MONASH University

MONASH BUSINESS SCHOOL

ISSN 1440-771X

Department of Econometrics and Business Statistics

http://business.monash.edu/econometrics-and-business-statistics/research/publications

On the evaluation of hierarchical forecasts

George Athanasopoulos and Nikolaos Kourentzes

August 2021

Working Paper 10/21
(Revised version of paper no. 02/20)

AACSB ACCREDITED EQUIS ACCREDITED AMBA ACCREDITED

On the evaluation of hierarchical forecasts

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

[a]*Department of Econometrics and Business Statistics, Monash University, Australia*
[b]*Skövde Artificial Intelligence Lab, School of Informatics, University of Skövde, Sweden.*
[c]*Department of Management Science, Lancaster University Management School, UK.*

Abstract

The aim of this paper is to provide a thinking road-map and a practical guide to researchers and practitioners working on hierarchical forecasting problems. Evaluating the performance of hierarchical forecasts comes with new challenges stemming from both the structure of the hierarchy and the application context. We discuss several relevant dimensions for researchers and analysts: the scale and units of the time series, the issue of sparsity, the forecast horizon, the importance of multiple evaluation windows and the multiple objective decision context. We conclude with a series of practical recommendations.

Keywords: Aggregation, coherence, hierarchical time series, reconciliation.
JEL Classifications: C18, C53, C55

1. Introduction

Evaluating hierarchical forecasting problems introduces new complications that are not relevant or critical in the standard case. In contrast to

*Correspondance: G Athanasopoulos, Department of Econometrics and Business Statistics, Monash University, Australia.
Email address: George.Athanasopoulos@monash.edu (George Athanasopoulos)

general time series modelling and forecasting, hierarchical time series by their very nature provide a specific problem context. As we demonstrate in what follows, the hierarchical time series decision making context requires different types of forecasts, supporting different types of decisions, on different levels of aggregation, all within a hierarchy. As such evaluating the accuracy of hierarchical forecasts requires considering accuracy measures on multiple dimensions, with forecasts servicing multiple objectives, something so far missing from the literature. Although many of the topics we discuss are independently relevant to the general forecast evaluation discourse, the hierarchical forecast context means that many of the dimensions of the evaluation setup may change across time series, requiring a framework that permits a holistic evaluation. To this end we propose a multi-objective forecast evaluation scheme.

Hierarchical forecasting has evolved over the decades to include different types. Depending on the nature of the time series included, a hierarchy can be cross-sectional, temporal, or cross-temporal. Cross-sectional refers to hierarchies that include different time series of the same sampling frequency across various demarcations, such as product categories, geographical regions, variable components, etc. (Athanasopoulos et al., 2009, 2019; Wickramasuriya et al., 2019). Temporal corresponds to hierarchies where series that measure the same object are aggregated at different frequencies (Athanasopoulos et al., 2017; Nystrup et al., 2019; Jeon et al., 2019). Cross-temporal joins both in a common structure, containing time series of various demarcations and different sampling frequencies (Kourentzes and Athanasopoulos, 2019; Di Fonzo and Girolimetto, 2020).

Another distinction is related to whether there is a unique mapping from the bottom-level of the hierarchy to the aggregate total or multiple alternative mappings, the latter case referred to as grouped series (Wickramasuriya et al., 2019). In fact, most cases belong to the second category, where aggregating factors are both nested and crossed (Hyndman and Athanasopoulos, 2021, Chapter 11). For instance one could aggregate bottom-level series across product categories but also market segments to form multiple mappings to the top-level. Cross-temporal structures are always grouped. Often we do not need to consider all alternative mappings and restrict our analysis to those that are relevant to the supported decisions.

A key characteristic of hierarchical forecasting is the requirement for coherent forecasts, where lower level forecasts must add up to levels above. For example, in a retailing scenario, the sales forecasts at product/store level must add up to sales at store level. Coherence in hierarchical forecasts was historically achieved by aggregating and/or disaggregating forecasts of a single level of the hierarchy to the rest of the structure. Over the last few years the dominant methodology has become one of forecast reconciliation. In this setting an initial set of forecasts are generated for each series in the hierarchical structure, without imposing any aggregation or coherence constraints, which are referred to as base forecasts. These are then reconciled so that they then become coherent (Hyndman and Athanasopoulos, 2021, Chapter 11). This is also connected to the defining objective of hierarchical forecasting: provide forecasts at different levels of the hierarchy that can support aligned decisions (Kourentzes and Athanasopoulos, 2019; Ord et al., 2017, Chapter 10). We note that coherency may introduce minimal differences in terms of accuracy, yet it can change the forecasts qualitatively. This aspect may be better captured by the supported decisions, rather than by error metrics.

