Annals of Tourism Research 75 (2019) 393–409

Contents lists available at ScienceDirect

Annals of Tourism Research

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

Cross-temporal coherent forecasts for Australian tourism

Nikolaos Kourentzes[a,*], George Athanasopoulos[b]

[a] Lancaster University Management School, Department of Management Science, Lancaster LA1 4YX, UK
[b] Department of Econometrics and Business Statistics, Monash University, Australia

ARTICLE INFO

Associate editor: Haiyan Song

Keywords:
Cross-sectional aggregation
Temporal aggregation
Forecast combinations
Spatial correlations

ABSTRACT

Key to ensuring a successful tourism sector is timely policy making and detailed planning. National policy formulation and strategic planning requires long-term forecasts at an aggregate level, while regional operational decisions require short-term forecasts, relevant to local tourism operators. For aligned decisions at all levels, supporting forecasts must be 'coherent', that is they should add up appropriately, across relevant demarcations (e.g., geographical divisions or market segments) and also across time. We propose an approach for generating coherent forecasts across both cross-sections and planning horizons for Australia. This results in significant improvements in forecast accuracy with substantial decision making benefits. Coherent forecasts help break intra- and inter-organisational information and planning silos, in a data driven fashion, blending information from different sources.

This article also launches the Annals of Tourism Research Curated Collection on Tourism Demand Forecast, a special selection of research in this field.

Introduction and background

The tourism sector is of vital importance to Australia.[1] In 2016–2017 tourism contributed $55.3 billion to Australia's economy, accounting for 3.2% of Australian GDP. It is the sixth largest sector in Australia directly employing 598,200 persons, accounting for 4.9% of the national workforce. Domestic tourism contributed an estimated $38.6 billion or approximately 70% of total tourism. Moreover, domestic consumers dispersed outside capital cities, much more than international arrivals, visiting regional Australia for 63% of their trips. This is extremely positive for the support and economic development of smaller regional areas, among them Indigenous communities in remote areas, an integral part of Australian society (see Mahadevan, 2018; Abascal, Fluker, & Jiang, 2016, and references therein for studies related to Australian indigenous tourism).

Sustaining a healthy, diverse and dynamic tourism sector that can meet demand is a very costly exercise. For example, in 2016–2017 there were 204 projects in the tourism investment pipeline for Australia, valued at $37.8 billion across the whole country. Such investment decisions to be successful require the support of accurate, detailed but also coherent forecasts. Forecasts are coherent when the predicted values at the disaggregate and aggregate scales are equal when brought to the same level. For example, monthly predictions sum up to annual ones and similarly regional predictions are add up to country level ones. This is an important qualifier for forecasts, so as to support aligned decision making across different planning units and horizons. Otherwise, different decision making units plan on different views of the future. In this paper we generate forecasts for Australian domestic tourism that are coherent across multiple geographical divisions, but are also coherent across time, i.e., the planning horizon.

* Corresponding author at: Department of Management Science, Lancaster University Management School, Lancaster, Lancashire LA1 4YX, UK.
E-mail address: nikolaos@kourentzes.com (N. Kourentzes).

[1] The source for all following figures is Tourism Research Australia (2018).

https://doi.org/10.1016/j.annals.2019.02.001
Received 21 November 2018; Received in revised form 30 January 2019; Accepted 1 February 2019

Table 1
Number of time series per level of hierarchy.

Hierarchy	Number of series
Level 0 (top-level)	1
Level 1	7
Level 2	27
Level 3 (bottom-level)	76
Total	111

As is the case with many tourism sectors worldwide, Australian tourist flows can be disaggregated geographically. A collection of time series adhering to such aggregation constraints, is referred to as a 'hierarchical time series' (Chapter 10 of Hyndman & Athanasopoulos, 2018, provides a detailed introduction to forecasting such structures). For the case of Australian domestic tourism this is a naturally formed geographical hierarchy. The most aggregate level of the hierarchy, referred to as level 0, comprises the total aggregate flows at the national level for Australia. Level 1, comprises flows disaggregated by the 7 states and territories, which are further disaggregated at level 2 into 27 zones and at level 3 into 76 regions. In total, the Australian tourism hierarchy, based on the geographical divisions, is constituted by 111 series that record tourism flows. These are summarised in Table 1. Table A.4 in the Appendix shows full details of the Australian tourism geographical divisions.

With such a structure each time series represents a different geographical component of the tourism sector and hence they naturally vary in nature, both in terms of scale, but also time series features. This is demonstrated in Fig. 1. The top-left panel shows total visitor nights, the proxy used in this paper for tourism flows, at the aggregate national level. The first prominent feature of this aggregate series is the strong seasonal component, with visitor nights spiking every January as this includes the summer vacations in Australia. There is also a notable upward trend, starting from 2010 until the end of the sample. As we move down the hierarchy, these features become less prominent. Although they may still exist, they are more challenging to identify and model, as the signal to noise ratio of the series decreases. Therefore, some of the series at Levels 2 and 3, illustrated at the bottom panels, show a lot more random variation and less pronounced features compared to the levels above.

Generating accurate forecasts for each component of such hierarchy is key to successful planning at all levels. Given that the data adheres to aggregation constraints, i.e., the data by nature is coherent, it is necessary that forecasts also adhere to these.

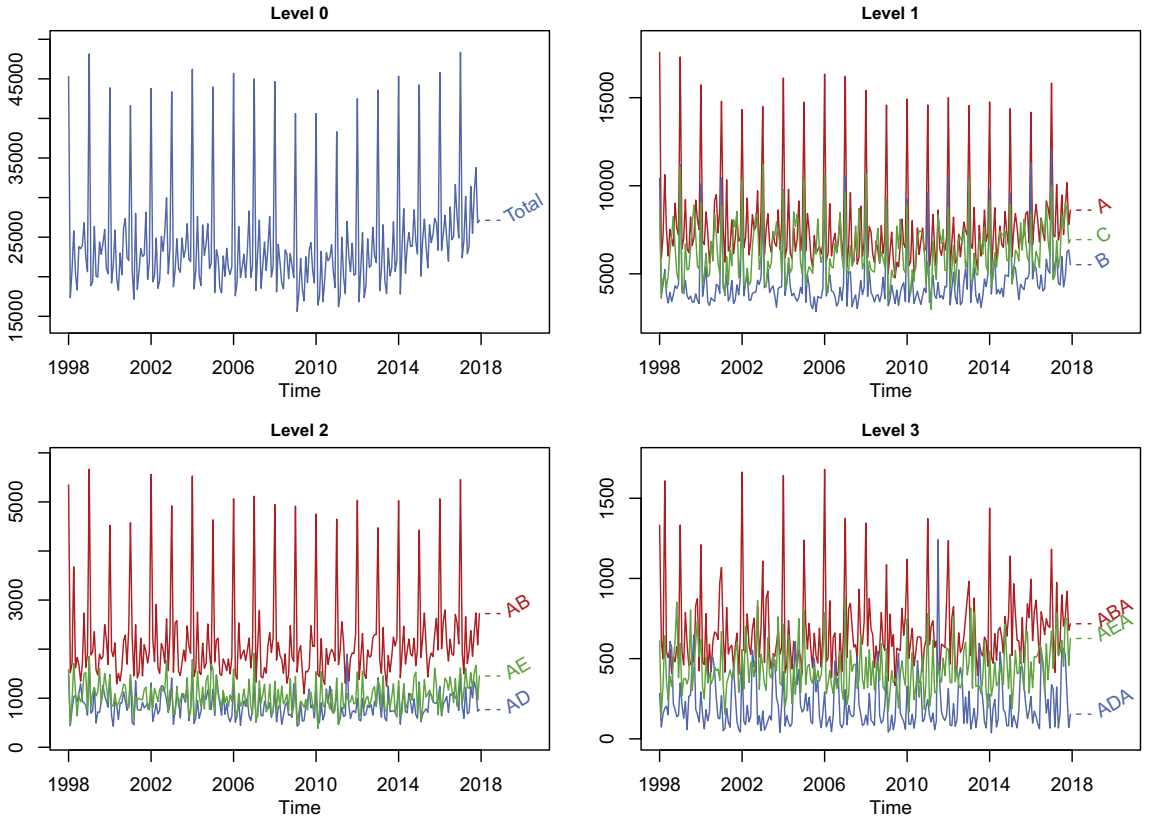


Fig. 1. Total domestic visitor nights in millions for selected geographical divisions of Australia. See Table A.4 for further details.

Traditionally, in order to achieve coherent forecasts, such structures were forecasted by first selecting a specific level of aggregation, generating forecasts at that level and then either: aggregating these up using a 'bottom-up' approach; disaggregating these down using a 'top-down' approach; or using a combination of these and implementing a 'middle-out' approach. Athanasopoulos, Ahmed, and Hyndman (2009) provide a detailed discussion and critical evaluation of the traditional approaches and their first implementation for forecasting tourism data. A later example of the traditional approach applied to tourism data is by Wan, Wang, and Woo (2013), who analyse aggregate versus disaggregate forecasts for international arrivals into Hong Kong. They consider alternative bottom-up approaches, arguing that these take advantage of the heterogeneity across the disaggregate series, and show that the traditional bottom-up approach is more accurate compared to directly forecasting at the aggregate level. A significant drawback of the traditional approaches is that they use limited information from the data, as only one level of the hierarchy is modelled and forecasted, ignoring valuable information from all other levels. The deeper the hierarchy, i.e., the more levels of aggregation, the more information is ignored. This also increases the model selection risk, where the modeller relies on a single model for all forecasts, which may be misspecified and inaccurate (Kourentzes, Barrow, & Petropoulos, 2018). Finally, they also ignore any correlations across the series.

To overcome the limitations of the traditional approaches for forecasting hierarchical times series the concept of forecast reconciliation has been developed over a sequence of papers (see among them Wickramasuriya, Athanasopoulos, & Hyndman, 2018; Hyndman, Ahmed, Athanasopoulos, & Shang, 2011; Athanasopoulos et al., 2009). The idea of forecast reconciliation works in the following way. First, a set of forecasts is generated independently for each time series in the structure. These forecasts are referred to as 'base'. Subsequently, base forecasts are reconciled so that they become coherent, while also accounting for any correlations across the series. The aforementioned papers show ample empirical evidence that forecast reconciliation does not only guarantee coherent forecasts, but also that it improves forecast accuracy. The flexibility of the approach is one of its main features. Generating base forecasts for each series means that different models can be used for different parts of the hierarchy, depending on the information set available. For example, for strategic decisions at the national or state levels, causal models may be most suitable (Sagaert, Aghezzaf, Kourentzes, & Desmet, 2018). At these levels leading indicators, such as variables capturing economic conditions and future advertising expenditure, are available to policy makers. Exploring future scenarios and the trickle down effect of these throughout the tourism sector is of interest. For the levels below, with either limited information, or when capturing the effects of explanatory variables becomes very difficult, due to the low signal-to-noise ratio, pure times series approaches may be the preferable, if not the only, choice.

