Forecast Accuracy

Fanciful Aspiration or False Advertising?

Stephan Kolassa, SAP

March 26, 2021

PUBLIC

Lancaster University
Management School

THE BEST RUN SAP

Agenda

Point (including quantile) forecast evaluation

Significance checking

Accuracy benchmarks

Agenda

Point (including quantile) forecast evaluation

Significance checking

Accuracy benchmarks

Point forecasts

- Consider *point* forecasts $\hat{y}_1, \dots, \hat{y}_T$ and corresponding actuals $y_1, \dots, y_T$ for a single time series
- Common point forecast error measures (PFEMs):
  - Mean squared error: $\mathrm{MSE} = \frac{1}{T}\sum_{t=1}^{T}(\hat{y}_t - y_t)^2$ (or root mean squared error, $\mathrm{RMSE} = \sqrt{\mathrm{MSE}}$)
  - Mean absolute error: $\mathrm{MAE} = \frac{1}{T}\sum_{t=1}^{T}|\hat{y}_t - y_t|$
  - Mean absolute scaled error (Hyndman & Koehler, 2006): $\mathrm{MASE} = \frac{\mathrm{MAE}}{s}$, with a scaling factor s, which is the in-sample error of the random walk one-step-ahead forecast
  - Mean absolute percentage error: $\mathrm{MAPE} = \frac{1}{T}\sum_{t=1}^{T}\frac{|\hat{y}_t - y_t|}{y_t}$
  - Weighted mean absolute percentage error: $\mathrm{wMAPE} = \frac{\sum_{t=1}^{T}|\hat{y}_t - y_t|}{\sum_{t=1}^{T} y_t}$
- Can be scaled, e.g., MASE, or scaled RMSE
- Aggregate over time series, e.g., using weights
- Always look at “bad guys” – e.g., the 10 worst forecasts
- Often, people will report multiple PFEMs (cf. link at bottom)

Hyndman, R. J. & Koehler, A. B. Another look at measures of forecast accuracy. *International Journal of Forecasting,* **2006**, *22*, 679-688

Why use a certain measure of forecast error (e.g. MAD) as opposed to another (e.g. MSE)? https://stats.stackexchange.com/q/45875/1352

The $\mathbf{MAPE} = \frac{1}{T}\sum_{t=1}^{T}\frac{|\hat{y}_t - y_t|}{y_t}$

- Undefined if any $y_t = 0$ – consider using the $\mathrm{wMAPE} = \frac{\sum_{t=1}^{T}|\hat{y}_t - y_t|}{\sum_{t=1}^{T} y_t}$ (Kolassa & Schütz, 2007)

- Only makes sense for strictly positive data (not temperatures)

- Not differentiable for minimization (can be smoothly approximated)

- “Interpretable”
  - Can be above 100%, which may be confusing (and then a constant zero forecast $\hat{y}_t = 0$ yields a lower MAPE of 100%)
  - Worse: may lead to strongly biased forecasts (see next slides)

- *Much* more information at Kolassa (2017)

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

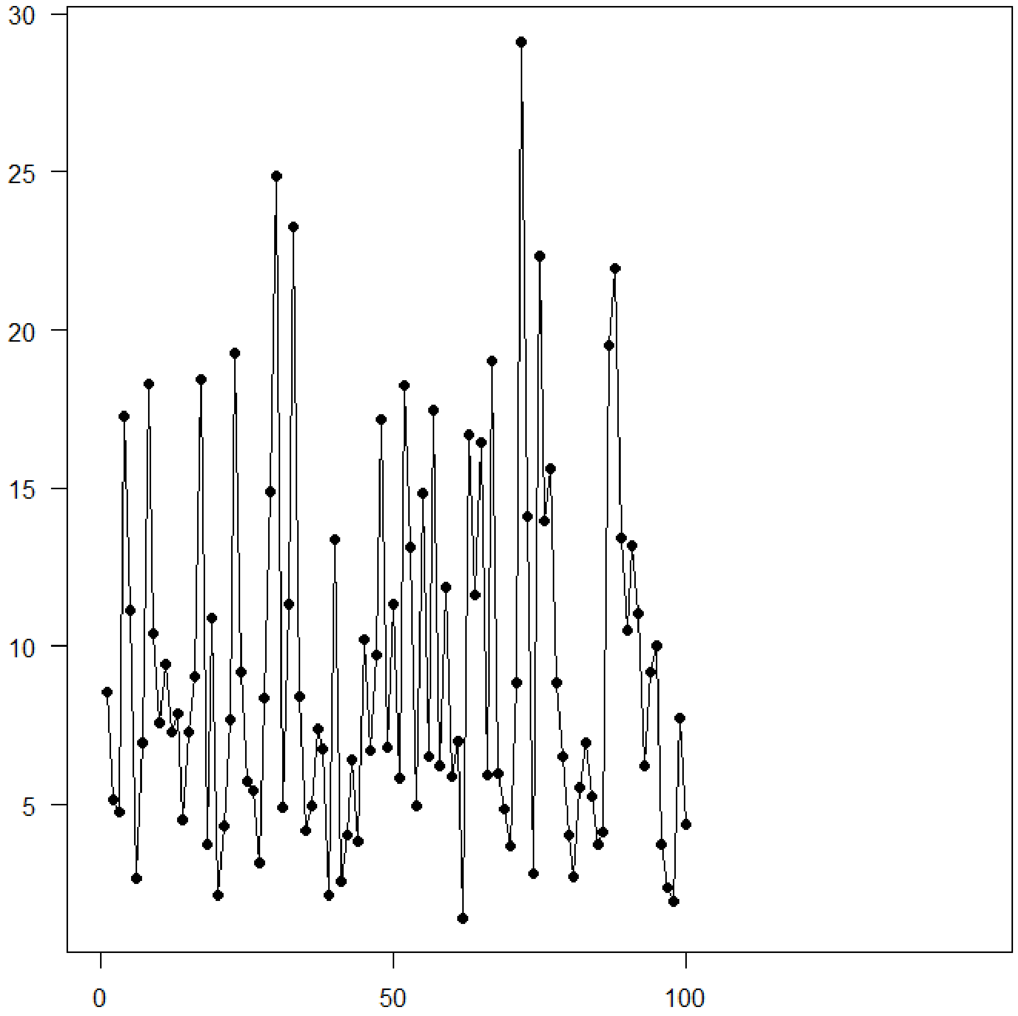
Kolassa, S. What are the shortcomings of the Mean Absolute Percentage Error (MAPE)? https://stats.stackexchange.com/q/299712/1352, **2017**