When we consider the different levels of a hierarchy, some are tightly connected with supported decisions, but many may act as a statistical device to enrich the resulting forecasts, through the concept of forecast reconciliation. Note that forecast reconciliation implies some form of forecast combination. This is apparent if we consider purely cross-sectional or temporal hierarchies. Suppose that at the lowest level we record daily sales of a specific ice cream at store level. Forecasts at this level are important for supporting inventory management and shelf filling decisions, at store level. Aggregating, we can produce daily forecasts for total sales of that specific product across stores in a region, supported by a regional depot. These forecasts can support replenishment and inventory decisions.

Aggregating further can be beneficial for elucidating additional information from the behaviour of consumers, across products and/or regions but forecasts at these levels are not typically required at a daily frequency and hence are not directly connected to decision making. However, forecasts at higher or other levels of aggregation, although not directly useful in supporting decision, can be beneficial in a statistical sense. More specifically, as we impose coherency, including such forecasts in the process provides additional information through the process of forecast reconciliation, which has proven to also lead to forecast accuracy gains (Panagiotelis et al., 2021). Similarly, temporal hierarchies have been shown to be very beneficial in terms of accuracy, but many levels are disconnected from decision making (Kourentzes et al., 2014; Kourentzes and Athanasopoulos, 2021; Kourentzes et al., 2021).

This has been one of the motivating arguments for cross-temporal hierarchies that attempt to pair the aggregation across scale and time (Kourentzes and Athanasopoulos, 2019). We present a simple example in Figure 1 by considering two hierarchies, a cross-sectional and a temporal, and showing the possible time series combinations in a matrix. Of course the dimensions of the matrix would be much greater for more realistic scenarios such as the one explored by Kourentzes and Athanasopoulos (2019). Arguably, the shaded cells represent decision relevant levels, while the white cells represent levels that serve as statistical devices. Naturally, different applications will correspond to different decision relevant nodes. Hierarchical forecasting is motivated by mirroring hierarchical structures within organisations to support and monitor decisions. Across a hierarchy there may be multiple different decisions with different performance criteria. This suggests two important issues in evaluating hierarchical forecasts: the connection to the decisions, and the different objectives that may be present across the hierarchical levels.

Such characteristics must be reflected in the evaluation of hierarchical forecasts. We suggest that the modeller should choose the appropriate error metrics by considering: the scale of the time series, any potential sparsity, the forecast horizon and should evaluate on multiple forecast windows. We discuss these in Section 2, where we summarise the relevant evaluation discourse. In Section 3 we introduce the multi-objective evaluation scheme. This is followed by our practical recommendations for setting up empirical evaluations for hierarchical forecasting in Section 4, albeit attempting to do so without a particular application in mind.

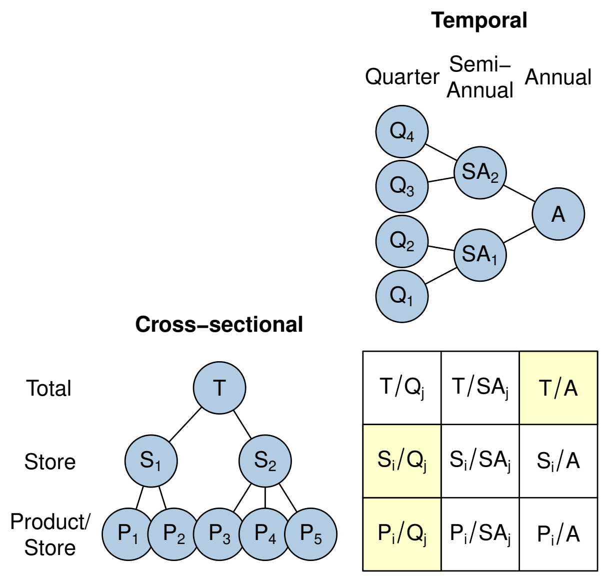

Figure 1: An example of cross-sectional and temporal hierarchies combined to provide all possible scale/time options. The shaded cells represent decision making relevant levels.

2. Error metrics

In the selection of error metrics two complementary sides need to be considered: the evaluation of point forecasts and interval forecasts. We refer to point forecasts as predicted estimates of the central tendency. Measuring the impact of improving forecasts on the decision variable is often a complex task, and improving the accuracy of the point forecasts is seen as a good proxy (Ord et al., 2017, Chapter 12). The accuracy of interval forecasts however, is generally seen as a better proxy for a large number of decisions, such as inventory management. Figure 2 illustrates the point, where we order metrics according to how well they connect to supporting decision making.

Ideally we would like to evaluate using the decision metric directly. But as this is rarely available, we revert to interval or point forecast metrics, with an increasingly weaker connection to the decision. Naturally, there are cases where the evaluation on these different levels provides equivalent results, and therefore we can rely on the simpler one. Irrespectively, the modeller has to account for factors such as scale, sparsity and decision relevant horizon.

Point forecast metric $\leqslant$ Interval forecast metric $\leqslant$ Decision metric

Figure 2: The further a metric is from the decision, usually the weaker the connection becomes between the outcome of the forecast and the decision. In ideal circumstances equality holds.