Forecast reconciliation, as implemented by the papers above, will generate coherent forecasts across a hierarchy spanning the cross-sectional dimension. For the case of Australian tourism these forecasts will assist in aligning policy decisions across the geographical divisions. In the temporal dimension, forecasts supporting decisions for different planning horizons may also be generated using approaches that utilise different information sets. For example, long-term annual forecasts supporting strategic decisions typically involve high level unstructured information from multiple sources and judgement (Ord, Fildes, & Kourentzes, 2017), in this context from tourism experts. An example is the Tourism Forecasting Reference Panel comprising experts from industry and government, that was established by Tourism Research Australia. On the other hand, short-term monthly forecasts, supporting operational decisions, may be generated by only considering past tourism flows.

Andrawis, Atiya, and El-Shishiny (2011) find that combining forecasts from deseasonalised monthly and annual series is beneficial for forecasting international tourist arrivals to Egypt. Kourentzes, Petropoulos, and Trapero (2014) proposed using multiple temporal aggregation levels, instead of a single one, as is the conventional time series approach. They introduce the Multiple Aggregation Prediction Algorithm (MAPA) that uses multiple temporal aggregation with univariate exponential smoothing forecasts and find that this approach generates considerable forecast gains, particularly for long-term forecasts (further evidence by Kourentzes & Petropoulos, 2016), while mitigating the model selection uncertainty (Kourentzes, Rostami-Tabar, & Barrow, 2017). Athanasopoulos, Hyndman, Kourentzes, and Petropoulos (2017) extend this concept by introducing the notion of temporal hierarchies and forecast reconciliation in the temporal dimension. Similarly to cross-sectional forecast reconciliation, base forecasts are first generated independently for all temporal aggregation levels. Only levels that do not introduce non-integer seasonality are retained. For example, using monthly data, forecasts are generated at the monthly, bi-monthly, quarterly, four-monthly, semi-annual and annual frequencies. The processes that generate these forecasts capture different features of the times series as these are strengthened or attenuated across the different temporal aggregation levels. Fig. 2 plots the annual view of the series in Fig. 1 and demonstrates that the trending behaviour of the series becomes even more apparent, as any seasonality is filtered out and some of the noise is smoothed by the non-overlapping temporal aggregation.

The base forecasts are then reconciled, resulting to coherent forecasts across all forecast horizons and all temporal aggregation levels. Coherent forecasts across all horizons will lead to aligned decisions at different planning horizons. For example, the short-run seasonal variation which guides staffing for seasonal planning will be aligned with the longer-term trends which guide staff training. A considerable difference between MAPA and forecasting using temporal hierarchies is that the latter is independent of forecasting methodology and can incorporate statistical forecasts, generated using various methods and information sources, as well as expert judgement.

Both cross-sectional and temporal forecast reconciliation approaches have shown substantial forecast improvements empirically. A critical but intuitive reason for this improvement is that forecast reconciliation methods are forecast combination approaches dealing with parameter estimation errors and model misspecification. Forecast combinations have been regarded to be beneficial, as they reduce error variance (see for example Kourentzes et al., 2018; Barrow & Kourentzes, 2016; Elliott & Timmermann, 2013; Winkler & Clemen, 1992; Bates & Granger, 1969). Forecast combinations have also been shown to be successful within the tourism literature (see for example Wan & Song, 2018; Shen, Li, & Song, 2011; Coshall & Charlesworth, 2011).

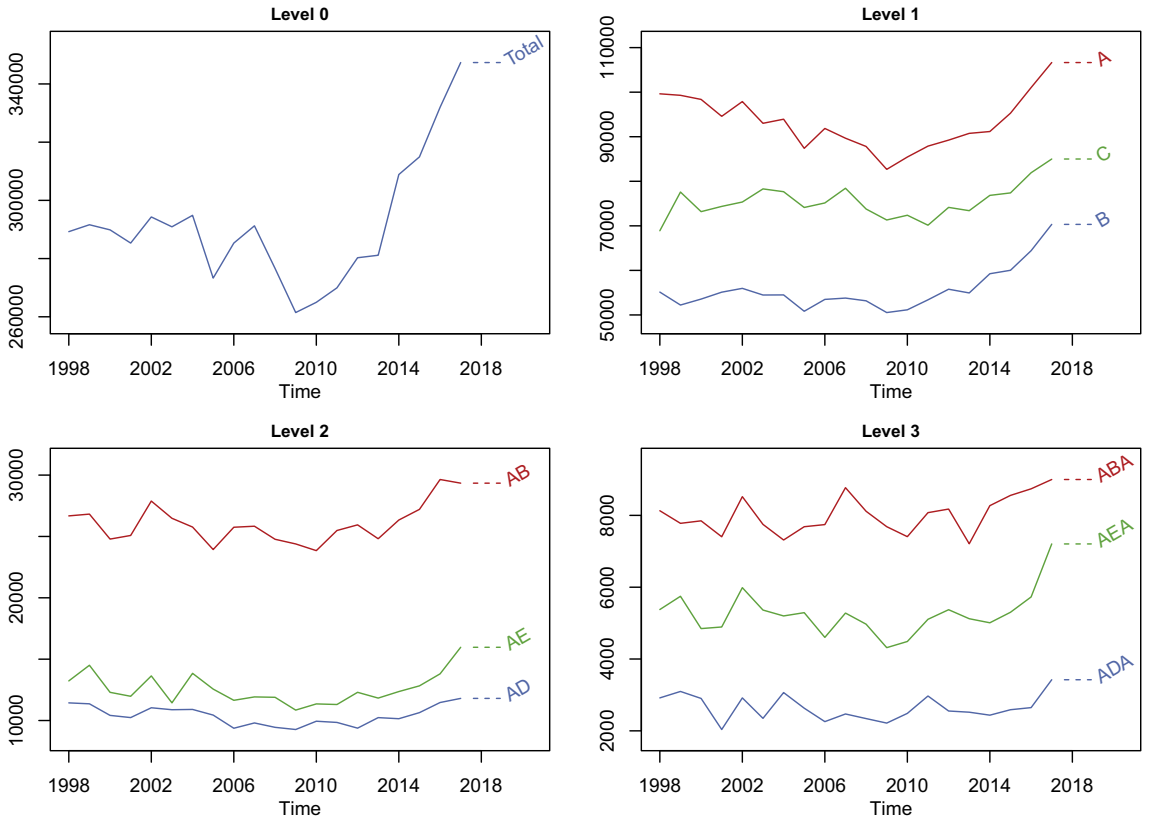


Fig. 2. Annual total domestic visitor nights in millions for selected geographical divisions of Australia. See Table A.4 for further details.

Although the appeal of cross-sectionally or temporally coherent forecasts, for decision making and improving forecast accuracy is evident, these approaches have so far only been used disjointedly. This introduces a key limitation. For example, considering forecasting tourism flows, there is limited use for high-frequency forecasts at a very aggregate geographical level. That could be weekly forecasts at the national level for Australia, if the cross-sectional reconciliation was done on weekly sampled series. On the other hand, one of the outcomes of the application of temporal forecast reconciliation could be very long-term forecasts at a very disaggregate level. Again, these are potentially of limited use to the decision makers of that level. Therefore, although using either cross-sectionally or temporally coherent forecasts offer benefits to decision making, not all outputs from these are directly useful. One would have to post-process forecasts further, for example by combining together multiple long-term disaggregate regional of temporally reconciled forecasts to produce long-term total tourism demand forecasts, which would then break the desired coherence across all levels and time periods.

In this paper we address this problem and propose a framework to generate cross-temporally coherent forecasts, supporting all levels of the hierarchy with short- to long-term forecasts. The outcome is a 'single number' forecast, where all decisions makers have a common view of the future, with apparent benefits for aligning decisions. Furthermore, we demonstrate empirically that cross-temporal reconciliation offers further accuracy gains to either cross-sectional or temporal reconciliation, as the forecasts are exposed to the complete information available to the problem domain.

The rest of the paper is structured as follows. 'Methodology: cross-temporal forecast reconciliation' section presents key concepts and insights of cross-sectional and temporal forecast reconciliation followed by our proposed approach for achieving cross-temporal forecast reconciliation. Empirical application section presents the empirical application results based on Australian tourism flows and Conclusions section discusses the managerial implications of the cross-temporally coherent forecasts and concludes.

Methodology: cross-temporal forecast reconciliation

As discussed in the introduction, forecast reconciliation so far has been applied to either the cross-sectional or temporal dimension. In this section we extend these approaches in order to achieve reconciliation in both dimensions. We start by presenting the general framework of reconciliation avoiding reference to cross-sectional or time indices where possible. We then discuss specific issues and solutions for each of these dimensions and present a process achieving reconciliation across both dimension. We refer to this as cross-temporal reconciliation.

For simplicity we demonstrate the methodology using the small hierarchy of Fig. 3. We label as y_{Tot} the observation at the most

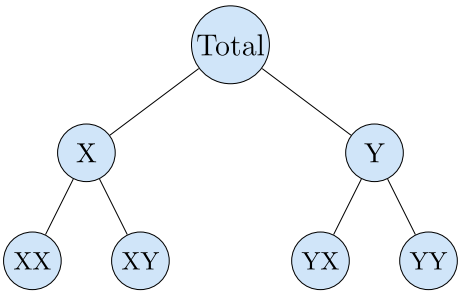


Fig. 3. A two-level hierarchical tree diagram.

aggregate (Total) level; and as y_j the observation corresponding to node j below the total. Aggregation constraints dictate that:

$$y_{Tot} = y_X + y_Y, \tag{1}$$

$$y_{Tot} = y_{XX} + y_{XY} + y_{YX} + y_{YY}, \tag{2}$$

$$y_X = y_{XX} + y_{XY}, \tag{3}$$

$$y_Y = y_{YX} + y_{YY}. \tag{4}$$

There are two important dimensions in a hierarchical setting. We denote as m the number of nodes in the bottom of the hierarchical tree, referred to as the bottom-level of the hierarchy; and as n the total number of nodes on the tree. In this simple example $n = 7$ and $m = 4$.

Stacking all the observations of the hierarchy in a n-dimensional vector $\boldsymbol{y} = (y_{Tot}, y_X, y_Y, y_{XX}, y_{XY}, y_{Y\ X}, y_{Y\ Y})'$, and similarly the bottom-level observations in an m-dimensional vector $\boldsymbol{b} = (y_{XX}, y_{XY}, y_{Y\ X}, y_{Y\ Y})'$ we can write

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

where

$$\boldsymbol{S} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ & \boldsymbol{I}_m & & \end{bmatrix} \tag{5}$$

has dimension $n \times m$ and is referred to as the 'summing' matrix. I_m is a m-dimensional identity matrix. $\boldsymbol{S}$ maps the hierarchical structure, where from the bottom level $\boldsymbol{b}$ we can construct the complete hierarchy $\boldsymbol{y}$. Observe that $\boldsymbol{S}$ captures the aggregation constraints within the hierarchy, reflected in the linear summations of the bottom-level observations.

