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

January 2020

Working Paper 02/20

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 note 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 statistical structure of the hierarchy and the application context. We discuss four relevant dimensions for researchers and analysts: the scale and units of time series, the issue of sparsity, the decision context and the importance of multiple evaluation windows. We conclude with a series of practical recommendations.

Keywords: Aggregation, coherence, hierarchical time series, reconciliation.

1. Introduction

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

*Correspondance: G Athanasopoulos, Department of Econometrics and Business Statistics, Monash University, Australia.

Email address: George.Athanasopoulos@monash.edu (George Athanasopoulos)

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), while temporal corresponds to hierarchies where series that measure the same object are sampled 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, 2019a).

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, 2018). 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, 2018). This is also connected to the defining objective of hierarchical forecasting: provide forecasts at different levels of the hierarchy that can support aligned decisions (Ord et al., 2017; Kourentzes and Athanasopoulos, 2019a).

When we consider the different levels in a hierarchy, some are tightly connected with supported decisions, but many act as a statistical device to enrich the resulting forecasts. 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. This decision is meaningful for inventory management and shelf filling at store level. Aggregating further, we can produce daily forecasts for total sales of that specific product across stores in a region, supported by a given depot. This forecast 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 that forecast is typically not needed at a daily frequency, making the remaining hierarchical forecasts useful in the statistical sense, but not for decision making. 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, 2019b).

This has been one of the motivating arguments for cross-temporal hierar-

chies that attempt to pair the aggregation across scale and time (Kourentzes and Athanasopoulos, 2019a). 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 (2019a). 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.

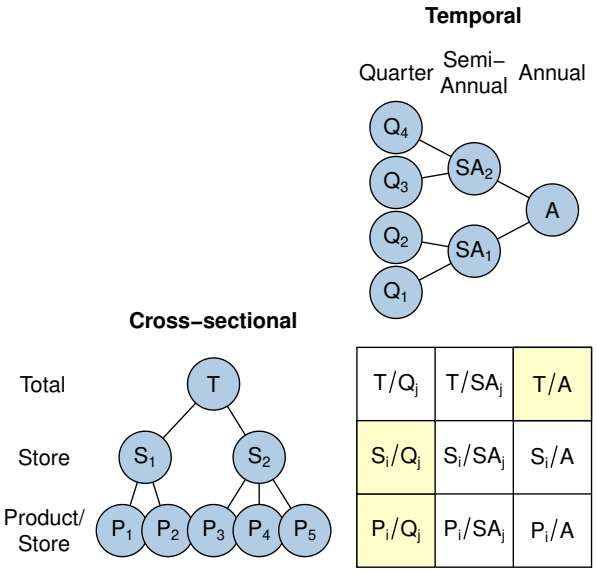

Figure 1: An example of cross-sectional and temporal hierarchies combined to provide all possible scale/time options. The shaded cells represent decision making relevant levels.

These characteristics must be reflected in the evaluation of hierarchical

forecasts. We argue that the modeller should choose the appropriate error metrics by considering three dimensions: the scale of the time series, any potential sparsity and the decision. We discuss these in Section 2, together with the importance of considering multiple evaluation windows. This is followed by our recommendation in Section 3 for an evaluation scheme, albeit attempting to do so without a particular application in mind.

2. Error metrics

In the selection of error metrics two complimentary sides need to be considered: the evaluation of point forecasts and quantiles. 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). The accuracy of quantile forecasts however, is generally seen as a better proxy, if not directly correlated, for a large number of decisions, such as inventory management. Figure 2 illustrates the point. Ideally we would like to evaluate using the decision metric directly. But as this is rarely available, we revert to quantile or point forecast metrics, with an increasingly weaker connection to the decision. Irrespectively, the modeller has to account for factors such as scale, sparsity and decision relevant horizon.

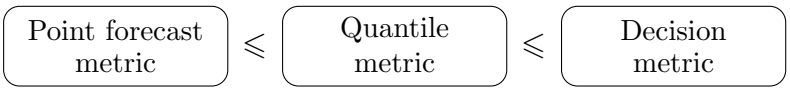

Figure 2: The further a metric is from the decision, the weaker the connection becomes. In ideal circumstances equality holds.

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 independence metrics (Hyndman and Koehler, 2006; Davydenko and Fildes, 2013). Using percentage errors, of various forms, is a popular approach. Yet we argue against it. Such metrics are asymmetric and can have computational problems 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, 2019b), but not exclusively there, as zero observations can appear at higher levels too.

