Commentary on the M5 forecasting competition

Stephan Kolassa

SAP Switzerland
Bahnstrasse 1
8274 Tägerwilen
Switzerland
Stephan.Kolassa@sap.com

Keywords: Comparative studies, Demand forecasting, Intermittent demand, M-competitions, Quantile forecasts

1. Introduction

The M5 forecasting competition is another milestone in the field of forecasting, especially so since it may have been the first exposure for the larger field of data scientists to probabilistic and quantile forecasting. Its importance to popularizing these important generalizations to more common expectation forecasting cannot be overstated. Similarly, it focuses on one specific industry, retail, which very much increases its relevance to this particular sector, compared to earlier M competitions.

We give a few comments on both the M5 competition and the attendant papers (Makridakis et al., 2020,), focusing (1) on the under-appreciated performance of simple methods, (2) on more appropriate count models, (3) on the role of forecasters and explainability, and (4) on the return to investment for complexity.

2. With probability 92.5%, Exponential Smoothing will be best for you (but Walmart is not a typical retailer)

In the Accuracy track, Makridakis et al. (2020) note in passing that only 7.5% of the submitting teams outperformed the top performing benchmark, which was a very simple automatically chosen Exponential Smoothing method with equally simple bottom-up aggregation (`ES_bu`). Relatedly, an analysis of the Accuracy submissions finds that even the top performing method `YJ_STU` outperforms `ES_bu` only on 58.5% of series in terms of MSE, which is certainly statistically significant, but not overwhelming. The second place `Matthias` was better than `ES_bu` on only 6.7% of series, which definitely has to do with the fact that `Matthias` only has integer values in their Accuracy submission file. In 19.3% of series, `ES_bu` was better than *all* five top submissions.

We find this interesting, and under-appreciated when the rest of the paper focuses (understandably) selectively on the top contestants. Even in an age of Machine Learning (ML) and Data Science blossoming around the world, after years of Kaggle competitions, only 7.5% of teams manage to beat an extremely simple benchmark! Yes, those that do beat it do so by a goodly margin, but this observation to us suggests that it is still quite hard to outperform the simple benchmarks. One could interpret this provocatively as a very rough prior probability: *a priori*, you only have a 7.5% chance of outperforming bottom-up Exponential Smoothing.

Of course, this may be due to participants losing interest and not polishing their submissions after the first one. That is quite possible. However, on the one hand, we can't say whether this interpretation is correct, because that would require deeper analysis of whether participants' later submissions improved in accuracy on the holdout test set. And on the other hand, if this explanation holds, it already tells us something: outperforming bottom-up Exponential Smoothing requires more work than simply plugging together an ML pipeline. It is not automatic.

Thus, one piece of forecasting wisdom is again supported: if you are tempted to invest heavily in data scientists and expect wonders from them, make sure to compare their methods to simple benchmarks that you can probably implement at a fraction of the cost of an ML pipeline.

Now, just where does this surprising finding come from? We suspect this is a result of Walmart not being representative of many retailers. Specifically, Walmart uses a so-called "Everyday Low Price" (EDLP) strategy, with very low reliance on promotions to drive traffic and sales. This is much easier on the supply chain (and the forecasting team!) than a strongly promotion driven strategy – and of course, the Exponential Smoothing benchmark will have an easier time in forecasting series without strong promotional effects.

However, most retailers do rely on promotions. And in our experience, marketers are enormously creative at dreaming up new promotions, in terms of pricing, communication, tactics and conditions. "Buy three units of product X at 20% off and two units of product Y at full price, and present your app coupon at the checkout, to have a chance of winning two tickets to Disneyland!" – how will this drive sales of products X and Y, and of complementary or substitute products (and how many Disneyland tickets should we budget for)? Indeed, different promotions vary widely in uplift, they interact with calendar effects, seasonality and other dynamics, and promotions not only have an impact on *average* demand, but also on the *variance* of demand (Fildes et al., in press), and thus on the necessary safety stocks. Note also that promotional forecasts are even more important than forecasts for regular sales, as little is as annoying to the customer as not finding *promoted* items in stock. (Conversely, stock left over at the end of the promotion may clog the shelves for a long time – so we may well decide we want a *lower* service level during promotions than outside promotions.) Figure 1 shows sales time series of two SKUs at a European store with an indication of the variety of promotions.