The concept of forecast reconciliation starts by first generating an initial set of forecasts independently for each node in the hierarchy, referred to as 'base' forecasts. We denote these as $\hat{\boldsymbol{y}}_h$, a set of h-step ahead forecasts stacked in the same order as the data $\boldsymbol{y}$. In general base forecasts will not be coherent. For example, the base forecasts generated for each series for the simple hierarchy of Fig. 3 will in general not adhere to the aggregation constraints of Eqs. (1)–(4).

Forecast reconciliation of the base forecasts is achieved by

$$\tilde{\boldsymbol{y}}_h = \boldsymbol{S}\boldsymbol{G}\hat{\boldsymbol{y}}_h, \tag{6}$$

where $\boldsymbol{G}$ maps the base forecasts into the reconciled bottom-level ones and $\boldsymbol{S}$ sums these up to a set of coherent forecasts $\tilde{\boldsymbol{y}}_h$. $\boldsymbol{S}\,\boldsymbol{G}$ can be thought of as a reconciliation matrix, it takes the incoherent base forecasts, of all levels, and reconciles them. It is apparent that $\boldsymbol{G}$ linearly combines all $\hat{\boldsymbol{y}}_h$ to the reconciled bottom level forecasts, hence these blend information from all levels. A major drawback of traditional approaches is that the $\boldsymbol{G}$ used only considers information from a single level. There is now ample empirical evidence showing that using the full information set has substantial benefit in forecast accuracy (see for example Athanasopoulos et al., 2017; Wickramasuriya et al., 2018, and references therein). Gamakumara, Panagiotelis, Athanasopoulos, and Hyndman (2018) also present theoretical justifications. More importantly any decisions based on the reconciled forecasts have the ability to use all information available at different parts of the hierarchy. For example, as argued before, the top and the most disaggregate levels of the hierarchy have different information available, with the later being very close to the customer, and the former having a bird's-eye view. Therefore, the identification of $\boldsymbol{G}$ is critical for the success of hierarchical forecasting.

Wickramasuriya et al. (2018) show that

$$\boldsymbol{G} = (\boldsymbol{S}'\boldsymbol{W}_h^{-1}\boldsymbol{S})^{-1}\boldsymbol{S}'\boldsymbol{W}_h^{-1} \tag{7}$$

minimises the $tr[\boldsymbol{S}\ \boldsymbol{G}\ \boldsymbol{W}_h\boldsymbol{G}'\boldsymbol{S}']$ subject to $\boldsymbol{S}\ \boldsymbol{G}\ \boldsymbol{S} = \boldsymbol{S}$, where $\boldsymbol{S}\boldsymbol{G}\boldsymbol{W}_h\boldsymbol{G}'\boldsymbol{S}' = \mathrm{Var}(\boldsymbol{y} - \tilde{\boldsymbol{y}}_h)$, the variance covariance matrix of the h-step ahead coherent forecast errors and $\boldsymbol{W}_h = E(\hat{\boldsymbol{e}}_h\hat{\boldsymbol{e}}_h)'$ is a positive definite covariance matrix of the base forecast errors $\hat{\boldsymbol{e}}_h = \boldsymbol{y} - \hat{\boldsymbol{y}}_h$. The method is referred to as MinT as it minimises the trace of the covariance of the h-step ahead coherent forecast errors. The significance of the $\boldsymbol{S}\ \boldsymbol{G}\ \boldsymbol{S} = \boldsymbol{S}$ constraint is that the resulting coherent forecasts are unbiased, as long as the base forecasts that were used are

$$\begin{bmatrix} \hat{\sigma}^2_{Tot} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \hat{\sigma}^2_{X} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \hat{\sigma}^2_{Y} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \hat{\sigma}^2_{XX} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \hat{\sigma}^2_{XY} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \hat{\sigma}^2_{YX} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \hat{\sigma}^2_{YY} \end{bmatrix}$$

(a) Variance scaling

$$\begin{bmatrix} 4 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

(b) Structural scaling

$$\begin{bmatrix} \hat{\sigma}^2_{Tot} & \hat{\rho}_{Tot,X} & \hat{\rho}_{Tot,Y} & \hat{\rho}_{Tot,XX} & \hat{\rho}_{Tot,XY} & \hat{\rho}_{Tot,YX} & \hat{\rho}_{Tot,YY} \\ \hat{\rho}_{X,Tot} & \hat{\sigma}^2_{X} & \hat{\rho}_{X,Y} & \hat{\rho}_{X,XX} & \hat{\rho}_{X,XY} & \hat{\rho}_{X,YX} & \hat{\rho}_{X,YY} \\ \hat{\rho}_{Y,Tot} & \hat{\rho}_{Y,X} & \hat{\sigma}^2_{Y} & \hat{\rho}_{Y,XX} & \hat{\rho}_{Y,XY} & \hat{\rho}_{Y,YX} & \hat{\rho}_{Y,YY} \\ \hat{\rho}_{XX,Tot} & \hat{\rho}_{XX,X} & \hat{\rho}_{XX,Y} & \hat{\sigma}^2_{XX} & \hat{\rho}_{XX,XY} & \hat{\rho}_{XX,YX} & \hat{\rho}_{XX,YY} \\ \hat{\rho}_{XY,Tot} & \hat{\rho}_{XY,X} & \hat{\rho}_{XY,Y} & \hat{\rho}_{XY,XX} & \hat{\sigma}^2_{XY} & \hat{\rho}_{XY,YX} & \hat{\rho}_{XY,YY} \\ \hat{\rho}_{YX,Tot} & \hat{\rho}_{YX,X} & \hat{\rho}_{YX,Y} & \hat{\rho}_{YX,XX} & \hat{\rho}_{YX,XY} & \hat{\sigma}^2_{YX} & \hat{\rho}_{YX,YY} \\ \hat{\rho}_{YY,Tot} & \hat{\rho}_{YY,X} & \hat{\rho}_{YY,Y} & \hat{\rho}_{YY,XX} & \hat{\rho}_{YY,XY} & \hat{\rho}_{YY,YX} & \hat{\sigma}^2_{YY} \end{bmatrix}$$

(c) Shrinkage MinT

Fig. 4. Examples of the three alternative estimators for $\boldsymbol{W}_h$ for the hierarchy in Fig. 3.

unbiased.

A challenge with the $\boldsymbol{G}$ matrix of the MinT approach specified in (7) is that it requires an estimate of $\boldsymbol{W}_h$, which is of dimension $n \times n$ and hence this can be potentially very large. A simplifying assumption imposed by Hyndman et al. (2011), and also implemented by Athanasopoulos et al. (2009), was to set $\boldsymbol{W}_h = k_h \boldsymbol{I}_n$ for all h, and $k_h > 0$ is a proportionality constant. This simplifying assumption has been shown to work well in practice (as shown in the aforementioned references) and also makes the approach trivial to use, as no further estimation of a covariance matrix is required and $\boldsymbol{G}$ depends only on $\boldsymbol{S}$ that is always known. However, it does ignore valuable information about the scale differences (captured by the variances) and the interrelations (captured by the covariances) of the observations within the hierarchical structure.

In this paper we consider three alternative estimators. The first two are diagonal covariance estimators accommodating for the scale differences across the hierarchical levels. The third one is a shrinkage estimator accommodating for both. Note in the estimators that follow k_h is a proportionality constant and does not need to be estimated.

Variance scaling

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

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

where $\boldsymbol{e}_t$ are in-sample residuals of the base forecasts stacked the same way as the data. This specification scales the base forecasts using the variance of the residuals. For example, the resulting estimator for the simple hierarchy of Fig. 3 is provided in Fig. 4a, where $\hat{\sigma}^2_j$ are the estimated variances of the in-sample residuals corresponding to each time series. We refer to this as Var in the results that follow.

Structural scaling

Set $\boldsymbol{W}_h = k_h \boldsymbol{\Lambda}$ for all h, where $k_h > 0$, $\boldsymbol{\Lambda} = \text{diag}(\boldsymbol{S}\,\boldsymbol{1})$, and $\boldsymbol{1}$ is a unit vector of dimension n. This specification assumes that each of the bottom-level base forecasts has errors with equal variance k_h and these are uncorrelated between nodes. Therefore, higher level error variances are the sum of the error variances of the lower level series that belong to that part of the hierarchy. Hence, each element of the diagonal matrix contains the number of forecast error variances contributing to each node. Fig. 4b provides the resulting matrix for the simple hierarchical structure of Fig. 3, where, for instance, the 4 at the top level signifies that four bottom level series are used to construct it. This estimator only depends on the structure of the aggregations, and not on the actual data. It is therefore referred to as structural scaling and we denote this as *Struc* in the results that follow.

Applying structural scaling is particularly useful in cases where residuals are not available, and so the variance scaling cannot be

applied; for example, in cases where the base forecasts are generated by judgemental forecasting.

A shrinkage covariance estimator for MinT

Set $\boldsymbol{W}_h = k_h \hat{\boldsymbol{W}}^*_{1,D}$ for all h, where $k_h > 0$ and $\hat{\boldsymbol{W}}^*_{1,D} = \lambda \hat{\boldsymbol{W}}_{1,D} + (1-\lambda)\hat{\boldsymbol{W}}_1$ is a shrinkage estimator with diagonal target $\hat{\boldsymbol{W}}_{1,D}$, a diagonal matrix comprising the diagonal entries of $\hat{\boldsymbol{W}}_1$, and λ the shrinkage intensity parameter. Schäfer and Strimmer (2005) proposed to set the shrinkage intensity parameter to

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

where $\hat{r}_{ij}$ is the ijth element of $\hat{\boldsymbol{R}}_1$, the 1-step-ahead in-sample correlation matrix. As the resulting shrinkage estimator is parametrised in terms of variances and correlations it is a scale and location invariant shrinkage estimator. The effect is to shrink off-diagonal elements of $\hat{\boldsymbol{W}}_1$ towards zero, while diagonal elements (variances) remain unchanged. Therefore, in contrast to the previous variance and structural scaling estimators, this allows for strong interrelations between time series in the hierarchy to be captured, while the shrinkage alleviates the complexity of the estimation due to the size of $\boldsymbol{W}_h$. Fig. 4c exemplifies this for the hierarchy of Fig. 3. The diagonal elements are the same as for the variance scaling case (Fig. 4a), while the off-diagonal elements, $\hat{\rho}_{ij}$, are the resulting empirical covariances as shrunk towards zero, according to the prescribed shrinkage intensity parameter $\hat{\lambda}$. We denote the results associated with this estimator as *MinT*.