“… the “best” point forecast depends on the error or accuracy measure”

Kolassa, S. Why the “best” point forecast depends on the error or accuracy measure (invited commentary on the M4 Competition) *International Journal of Forecasting,* **2020***, 36*, 208-211

The “best” point forecast depends on the error or accuracy measure

- Simulated data – independent & identically distributed (iid)
- What is the best point forecast?

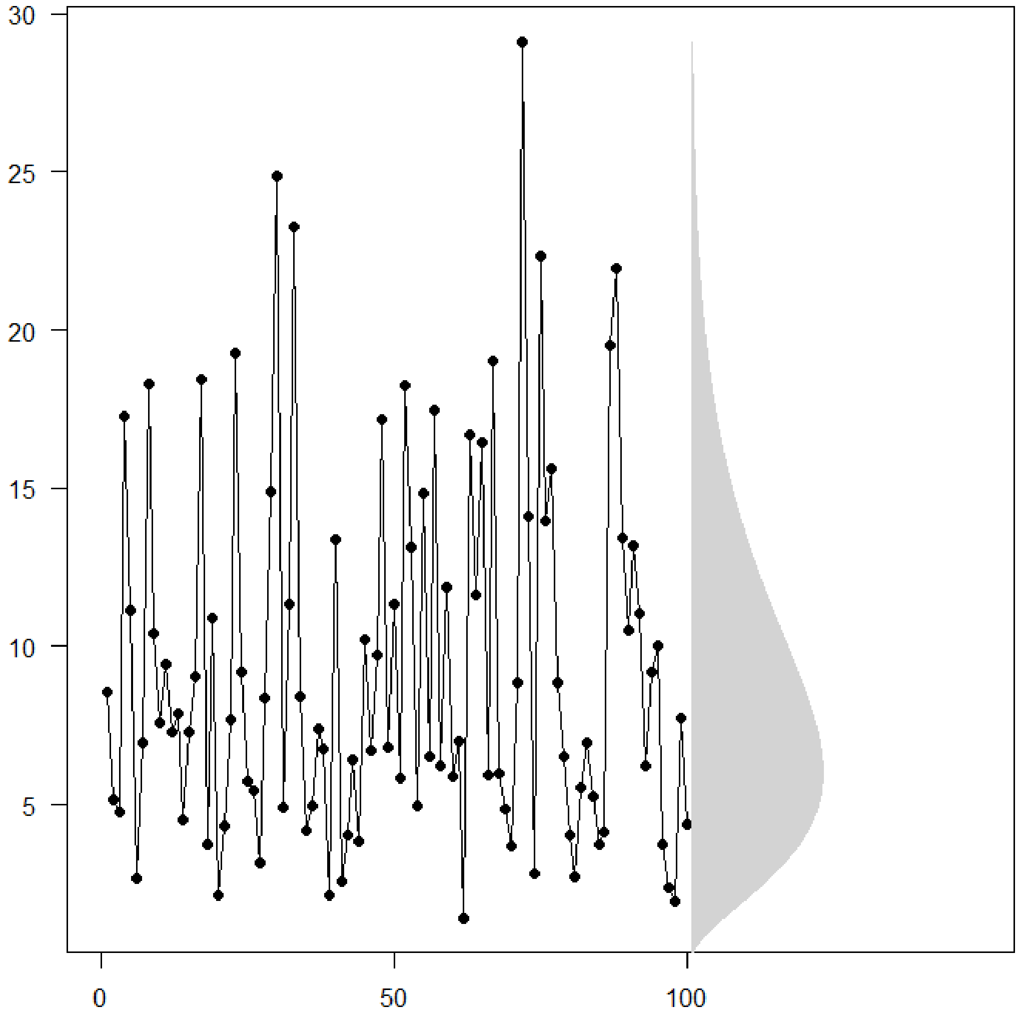


Kolassa, S. Why the “best” point forecast depends on the error or accuracy measure (invited commentary on the M4 Competition) *International Journal of Forecasting,* **2020**, *36*, 208-211

The “best” point forecast depends on the error or accuracy measure

- Simulated data – independent & identically distributed (iid)
- What is the best point forecast?