At this point, it is useful to briefly consider the connection between the supported decisions, the loss function used in the construction of the forecasts, and the evaluation metrics used to assess them. Although in theory these three could match, in practice this is typically not the case. Forecasts often have multiple users and stakeholders, irrespective of any technical difficulties in calculating any decision metrics. In an organisational setting, a single forecast can be the basis of different decisions with different objectives, each one of which may be evaluated on different aspects. Kourentzes et al. (2020) explore this further for the case of inventory management and demonstrate substantial benefits if the loss function matches the decision metric. However, at the same time, they highlight multiple challenges in doing so. These result in simplifying assumptions yet again increasing the inevitable disconnect between metrics and the multiple dimensions considered in decision making. In the literature there is limited research on how the different metrics may diverge, with examples supporting both the inequality and the equality (Sanders and Graman, 2009; Kourentzes et al., 2019, 2020). Hence, not only these three may not match, but that it is essential to consider multiple metrics as proxies for the needs of the different users.

2.1. Scale

In hierarchical forecasting time series will have by definition different scales and units. This has to be reflected in any error metrics employed. There are various approaches to attain scale independent metrics (Hyndman and Koehler, 2006; Davydenko and Fildes, 2013). Using percentage errors, of various forms, is a popular approach. Using percentage errors, of various forms, is a popular approach. However, percentage errors do have significant deficiencies and great caution needs to be exerted if one decides to consider these. Such metrics are asymmetric and can have computational issues when the actuals are zero or close to zero, something that can often occur at the lower levels of hierarchical time series (Kourentzes and Athanasopoulos, 2021), but not exclusively there, as zero observations can appear at higher levels too.

The second approach is to scale errors. This can be achieved by dividing the forecast errors by the standard deviation of the time series of interest (e.g., Trapero et al., 2013). An alternative popular approach is to scale by some reference error distribution, as is the case with scaled error measures such as the Mean Absolute Scaled Error (MASE, Hyndman and Koehler, 2006) or the Root Mean Squared Scaled Error (RMSSE, recently implemented in the M5 forecasting competition, see Hyndman and Athanasopoulos, 2021, Chapter 5). Hence, our attention is drawn to two aspects: the choice of the dispersion measure (the numerator of the scaled error measure) and the reference distribution (the denominator of the scaled error measure). The dispersion measure should match the loss function of the hierarchical forecasts. For instance, if a quadratic loss is used, then the squared deviation should be employed. For absolute loss the absolute deviation would be appropriate.

In terms of the reference distribution one can use the forecast errors of a reference forecast. In a hierarchical setting a natural choice would be the base incoherent forecasts, used to generate reconciled ones, if these are available. If these are not available, the dispersion of the target time series is a reasonable alternative. However, in this case one needs to consider the nature of the time series. For example, if the series is stationary the dispersion would be appropriate. However, if the series is non-stationary the dispersion of the differenced series is more appropriate, and seasonality would suggest seasonal differencing, and so on. MASE (or RMSSE) consider these, and therefore use as a scaling factor the absolute deviation of appropriately differenced series (Hyndman and Athanasopoulos, 2021, Chapter 5, use the naïve and the seasonal naïve as the scaling factors respectively for non-seasonal and seasonal time series). However, unless the analyst performs some rudimentary analysis to ensure that the dispersion is calculated on stationary data, using such scaling factors for all time series in the hierarchy can potentially be statistically improper. However, these have been shown to be reasonable and also stable in practice.

Note that an in-between percentage errors and scaled errors is to divide standard scale dependent errors by the mean of the time series (Kolassa and Schütz, 2007). Although the mean carries both the scale and units, so it is fit for purpose, it is only meaningful for stationary time series.

A third option is to use relative errors, such as the Geometric Mean Relative Absolute Error (Armstrong and Collopy, 1992; Fildes, 1992) that produces relative errors per period, or the Average Relative Mean Absolute Error (Davydenko and Fildes, 2013) that constructs the relative errors from other summary error metrics like the Mean Absolute Error. Quadratic versions of the metric, based on the Root Mean Squared Error, have been used in hierarchical forecasting as well (e.g., see Kourentzes and Athanasopoulos, 2019).

As the errors are relative, both scale and units become irrelevant. Producing the relative metric over other summary error metrics, instead of per period, makes them more widely applicable, especially when time series observations may contain zeros, which is probable at the lower levels of a hierarchy. Relative metrics have the further advantage that the forecast horizon can be considered in both the evaluated and benchmark forecasts. Finally, they are very easy to interpret, in contrast to scaled errors. However, they require the existence of a benchmark forecast. Again, given the hierarchical context, base incoherent forecasts can serve as a good evaluation benchmark, although one should consider the objective of the evaluation that may dictate use of a set of coherent forecasts as a preferable benchmark.

We note that all scale independent metrics suffer from various limitations. For example, scaled errors can be hard to interpret, while relative errors do not provide an absolute reference for the magnitude of the errors. This may often suggest that multiple metrics may be necessary, particularly across different levels of the hierarchy. For example, in constructing the denominator of scaled errors at disaggregate levels of a temporal hierarchy the seasonal naïve may be more apt, while at aggregate levels, where seasonality is filtered, the naïve may be best. We address this issue in Section 3.