Cross-sectional forecast reconciliation

In the cross-sectional setting the nodes of a hierarchical tree, such as the simple tree in Fig. 3, represent observations at time t of a collection of time series connected by the aggregation constraints. The base forecasts $\hat{\boldsymbol{y}}_h$ are h-step ahead forecasts for each time series, which in this context may represent different geographical regions, market segments, etc.

In this setting, there are two main challenges: (i) the size of the cross-sectional dimension of the hierarchy; and (ii) the heterogeneity of the series across, but also within levels. The size relates directly with estimation of $\boldsymbol{W}_h$, and therefore very large hierarchies, potentially with limited history, introduce potential estimation and computational cost challenges. Given that the time series across each level can represent very different entities, the expectation is that there will be substantial heterogeneity between them. Assuming a common variance across all bottom-level series is not suitable and therefore we choose to not apply structural scaling. Hence, in the cross-sectional setting we only apply variance scaling and the shrinkage MinT estimator for reconciling the base forecasts.

Temporal forecast reconciliation

Athanasopoulos et al. (2017) proposed that in analogy to cross-sectional hierarchies, one can specify hierarchies that span the time dimension and are therefore referred to as 'temporal hierarchies'. The bottom-level of a temporal hierarchy comprises a time series observed at its highest frequency. Aggregation levels above are generated by non-overlapping temporal aggregation for all frequencies that do not introduce non-integer seasonality. For example, the hierarchical tree of Fig. 3 can be seen to represent a temporal hierarchy constructed for a quarterly series as shown in Fig. 5. The bottom-level comprises of the four quarterly observations (Q_j, with $j = 1,\ldots,4$), the middle level of two semi-annual observations (SA_j, with $j = 1,\ldots,2$) and the top level a single annual observation (A).

Base forecasts for temporal hierarchies are generated for each time series across the temporal aggregation levels, for forecast horizon h. Let $h^* = \lceil h/M_1 \rceil$ be the horizon at the most aggregate (annual) level, where M_1 is the number of observations within a year at the data sampling frequency and $\lceil x \rceil$ is the ceiling function that returns the least integer greater than or equal to x. Then, $h^* M_\ell$ steps ahead forecasts are generated for each temporal aggregation level, where M_ℓ depicts the number of observations per year for

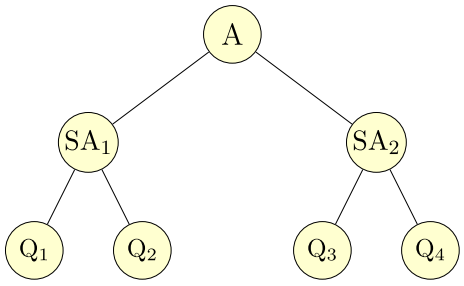


Fig. 5. A temporal hierarchy for quarterly data. Q_j with $j = 1,\ldots,4$, denote quarters, SA_j with $j = 1,2$, semi-annual observations, and A the annual observation.

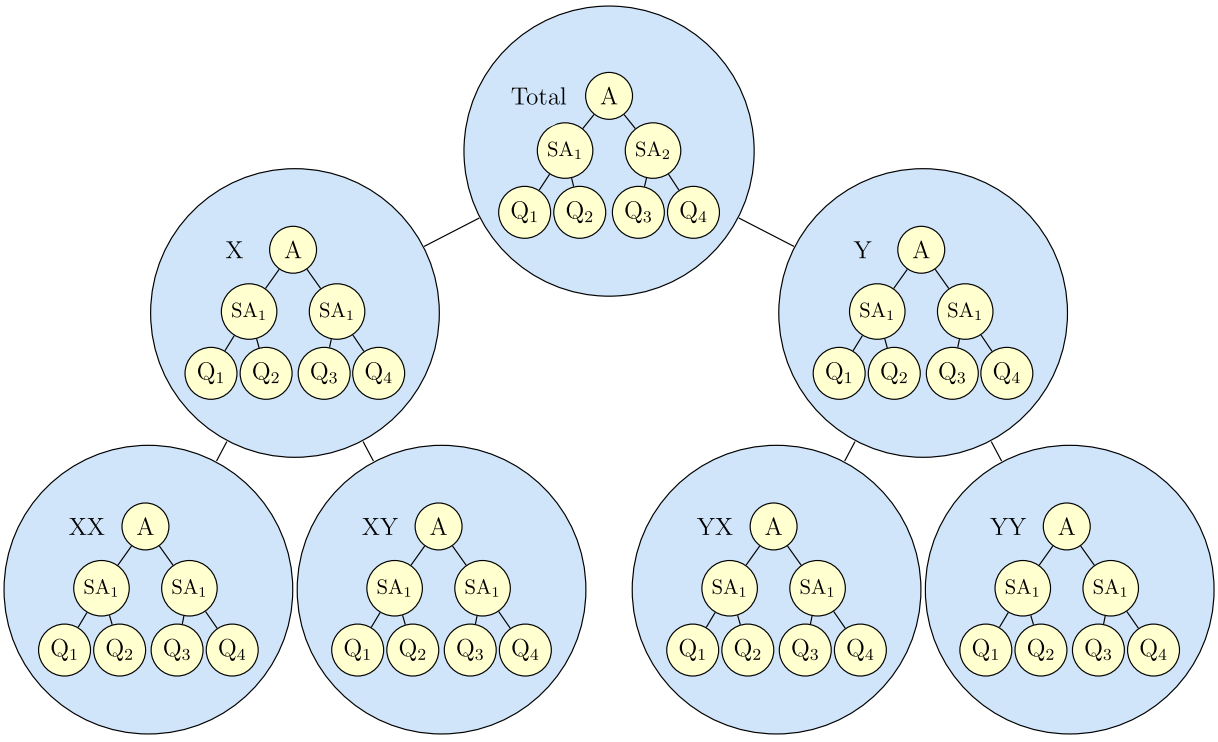


Fig. 6. A two-level cross-sectional hierarchy with base THieFs, assuming quarterly data.

aggregation level ℓ. Using as an example the temporal hierarchy in Fig. 5 and assuming that the target forecast horizon is $h = 6$ quarters, then $M_1 = 4$ (4 quarters in a year), $h^* = \lceil 6/4 \rceil = 2$ (forecast two complete years) and for the quarterly, semi-annual and annual levels $h^*M_1 = 2 \cdot 4 = 8$, $h^*M_2 = 2 \cdot 2 = 4$, $h^*M_3 = 2 \cdot 1 = 2$ steps-ahead forecasts are generated respectively. Therefore, when forecasting with temporal hierarchies we need to produce forecasts for complete hierarchical trees and then use as many as needed for the forecasting problem at hand.

In contrast to the cross-sectional case, since the forecasts for each level are for the same series, assuming homogeneous forecast errors within each level is reasonable. On the other hand, following the arguments by Athanasopoulos et al. (2017), since the covariances in $\boldsymbol{W}_h$ would be between series of different sampling frequencies due to the temporal aggregation, we do not implement the MinT shrinkage estimator. We refer to forecasts generated using temporal reconciliation as THieFs (Temporal Hierarchy Forecasts).

Cross-temporal forecast reconciliation

Intuitively, one could apply temporal and cross-sectional forecast reconciliation sequentially, aiming to achieve cross-temporally coherent forecasts. However, this does not guarantee the desired outcome. Suppose that for each node of a cross-sectional hierarchy, we consider a temporal hierarchy and generate THieFs for each node as base forecasts. We illustrate this cross-temporal combination in Fig. 6 by combining the trees of Figs. 3 and 5. Using reconciliation matrix $\boldsymbol{S}\,\boldsymbol{G}$, where $\boldsymbol{G}$ is specified as in (7), we can apply cross-sectional reconciliation for each node across the temporal hierarchies. However, although the summing matrix $\boldsymbol{S}$ will be common across each node, $\boldsymbol{W}_h$ will not, as the in-sample residuals differ across these. Hence, although we will achieve cross-sectional reconciliation, we will no longer have temporally coherent forecasts within each cross-sectional node. In the literature there have been some attempts to apply cross-sectional and temporal hierarchical forecasting sequentially (for example, see Spiliotis, Petropoulos, Kourentzes, & Assimakopoulos, 2018), which due to the sequential nature do not ensure coherence across all dimensions.

In principle it is possible to design a summing matrix $\boldsymbol{S}$ that would simultaneously consider both dimensions of reconciliation. However, its size will become prohibitively large very quickly. Each element in the cross-sectional summing matrix will need to be replaced by the complete temporal summing matrix. Furthermore, the estimation of the cross-temporal $\boldsymbol{W}_h$ will not be trivial. First its size will be equally large. Furthermore, as we argue above, the shrinkage MinT estimator is not suitable for the temporal dimension, and the structural scaling estimator is not suitable for the cross-sectional dimension. Hence designing an estimator to fully capture scaling issues and cross-sectional interdependencies is not straight forward. Instead, we propose an alternative approach to achieve cross-temporally coherent forecasts.

Given THieFs, which are independently generated for each node of the cross-sectional hierarchy, we use (7) and generate k cross-sectional reconciliations, setting $\boldsymbol{W}_h = \hat{\boldsymbol{W}}_{h,\ell}$, for each $\ell = 1,\ldots,k$, where k denotes the number of temporal aggregation levels. This

results in a respective reconciliation matrix $\boldsymbol{S}\,\boldsymbol{G}_\ell$ for each temporal aggregation level. Averaging across these we compose a consensus reconciliation matrix $\boldsymbol{SG}$, where $\boldsymbol{G} = 1/k \sum_{\ell=1}^{k} \boldsymbol{G}_\ell$, capturing the reconciliation consensus across all k temporal aggregation levels.

This fairly simple approach has the benefit of using equal weights to obtain $\boldsymbol{G}$, which eliminates any further estimation issues. Furthermore, using fixed weights has been shown to result in reliable and accurate forecast combinations (Smith & Wallis, 2009). The outcome are cross-temporally reconciled forecasts, which are coherent across both dimensions, at all scales. In the empirical evaluation that follows we denote as *Var-A* and *MinT-A* the cross-temporally coherent forecasts using respectively variance and MinT scaling in the cross-sectional dimension. We provide evidence of the magnitude of the coherency violation that occurs when $\boldsymbol{G}$ is not used.

Empirical application

Case study data

We consider 'visitor nights', the total number of nights spent by Australians away from home, as a proxy of domestic tourism flows.[2] The total number of time series considered are 111, and their split in the different levels of the hierarchy is summarised in Table 1. Total details of the geographical divisions are shown in Table A.4. Fig. 1 illustrates example series from different levels of the hierarchy, where it can be observed that series can exhibit local trend and seasonality. The data are monthly and span the period January 1998 to December 2017.

