

Do we want coherent hierarchical forecasts, or minimal MAPEs or MAEs? (We won't get both!)

Keywords: Copulas, Density forecasts, Error measures, Evaluating forecasts, Forecasting education, Hierarchical forecasting, Loss function

1. Introduction

Hierarchical forecasting has become a research focus in recent years, driven on the one hand by advances in our understanding of optimal reconciliation (Athanasopoulos et al., 2009, 2015; Hyndman et al., 2011; Athanasopoulos et al., 2017; Wickramasuriya et al., 2019; Panagiotelis et al., 2021, in press), and on the other hand by the increasing availability of hierarchical data to be forecasted. For example, supply chain applications such as retail sales (as used in the recent M5 forecasting competition, Makridakis et al., 2022) routinely require hierarchical forecasts along multiple crossed hierarchies (Fildes et al., 2022; Syntetos et al., 2016): forecasts for retail store replenishment are based on forecasts on SKU \times store \times day granularity, whereas regional distribution center replenishment is driven by forecasts on SKU \times region \times week granularity, and promotional plans are made based on (explicit or implicit) forecasts on brand \times chain \times week granularity.

Against this backdrop, it becomes important to investigate whether there is anything specific about *evaluating* hierarchical forecasts. Can we use the standard forecast accuracy measurements (see any forecasting textbook, such as section 3.4 in Hyndman & Athanasopoulos, 2018) “out of the box”, or are there hidden pitfalls?

Athanasopoulos & Kourentzes (in press) argue that there indeed are. While their arguments are **mostly** correct and convincing, many of them are not specific to the *hierarchical* case but apply equally to situations where general dis-

parate time series are to be forecasted. We would like to draw additional attention here to one aspect: the only forecast accuracy measure that makes sense to evaluate coherent hierarchical point forecasts with is the (R)MSE, or more generally, monotonic functions of any weighted sum of *squared* errors (cf. Panagiotelis et al., 2021, especially Theorems 3.1 and 3.2). ~~whereas using the MAPE – and potentially the MAE, wMAPE or MASE, in contradiction to one of the recommendations given by Athanasopoulos & Kourentzes (in press) – is inherently nonsensical.~~ **In contrast, using the MAPE (which, to be clear, Athanasopoulos & Kourentzes, in press do *not* recommend) – and potentially the MAE, wMAPE or MASE, in contradiction to one of the recommendations given by Athanasopoulos & Kourentzes (in press) – is inherently nonsensical.** We investigate the issue in section 2, discuss examples in section 3, draw conclusions (and suggest ways forward) in section 4 and conclude in section 5.

2. The issue

The “best” point forecast depends on the error or accuracy measure: given a – possibly only implicit – predictive density that encodes our probabilistic belief about the outcome to be forecasted, different error measures will be minimized in expectation by different “one number summaries” of this density (Kolassa, 2020). Specifically, the point forecast minimizing the expected MAE or MASE will only be identical to the MSE-optimal point forecast if the predictive density is symmetric, and the MAPE-optimal forecast will be different from both.

Athanasopoulos & Kourentzes (in press) write that “[a] key characteristic of hierarchical forecasting is the requirement for coherent forecasts, where lower level forecasts must add up to levels above.” Indeed, the requirement of coherence seems to be accepted as self-evident in the hierarchical forecasting literature.

But does this requirement **always** make sense? Few people with any statistical knowledge would argue that *quantile* forecasts should be coherent. The 95% quantile forecast for total demand at multiple sites will almost always be smaller

than the sum of the *separate* 95% quantile forecasts for each site. And note that quantile forecasts seamlessly fit into the “different error measures elicit different one number summaries of predictive densities” argument above, because they minimize a very specific error measure: the pinball loss (Gneiting, 2011b). Thus, calculating quantile forecasts (and evaluating them using the pinball loss) and requiring these to be coherent would be nonsensical.

It turns out that the very same effect occurs for most other common error measures.