2.2. Sparsity

Depending on the context, the more disaggregate levels of the hierarchy can exhibit multiple periods of zero observations. We call time series with multiple periods of zero valued observations as sparse. Intermittent demand time series are a sub-case of sparse, as there is variability in the timing of the occurrence of zeros, which may not be the case for sparse series where these observations can have canonical timing (e.g., solar irradiance). Nonetheless, both introduce a series of complications for the setup of the evaluation, due to the zero valued observations. Intermittency can introduce further complications due to the nature of the forecasting methods that are commonly used (Kourentzes, 2014).

On top of the previous considerations for the selection of error metrics, we need to consider the loss function. Gneiting (2011a) discusses this further, pointing out that the selection of absolute or quadratic errors is associated with the different parts of the predictive distribution. Absolute errors are appropriate if we are interested in the median of the predictive distribution, which for sparse time series may end up being zero, depending on the degree of sparsity. In that case the evaluation will suggest that a zero forecast is best, while this may have no practical usefulness (Kolassa, 2016).

The analyst needs to consider the limitations of absolute errors and also the potential computational problems that arise from zero valued observations, for instance for percentage metrics. The use of cumulative errors is a potential remedy, where the accuracy is tracked over a number of periods. In some settings, such as inventory management, this is natural as it connects with the demand over the lead time. Note that cumulative errors may not fully resolve the computational challenges of sparsity, as this depends on the number of zeros and the window of the cumulation.

2.3. Forecast horizon

We do not forecast for the sake of forecasting, but to support organisational decisions that require insights about the future. Different decisions are associated with different planning horizons, and require different types of forecasts. Given a forecast horizon h, we distinguish three cases (i) produce an accurate forecast for specific period $t + h$; (ii) produce accurate forecasts for all periods from $t + 1$ to $t + h$; and (iii) produce an accurate cumulative forecast for periods $t + 1$ to $t + h$, where now we are interested in the total, rather than the forecast per period. Different decisions will require different types of forecasts. For example, capacity or scheduling decisions are typically of type (i) or (ii), while inventory decisions typically fall under (iii).

In the hierarchical context this will often necessitate the use of different horizons across the various levels of aggregation. Given the reasonable expectation that longer-term forecasts are more difficult, arguably the outcomes of these measurements are not comparable (Tashman, 2000) and may not exhibit strong correlations, substantially complicating the evaluation. We expand on this in the next section, with an example that uses temporal hierarchies.

2.4. The importance of multiple evaluation windows

It is well accepted that reliable evaluation of forecasts should be based on multiple measurements of the forecast errors, often achieved by the use of a rolling origin evaluation scheme (Tashman, 2000; Ord et al., 2017, Chapter 3). However, when a large number of time series, of similar nature, is available, one can evaluate forecast accuracy across the time series, instead of across forecast origins. This has been a principal argument in the design of many influential forecasting competitions, such as the M-competitions (Makridakis and Hibon, 2000; Makridakis et al., 2019).

With hierarchical structures this cannot be the case, even though they comprise a large number of time series. Every level of aggregation reflects a different part of the problem space, supporting a different decision, potentially associated with a different forecast horizon and evaluation metric. Therefore, in the hierarchical context one cannot make use of the sheer volume of time series to avoid measuring the performance from multiple forecast origins. Furthermore, for any one forecast origin the forecast errors across the series within a hierarchy are highly likely to be correlated, further reducing the effective sample size in such an evaluation.

3. Multiple evaluation objectives

Depending how the forecast is translated into a decision we may be interested in accurate point or interval forecasts (Ord et al., 2017, Chapter 12). Note that accurate point forecasts do not necessarily mean superior interval forecasts and vice-versa (Kourentzes et al., 2020). Forecast bias is another dimension of interest, again depending on the nature of the decision (Sanders and Graman, 2009).