We retain the last 72 months (6 years) as a test set, which will be used to assess the performance of the competing forecasts. We choose a relatively long test set to facilitate the use of rolling origin evaluation. This allows us to generate a distribution of forecast errors for each case, so as to increase the confidence in our findings. For each time series, we consider 12 months ahead forecasts. The rolling origin is implemented in the following way. For each forecast origin, all models for the base forecasts are re-specified, i.e. re-selected and re-optimised, and the corresponding forecasts are generated and reconciled. The training data are expanded by one observation and the process is repeated for the next forecast origin, until the complete test set is exhausted. Therefore, for each time series there are $q = 61$ forecast origins. From each, we generate forecasts and calculate forecast errors.

Forecasting models

We consider two alternative forecasting model families for generating the independent base forecasts, namely the ExponenTial Smoothing (ETS) and AutoRegressive Integrated Moving Average (ARIMA) families. Both model families have been shown to perform well on tourism data (Athanasopoulos, Hyndman, Song, & Wu, 2011).

ETS captures time series as the total of four fundamental time series components: level, trend, seasonality and the error process. These components can interact in an additive or multiplicative way, in principle producing up to 30 different models, some of the most well known ones being the local level (single exponential smoothing), local trend (Holt's exponential smoothing) and the trend-seasonal (Holt-Winter's exponential smoothing) models. ETS is widely used in research and practice, due to its relatively good forecast accuracy, simplicity and minimal data requirements (Gardner, 2006). Hyndman, Koehler, Snyder, and Grose (2002) embedded exponential smoothing in the state space modelling framework, providing the statistical rationale for automatic parameter specification and model selection, greatly improving the automation and accuracy of ETS (Hyndman, Koehler, Ord, & Snyder, 2008).

The complete ETS taxonomy includes 30 alternative model specifications that correspond to archetypical time series that cover a very wide range of real time series. Hyndman, Akram, et al. (2006) shows that 11 of these specification are unstable, limiting their usefulness and Hyndman, Athanasopoulos, Bergmeir, Caceres, Chhay, and O'Hara-Wild (2018) further restrict models that have multiplicative trends, on grounds of weak forecasting performance, leaving the 15 potential models listed in Table 2. Given the widespread use of ETS, we refer the reader to standard textbooks for the formulation of the models (Hyndman et al., 2008; Ord et al., 2017; Hyndman & Athanasopoulos, 2018).

ARIMA model time series as a collection of autoregressive and moving average components, where the former regress the forecast target on its past realisations, and the latter regress the forecast target on past errors, once the time series has been differenced as needed to become stationary. Intuitively, ARIMA models aim to capture habitual elements of demand in time series, through the autoregressive components, while smoothing out the inherent noise in the data, through the moving average components. Again, we refer the reader to standard textbooks (cited above) for the formulation of ARIMA models. Following the methodology proposed by Hyndman and Khandakar (2008) for each time series we identify the appropriate ARIMA components. This involves first selecting the orders of seasonal and non-seasonal integration and then selecting seasonal and non-seasonal ARMA components based on a model selection criterion.

For each time series, at each forecast origin, the appropriate ETS and ARIMA models are chosen by minimising the Akaike Information Criterion corrected for small sample sizes (AICc, Burnham & Anderson, 2003), as implemented in the forecast package (Hyndman et al., 2018) for R (R Core Team, 2018). Although ETS and ARIMA model families have some commonalities, generally these two families perform differently. First, although additive ETS models are encompassed by ARIMA, its mixed and multiplicative

[2] The data come from the National Visitor Survey, managed by Tourism Research Australia, and are collected throughout the year using computer assisted telephone interviews from nearly 120,000 Australian residents aged 15 years and over.

Table 2
Considered ETS models.

Model component	Specification									
	1	2	3/4	5/6	7	8	9/10	11/12	13	14/15
Error										
Additive	✓		✓		✓		✓			
Multiplicative		✓		✓		✓		✓	✓	✓
Trend										
None	✓	✓			✓	✓			✓	
Additive/damped			✓	✓			✓	✓		✓
Season										
None	✓	✓	✓	✓						
Additive					✓	✓	✓	✓		
Multiplicative									✓	✓

model forms are not. Second, ARIMA can potentially capture higher order time series dynamics than ETS. Third, the different model specification strategies can result in substantially different forecasts, even if in principle both families contain some mathematically equivalent models. Therefore, we use both families to generate base forecasts so as to investigate how enforcing forecast coherence influences the outcome.

Although we focus on two univariate families of models, a major advantage of the reconciliation approaches, is that the model choice or the forecast generating mechanism for each series or each level of the hierarchy is completely flexible. For example, it may be desirable that for aggregate levels such as national or state tourism flows or for the quarterly frequency for which economic predictors are available, regression type models incorporating predictors or even multivariate models may be a better choice.

To generate cross-temporally coherent forecasts, based on the base ETS and ARIMA predictions, we expand on the base implementations of cross-sectional and temporal reconciliations available in the hts (Hyndman, Lee, Wang, & Wickramasuriya, 2018) and thief (Hyndman & Kourentzes, 2018) packages in R.

Evaluation setup

To track the forecast accuracy we use the Average Relative Mean Squared Error (AvgRelMSE). For each time series we calculate

$$\mathrm{MSE}_{i,t} = \frac{1}{h}\sum_{j=1}^{h}(y_{i,t+j} - \hat{y}_{i,t+j})^2,$$
$$\mathrm{RelMSE}_{i,t} = \frac{\mathrm{MSE}^{A}_{i,t}}{\mathrm{MSE}^{B}_{i,t}},$$

where $y_{i,t+j}$ and $\hat{y}_{i,t+j}$ are the observed value and forecast for time series i in time period $t + j$, from forecast origin t and for forecast horizon $j = 1,\ldots,h$. We need to aggregate the accuracy measurement across multiple time series and therefore it is important to remove any scale and unit information from MSE. We construct $\mathrm{RelMSE}_{i,t}$ as the ratio of $\mathrm{MSE}^{A}_{i,t}$ of the forecast of interest over a benchmark $\mathrm{MSE}^{B}_{i,t}$. As a benchmark we use the incoherent base forecast for each time series, that is the selected ETS (or ARIMA) forecast for that time series and forecast origin, prior to any hierarchical reconciliation. We aggregate across all forecast origins using the geometric mean

$$\mathrm{AvgRelMSE}_{i} = \left(\prod_{t=1}^{q}\mathrm{RelMSE}_{i,t}\right)^{1/q}, \tag{8}$$

where q is the number of forecast origins. The AvgRelMSE has very intuitive interpretation, where if it is smaller than 1, then the evaluated forecast is better than the benchmark by (1 −AvgRelMSE)100%. Furthermore, AvgRelMSE has several attractive statistical properties, in being calculable in a very wide variety of scenarios and being symmetric to over and under-forecasting (Davydenko & Fildes, 2013).

To assess whether the reported forecast error differences are significant or not, we use the non-parametric Friedman and post-hoc Nemenyi tests (Hollander, Wolfe, & Chicken, 2013). The Friedman test first establishes whether at least one of the forecasts is significantly different from the rest. If this is the case, we use the Nemenyi test to identify groups of forecasts for which there is no evidence of statistically significant differences. The advantage of this testing approach is that it does not impose any distributional assumptions and does not require multiple pairwise testing between forecasts, which would distort the outcome of the tests.We use the implementation of the tests available in the tsutils (Kourentzes, 2019) package for R.

Table 3
AvgRelMSE for ETS and ARIMA for 'All' 116 and only the 76 'Bottom' level series.

Temporal	Cross-sectional	ETS		ARIMA	
		All	Bottom	All	Bottom
None	None	1.000	1.000	1.000	1.000
	Var	0.992	0.986	0.968	0.962
	MinT	0.982	0.976	0.934	0.930
	Var-A	0.992	0.986	0.970	0.963
	MinT-A	0.986	0.981	0.944	0.938
Var	None	0.982	0.978	0.969	0.971
	Var	0.976	0.967	0.944	0.939
	MinT	0.969	**0.961**	**0.919**	**0.916**
	Var-A	0.976	0.967	0.945	0.939
	MinT-A	**0.968**	0.961	0.921	0.919
Struc	None	0.983	0.979	0.969	0.972
	Var	0.977	0.968	0.945	0.939
	MinT	0.971	0.962	0.921	0.917
	Var-A	0.977	0.968	0.945	0.940
	MinT-A	0.970	0.963	0.923	0.920

Results

Table 3 presents the summary AvgRelMSE results, over all forecast origins, for both ETS and ARIMA forecasts. We provide results for the complete hierarchy and the bottom-level time series separately, where the errors across time series are summarised using the geometric mean. The results are grouped by reconciliation method. At each column, the lowest error is highlighted in boldface. Fig. 7 illustrates the forecast errors as different reconciliation methods are used. In both panels (a) and (b), which present the results for ETS and ARIMA respectively, the proposed cross-temporally coherent Var-A and MinT-A are highlighted with light grey background.

First, we focus on the case of no-temporal reconciliation. The case of no reconciliation (None), where forecasts are produced independently and there is no guarantee of coherence is reported in the first row of Table 3. As this is used to scale the errors, across all columns its value is equal to 1. This is the worst performing case, demonstrating that any partial or complete coherence always benefits forecast accuracy. The next two results refer to the cross-sectionally reconciled forecasts using Var and MinT scaling, with the later performing best across both ETS and ARIMA forecasts, for both the bottom-level and the complete hierarchy. This is due to Var ignoring any interrelationships between the cross-sections of the hierarchy. In this case, Var-A and MinT-A do not achieve cross-temporal coherence, as the forecasts are not temporally coherent. Due to this, their performance is somewhat inferior to Var and MinT respectively.

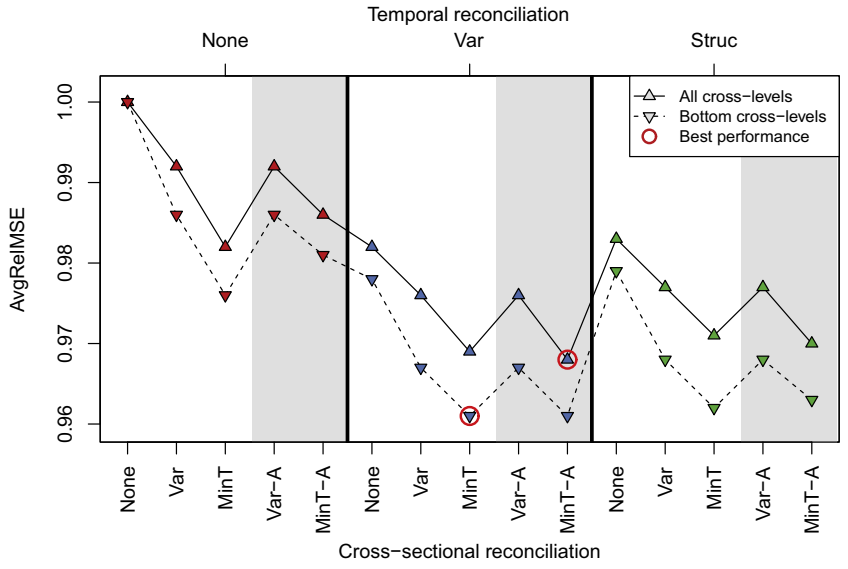
Next, we analyse the results when the forecasts are temporally reconciled, using Var scaling. It is interesting to observe that the temporal coherence, achieved by temporal reconciliation, decreases forecast errors almost uniformly, irrespective of the cross-sectional reconciliation used. This is evident in Fig. 7. For the temporally coherent forecasts, when no cross-sectional reconciliation is applied, we observe the lowest accuracy within the group. The differences between Var and Var-A, and MinT and MinT-A, are marginal, once the starting point is temporally reconciled forecasts. Across the board, the MinT variants perform better than the Var variants. The results are very similar when the temporal reconciliation uses structural scaling, albeit marginally inferior to the Var results.