- The MAPE is minimized in expectation by the (-1) -median of a given predictive density (Eq. (4), pages 748 and 752 with $\beta = -1$ in Gneiting, 2011a), and the (-1) -median of the density of a sum is not equal to the sum of the (-1) -medians of the separate densities. Thus, MAPE-optimal point forecasts in a hierarchy are never coherent.
- The MAE is minimized in expectation by the median of the given predictive density (Hanley et al., 2001), and the median of the density of a sum is usually not equal to the sum of the medians of the separate densities, unless the separate densities are symmetric. Thus, MAE-optimal point forecasts are usually not coherent.
- The same applies to the MASE (Hyndman & Koehler, 2006) or the wMAPE (Kolassa & Schütz, 2007), which are just scaled MAEs.
- The *only* error measure whose minimizing point forecasts are coherent is the squared error, and monotonic functions of weighted sums of squared errors, because it is minimized in expectation by the expected value of the predictive density, and the expectation is additive.

To summarize, it makes as much sense to require coherence from forecasts that aim at minimizing MAPE as to require it from quantile forecasts. That is: none.

KPI	\hat{x}_{KPI}	\hat{y}_{KPI}	$\hat{x}_{\text{KPI}+}$ \hat{y}_{KPI}	$\widehat{x + y_{\text{KPI}}}$		
				$\rho_{XY} = 0$	$\rho_{XY} \approx 0.784$	$\rho_{XY} \approx -0.655$
MSE	2	2	4	4	4	4
MAE	1.678	1.678	3.356	3.672	3.420	3.828
MAPE	0.693	0.693	1.386	2.674	1.667	3.556

Table 1: Optimal forecasts in the three examples discussed in section 3, depending on the KPI

3. Examples

As a very simple example, assume our hierarchy consists of just two bottom series X_t and Y_t and their sum $X_t + Y_t$. Assume further that $X_t, Y_t \sim \Gamma(2, 1)$ are identically distributed, so for simplicity, we will omit the subscript t from now on.

We will investigate optimal forecasts for X and Y in subsection 3.1. Subsequently, we consider optimal point forecasts for $X + Y$ in the case where X and Y are uncorrelated (subsection 3.2), where they are positively correlated (subsection 3.3), and finally where they are negatively correlated (subsection 3.4). We collect our results in Table 1.

In our analyses, we use R (R Core Team, 2019) and the `copula` (Hofert et al., 2018; Yan, 2007; Kojadinovic & Yan, 2010) and the `hexbin` packages (Carr, 2019).

3.1. The bottom level

First of all, let us investigate the bottom level series (Kolassa, 2020).

- Their MSE-optimal forecasts are their expectations, which is $2 \times 1 = 2$.
- Their MAE-optimal forecasts are their medians (Hanley et al., 2001), which for a $\Gamma(2, 1)$ distributed variable can be approximated (Berg & Pedersen, 2006) to be about 1.678.