In the hierarchical context, different nodes in the hierarchy will be connected with different decisions and require different types of forecasts. This must be reflected in the evaluation setup. Collecting all metrics together results in a multidimensional evaluation metric. We argue that we should not simply average or otherwise simplistically combine the different dimensions into a single composite metric. Each forecast is associated with a different decision that has its own importance and impact on the organisation. Effectively we are dealing with a multi-objective loss function. In this setting we need to identify forecasts that form the Pareto frontier of competing forecasts across the different objectives. Suppose we are considering different forecasts f_i, $i = 1, \ldots, k$, across different performance metrics $p_j(f_i)$, $j = 1, \ldots, l$, where a lower p_j denotes a better performance. The p_j could be a metric on point or interval forecasts, or decisions. Then a forecast f_m belongs to the Pareto frontier when $p_j(f_m) < p_j(f_i)$ for $i \neq m$ and at least one $j \in 1, \ldots, l$. Figure 3 provides examples for three forecasts, across two metrics. These two metrics can be summary statistics for two different levels of interest in a hierarchy. For example, the first may be the average over the lead time accuracy, across products and stores, to support inventory management, while the second may be the average per period accuracy at the brand level, to support marketing expenditure decisions. In the same logic, the three forecasts would be outcomes of alternative forecasting approaches or methodologies across the hierarchy. The hierarchy itself maybe much larger, but for this example we consider the other levels to be of no or limited importance. Forecasts that are on the Pareto frontier are highlighted with filled markers. In panel (i) Forecast 1 dominates the competing forecasts across both metrics p_1 and p_2 and would be the apparent choice. In panel (ii) Forecast 1 dominates on p_1, while Forecast 2 dominates on p_2. Although both are superior to Forecast 3, it is not possible to choose from the two, unless we can assign weights to p_1 and p_2. These importance weights are related to the decision supported by the forecasts and are dependent on the application setting. We revisit this in the multidimensional case example in Section 3.1.

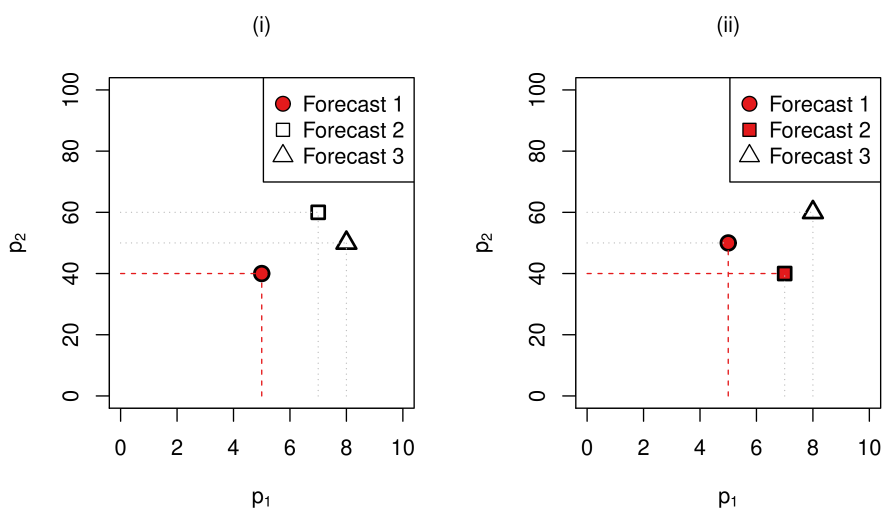

Figure 3: An illustration of the multi-objective evaluation for two decisions/levels of a hierarchy. Forecasts that form the Pareto frontier are filled. In panel (i) Forecast 1 dominates the competing forecasts. In panel (ii) both Forecasts 1 and 2 belong to the Pareto frontier, as each is better in a different objective.

There are various strategies to identify a single best forecast by scalarising the multiple objectives into a single one, originating from the multi-objective optimisation literature (Keeney and Raiffa, 1976; Hwang and Masud, 2012; Abbas, 2018). The user can provide importance weights, examples of preferred solutions, or other ways to guide the combination of the different objectives. Naturally, it is not possible to provide a general scalarisation for all applications of hierarchical forecasting. As an example, consider cross-temporal forecasts for a retailer. At the lower levels of the hierarchy the forecasts support inventory decisions. At the product level and longer forecast horizons pricing decisions are made. While at a more aggregate level forecasts are required for supporting budgeting decisions. When a forecasting approach does not provide a globally dominant result, the analyst has to weight the importance of improving (or not) the different forecasts at the different levels of the hierarchy.

In this context, we can understand better what the average performance (for a common horizon) across all levels of the hierarchy means, which has been often provided in the literature. Drawing on the geometrical interpretation by Panagiotelis et al. (2021), that is a statement about the quality of the coherency of the forecasts. Different hierarchical forecasts will provide coherent results, yet some will have lower total error variance. However, this may mean little for the decision relevant levels, and hence when an application context is available providing an average metric across all levels is of little use as a proxy to the decision variables.

3.1. An example evaluation with multiple targets

We use the application by Athanasopoulos et al. (2017) to demonstrate the discussed evaluation approach. The authors in Section 7, look at predicting different components of accident and emergency services demand in England, which are important to management. We use the same 13 weekly-sampled time series that span from 7 November 2010 to 7 June 2015. We withhold the last year as a test set. We are interested in supporting a number of decisions as presented in Table 1.

Table 1: Forecast horizons corresponding to important managerial decisions for accidents and emergency services in England.

Managerial decision	Frequency	Forecast horizon
Budget	Yearly	1-step ahead
Supply of material	Quarterly	1-step ahead
Staffing needs	4-Weekly	1-step ahead
Scheduling permanent and temporary staff	Weekly	1- to 4-steps ahead
Staff timetabling	Weekly	1-step ahead

We model the time series at all frequencies independently and generate (Base) forecasts employing exponential smoothing (ETS) and ARIMA models. The Base forecasts are then reconciled using temporal hierarchies (with Variance and Structural scaling). This results in six alternative sets of forecasts.