Fig. 8 helps us visualise the results of the statistical comparisons between the alternative forecasts. The figure has four panels: (a) and (c) provide the results for ETS across all levels of the hierarchy and for the bottom-level only; and (b) and (d) provide the respective results for ARIMA forecasts. On the vertical axis of each panel, alternative forecasts are sorted according to their mean MSE rank. Hence, the top row shows the best performing cross-temporal combination. For example panel (c) shows that, MinT-A-Var, i.e., the combination of Mint-A for cross-sectional and Var for temporal reconciliation produces the most accurate forecasts. On the horizontal axis the forecasts are grouped by the temporal reconciliation approach.

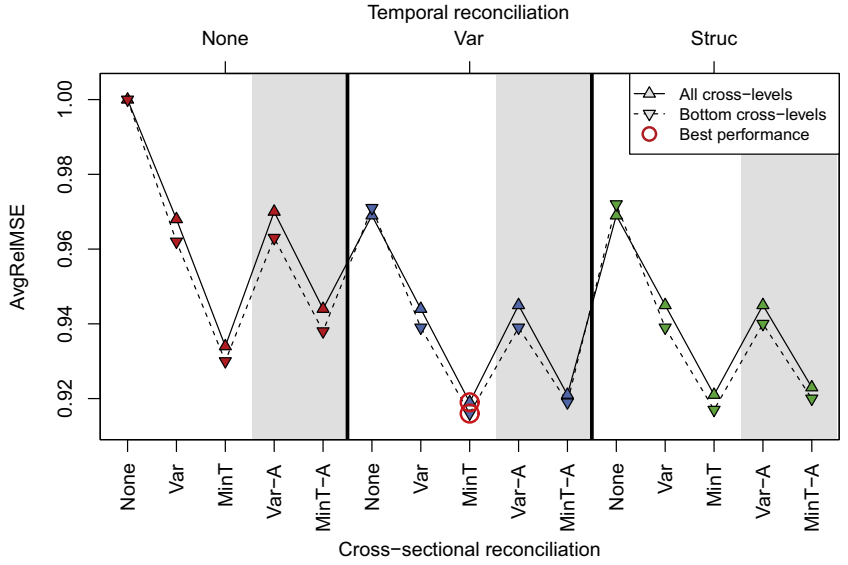
The black cell in each row indicates the tested forecast, while any blue cells in the corresponding row or column indicate forecasts for which there is no evidence of statistically significant differences, at a 5% level. Hence, any filled cells in the corresponding row or column indicate forecasts that can be grouped together as having similar forecast accuracy in statistical terms. For example, in panel (c) the first row tests the MinT-A-Var forecasts. These are found to be statistically indifferent to MinT-Var and MinT-Struc. Note that the columns shaded in light grey correspond to the cross-temporally coherent forecasts (Var-A and MinT-A).

First, we explore the results for ETS presented in panels (a) and (c). Across all levels of the hierarchy MinT-A-Var is performing significantly better than all alternatives. This is followed by MinT-Var, which is grouped together with MinT-A-Struc. When we solely consider bottom-level series in panel (c), we observe that MinT-A-Var, MinT-A-Struc and MinT-Var are grouped together as top performing approaches. On the other extreme, not implementing any reconciliation is significantly worst than all other forecasts, across the board. Analysing the results for ARIMA, panels (b) and (d), we observe that for both the bottom-level series and the complete hierarchy, MinT-A-Var, MinT-Var, MinT-A-Struc and MinT-Struc belong to the top performing group of forecasts. Similarly to the ETS case, using no reconciliation results in a significantly worse performance.

Note that for both ETS and ARIMA, the Var and Var-A variants for the cross-sectional reconciliation are similarly grouped. At this



(a) ETS results



(b) ARIMA results

Fig. 7. AvgRelMSE results for ETS and ARIMA. Methods are grouped per temporal reconciliation type. Methods with light greyed background generate cross-temporally coherent forecasts.

point we can draw some conclusions with respect to the performance of the different forecasts. Overall, temporal reconciliation provides substantial accuracy benefits over forecasts that are not temporally coherent, irrespectively of whether Var or Struc is used. The benefits are not restricted to accuracy gains, but also to aligned plans across different horizons, as short- and long-term forecasts are coherent. Cross-sectional reconciliation offers further advantage, with MinT being more accurate than Var. The cross-temporally coherent MinT-A forecasts are either best overall, or within the top performing group of forecasts, depending on the case. Therefore, the proposed cross-temporal schemes offer small, yet significant accuracy gains. However, a paramount of cross-temporally coherent forecasts is that across all levels and scales of decision making, from the operational micro to strategic macro level, from the short-term to the long-term, from the disaggregate regional to the aggregate country level, all decision makers have a coherent view of the future and are therefore able to make consistent decisions and implement aligned policy.

Fig. 9 plots the magnitude of the temporal reconciliation errors for ETS forecasts from a single forecast origin across all time series of the hierarchy, for the various reconciliation approaches. A temporal reconciliation error is defined and calculated as the difference

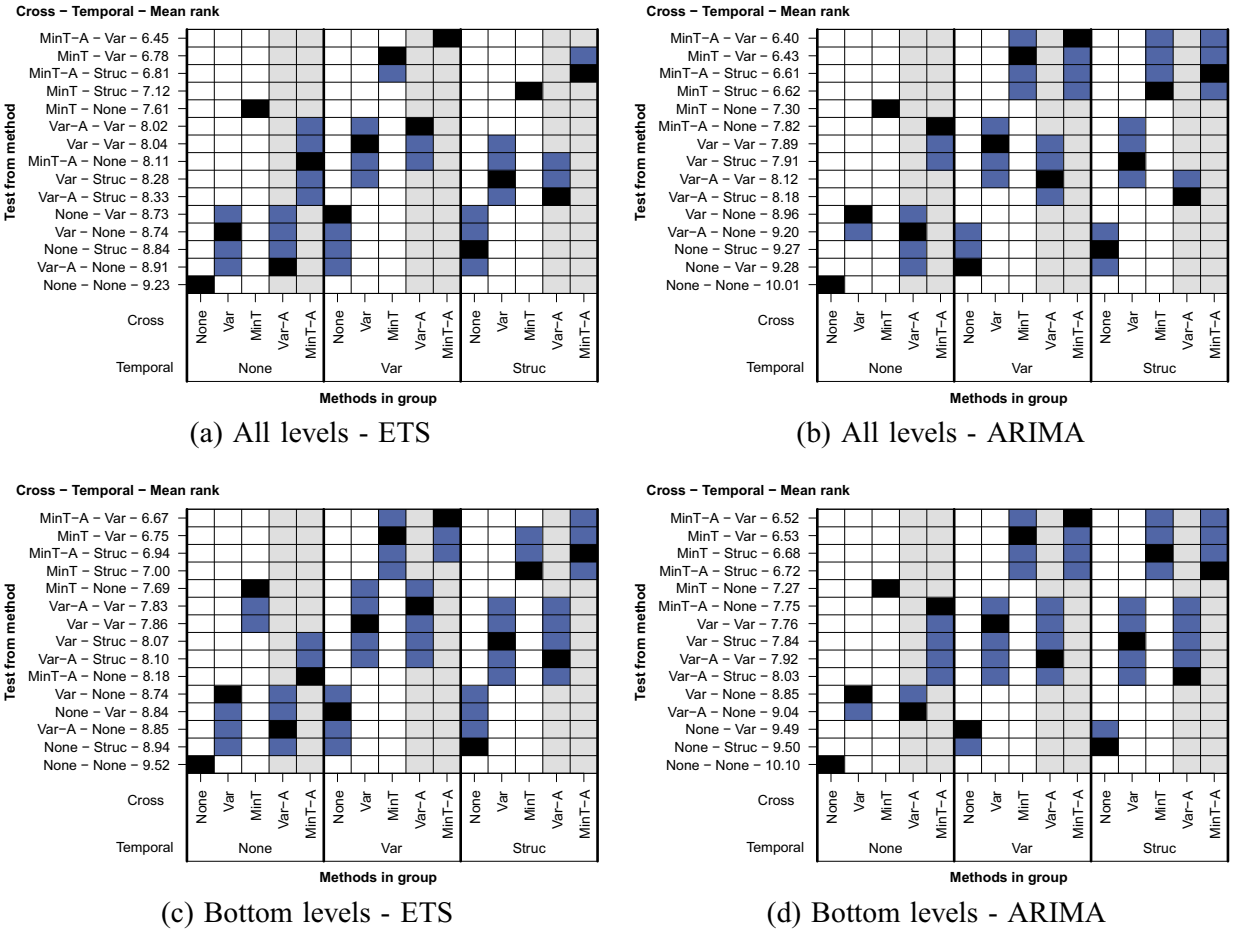


Fig. 8. Nemenyi test results at 5% significance level. Methods are sorted vertically according to MSE mean rank. Horizontally they are grouped by modelling regime. At each row, the test method is in black and any methods with no evidence of significant differences are in blue. The light grey columns refer to cross-temporally consistent methods, Var-A and MinT-A. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

between an aggregate forecast at the top (annual) temporal aggregation level and the sum of the bottom-level (monthly) forecasts. The columns of Fig. 9 show the reconciliation errors across each of the 111 series, across the four cross-sectional hierarchical levels.

In the top panel, no temporal reconciliation is applied and we observe discrepancies across all series. Note that cross-sectional reconciliation, using either Var or MinT, somewhat mitigates these differences. When temporal reconciliation is applied, using either Var or Struc scaling shown by the middle and bottom panels, the reconciliation errors become much smaller. This is evident by the significant decrease in the scale of the legend on the right side of the panels.