The second approach is to normalise errors. For example dividing by the standard deviation of the time series of interest (this has been a popular approach in the behavioural forecasting research, Trapero et al., 2013), or by some reference error distribution, as is the case with the Mean Absolute Scaled Error (MASE, Hyndman and Koehler, 2006). We draw the attention of the modeller to two aspects: the choice of the dispersion measure and the reference distribution. The dispersion measure should match the loss function of the hierarchical forecasts. For instance, if a quadratic loss is used, then the standard deviation should be employed for normalising the errors. 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, which in this case could be the base incoherent forecasts that are used to produce the hierarchical ones, if they are available. Otherwise, the dispersion of the target time series is a reasonable choice. However, in that case one has to consider the nature of the time series. If it is station-

ary, the dispersion would be appropriate. However, if it is non-stationary the dispersion of the differenced series is more appropriate, and seasonality would suggest seasonal differencing, and so on. MASE considers this and uses for scaling the absolute deviation of the appropriately differenced series (Hyndman and Athanasopoulos, 2018, use the naïve and the seasonal naïve as the scaling factors respectively for non-seasonal and seasonal time series). We argue that unless the analyst performs some rudimentary analysis to ensure that the dispersion is calculated on stationary data, using the MASE scaling factor for all time series in the hierarchy will potentially be statistically improper, but practically reasonable.

Note that an in-between percentage errors and normalisation is to divide standard scale dependent errors with the mean of the time series (Kolassa et al., 2007). Although the mean carries both the scale and units, so it is fit for purpose, we advise against this, as the mean is meaningful only 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 construct 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 appeared in the literature as well (Kourentzes and Athanasopoulos, 2019a).

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 ob-

servations 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 normalised 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.

2.2. Sparsity

Depending on the context, the more disaggregate levels of the hierarchy can exhibit multiple periods of zero observations. We distinguish sparsity from intermittency, as the latter assumes uncertainty in the inter-demand periods as well, which may not be the case for sparse demand patterns. Intermittent demand is a sub-case of sparse, with increased uncertainty. 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 and Raftery (2007) discuss 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).

Therefore, if sparsity is an issue, the modeller is advised to consider this in the selection of the error measure.

2.3. Decision parameters

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, scheduling decisions are typically of type (i) or (ii), while inventory decisions typically fall under (iii).

Depending how the forecast is translated into a decision we may be interested in accurate point or quantile forecasts (Ord et al., 2017). Note that accurate point forecasts do not necessarily mean superior quantile predictions and vice-versa. 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 forecasting. 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. Effec-

tively 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 $p_j(f_i),\ j = 1, \ldots, l$. 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 average over the lead time accuracy across products and stores, to support inventory management, and the average per period accuracy at 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 2.3.1.

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 (Hwang and Masud, 2012). The user can provide

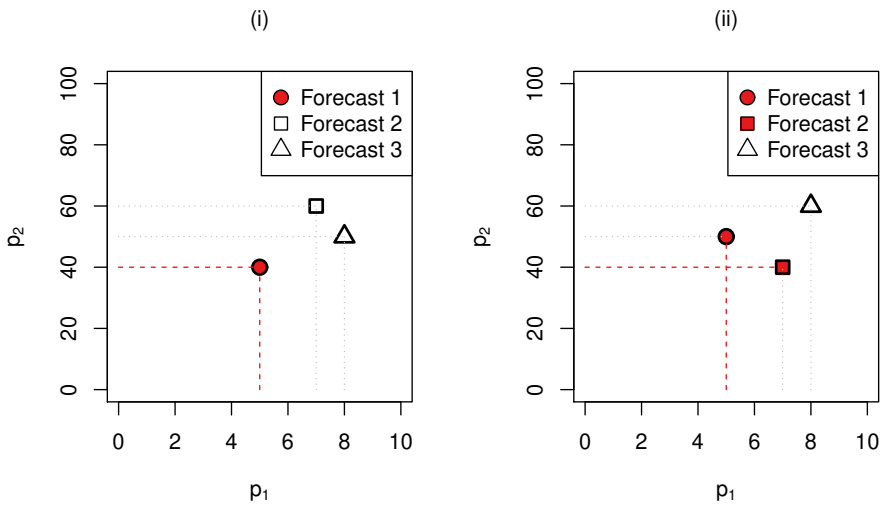

Figure 3: An illustration of the multi-objective evaluation for two decisions/levels of a hierarchy. Forecasts that form the Pareto frontier are filled. In panel (i) forecast 1 dominates the competing forecasts. In panel (ii) both forecasts 1 and 2 belong to the Pareto frontier, as each is better in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

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 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. (2019), 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 we argue that when an application context is available providing an average metric across all levels is of little use as a proxy to the decision variables.