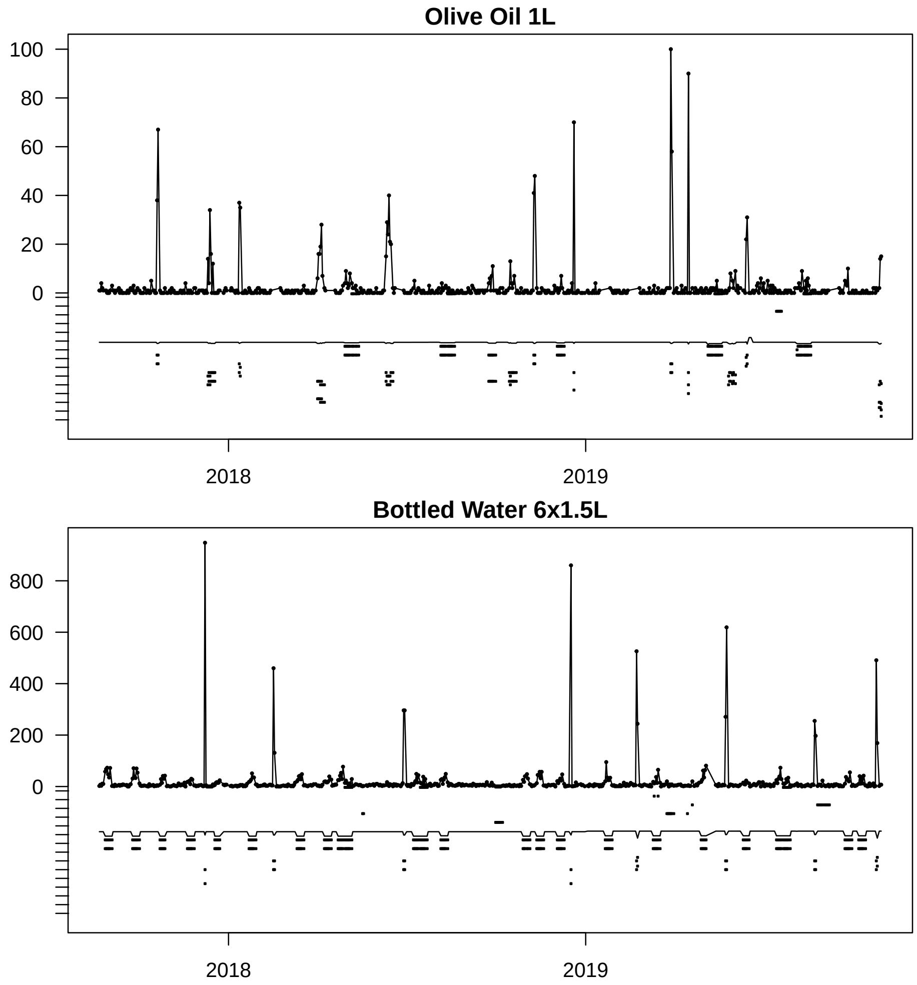


Figure 1: Daily sales of two products in one European store, with scaled prices and Boolean promotion predictors (promotion types, display types etc.) at the bottom. White space would be filled with *other* predictors for other products (note tickmarks). Note the varying length and impact of promotions, and, e.g., that the highest selling promotions for the bottom SKU are *not* the ones with the largest price reduction. Compare Fildes et al. (in press, Fig. 9).

Incidentally, we should also keep the issue of variability of forecasts in mind. Competitions, and especially the Kaggle format with continuously updated leaderboards, incentivize participants to submit more variable forecasts, since it does not matter whether they land in 20th or in 50th place – but with a highly variable forecast, they have a chance of coming in among the coveted top 10. One could call this “overfitting to the leaderboard,” and the effect has first been pointed out and analyzed by Ma & Fildes (submitted). We emphasize that this overfitting is probably unconscious, and that *repeated* strong performance would increase our trust in the quality of a forecasting method. In particular, the *consistent* strong showing of Gradient Boosted Machines (GBM) can in our opinion not be due to overfitting alone.