When no cross-sectional reconciliation is applied, as shown in the first row of the middle and bottom panels, temporal reconciliation holds. However, when applying cross-sectional reconciliation using either Var or MinT, the resulting forecasts become decoherent, with the temporal coherence violated so as to satisfy cross-sectional coherence. MinT-A and Var-A, which both offer cross-temporal coherence, avoid this problem and the reconciliation errors remain zero, for all time series. Fig. 9 is illustrative of the behaviour of the errors for other forecast origins and for ARIMA.

Conclusions

It is worthwhile to reflect on the development of hierarchical forecasting over the recent years. Athanasopoulos et al. (2009) demonstrated the benefits of hierarchical approaches for forecasting tourism flows. In that evaluation traditional approaches, such as bottom-up and top-down, that lacked the theoretical grounding of current approaches, performed competitively with the then fledgling reconciliation framework used here, and offered only cross-sectional coherence. With the development of the methodology, and specifically with the work by Wickramasuriya et al. (2018) that introduced MinT and its theoretical grounding, substantial accuracy gains became achievable. Indeed, in our evaluation, MinT provides the best accuracy. This is due to its blending of information available at different levels of the hierarchical tree and capturing any interconnections between the time series.

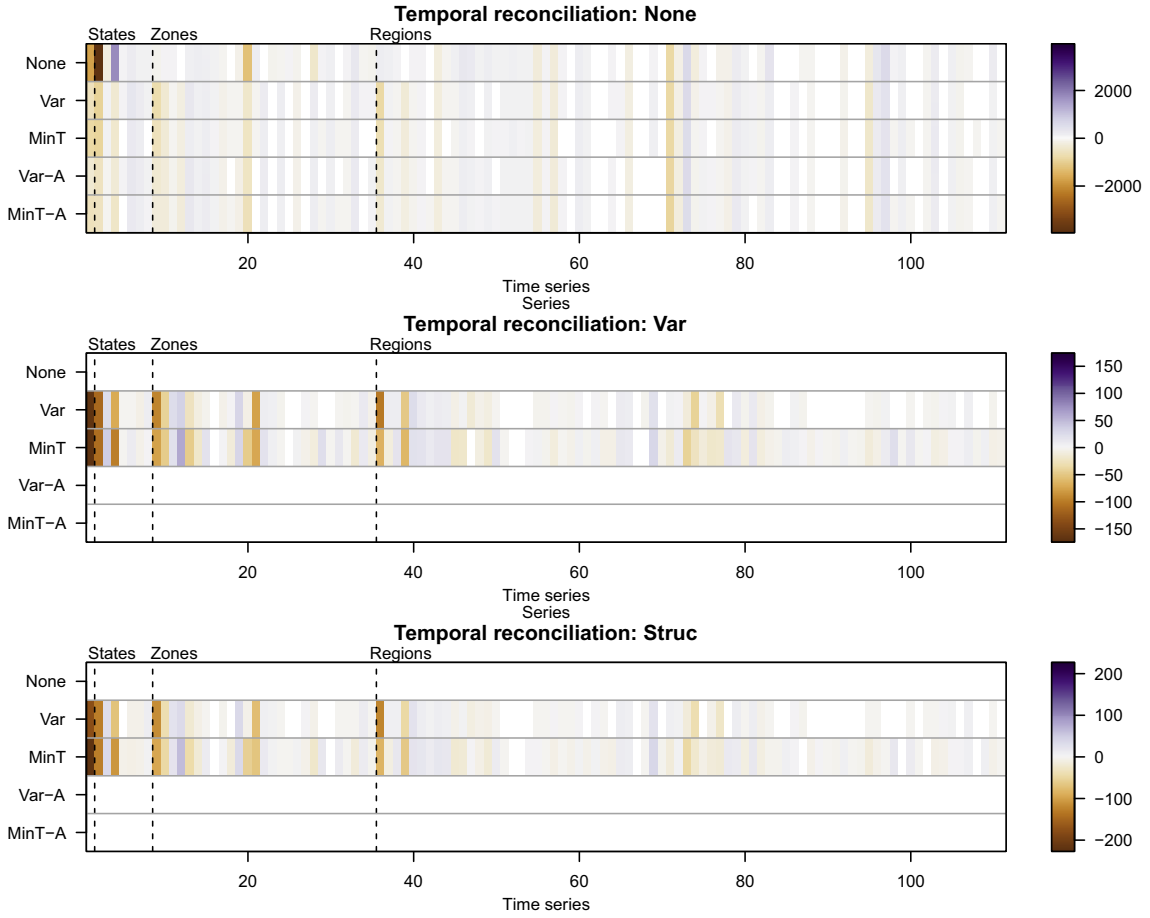


Fig. 9. Reconciliation errors after cross-sectional reconciliation, for a single series and forecast origin, between temporally disaggregate and aggregate views of the data, for the different temporal reconciliation methods. Time series are ordered in the horizontal axis as in Table A.4.

Another substantial innovation came with the idea of temporally reconciling forecasts, introduced by Kourentzes et al. (2014) and generalised in the THieF framework by Athanasopoulos et al. (2017). This added a new dimension of forecast coherence, aligning different planning levels and forecast horizons. In our empirical results, we attributed to temporal reconciliation, irrespective of the approach followed, the biggest gains in accuracy. It should be noted that THieF and MinT share a common mathematical framework, originally introduced by Hyndman et al. (2011) and subsequently refined towards these two directions: cross-sectional and temporal reconciliation, nowadays both routinely offering substantive and consistent forecast improvements. Yet, the two approaches had remained so far completely disjointed.

This paper proposes a solution to the 'decoherence' between the two reconciliation approaches, offering cross-temporally reconciled forecasts. We found that this offered small, yet statistically significant, accuracy gains. This is to be expected, as the temporal and cross-sectional reconciliations had already separately eliminated many of the incoherences of the forecasts. Nonetheless, as these reconciled forecasts had access to information only across a single dimension of the hierarchical tree, they could not achieve the maximum benefits. However, we argue that the accuracy improvements are in fact secondary to the managerial implications.

Cross-temporal reconciliation offers a single view of the future to all decision makers, removing any organisational friction from misaligned decisions. More crucially, it offers a data driven way to break within and between organisations information silos. Cross-temporal reconciliation blends information from different sources, levels of the hierarchy and scales. Base forecasts can be operational, short term, taking into consideration information directly sourced from consumers, or long term strategic ones for a whole company, sector or country, taking into consideration macro-economic and soft-information, such as technological innovation, that is infeasible to consider for very detailed disaggregate forecasts, as well as all in-between information sources. All these forecasts, conditional on different information, and generated with a variety of statistical and judgemental methods, can be blended together with the proposed approach to provide a common view of a 'single-number forecast'. If decisions and plans are based on this common view of the future, these will be already aligned, even with limited interaction between different functions of an organisation, or actors in a sector, here tourism. This is achieved without requiring changing the organisational culture or the collaboration between different allied actors in a sector, which could be the different tiers of a supply chain, or private and public organisations aiming to offer high quality service to consumers. Such managerial changes are time consuming and expensive, while the data driven nature of

cross-temporal reconciliation makes it possible to automate and frictionless.

Specifically for tourism flows, the different decisions in managing a hotel, with different scope and planning horizons, can become aligned, but at the same time decisions can become coherent with different hotels within a hotel chain, with the various supporting tourist attractions and services that benefit from and require the presence of well functioning hotels, and with the bird's eye view that a tourism board has at a regional or country level. The extend of the alignment depends solely on the scope of the cross-sectional reconciliation. This becomes particularly relevant given the availability of multiple traditional and novel data sources that lie beyond the boundaries of an organisation, such as publicly available macro-economic indicators (Sagaert, Aghezzaf, Kourentzes, & Desmet, 2017; Sagaert et al., 2018) or online consumer behaviour and social media interactions (Schaer, Kourentzes, & Fildes, 2018). We aim to explore the use of such information sources within the cross-temporal context in future work.

This raises a related question. Can we inform the analysts about how many different hierarchical levels are enough? From a statistical standpoint, additional information is beneficial. However this raises a practical consideration of how expensive is the associated data gathering and analytics infrastructure. Depending on the decision making and planning scope, forecasts at different levels need to be coherent and at minimum data for these levels needs to be collected. Nonetheless, more levels can be beneficial, as the hierarchical framework can make use of the additional information. Therefore, the analyst must balance the data gathering associated costs with the cross-temporal hierarchical forecasting benefits. As the analytics capabilities in organisations increase and data sources become increasingly interconnected the cost is expected to reduce. Furthermore, a major advantage of cross-temporal hierarchies is that they are modular in nature. When an additional layer of information becomes available, it is trivial to extend the hierarchy. Therefore, the analyst can rely on existing data structures and expand these as more data becomes available.

Furthermore, the proposed cross-temporal reconciliation revealed some of the limitations of either sides of the reconciliation, in that the generation of a global summing matrix or the estimation of respective $\boldsymbol{W}_h$ is neither trivial, not feasible by simply expanding the size of current matrices, providing a fertile route of investigation for further refining the forecast reconciliation theory.

Statement of contribution

Our paper develops the theory for improving tourism statistical forecasting and we demonstrate the value of our proposed approach on a real case predicting domestic tourism flows for Australia. We look at the hierarchical aspect of the problem, where forecasts for different geographical regions or segments, or of different planning horizons, must adhere to aggregation constraints. In the tourism literature the cross-sectional dimension of the hierarchy has been shown to provide accuracy benefits. Our work introduces the temporal aspect, for different planning horizons, and provides a novel approach for merging the two into cross-temporally coherent forecasts. We demonstrate empirically forecast accuracy gains. These forecasts provide the decision makers with a common view from the long-term aggregate strategic level, to the detailed disaggregate, supporting short term operations. This ensures aligned decision making and assists to overcome organisational and information silos in a data driven fashion.

Appendix A. Australian tourism flow time series

Table A.4 provides a detailed breakdown of the 111 time series used in this study into different geographical divisions. These report visitor nights and are used as a proxy for tourism flows.

Table A.4
Geographical divisions of Australia.