- The density may help: Gamma (shape 3, scale 3)



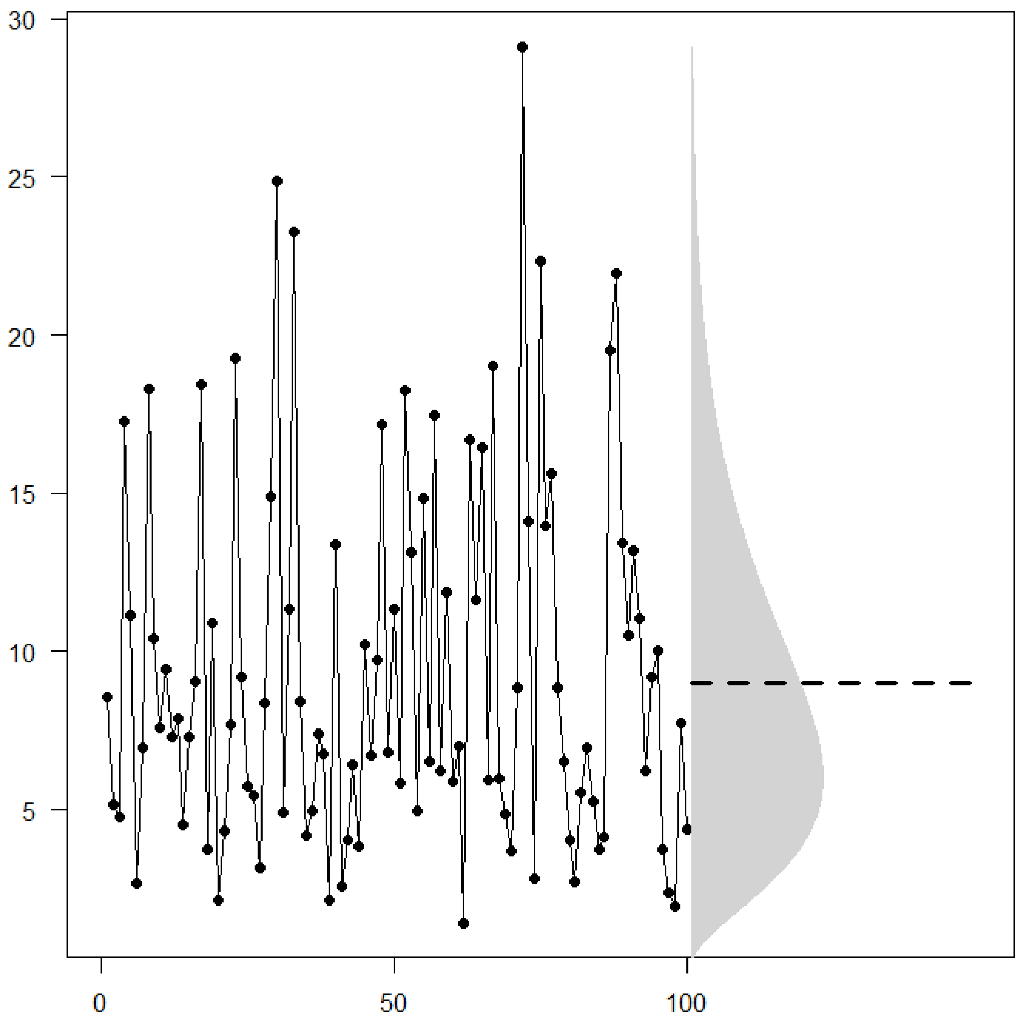
Kolassa, S. Why the “best” point forecast depends on the error or accuracy measure (invited commentary on the M4 Competition) *International Journal of Forecasting,* **2020**, *36*, 208-211

The “best” point forecast depends on the error or accuracy measure

- Simulated data – independent & identically distributed (iid)
- What is the best point forecast?

- The density may help: Gamma (shape 3, scale 3)

- Option 1: the expectation ($\hat{y} = 9.00$)



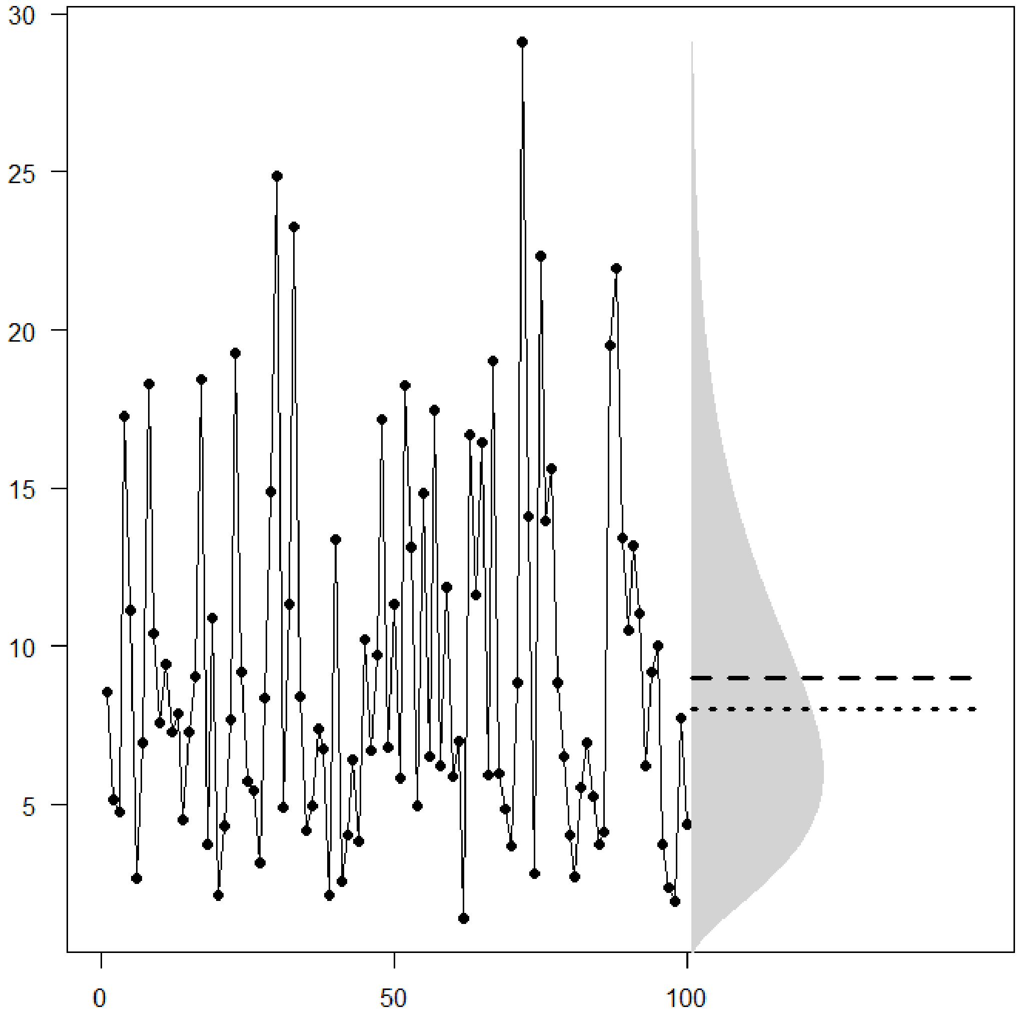
Kolassa, S. Why the “best” point forecast depends on the error or accuracy measure (invited commentary on the M4 Competition) *International Journal of Forecasting*, **2020**, *36*, 208-211