3. Benchmarking against *appropriate* models

With regard to the M5 Uncertainty track (Makridakis et al., 2020), we have some reservations about the benchmark methods used. The dataset consisted of a hierarchy of supermarket sales, with the bottom level being on stock keeping unit (SKU) × store × day granularity. Such data are invariably low counts and usually intermittent (Fildes et al., in press). “Classical” forecasting methods like Exponential Smoothing and ARIMA(X) presuppose homoskedastic normally distributed errors, and any quantile forecasts from them will also use this normal distribution assumption – as do the benchmark quantile calculation methods used by Makridakis et al. (2020). A quantile forecast formula of the form “$\widehat{\mu} \pm z_\alpha \widehat{\sigma}$” with an expectation forecast $\widehat{\mu}$, a standard normal quantile z_α and a forecasted standard deviation $\widehat{\sigma}$ does not even make sense for count data: it outputs noninteger or even negative quantiles. (We note, however, that a normal distribution assumption on the bottom level, possibly together with forecasts of product demand covariances, allows an easy derivation of distributional and quantile forecasts on higher hierarchical levels.)

We thus re-ran the M5 uncertainty analysis with benchmark methods tailored to count data, as inspired by Kolassa (2016), whose Retailer A incidentally also uses an EDLP strategy. For simplicity in benchmarking, methods were applied to all series separately, without accounting for the hierarchical structure.

Empirical (Emp): The empirical quantiles of the entire historical series were used, calculated using R’s `quantile()` function with `type=8`, as recommended by Hyndman & Fan (1996), and rounding to the nearest integer.

Empirical with Weekdays (Emp-Wd): As in the Emp method, but quantiles were calculated separately for each weekday, using historical observations on the same weekday.

Poisson (Pois): Quantiles were taken from a Poisson distribution fitted to the entire historical series via moment matching.

						Aggregation level							
	1	2	3	4	5	6	7	8	9	10	11	12	Average
Emp	0.501	0.465	0.484	0.486	0.502	0.451	0.458	0.465	0.464	0.387	0.339	0.312	0.443
Emp-Wd	0.507	0.449	0.473	0.475	0.477	0.424	0.428	0.444	0.439	0.378	0.334	0.310	0.428
Pois	1.074	0.953	0.941	1.025	1.007	0.899	0.869	0.863	0.813	0.529	0.425	0.360	0.813
NB-CMP	3.034	2.632	2.309	2.733	2.504	2.283	2.065	1.885	1.635	0.594	0.437	0.349	1.872
ZIP	1.074	0.953	0.938	1.022	1.002	0.896	0.864	0.859	0.808	0.460	0.375	0.327	0.798
ZINB	3.034	2.632	0.858	1.940	1.329	1.627	1.135	0.657	0.577	0.396	0.341	0.312	1.236

Table 1: The performance of the benchmark methods considered in terms of Weighted Scaled Pinball Loss (WPSL). Compare Table 2 of Makridakis et al. (2020).

Negative Binomial/Conway-Maxwell-Poisson (NB-CMP): Quantiles were taken from

- either a Negative Binomial distribution fitted to the entire historical series via moment matching, if the series was overdispersed,
- or a Conway-Maxwell-Poisson distribution fitted to the entire historical series using the `glm.cmp()` function in the `COMPoissonReg` package (Sellers et al., 2019), if the series was equi- or underdispersed.

Zero-Inflated Poisson (ZIP): Quantiles were taken from a Zero-Inflated Poisson distribution fitted to the entire historical series using the `zeroinfl()` function in the `pscl` package (Jackman, 2020; Zeileis et al., 2008). If the minimum of the historical series was greater than zero, we fell back to the Pois method.

Zero-Inflated Negative Binomial (ZINB): Quantiles were taken from a Zero-Inflated Negative Binomial distribution fitted to the entire historical series using the `zeroinfl()` function in the `pscl` package (Jackman, 2020; Zeileis et al., 2008). If the minimum of the historical series was greater than zero, we fell back to the NB-CMP method. For series where the fitting routine threw an error because of a numerical singularity, we fell back to the ZIP method.

We downloaded the Rdata file from the Google Drive folder linked from the M5 competition GitHub site at `https://github.com/Mcompetitions/M5-methods` on January 6, 2021. All methods were implemented in R (R Core Team, 2020).