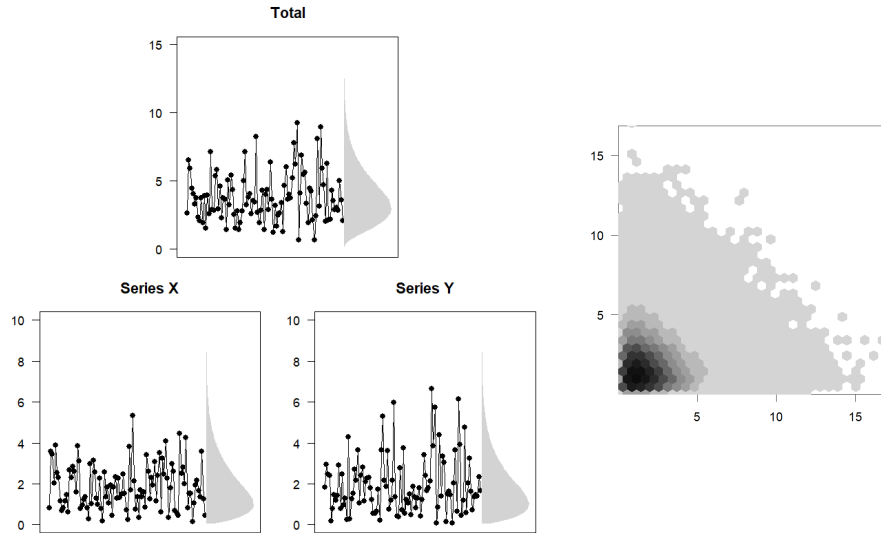


Figure 1: Left: 100 realizations of uncorrelated simulated time series, $X, Y \sim \Gamma(2, 1)$, and their total, together with rotated histograms. Right: hexbinplot of X vs. Y . Histograms and the hexbinplot are based on 10^6 simulations.

- Their MAPE-optimal forecasts are their (-1) -medians (see above), which for a $\Gamma(k, \theta)$ distributed variable with $k \geq 2$ is the median of a $\Gamma(k - 1, \theta)$ distribution (Kolassa, 2019), which in turn in our case of $k = 2$ and $\theta = 1$ is straightforwardly calculated to be $\log 2 \approx 0.693$.

Table 1 also shows the sums $\hat{x}_{\text{KPI}} + \hat{y}_{\text{KPI}}$ of these optimal bottom level forecasts \hat{x}_{KPI} and \hat{y}_{KPI} . How these compare to the optimal total forecast $\widehat{x + y}_{\text{KPI}}$ depends on the KPI and on the correlation structure. We now turn to this analysis.

3.2. X and Y are uncorrelated

Figure 1 shows a simulation example. Since X and Y have the same scale parameter of 1, we have $X + Y \sim \Gamma(4, 1)$, so we can follow the same logic as for

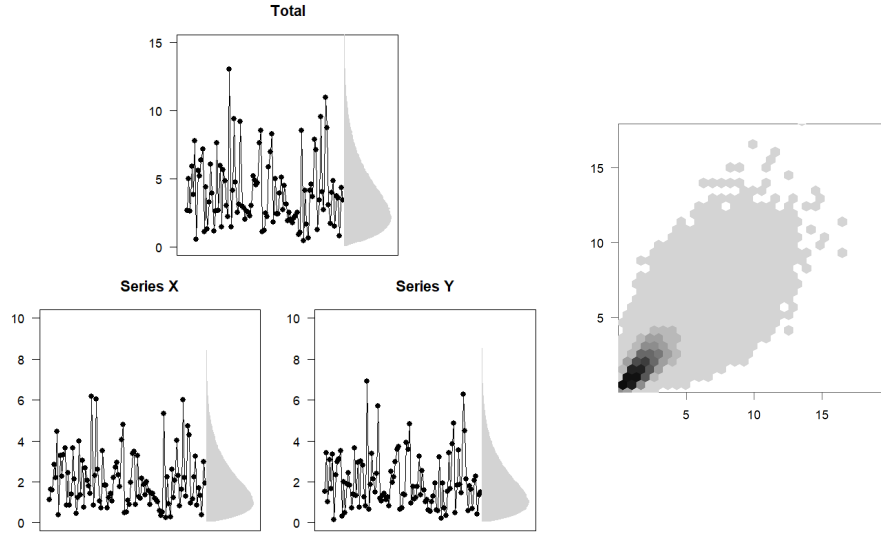


Figure 2: Left: 100 realizations of positively correlated simulated time series, $X, Y \sim \Gamma(2, 1)$, and their total, together with rotated histograms. Right: hexbinplot of X vs. Y . Histograms and the hexbinplot are based on 10^6 simulations.

the bottom level series (see above) to obtain the following optimal forecasts:

$$\begin{aligned}\widehat{x + y}_{\text{MSE}} &= 4 \times 1 = 4 \\ \widehat{x + y}_{\text{MAE}} &\approx 3.672 \\ \widehat{x + y}_{\text{MAPE}} &\approx 2.674.\end{aligned}$$

We note in particular that the MAPE-optimal sum forecast is almost twice the sum of the MAPE-optimal bottom level forecasts.