Series	Name	Label	Series	Name	Label
Total			Regions continued		
1	Australia	Total	55	Gippsland	BCB
States			56	Phillip Island	BCC
2	NSW	A	57	Central Murray	BDA
3	VIC	B	58	Goulburn	BDB
4	QLD	C	59	High Country	BDC
5	SA	D	60	Melbourne East	BDD
6	WA	E	61	Upper Yarra	BDE
7	TAS	F	62	Murray East	BDF
8	NT	G	63	Mallee	BEA
Zones			64	Wimmera	BEB
9	Metro NSW	AA	65	Great Ocean Road	BEC
10	Nth Coast NSW	AB	66	Bendigo Loddon	BED
11	Sth Coast NSW	AC	67	Macedon	BEE
12	Sth NSW	AD	68	Spa Country	BEF
13	Nth NSW	AE	69	Ballarat	BEG
14	ACT	AF	70	Central Highlands	BEH
15	Metro VIC	BA	71	Gold Coast	CAA
16	West Coast VIC	BB	72	Brisbane	CAB
17	East Coast VIC	BC	73	Sunshine Coast	CAC
18	Nth East VIC	BD	74	Central Queensland	CBA

(*continued on next page*)

Table A.4 (*continued*)

Series	Name	Label	Series	Name	Label
19	Nth West VIC	BE	75	Bundaberg	CBB
20	Metro QLD	CA	76	Fraser Coast	CBC
21	Central Coast QLD	CB	77	Mackay	CBD
22	Nth Coast QLD	CC	78	Whitsundays	CCA
23	Inland QLD	CD	79	Northern	CCB
24	Metro SA	DA	80	Tropical North Queensland	CCC
25	Sth Coast SA	DB	81	Darling Downs	CDA
26	Inland SA	DC	82	Outback	CDB
27	West Coast SA	DD	83	Adelaide	DAA
28	West Coast WA	EA	84	Barossa	DAB
29	Nth WA	EB	85	Adelaide Hills	DAC
30	Sth WA	EC	86	Limestone Coast	DBA
31	Sth TAS	FA	87	Fleurieu Peninsula	DBB
32	Nth East TAS	FB	88	Kangaroo Island	DBC
33	Nth West TAS	FC	89	Murraylands	DCA
34	Nth Coast NT	GA	90	Riverland	DCB
35	Central NT	GB	91	Clare Valley	DCC
Regions			92	Flinders Range and Outback	DCD
36	Sydney	AAA	93	Eyre Peninsula	DDA
37	Central Coast	AAB	94	Yorke Peninsula	DDB
38	Hunter	ABA	95	Experience Perth	EAA
39	North Coast NSW	ABB	96	Australia's Coral Coast	EAB
40	South Coast	ACA	97	Australia's South West	EAC
41	Snowy Mountains	ADA	98	Australia's North West	EBA
42	Capital Country	ADB	99	Australia's Golden Outback	ECA
43	The Murray	ADC	100	Hobart and the South	FAA
44	Riverina	ADD	101	East Coast	FBA
45	Central NSW	AEA	102	Launceston, Tamar and the North	FBB
46	New England North West	AEB	103	North West	FCA
47	Outback NSW	AEC	104	Wilderness West	FCB
48	Blue Mountains	AED	105	Darwin	GAA
59	Canberra	AFA	106	Kakadu Arnhem	GAB
50	Melbourne	BAA	107	Katherine Daly	GAC
51	Peninsula	BAB	108	Barkly	GBA
52	Geelong	BAC	109	Lasseter	GBB
53	Western	BBA	110	Alice Springs	GBC
54	Lakes	BCA	111	MacDonnell	GBD

References

Abascal, T. E., Fluker, M., & Jiang, M. (2016). Domestic demand for indigenous tourism in Australia: Understanding intention to participate. *Journal of Sustainable Tourism, 24*(8–9), 1350–1368. https://doi.org/10.1080/09669582.2016.1193187.

Andrawis, R. R., Atiya, A. F., & El-Shishiny, H. (2011, jul). Combination of long term and short term forecasts, with application to tourism demand forecasting. *International Journal of Forecasting, 27*(3), 870–886. http://linkinghub.elsevier.com/retrieve/pii/S0169207010001147.

Athanasopoulos, G., Ahmed, R. A., & Hyndman, R. J. (2009). Hierarchical forecasts for Australian domestic tourism. *International Journal of Forecasting, 25*, 146–166.

Athanasopoulos, G., Hyndman, R. J., Kourentzes, N., & Petropoulos, F. (2017). Forecasting with temporal hierarchies. *European Journal of Operational Research, forthcomin*(September), 1–25.

Athanasopoulos, G., Hyndman, R. J., Song, H., & Wu, D. C. (2011, jul). The tourism forecasting competition. *International Journal of Forecasting, 27*(3), 822–844. http://linkinghub.elsevier.com/retrieve/pii/S016920701000107X.

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

Bates, J. M., & Granger, C. W. J. (1969). The combination of forecasts. *Operations Research, 20*(4), 451–468.

Burnham, K. P., & Anderson, D. R. (2003). *Model selection and multimodel inference: A practical information-theoretic approach.* Springer Science & Business Media.

Coshall, J. T., & Charlesworth, R. (2011). A management orientated approach to combination forecasting of tourism demand. *Tourism Management, 32*(4), 759–769. https://doi.org/10.1016/j.tourman.2010.06.011.

Davydenko, A., & Fildes, R. (2013). Measuring forecasting accuracy: The case of judgmental adjustments to Sku-level demand forecasts. *International Journal of Forecasting, 29*(3), 510–522.

Elliott, G., & Timmermann, A. (2013). *Handbook of economic forecasting.* Elsevier.

Gamakumara, P., Panagiotelis, A., Athanasopoulos, G., & Hyndman, R. J. (2018). *Probabilistic forecasts in hierarchical time series.* Melbourne, Australia: JMonash University.

Gardner, E. S., Jr (2006). Exponential smoothing: The state of the art — Part II. *International Journal of Forecasting, 22*, 637–666.

Hollander, M., Wolfe, D. A., & Chicken, E. (2013). *Nonparametric statistical methodsVol. 751.* John Wiley & Sons.

Hyndman, R., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M.., ... Yasmeen, F.. forecast: Forecasting functions for time series and linear models. (2018). http://pkg.robjhyndman.com/forecast R package version 8.4.

Hyndman, R., Lee, A., Wang, E., & Wickramasuriya, S.. hts: Hierarchical and grouped time series. (2018). https://CRAN.R-project.org/package=hts R package version 5.1.5.

Hyndman, R. J., Ahmed, R. A., Athanasopoulos, G., & Shang, H. L. (2011, sep). 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., & Akram, M. (2006). Some nonlinear exponential smoothing models are unstable. *Tech. rep.*. Monash University, Department of Econometrics and

Business Statistics.
Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and practice*. OTextshttp://otexts.com/fpp/.
Hyndman, R. J., & Khandakar, Y. (2008). Automatic time series forecasting: The forecast package for R. *Journal of Statistical Software, 26*(3), 1–22.
Hyndman, R. J., Koehler, A. B., Ord, J. K., & Snyder, R. D. (2008). *Forecasting with exponential smoothing: The state space approach*. Berlin-Heidelberg: Springer-Verlag. www.exponentialsmoothing.net.
Hyndman, R. J., Koehler, A. B., Snyder, R. D., & Grose, S. (2002). A state space framework for automatic forecasting using exponential smoothing methods. *International Journal of Forecasting, 18*, 439–454.
Hyndman, R. J., & Kourentzes, N. (2018). thief: Temporal hierarchical forecasting. R package version 0.3 http://pkg.robjhyndman.com/thief.
Kourentzes, N. (2019). tsutils: Time series exploration, modelling and forecasting. R package version 0.9.1 https://github.com/trnnick/tsutils/.
Kourentzes, N., Barrow, D., & Petropoulos, F. (2018). Another look at forecast selection and combination: Evidence from forecast pooling. *International Journal of Production Economics*.
Kourentzes, N., & Petropoulos, F. (2016). Forecasting with multivariate temporal aggregation: The case of promotional modelling. *International Journal of Production Economics, 181*, 145–153.
Kourentzes, N., Petropoulos, F., & Trapero, J. R. (2014, Apr). Improving forecasting by estimating time series structural components across multiple frequencies. *International Journal of Forecasting, 30*(2), 291–302. http://linkinghub.elsevier.com/retrieve/pii/S0169207013001477.
Kourentzes, N., Rostami-Tabar, B., & Barrow, D. K. (2017). Demand forecasting by temporal aggregation: Using optimal or multiple aggregation levels? *Journal of Business Research, 78*, 1–9.
Mahadevan, R. (2018). Examining domestic and international visits in Australia's Aboriginal tourism. *Tourism Economics, 24*(1), 127–134. https://doi.org/10.1177/1354816617701440.
Ord, J. K., Fildes, R., & Kourentzes, N. (2017). *Principles of business forecasting*. Wessex Press Publishing Co.
R Core Team. (2018). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/
Sagaert, Y. R., Aghezzaf, E.-H., Kourentzes, N., & Desmet, B. (2017). Temporal big data for tactical sales forecasting in the tire industry. *Interfaces, 48*(2), 121–129.
Sagaert, Y. R., Aghezzaf, E.-H., Kourentzes, N., & Desmet, B. (2018). Tactical sales forecasting using a very large set of macroeconomic indicators. *European Journal of Operational Research, 264*(2), 558–569.
Schaer, O., Kourentzes, N., & Fildes, R. (2018). Demand forecasting with user-generated online information. *International Journal of Forecasting, 35*(1), 197–212.
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), 1–29.
Shen, S., Li, G., & Song, H. (2011). Combination forecasts of International tourism demand. *Annals of Tourism Research, 38*(1), 72–89. https://doi.org/10.1016/j.annals.2010.05.003.
Smith, J., & Wallis, K. F. (2009). A simple explanation of the forecast combination puzzle. *Oxford Bulletin of Economics and Statistics, 71*(3), 331–355.
Spiliotis, E., Petropoulos, F., Kourentzes, N., & Assimakopoulos, V. (2018). Cross-temporal aggregation: Improving the forecast accuracy of hierarchical electricity consumption. *Tech. rep.*National Technical University of Athens, Forecasting and Strategy Unithttps://mpra.ub.uni-muenchen.de/91762/.
Tourism Research Australia (2018). State of the industry 2016–17. *Tech. Rep.* February.
Wan, S. K., & Song, H. (2018). Forecasting turning points in tourism growth. *Annals of Tourism Research, 72*(January), 156–167. https://doi.org/10.1016/j.annals.2018.07.010.
Wan, S. K., Wang, S. H., & Woo, C. K. (2013). Aggregate vs. disaggregate forecast: Case of Hong Kong. *Annals of Tourism Research, 42*, 434–438. https://doi.org/10.1016/j.annals.2013.03.002.
Wickramasuriya, S. L., Athanasopoulos, G., & Hyndman, R. J. (2018). Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization. *Journal of the American Statistical Association, forthcoming*.
Winkler, R. L., & Clemen, R. T. (1992). Sensitivity of weight in combining forecasts. *Operations Research, 40*(3), 609–614. https://doi.org/10.1287/opre.40.3.609.

George Athanasopoulos is a Professor at the Department of Econometrics and Business Statistics at Monash University, Australia. His research interests include forecasting hierarchical and grouped times series, multivariate time series analysis and tourism economics.

Nikolaos Kourentzes is a Professor at the Department of Management Science at Lancaster University Management School, UK. His research interests are in business forecasting, model selection and specification uncertainty, judgement in forecasting and artificial intelligence.