Table 1 contains the results, in an analogous format to Table 2 of Makridakis et al. (2020). Figure 2 shows the performance of the top 50 submissions and the benchmarks considered for aggregation level 12 (SKU × store) only. Both are given in terms of Weighted Scaled Pinball Loss (WSPL).

Results show that the proposed discrete benchmarks are competitive with the top 50 M5 submissions on the most granular aggregation level 12, which is most relevant for operational store replenishment. On higher levels – where count series behave more like continuous series – they do not perform well. Particularly notable are the abysmal results of NB-CMP on levels 1-9 and ZINB on levels 1-6. It may well be that the high volumes at these aggregation levels led to numerical

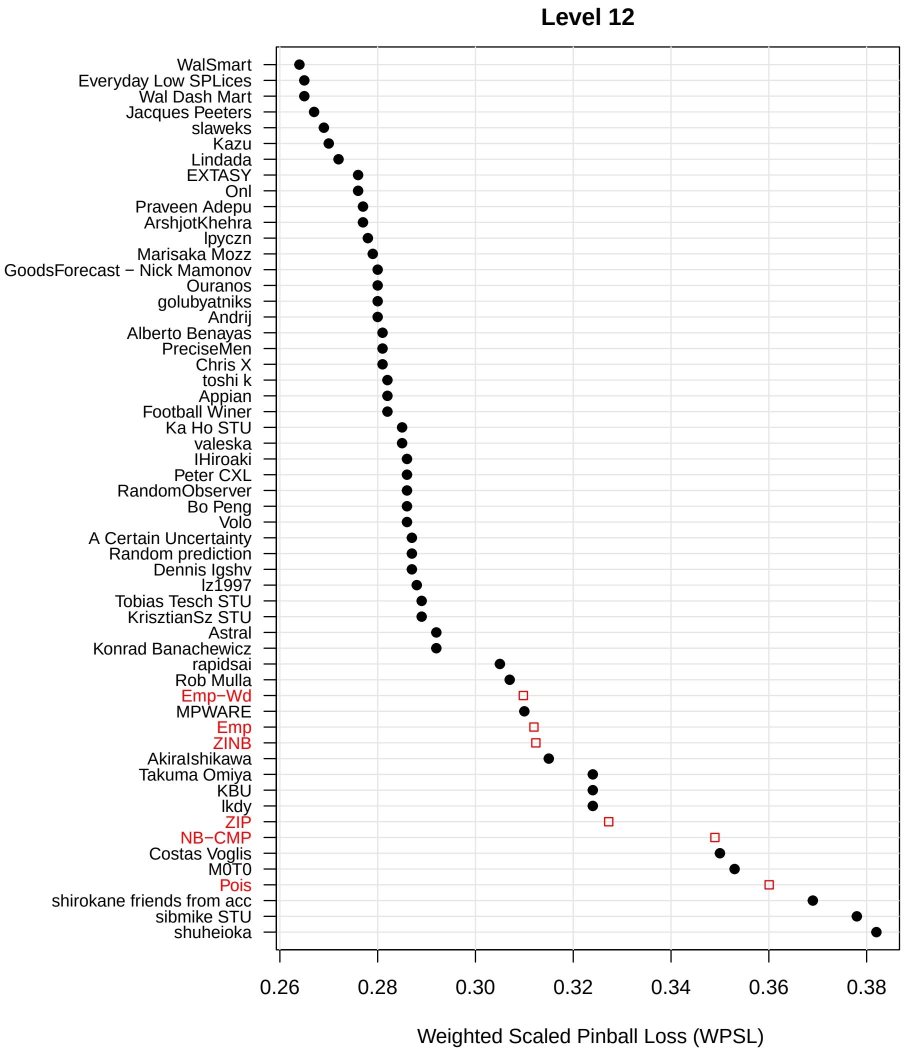


Figure 2: The performance of the top 50 submissions and the benchmark methods considered on aggregation level 12, in terms of Weighted Scaled Pinball Loss (WPSL).

instabilities in model estimation, but we did not investigate this more deeply, since our main focus is on the more disaggregate data on level 12.