We use the ETS–Base forecasts as the benchmark, from which we calculate the RelRMSE for the different targets, which is summarised across time series using the geometric mean, resulting in the AvgRelRMSE. The forecasts are generated using the `es` and `auto.arima` functions from the smooth (Svetunkov, 2019) and forecast (Hyndman et al., 2019) packages for R respectively (R Core Team, 2021).

Table 2 summarises the errors across all 13 time series. Each column corresponds to a different forecast target, and the best forecast for each is highlighted in boldface. We can observe that no forecast is best across all targets, which match different levels in the temporal hierarchy, yet most

Table 2: AvgRelRMSE for accident and emergency services in England

Forecast	Target				
	Weekly t+1	Weekly t+1 to t+4	4-Weekly t+1	Quarterly t+1	Yearly t+1
ETS - Base	1.000	1.000	1.000	1.000	1.000
ETS - Variance	**0.777**	1.047	0.736	0.803	0.328
ETS - Structural	1.887	1.129	0.843	**0.537**	**0.323**
ARIMA - Base	2.848	1.086	1.045	1.304	0.987
ARIMA - Variance	1.030	**0.865**	**0.720**	0.719	0.440
ARIMA - Structural	3.207	1.182	1.211	0.959	0.477

outperform the ETS–Base benchmark. To simplify the discussion, we first focus on two targets: the Weekly and 4-Weekly 1-step ahead forecasts. Their errors are plotted in Figure 4. The light-shaded areas partially dominate the benchmark, i.e., outperform it in one of the targets. The dark-shaded area includes forecasts that are fully dominated by the benchmark, while the white area includes the opposite, i.e., forecasts that dominate the benchmark. Hence, we observe that the ETS–Variance fully dominates the benchmark, but only partially dominates ARIMA–Variance which has a marginally lower 4-Weekly error. ETS–Structural outperforms the benchmark on the 4-Weekly forecast, but is substantially worse on the Weekly one. It is not possible to identify a dominant solution, as none of the forecasts is best across both targets, although ETS–Variance seems to be a better option, depending on the weighting between the two targets.

Assuming equally weighted importance for the different targets, we can linearise the multiple objectives by calculating the unweighted average, and

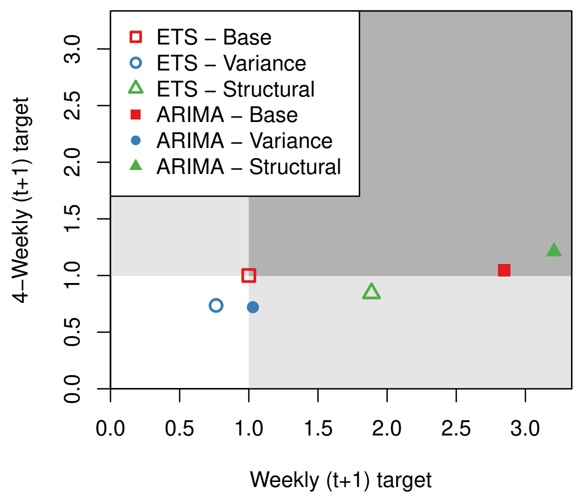

Figure 4: Weekly and 4-weekly 1-step ahead forecast errors. Forecasts in white area dominate the benchmark (ETS–Base) while forecasts in the dark grey area are dominated by the benchmark. Forecasts in the light grey areas partially dominate the benchmark.

also calculate whether there is a forecast that dominates all others or not. This is provided in Table 3. We can observe that none of the forecasts fully dominates all others, i.e., there is at least one forecast target for which it does not rank first. We can also see that the Base forecasts are partially dominated by the hierarchical forecasts. That is, there is no single forecast target that the Base forecasts rank first. Similarly the ARIMA–Structural is dominated by other hierarchical forecasts. This informs us that none of the ETS–Base, ARIMA–Base and ARIMA–Structural need to be considered further. This is also reflected in the (unweighted) average provided in the last column of the table, which ranks ETS–Variance first.

It is apparent that choosing a single method is not trivial, as we need to

Table 3: Dominance between forecasts

Forecast	Dominance		Average
	Full	Partial	
ETS–Base	✗	✗	1.000
ETS–Variance	✗	✓	0.738
ETS–Structural	✗	✓	0.944
ARIMA–Base	✗	✗	1.454
ARIMA–Variance	✗	✓	0.755
ARIMA–Structural	✗	✗	1.407

take into consideration the application for which the forecasts are made. If we weigh the forecast targets differently, then the results in Table 3 changes accordingly. For instance, if we weigh the budget forecast very heavily (yearly 1-step ahead) then ETS–Structural can potentially dominate all other alternatives.