The “best” point forecast depends on the error or accuracy measure

- Simulated data – independent & identically distributed (iid)
- What is the best point forecast?

- The density may help: Gamma (shape 3, scale 3)

- Option 1: the expectation ($\hat{y} = 9.00$)
- Option 2: the median ($\hat{y} = 8.02$)



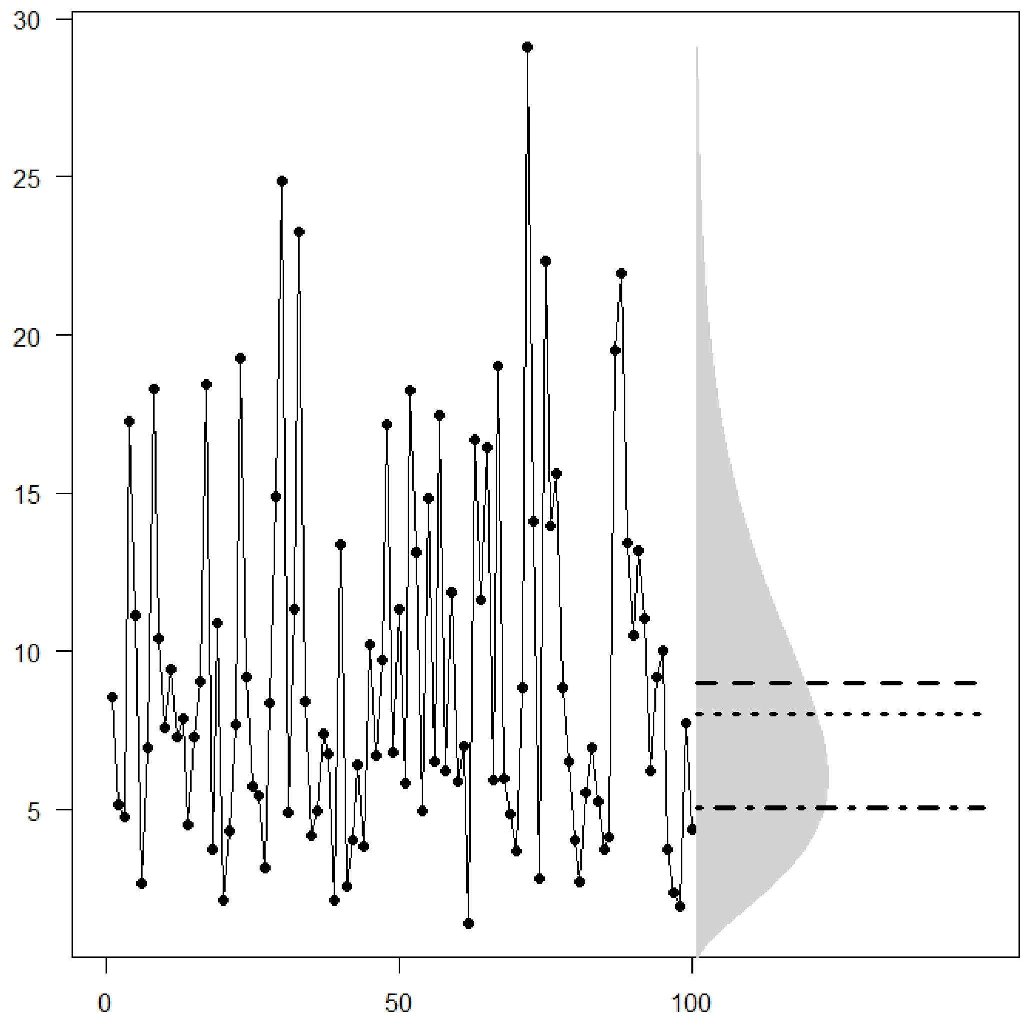
Kolassa, S. Why the “best” point forecast depends on the error or accuracy measure (invited commentary on the M4 Competition) *International Journal of Forecasting,* **2020**, *36*, 208-211

The “best” point forecast depends on the error or accuracy measure

- Simulated data – independent & identically distributed (iid)
- What is the best point forecast?