It was surprising to us that judging from Figure 2 in Makridakis et al. (2020), the ARIMA benchmark appears to perform better than our count data benchmarks even on level 12. Capturing time series dynamics (as ARIMA does, but not our discrete benchmarks) may be more important than modeling count data as such (as our benchmarks do, but not ARIMA). It may be interesting to model these series using dedicated methods like Integer Autoregression (INAR; Mohammadipour & Boylan, 2012; Weiß, 2018), which unfortunately have not received the attention they deserve in the forecasting community.

Two extremely simple benchmarks – Empirical and Empirical with Weekdays – outperformed the "statistically more sophisticated" benchmarks, and that across all aggregation methods. This is in agreement with the findings of Kolassa (2016).

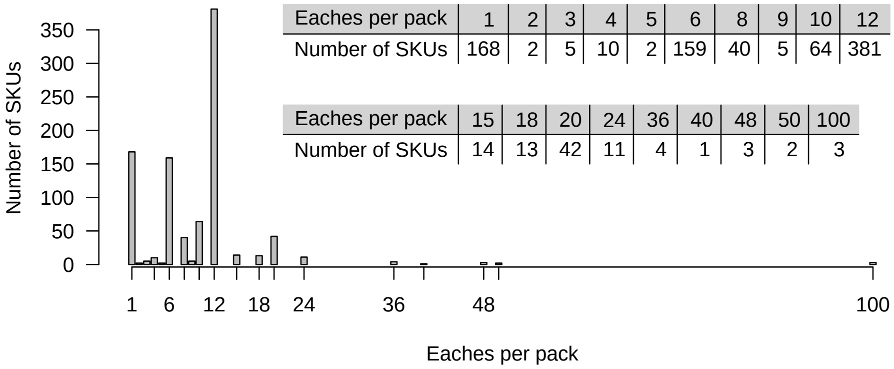
Thus, our findings indicate that the results of the M5 uncertainty competition hold even when submissions are judged against benchmarks that observe the integer character of the bottom level series, at least in terms of forecast accuracy as measured by the WSPL.

4. The role of forecasters and explainability

Do the results of the M5 competition sound the death knell for trained forecasters? "Will these results lead to the demise of forecasters as we know them and the ascent of data scientists who take their place?" (Makridakis et al., 2020) Will ML tools, plugged together by an IT consultant with no scientific training in forecasting, become the norm and consign us to the dustbin of history? We do not believe so.

First, even with great forecasts, the point will come when someone questions them. If you run millions of forecasts every day (Fildes et al., in press), a few will be off. And then, someone *will* complain. The forecaster had better have a response to these complaints, and ideally to analyze and improve the forecast, not just say that it's what the black box spit out. In practice, even a tiny number of truly bad forecasts in a vast sea of reasonable or even great ones will lead to a lack of trust in the entire forecasting system (Dietvorst et al., 2015; Prahl & Van Swol, 2017, and literature cited therein). And if the end users do not trust the forecasting system, they will start modifying the forecasts or creating their own forecasts from scratch.

Second, recall the discussion of retailers' promotional strategies above. Retailers' promotions change constantly, and forecasting their impact requires understanding both the logic of the promotion, and how the forecasting method will deal with predictors, so that the business logic can be translated into predictors which are modeled in a reasonable way by the forecasting tool.



Eaches per pack	1	2	3	4	5	6	8	9	10	12
Number of SKUs	168	2	5	10	2	159	40	5	64	381

Eaches per pack	15	18	20	24	36	40	48	50	100
Number of SKUs	14	13	42	11	4	1	3	2	3

Figure 3: Distribution of logistical units for 929 SKUs at a European retailer. For instance, 168 SKUs are replenished in pack sizes of 1, but 381 SKUs can only be replenished in cartons containing 12 eaches.

Thus, we believe that forecasting experts with domain knowledge will continue to play an important role. One task will be troubleshooting problematic forecasts (or indeed, explaining statistical variation to non-experts) and improving existing models, and another one will be expanding models to cover new business requirements. These tasks will require statistical and business knowledge – as well as programming and communication skills (Kolassa, 2016).