Finally, note that we did not provide the errors for all levels and horizons of the temporal hierarchy (e.g., 2-weekly, 3-weekly, half-yearly, etc.), as many are not relevant to the decision makers. Nor did we include these in the calculation of the average in Table 3. In the absence of a weighting strategy, so that dominating forecasts can be calculated, providing results as in Table 2 remains the most informative, where all decision relevant forecasts are evaluated.

4. A practical recommendation

Often forecasting research is not closely tied to a specific application, yet we need to devise reliable and robust evaluation schemes. Research in hierarchical forecasting is no different. In this section we attempt to use the above as a guideline to recommend an evaluation scheme.

First, we need to consider whether sparsity is an issue in the hierarchy. If that is the case we propose using cumulative errors, as this can limit computational problems. For levels of the hierarchy with no sparsity this is not necessary, although this may still be relevant if the forecasts inform decisions that rely on a cumulative view such as inventory management. Nonetheless, the analyst needs to carefully consider the degree of remaining sparsity and the applicability of various metrics (e.g., Kourentzes, 2014, directly evaluates on inventory performance). In terms of error metrics, depending on the context, we propose considering at minimum (i) Mean Error, to measure forecast bias; (ii) Root Mean Squared Error, to assess the variance of forecast errors; and (iii) Pinball Loss for a number of target quantiles (Gneiting, 2011b).

All these error metrics are scale dependent and therefore using either relative or scaled versions of these is imperative. To calculate relative errors, as in Davydenko and Fildes (2013) we recommend using either a set of naïve forecasts (which by default are coherent) or base incoherent forecasts, if these are available. Note that this choice has implications on what is being evaluated, where the forecast performance or the impact of forecast reconciliation may be of interest and therefore guide the selection of the reference forecasts. In some cases replacing the Pinball Loss with the Mean Interval Score (Gneiting and Raftery, 2007) might be preferable, as it considers both the upper and lower quantiles. Kourentzes and Athanasopoulos (2021) provide examples of relative metrics for accuracy, bias, and intervals. For scaled errors, we recommend MASE and RMSSE. These are particularly useful when reasonable benchmark forecasts are not available for the calculation of the relative metrics.

Finally, in contrast to common practice, we believe that there is limited benefit in an empirical evaluation setting, to report average accuracy measures across all levels of the hierarchy (although we have been culprits of doing so ourselves). It is highly improbable that this reflects a realistic situation and therefore it is paramount that the modeller attempts to establish a strong connection between the objectives of the forecasts and the evaluation. A more realistic alternative is presented in Panagiotelis et al. (2021) where forecast errors across a tourism geographical hierarchy are weighted by the average tourism expenditure in each region. Another plausible alternative may be to focus at the most disaggregate level or sample some important levels, tied to appropriate forecast horizons (e.g., Kourentzes et al., 2021). In many cases focusing at the most disaggregate level may be more meaningful, and arguably improving bottom-level forecasts implicitly supports the whole hierarchy. Note that we do not suggest that looking solely at the bottom-level forecasts is sufficient, rather than that this may be a compromise if the forecast evaluation is very disjoint from a practical application.

5. Conclusions

Research in hierarchical forecasting has exploded over the last decade, matching an increasing interest in the field from practice. This makes rigorous and valid evaluation of hierarchical forecasts crucial. Given the wide range of applications that use hierarchical forecasting techniques, it is impossible to provide a framework to accommodate all settings. Instead, we discuss important considerations and attempt to provide a set of practical recommendations when a specific application is not available.

One of our contributions is to connect the evaluation of hierarchical forecasts with multi-objective optimisation. We believe that the use of multi-objective optimisation tools and learnings can be beneficial for predictive modelling evaluation more generally and we call for more interaction between the disciplines.

In our attempt to highlight the importance of the decision metrics in evaluating forecasts, we recognise that there is limited research in the literature addressing this, with only a handful of contributions looking at the correlation between the two (e.g., Sanders and Graman, 2009; Fildes and Kingsman, 2011) or specifying forecasting models to better match the decision context (e.g., Kourentzes et al., 2019, 2020). Given the relevance of this to hierarchical forecasting, we call for more research in the area.

References

Abbas, A. E., 2018. Foundations of Multiattribute Utility. Cambridge University Press.

Armstrong, J. S., Collopy, F., 1992. Error measures for generalizing about forecasting methods: Empirical comparisons. International Journal of Forecasting 8 (1), 69–80.

Athanasopoulos, G., Ahmed, R. A., Hyndman, R. J., 2009. Hierarchical forecasts for Australian domestic tourism. International Journal of Forecasting 25 (1), 146–166.

Athanasopoulos, G., Gamakumara, P., Panagiotelis, A., Hyndman, R. J., Affan, M., 2019. Hierarchical Forecasting. In: Peter Fuleky (Ed.), Macroeconomic Forecasting in the Era of Big Data, 1st Edition. Springer, Honolulu, Ch. 21, pp. 703–733.

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

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.