- The density may help: Gamma (shape 3, scale 3)

- Option 1: the expectation ($\hat{y} = 9.00$)
- Option 2: the median ($\hat{y} = 8.02$)
- Option 3: the (-1)-median ($\hat{y} = 5.04$)



Kolassa, S. Why the “best” point forecast depends on the error or accuracy measure (invited commentary on the M4 Competition) *International Journal of Forecasting*, **2020**, *36*, 208-211

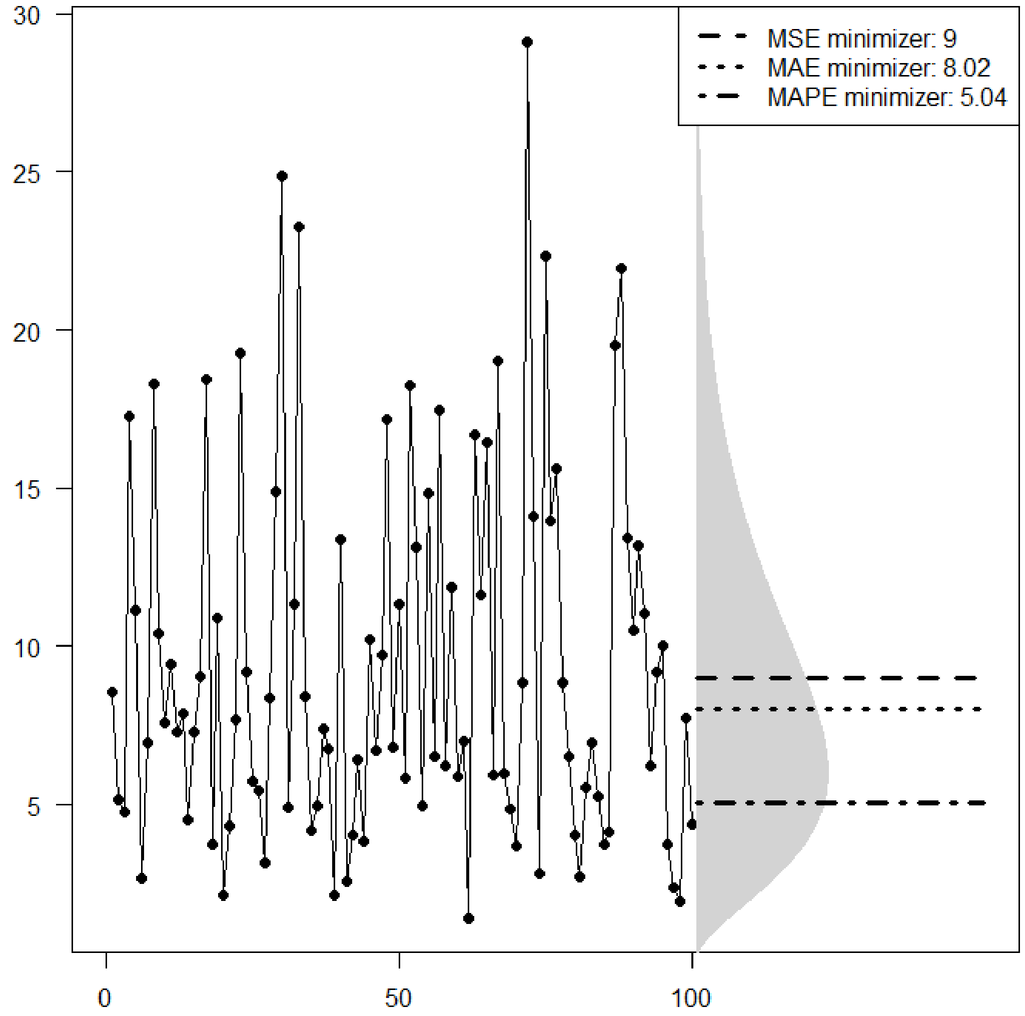
The “best” point forecast depends on the error or accuracy measure

- Simulated data – independent & identically distributed (iid)
- What is the best point forecast?

- The density may help: Gamma (shape 3, scale 3)

- Option 1: the expectation ($\hat{y} = 9.00$) → **minimizes MSE**
- Option 2: the median ($\hat{y} = 8.02$) → **minimizes MAE/wMAPE**
- Option 3: the (-1)-median ($\hat{y} = 5.04$) → **minimizes MAPE**