3.3. X and Y are positively correlated

We generate positively correlated random variables with $\Gamma(2, 1)$ marginals using a Gaussian copula (Dave, 2019) with an input correlation of 0.80, resulting in a correlation of $\rho_{XY} \approx 0.784$. Figure 2 shows our simulated data. We empirically derive the optimal sum forecasts $\widehat{x + y}_{\text{KPI}}$ by minimizing the KPI over 10^6

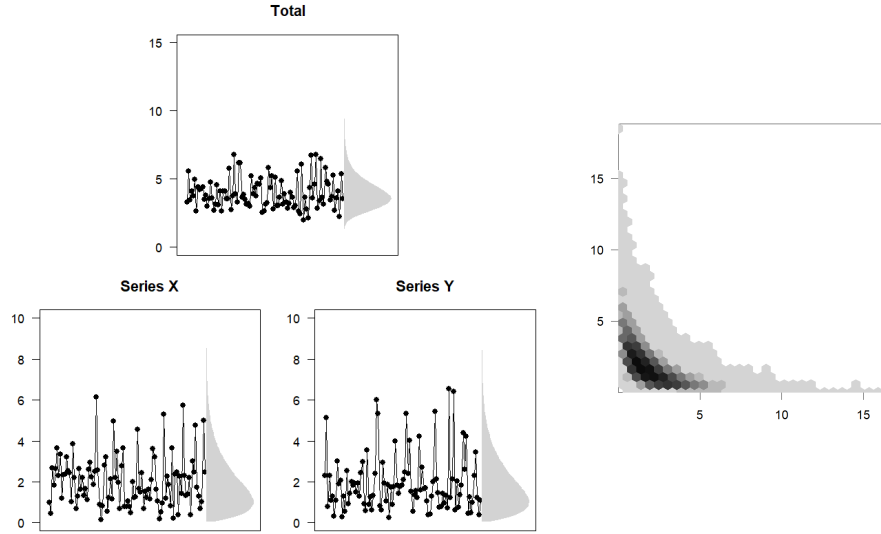


Figure 3: Left: 100 realizations of negatively correlated simulated time series, $X, Y \sim \Gamma(2, 1)$, and their total, together with rotated histograms. Right: hexbinplot of X vs. Y . Histograms and the hexbinplot are based on 10^6 simulations.

simulations, obtaining

$$\begin{aligned}\widehat{x + y}_{\text{MSE}} &= 4 \times 1 = 4 \\ \widehat{x + y}_{\text{MAE}} &\approx 3.420 \\ \widehat{x + y}_{\text{MAPE}} &\approx 1.667.\end{aligned}$$

3.4. X and Y are negatively correlated

We generate negatively correlated random variables with $\Gamma(2, 1)$ marginals using a Gaussian copula (Dave, 2019) with an input correlation of -0.80 , resulting in a correlation of $\rho_{XY} \approx -0.655$. Figure 3 shows our simulated data. We again empirically derive the optimal sum forecasts $\widehat{x + y}_{\text{KPI}}$ by minimizing the KPI over 10^6 simulations, obtaining

$$\begin{aligned}\widehat{x + y}_{\text{MSE}} &= 4 \times 1 = 4 \\ \widehat{x + y}_{\text{MAE}} &\approx 3.828 \\ \widehat{x + y}_{\text{MAPE}} &\approx 3.556.\end{aligned}$$