Di Fonzo, T., Girolimetto, D., 2020. Cross-temporal forecast reconciliation: Optimal combination method and heuristic alternatives.
URL http://arxiv.org/abs/2006.08570

Fildes, R., 1992. The evaluation of extrapolative forecasting methods. International Journal of Forecasting 8 (1), 81–98.

Fildes, R., Kingsman, B., 2011. Incorporating demand uncertainty and forecast error in supply chain planning models. Journal of the Operational Research Society 62 (3), 483–500.

Gneiting, T., 2011a. Making and evaluating point forecasts. Journal of the American Statistical Association 106 (494), 746–762.

Gneiting, T., 2011b. Quantiles as optimal point forecasts. International Journal of forecasting 27 (2), 197–207.

Gneiting, T., Raftery, A. E., 2007. Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association 102 (477), 359–378.

Hwang, C.-L., Masud, A. S. M., 2012. Multiple objective decision making-methods and applications: a state-of-the-art survey, 1st Edition. Vol. 164 of Lecture Notes in Economics and Mathematical Systems. Springer Science & Business Media.

Hyndman, R., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., Petropoulos, F., Razbash, S., Wang, E., Yasmeen, F., 2019. forecast: Forecasting functions for time series and linear models. R package version 8.9.
URL http://pkg.robjhyndman.com/forecast

Hyndman, R. J., Athanasopoulos, G., 2021. Forecasting: Principles and Practice, 3rd Edition. OTexts, Melbourne, Australia.
URL http://otexts.com/fpp3/

Hyndman, R. J., Koehler, A. B., 2006. Another look at measures of forecast accuracy. International Journal of Forecasting 22 (4), 679–688.

Jeon, J., Panagiotelis, A., Petropoulos, F., 2019. Probabilistic forecast reconciliation with applications to wind power and electric load. European Journal of Operational Research 279 (2), 364–379.

Keeney, R., Raiffa, H., 1976. Decisions with Multiple Objectives. New York: Wiley.

Kolassa, S., 2016. Evaluating predictive count data distributions in retail sales forecasting. International Journal of Forecasting 32 (3), 788–803.

Kolassa, S., Schütz, W., 2007. Advantages of the MAD/MEAN ratio over the MAPE. Foresight: The International Journal of Applied Forecasting 6, 40–43.

Kourentzes, N., 2014. On intermittent demand model optimisation and selection. International Journal of Production Economics 156, 180–190.

Kourentzes, N., Athanasopoulos, G., 2019. Cross-temporal coherent forecasts for Australian tourism. Annals of Tourism Research 75, 393–409.

Kourentzes, N., Athanasopoulos, G., 2021. Elucidate structure in intermittent demand series. European Journal of Operational Research 288 (1), 141–152.

Kourentzes, N., Li, D., Strauss, A. K., 2019. Unconstraining methods for revenue management systems under small demand. Journal of Revenue and Pricing Management 18, 27–41.

Kourentzes, N., Petropoulos, F., Trapero, J. R., 2014. Improving forecasting by estimating time series structural components across multiple frequencies. International Journal of Forecasting 30 (2), 291–302.

Kourentzes, N., Saayman, A., Jean-Pierre, P., Provenzano, D., Sahli, M., Seetaram, N., Volo, S., 2021. Visitor arrivals forecasts amid covid-19: A perspective from the africa team. Annals of Tourism Research 88, 103197.

Kourentzes, N., Trapero, J. R., Barrow, D. K., 2020. Optimising forecasting models for inventory planning. International Journal of Production Economics, 107597.

Makridakis, S., Hibon, M., 2000. The M3-competition: results, conclusions and implications. International Journal of Forecasting 16 (4), 451–476.

Makridakis, S., Spiliotis, E., Assimakopoulos, V., 2019. The M4 competition: 100,000 time series and 61 forecasting methods. International Journal of Forecasting.

Nystrup, P., Lindström, E., Pinson, P., Madsen, H., 08 2019. Temporal hierarchies with autocorrelation for load forecasting. European Journal of Operational Research 280.

Ord, J. K., Fildes, R., Kourentzes, N., 2017. Principles of Business Forecasting, 2nd Edition. Wessex Press Publishing Co.

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

R Core Team, 2021. R: A Language and Environment for Statistical Computing, version 4.1.0. R Foundation for Statistical Computing, Vienna, Austria.
URL https://www.R-project.org/

Sanders, N. R., Graman, G. A., 2009. Quantifying costs of forecast errors: A case study of the warehouse environment. Omega 37 (1), 116–125.

Svetunkov, I., 2019. smooth: Forecasting Using State Space Models. R package version 2.5.4.
URL https://CRAN.R-project.org/package=smooth

Tashman, L. J., 2000. Out-of-sample tests of forecasting accuracy: an analysis and review. International Journal of Forecasting 16 (4), 437–450.

Trapero, J. R., Pedregal, D. J., Fildes, R., Kourentzes, N., 2013. Analysis of judgmental adjustments in the presence of promotions. International Journal of Forecasting 29 (2), 234–243.

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