Both tasks hinge crucially on the explainability and analyzability of forecasts, which have indeed been noted by retailers as key requirements of a forecasting system (Yelland et al., 2019; Ulrich et al., 2021) and should be the topic of more future research. Shapley values, which have been used for GBMs (Antipov & Pokryshevskaya, 2020) and can also be applied for other ML methods, may be helpful but are still not very interpretable for non-expert end users, who typically prefer an additive decomposition of the impact of predictors on forecasts.

5. Return on investment for complexity

Finally, is the added complexity in ML methods worth the improvement in (quantile) forecasts? In contrast to some of the papers cited by Makridakis et al. (2020), we usually find that improvements in retail forecast accuracy have a surprisingly low impact on stock performance. Suppose you sell 3 units in a week. Then a forecast of 4 is indubitably better than one of 6. But if the store can only order in pack sizes of 8 and have no stock on hand, this difference in forecast accuracy does not matter – we will order one pack of 8, no matter which of the forecasts we rely on. (Similarly, if we have 6 units on hand, neither forecast will cause us to order anything.)

How influential is this issue? As we are not aware of published research or open datasets on common retailers' logistical units, Figure 3 provides the distribution of logistical pack sizes for 929 SKUs of a European retailer we are working with

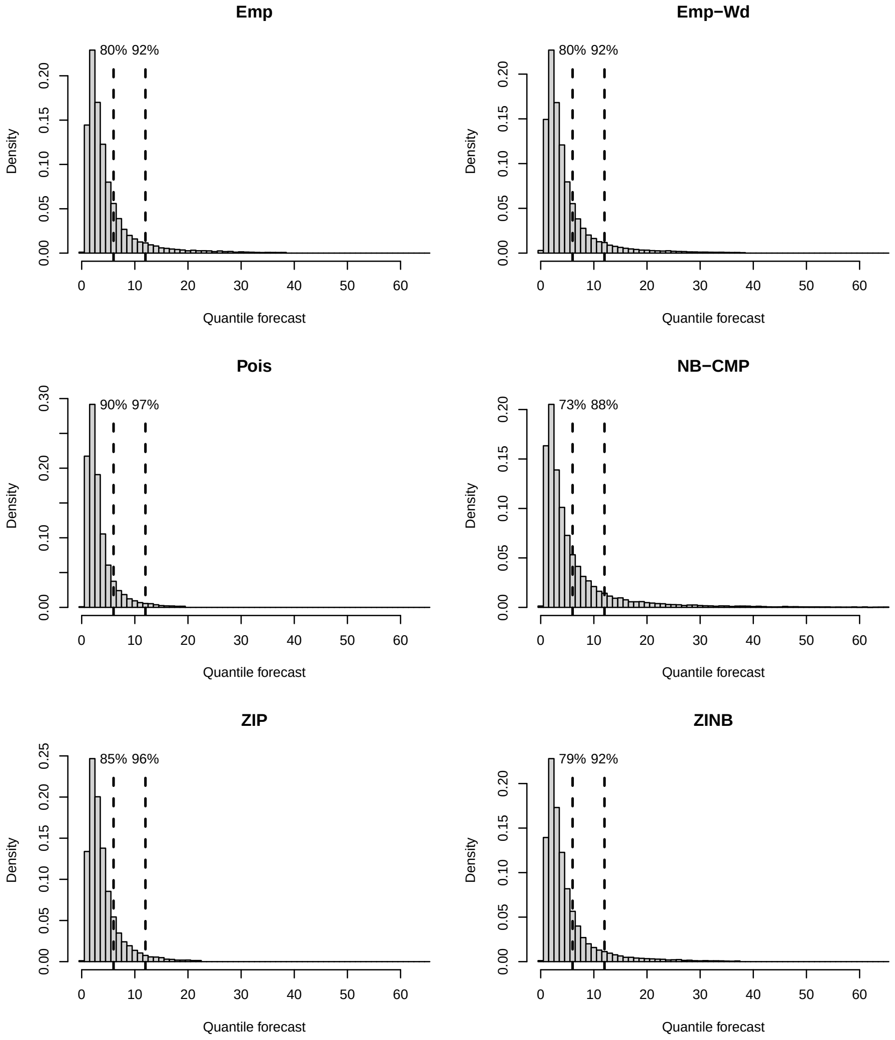


Figure 4: Histograms of the 97.5% quantiles from the benchmark methods defined in section 3 (horizontal axis truncated at the 99th percentile of forecasts for clarity). Vertical lines indicate the proportion of these forecasts that are at most 6 or 12 units.

at the moment. We note that only 168 out of the 929 SKUs (18%) have a logistical pack size of 1, that the most common pack size is 12 and that pack sizes range up to 100.