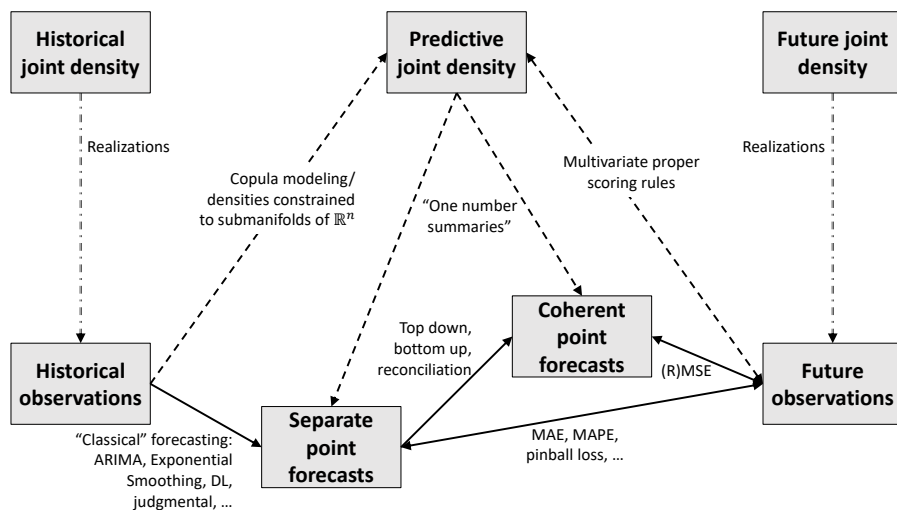


Figure 4: A comparison of current (solid lines) and proposed (dashed lines) research directions on hierarchical forecasting. See section 4 for details.

4. Takeaways

What are the implications of the argument above?

- Not all hierarchical point forecasts need to be coherent. This is obvious for quantile forecasts, and it also holds for other forecasts, since what functional of the predictive distribution we use as a point forecast, as elicited by our error measure, depends on the *decision* we want to take based on this forecast.
- Do use the (R)MSE to evaluate coherent forecasts. Or the MAE or MASE if your predictive densities are symmetric. (Note that, e.g., low volume count data series, as found in retail demand on $\text{SKU} \times \text{store} \times \text{day}$ granularity, are *not* symmetric.)
- If you are tempted to evaluate coherent forecasts using the MAPE, **MAE or MASE**, remember that the MAPE-optimal forecasts are almost certainly not coherent, **and MAE-/MASE-optimal ones likely not**. Think

about what you *really* want, and whether a low MAPE/MAE/MASE or a coherent forecast is more important.

- The recent emphasis on coherent *point* forecasts in hierarchical settings has led to important advances. However, more emphasis should be paid to *distributional* forecasting in hierarchical contexts. See below for details.
- However, although the evaluation of marginal predictive densities has seen a lot of development, e.g., by using proper scoring rules, the development of analogous accuracy measures for *multivariate* predictive densities is in its infancy. Panagiotelis et al. (in press) is a pioneering article in this direction. More research is clearly needed in this area.

Let us consider the last two bullet points in more detail. Figure 4, which is consciously modeled on Figure 2 in Kolassa (2020), may serve as an illustration. Figure 4 shows unobservable and unknowable historical and future joint densities, from which we observe actuals, which are the only realizations we can use to build or evaluate our models. Most current research and practice for hierarchical forecasting proceeds along the solid lines: first, we calculate separate point forecasts using our point forecasting toolbox (containing methods like ARIMA, Exponential Smoothing, Deep Learning, judgmental forecasting etc.). Then, these separate point forecasts are turned into coherent (i.e., sum-consistent) point forecasts through top down disaggregation, bottom up aggregation or (optimal) reconciliation. Finally, the point forecasts are evaluated using the (R)MSE or other accuracy measures. (By the argument above, the evaluation of coherent forecasts using the MAPE etc. is inappropriate.)

We propose to shift our attention, at least in part, to a “coherent density approach” to hierarchical forecasting, shown in Figure 4 by dashed lines. We should use historical observations to derive predictive joint densities, e.g., through copula modeling of the base time series, or by considering densities restricted to submanifolds of \mathbb{R}^n in general. It is gratifying to see initial steps being taken in

this direction (Ben Taieb et al., 2020; Jeon et al., 2019; Panagiotelis et al., in press).