2.3.1. 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 the accident and emergency services demand in England. 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	Quartely	1-step ahead
Staffing needs	4-Weekly	1-step ahead
Staff scheduling including temps	Weekly	1- to 4-steps ahead
Staff timetabling	Weekly	1-step ahead

We model the time series independently (Base) and using temporal hierar-

chies (with Variance and Structural scaling), employing exponential smoothing (ETS) and ARIMA forecasts. This results in six alternative sets of forecasts. We use the ETS–Base forecasts as the benchmark, from which we calculate the AvgRelRMSE for the different targets. 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, 2019).

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 - Variace	**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 - Variace	1.030	**0.865**	**0.720**	0.719	0.440
ARIMA - Structural	3.207	1.182	1.211	0.959	0.477

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 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.

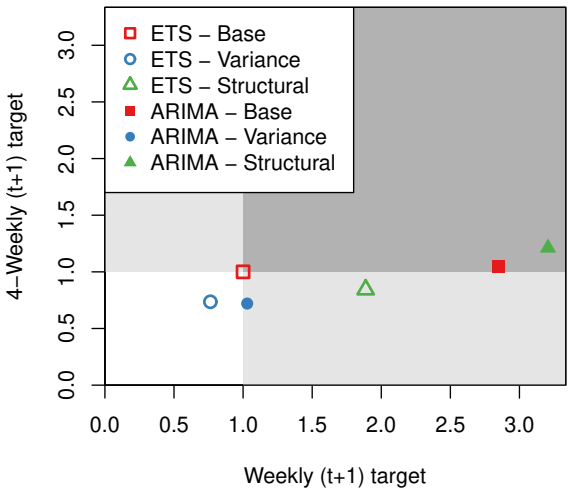

Figure 4: Weekly and 4-weekly 1-step ahead forecast errors. Forecasts in white area dominate the benchmark (ETS–Base) while forecasts in the dark grey area are dominated by the benchmark. Forecasts in the light grey areas partially dominate the benchmark.

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

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.

Table 3: Dominance between forecasts

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

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

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). 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).

However, 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.

3. 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 limits substantially any 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. In terms of error metrics we advise considering at minimum the (i) Mean Error, to measure forecast bias; (ii) Root Mean Squared Error, to assess the variance of forecast errors; and (iii) the Pinball Loss for a number of target quantiles (Gneiting, 2011). All these error metrics are scale dependent and therefore we recommend using relative versions of these, as in Davydenko and Fildes (2013), to either a set of naïve coherent forecasts or base incoherent forecasts if these are available. Particularly for the bias, to calculate the relative error we suggest using its absolute value, as in Kourentzes et al. (2019). 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. Furthermore, MASE is a useful alternative 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 mea-

sures across all levels of the hierarchy (although we have been corporates of doing so ourselves). It is very improbable that this reflects a realistic situation. Hence, it is paramount that the modeller attempts to establish a strong connection between the objectives of the forecasts and the evaluation. For example, one may want to focus at the most disaggregate level or sample some important levels, tied to appropriate forecast horizons. 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.

4. 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.

References

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.

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

Gneiting, T., 2011. 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. Vol. 164. 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., 2018. Forecasting: Principles and Practice, 2nd Edition. OTexts, Melbourne, Australia.
URL http://otexts.com/fpp2/

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.

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

Kolassa, S., Schütz, W., et al., 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., 2019a. Cross-temporal coherent forecasts for australian tourism. Annals of Tourism Research 75, 393–409.

Kourentzes, N., Athanasopoulos, G., 2019b. Elucidate structure in intermittent demand series. Monash University, Work Paper 27/19.

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., Trapero, J. R., Barrow, D. K., 2019. 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., Gamakumara, P., Athanasopoulos, G., Hyndman, R., et al., 2019. Forecast reconciliation: A geometric view with new insights on bias correction. Tech. rep., Monash University, Department of Econometrics and Business Statistics.

R Core Team, 2019. R: A Language and Environment for Statistical Computing. 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.