding (POET) by Fan et al. (2013)

  A Low-rank + Sparse method.
  It decompose the covariance matrix into a prominent principle components part (Low-rank) and a orthogonal complement part R_K. Then apply thresholding to R_K.
  This is similar to the NOVELIST estimator with PC-adjusted, but the difference is that POET apply thresholding to R_K, not a NOVELIST function.

5 Simulation

- Talk about description of the hierarchical time series data set (e.g., economic, financial, or synthetic data).
- Characteristics such as dimensionality, frequency, and hierarchical structure.

Experimental Design

- Design different cases of simulation studies and real-data experiments.
- Metrics: Forecast accuracy (e.g., RMSE, MAE), reconciliation error reduction, and matrix stability...
- Visualisation...

General Set-up I'm currently working on

The designed 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 $\boldsymbol{A} = diag(\boldsymbol{A}_1, \dots, \boldsymbol{A}_m)$, 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 better performed than the usual base forecasts.

The model is added with a Gaussian innovation process $\boldsymbol{\epsilon}_t$, with covariance matrix Σ. The covariance matrix Σ is generated specifically in the following way:

1. A compound symmetric correlation matrix is used for each block of size $n_{b,i}$ in $\boldsymbol{A}_i$, where the coefficients are sampled from a uniform distribution between 0 and 1.
2. The correlations between different blocks are imposed using the Algorithm 1 in Hardin et al. (2013).
3. The covariance matrix Σ is then constructed by uniform sampling of standard deviations, in a range of $[\sqrt{2}, \sqrt{6}]$, for all n_b series.

We have an option to 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 Σ by a random diagonal matrix $\boldsymbol{V}$ with entries sampled from $\{-1, 1\}$, yielding $\Sigma^* = \boldsymbol{V}\Sigma\boldsymbol{V}$.

For each series, $T = 116$ or 316 observations are generated. The first 100 or 300 observations are used for training, and the last 16 observations are used for testing. The remaining data is used to compute the best fitted ARIMA models by minimising the AICc criterion, in which we use the automatic algorithm from Fabletools R package (O'Hara-Wild et al., 2024). We refer to them as base models, and their base forecasts are then reconciled using the MinT with different covariance estimators. These include using the unbiased sample covariance matrix - MinT(Sample), the shrinkage estimator - MinT(Shrink), and the NOVELIST estimator - MinT(N). The Monte Carlo simulation is repeated $M = 10000$ times, in which the parameters for data generating process is fixed.

6 Empirical Analysis

- Comparative analysis of forecast performance using different covariance estimators on real-life dataset.

7 Discussion and Conclusion

- Evaluate how covariance estimation impacts the MinT reconciliation.
- Advantages and limitations of the NOVELIST estimator (with different ways of choosing threshold parameter).
- Practical considerations: Computational efficiency, robustness, and ease of implementation.
- Limitations of the research

References

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

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

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

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

Schäfer, J., & Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Stat. Appl. Genet. Mol. Biol.*, *4*(1), Article32. https://doi.org/10.2202/1544-6115.1175

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. *J. Am. Stat. Assoc.*, *114*(526), 804–819. https://doi.org/10.1080/01621459.2018.1448825