- → The “best” point forecast depends on the error or accuracy measure

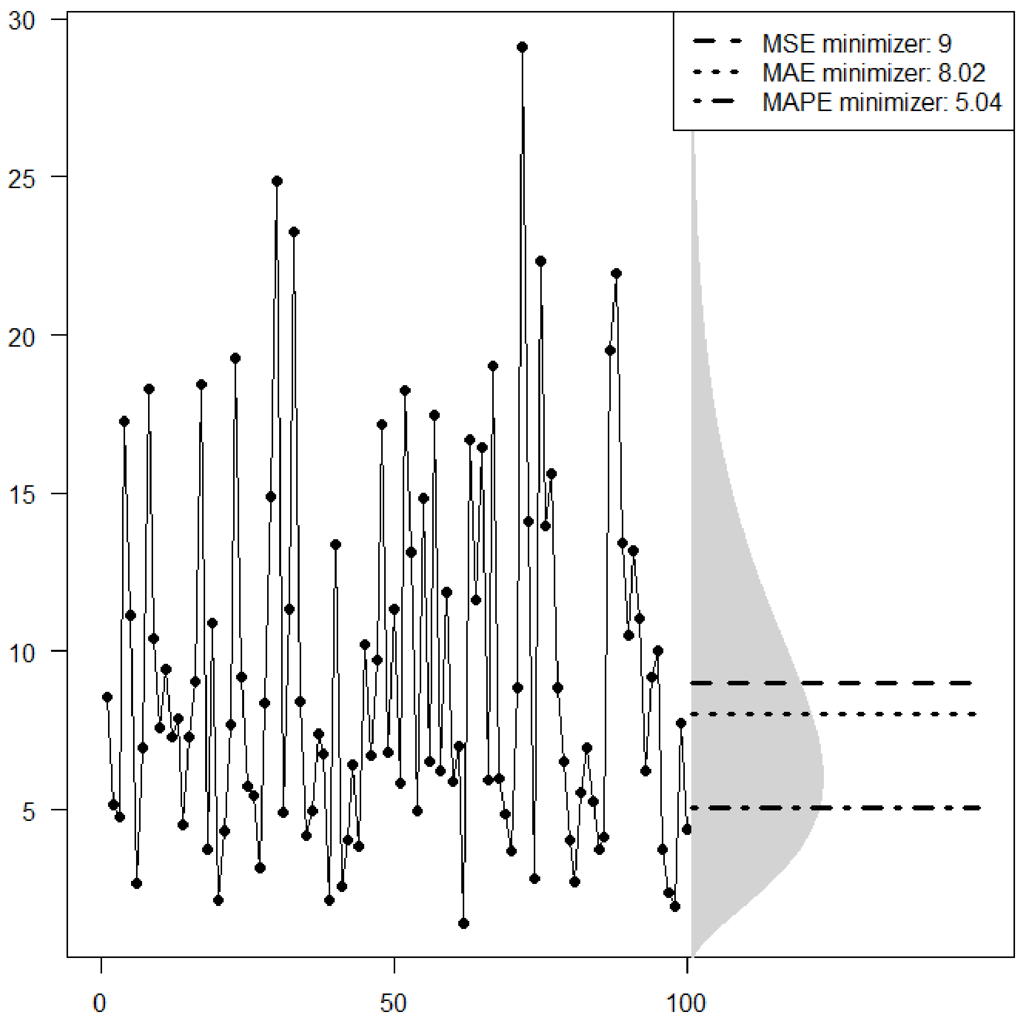


Kolassa, S. Why the “best” point forecast depends on the error or accuracy measure (invited commentary on the M4 Competition) *International Journal of Forecasting*, **2020**, *36*, 208-211

The “best” point forecast depends on the error or accuracy measure

“Interesting” consequences

- Typically, forecasting methods aim for an unbiased expectation forecast…
- … i.e., an MSE minimizing forecast
- Can we meaningfully evaluate this point forecast using the MAPE? Or even with multiple *different* error measures?
  – No!