From these predictive joint densities, we can derive point forecasts as optimal “one number summaries”, which aim to minimize an appropriately chosen accuracy measure in expectation. As Kolassa (2020) argues, these marginal functionals of the predictive joint density should usually be *different* for different accuracy measures. If we really want coherent point forecasts that minimize MAPE, nothing keeps us from seeing this as a constrained optimization problem, namely to minimize the expected MAPE *subject to coherence*. (We still maintain that this illustrates a deep confusion about what we want.)

However, we should really not be aiming at good point forecasts derived from the predictive joint density, but at good predictive joint densities themselves. The marginal analogue would be to evaluate marginal predictive densities through proper scoring rules, which are unintuitive, but at least a step in the right direction. In the multivariate case, the energy score (Gneiting & Raftery, 2007) and the variogram score (Scheuerer & Hamill, 2015) have been proposed, while it is known that the common log score is improper in the hierarchical case, in the sense that a probabilistically incoherent predictive density can yield a lower log score than the correct (probabilistically coherent) density (Panagiotelis et al., in press). Multivariate proper scoring rules and their properties clearly merit more research.

5. Conclusion

Hierarchical forecasting has been an important research topic, and it will get even more important in the future, as more and more hierarchical data becomes available. We need to make sure we evaluate hierarchical forecasts using appropriate accuracy measures, be open to asking whether point forecasts truly have to be coherent, and consider joint distributional forecasting in hierarchical contexts.

References

- Athanasopoulos, G., Ahmed, R. A., & Hyndman, R. J. (2009). Hierarchical forecasts for Australian domestic tourism. *International Journal of Forecasting*, *25*, 146–166. doi:10.1016/j.ijforecast.2008.07.004.
- Athanasopoulos, G., Ahmed, R. A., & Hyndman, R. J. (2015). Corrigendum to: “Hierarchical forecasts for Australian domestic tourism” [International Journal of Forecasting 25 (2009) 146-166]. *International Journal of Forecasting*, *31*, 585. doi:10.1016/j.ijforecast.2014.08.011.
- Athanasopoulos, G., Hyndman, R. J., Kourentzes, N., & Petropoulos, F. (2017). Forecasting with temporal hierarchies. *European Journal of Operational Research*, *262*, 60–74. doi:10.1016/j.ejor.2017.02.046.
- Athanasopoulos, G., & Kourentzes, N. (in press). On the evaluation of hierarchical forecasts. *International Journal of Forecasting*, .
- Ben Taieb, S., Taylor, J. W., & Hyndman, R. J. (2020). Hierarchical probabilistic forecasting of electricity demand with smart meter data. *Journal of the American Statistical Association*, . doi:10.1080/01621459.2020.1736081.
- Berg, C., & Pedersen, H. L. (2006). The Chen-Rubin conjecture in a continuous setting. *Methods and Applications of Analysis*, *13*, 63–88. doi:10.4310/MAA.2006.v13.n1.a4.
- Carr, D. (2019). *hexbin: Hexagonal Binning Routines*. URL: <https://CRAN.R-project.org/package=hexbin> R package version 1.28.0. Ported by Nicholas Lewin-Koh and Martin Maechler, containing copies of lattice functions written by Deepayan Sarkar.
- Dave (2019). Simulate a Gaussian copula with t margins. CrossValidated (version: 2019-08-21). URL: <https://stats.stackexchange.com/q/423189>.
- Fildes, R., Ma, S., & Kolassa, S. (2022). Retail forecasting: research and practice. *International Journal of Forecasting*, *38*, 1283–1318. doi:10.1016/j.ijforecast.2019.06.004.