Next, Figure 4 shows the distribution of the 97.5% quantiles from the benchmark methods defined in section 3 and indicates which proportion of these quantile forecasts is at most 6 or 12 (chosen arbitrarily, inspired by Figure 3). Depending on the method, up to 90% of quantile forecasts are no higher than 6 base units. Note that the (quantile) forecasts between different methods are usually highly correlated (after all, they are all aiming at the same actual value) and that forecasts will only result in an order being released if the *current* stock position is insufficent to satisfy the predicted demand. It thus stands to reason that two forecasting methods, even if their WSPLs are statistically significantly different, will not differ by much in the stock position they result in.

A detailed analysis of this effect would, in addition to the logistical pack sizes, need to take the replenishment schedule into account, i.e., when orders need to be released (one, two or more days before the delivery) and how frequently stock is delivered (every day vs. only on certain days in the week). Another potentially important aspect is how forecasts on daily levels are convolved for aggregation to multi-day periods between deliveries. Most published research on stock control in retail that we are aware of only uses *ad hoc* assumptions on these parameters, and studies that couple the state of the art in forecasting with stock simulations based on true logistical parameters would be welcome indeed.

We note that the papers Makridakis et al. (2020) cite to support the claim that "small improvements in accuracy lead to considerable inventory reductions" do not apply to operational daily retail replenishment: Syntetos et al. (2010) discuss *monthly* (not daily, as in the M5!) sales of an entire *pharmaceutical* company, not a single supermarket, which is where store replenishment takes place; Ghobbar & Friend (2003) are concerned with aviation spare parts, which are typically much more expensive and do not come in packs of eight; and Pooya et al. (2019) use model parameter values that are very unrealistic in retail.

As a matter of fact, this aspect is related to our point about promotions above. Forecast accuracy during promotions does matter indeed, since promotions typically involve much larger amounts of product that may dwarf logistical units.

We thus again need experts with knowledge both of forecasting and of the business domain. Specifically, we need to understand when our scarce resources are better invested in a quest for ever higher accuracy, and when we are better served by challenging logistical constraints in our supply chain.

6. Conclusion

In summary, we find multiple points on which the M5 competition stimulates further discussion and research. As such, these are important papers that will continue to inspire forecasters for years, and we are grateful to be able to participate in this discussion.

In contrast to the M5 authors, we still see an important role for forecasters even in the Data Science world. They may not be called "forecasters", but "data scientists", but that does not matter. What does matter is that their purview will need to expand beyond statistics and programming to encompass an understanding of business needs and processes — and communication skills to explain what they are doing to non-experts (Kolassa, 2016). In addition, the forecasting tools they use will need to be explainable in order to be debuggable, and to build trust by the end user, because otherwise, our beautiful models will simply be ignored.

We are less sanguine than Makridakis et al. (2020,) about the monetary improvement possible through improved quantile forecasts. Much of the literature that posits a direct relationship between (quantile) forecast accuracy and a better inventory position is simply not applicable to supermarket replenishment. Future research should make an effort to also obtain logistical information so the impact of forecasts on the stock position can be assessed.

Finally, we find it fascinating that quite simple benchmarks are still quite competitive. We hope that future retail forecasting competitions feature promotional data, on which more compliated causal methods should be more clearly superior to simple bottom-up Exponential Smoothing. However, in the light of the complexity of many retailers' promotions, adequately explaining the promotional data (and the logistical data, see above!) would be complex indeed, and likely undermine any required anonymity on the part of the retailer.

7. Acknowledgements

We thank Shari De Baets, who very kindly pointed us to pertinent references, and Robert Fildes for very helpful comments on an earlier draft. The `plotrix` package for R (Lemon, 2006) was used in the preparation of some figures.