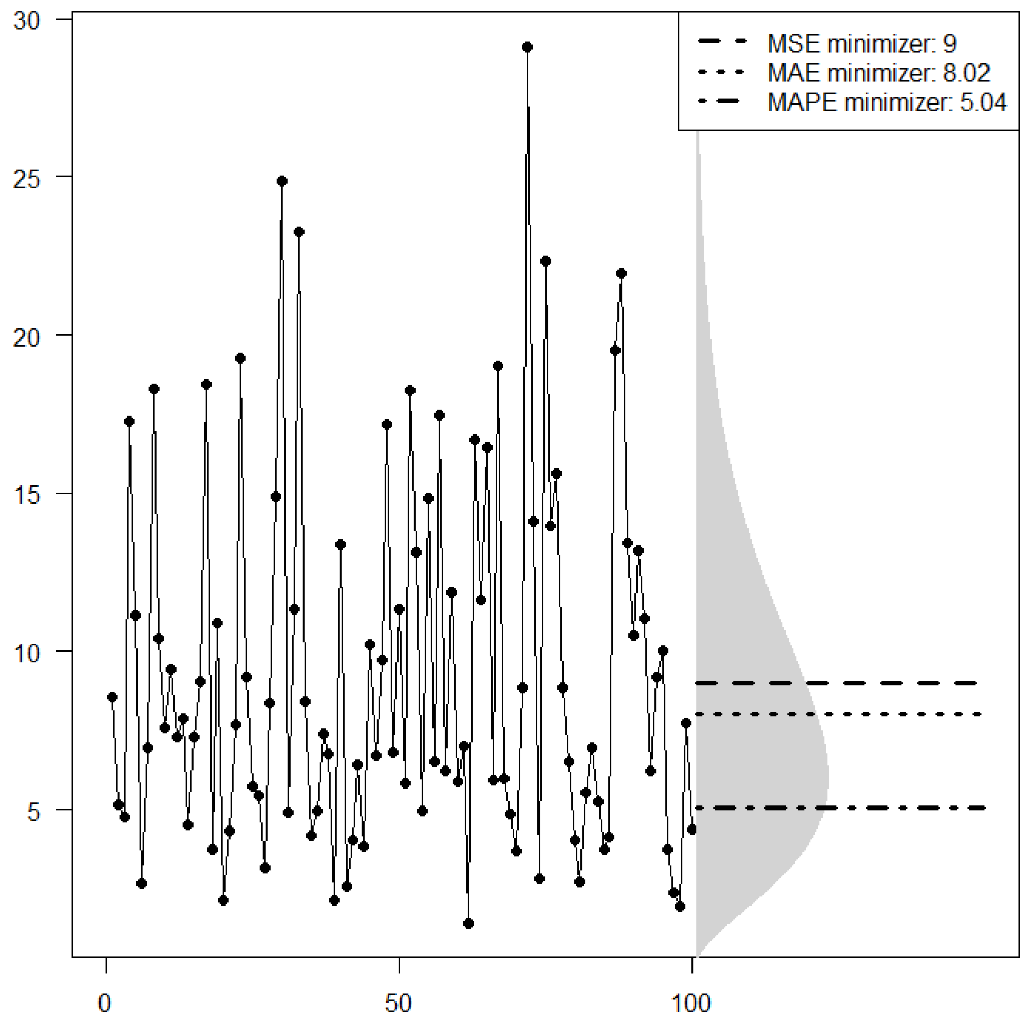


Kolassa, S. Why the “best” point forecast depends on the error or accuracy measure (invited commentary on the M4 Competition) *International Journal of Forecasting,* **2020**, *36*, 208-211

The “best” point forecast depends on the error or accuracy measure

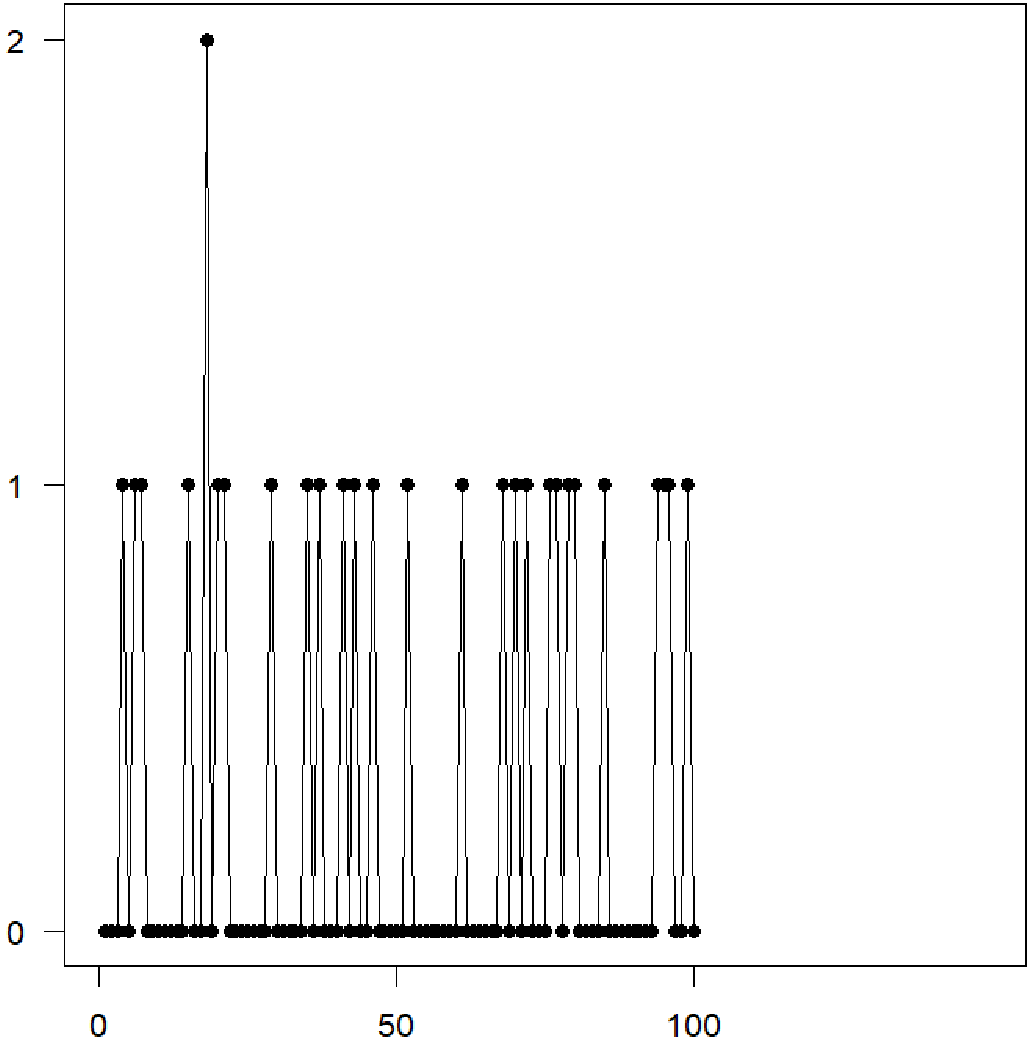
The dirty little secret

- If your forecasting software aims for an unbiased expectation forecast…
- … but your bonus depends on getting a low MAPE…
- … then just reduce the software-generated forecast by a few percent
- (Also, explain to your manager why tying your bonus to MAPE is not a good idea)



Kolassa, S. Why the “best” point forecast depends on the error or accuracy measure (invited commentary on the M4 Competition) *International Journal of Forecasting,* **2020**, *36*, 208-211