- Gneiting, T. (2011a). Making and evaluating point forecasts. *Journal of the American Statistical Association*, *106*, 746–762. doi:10.1198/jasa.2011.r10138.
- Gneiting, T. (2011b). Quantiles as optimal point forecasts. *International Journal of Forecasting*, *27*, 197–207. doi:10.1016/j.ijforecast.2009.12.015.
- Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, *102*, 359–378. doi:10.1198/016214506000001437.
- Hanley, J. A., Joseph, L., Platt, R. W., Chung, M. K., & Belisle, P. (2001). Visualizing the median as the minimum-deviation location. *The American Statistician*, *55*, 150–152. doi:10.1198/000313001750358482.
- Hofert, M., Kojadinovic, I., Maechler, M., & Yan, J. (2018). *copula: Multivariate Dependence with Copulas*. URL: <https://CRAN.R-project.org/package=copula> R package version 0.999-19.1.
- Hyndman, R. J., Ahmed, R. A., Athanasopoulos, G., & Shang, H. L. (2011). Optimal combination forecasts for hierarchical time series. *Computational Statistics & Data Analysis*, *55*, 2579–2589. doi:10.1016/j.csda.2011.03.006.
- Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and Practice*. (2nd ed.). Melbourne, Australia: OTexts. URL: <https://otexts.com/fpp2/> accessed on 2019-12-18.
- Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, *22*, 679–688. doi:10.1016/j.ijforecast.2006.03.001.
- Jeon, J., Panagiotelis, A., & Petropoulos, F. (2019). Probabilistic forecast reconciliation with applications to wind power and electric load. *European Journal of Operational Research*, *279*, 364–379. doi:10.1016/j.ejor.2019.05.020.

- Kojadinovic, I., & Yan, J. (2010). Modeling multivariate distributions with continuous margins using the copula R package. *Journal of Statistical Software*, *34*, 1–20. doi:10.18637/jss.v034.i09.
- Kolassa, S. (2019). What is the best point forecast for gamma distributed data? CrossValidated (version: 2019-02-21). URL: <https://stats.stackexchange.com/q/389318>.
- Kolassa, S. (2020). Why the “best” point forecast depends on the error or accuracy measure (invited commentary on the M4 competition). *International Journal of Forecasting*, *36*, 208–211. doi:10.1016/j.ijforecast.2019.02.017.
- Kolassa, S., & Schütz, W. (2007). Advantages of the MAD/Mean ratio over the MAPE. *Foresight: The International Journal of Applied Forecasting*, *6*, 40–43. URL: <https://ideas.repec.org/a/for/ijafaa/y2007i6p40-43.html>.
- Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2022). The m5 competition: Background, organization, and implementation. *International Journal of Forecasting*, *38*, 1325–1336. doi:10.1016/j.ijforecast.2021.07.007.
- 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*, 343–359. doi:10.1016/j.ijforecast.2020.06.004.
- Panagiotelis, A., Gamakumara, P., Athanasopoulos, G., & Hyndman, R. J. (in press). Probabilistic forecast reconciliation: Properties, evaluation and score optimisation. *European Journal of Operational Research*, . URL: <https://www.sciencedirect.com/science/article/pii/S0377221722006087>. doi:<https://doi.org/10.1016/j.ejor.2022.07.040>.
- 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/> version 3.6.2.

- Scheuerer, M., & Hamill, T. M. (2015). Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities. *Monthly Weather Review*, *143*, 1321 – 1334. URL: <https://journals.ametsoc.org/view/journals/mwre/143/4/mwr-d-14-00269.1.xml>. doi:10.1175/MWR-D-14-00269.1.
- Syntetos, A. A., Babai, Z., Boylan, J. E., Kolassa, S., & Nikolopoulos, K. (2016). Supply chain forecasting: Theory, practice, their gap and the future. *European Journal of Operational Research*, *252*, 1–26. doi:10.1016/j.ejor.2015.11.010.
- 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*, 804–819. doi:10.1080/01621459.2018.1448825.
- Yan, J. (2007). Enjoy the joy of copulas: With a package copula. *Journal of Statistical Software*, *21*, 1–21. doi:10.18637/jss.v021.i04.