References

Antipov, E. A., & Pokryshevskaya, E. B. (2020). Interpretable machine learning for demand modeling with high-dimensional data using gradient boosting machines and Shapley values. *Journal of Revenue and Pricing Management*, *19*, 355–364. doi:10.1057/s41272-020-00236-4.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*, 114–126. doi:10.1037/xge0000033.

Fildes, R., Ma, S., & Kolassa, S. (in press). Retail forecasting: research and practice. *International Journal of Forecasting*, . doi:10.1016/j.ijforecast.2019.06.004.

Ghobbar, A. A., & Friend, C. H. (2003). Evaluation of forecasting methods for intermittent parts demand in the field of aviation: a predictive model. *Computers & Operations Research*, *30*, 2097–2114.

Hyndman, R. J., & Fan, Y. (1996). Sample quantiles in statistical packages. *The American Statistician*, *50*, 361–365. doi:10.2307/2684934.

Jackman, S. (2020). *pscl: Classes and Methods for R Developed in the Political Science Computational Laboratory*. United States Studies Centre, University of Sydney Sydney, New South Wales, Australia. URL: https://github.com/atahk/pscl/ R package version 1.5.5.

Kolassa, S. (2016a). Commentary: That feeling for randomness. *Foresight: The International Journal of Applied Forecasting*, *42*, 44–47.

Kolassa, S. (2016b). Evaluating predictive count data distributions in retail sales forecasting. *International Journal of Forecasting*, *32*, 788–803. doi:10.1016/j.ijforecast.2015.12.004.

Lemon, J. (2006). Plotrix: a package in the red light district of R. *R-News*, *6*, 8–12. R package version 3.8_1.

Ma, S., & Fildes, R. (submitted). The performance of the global bottom-up approach in the M5 accuracy competition: a robustness check. *International Journal of Forecasting*, .

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2020a). The M5 accuracy competition: Results, findings and conclusions.

Makridakis, S., Spiliotis, E., Assimakopoulos, V., Chen, Z., Gaba, A., Tsetlin, I., & Winkler, R. (2020b). The M5 uncertainty competition: Results, findings and conclusions.

Mohammadipour, M., & Boylan, J. E. (2012). Forecast horizon aggregation in integer autoregressive moving average (INARMA) models. *Omega*, *40*, 703–712. doi:10.1016/j.omega.2011.08.008.

Pooya, A., Pakdaman, M., & Tadj, L. (2019). Exact and approximate solution for optimal inventory control of two-stock with reworking and forecasting of demand. *Operational Research*, *19*, 333–346. doi:10.1007/s12351-017-0297-6.

Prahl, A., & Van Swol, L. (2017). Understanding algorithm aversion: When is advice from automation discounted? *Journal of Forecasting*, *36*, 691–702. doi:10.1002/for.2464.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing Vienna, Austria. URL: https://www.R-project.org/ version 4.0.3.

Sellers, K., Lotze, T., & Raim, A. (2019). *COMPoissonReg: Conway-Maxwell Poisson (COM-Poisson) Regression*. URL: `https://CRAN.R-project.org/package=COMPoissonReg` R package version 0.7.0.

Syntetos, A. A., Nikolopoulos, K., & Boylan, J. E. (2010). Judging the judges through accuracy-implication metrics: The case of inventory forecasting. *International Journal of Forecasting*, *26*, 134–143.

Ulrich, M., Jahnke, H., Langrock, R., Pesch, R., & Senge, R. (2021). Distributional regression for demand forecasting in e-grocery. *European Journal of Operational Research*, *294*, 831–842. doi:`10.1016/j.ejor.2019.11.029`.

Weiß, C. (2018). *An Introduction to Discrete-Valued Time Series*. John Wiley & Sons, Ltd. doi:`10.1002/9781119097013`.

Yelland, P., Baz, Z. E., & Serafini, D. (2019). Forecasting at scale: The architecture of a modern retail forecasting system. *Foresight: The International Journal of Applied Forecasting*, *55*, 10–18.

Zeileis, A., Kleiber, C., & Jackman, S. (2008). Regression models for count data in R. *Journal of Statistical Software*, *27*, 1–25. URL: `http://www.jstatsoft.org/v27/i08/`.