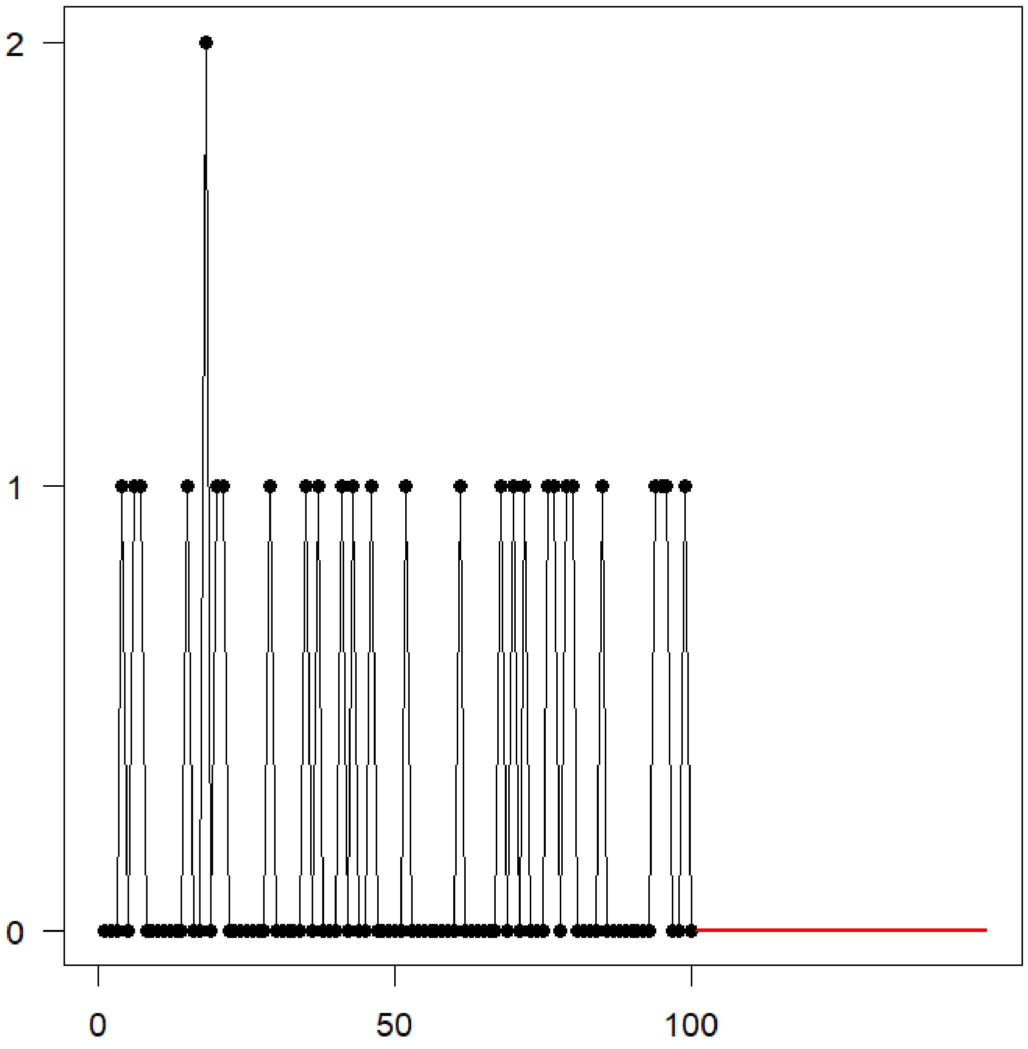
A little test



- Independent identically distributed *intermittent* data – most observations are zero

- What do you forecast to minimize the expected MAE?

A little test



- Independent identically distributed *intermittent* data – most observations are zero
- What do you forecast to minimize the expected MAE?
- Recall that the expected MAE is minimized by the median of the distribution…
- … if *most observations are zero*, the median is zero…
- … so forecast a flat zero!
- Same for wMAPE and MASE (which are just scaled MAEs)

Quantile forecast PFEMs

- Quantile forecasts for a level α: a forecast $\hat{y}_t^{(\alpha)}$ such that $P\left(y_t \leq \hat{y}_t^{(\alpha)}\right) = \alpha$, or equivalently $F\left(\hat{y}_t^{(\alpha)}\right) = \alpha$ for the CDF F (which is unknown)

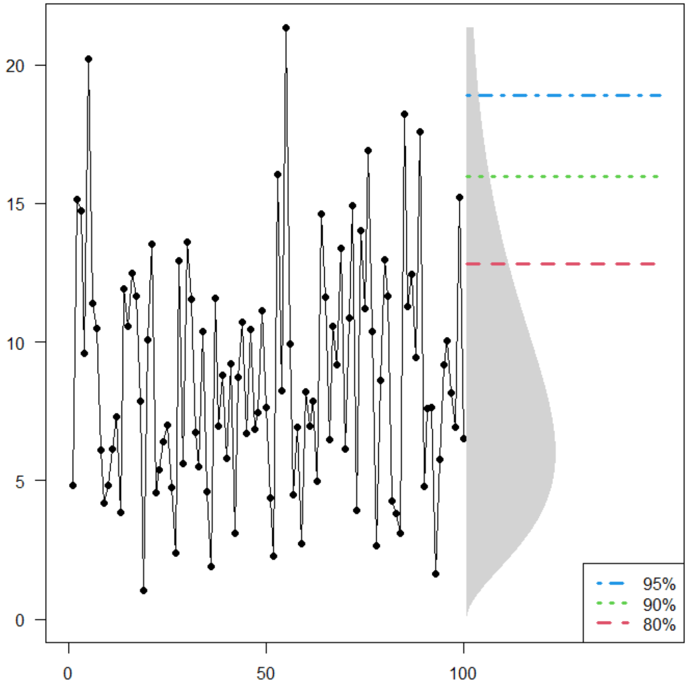
- Extremely important for safety stock calculation or other capacity planning

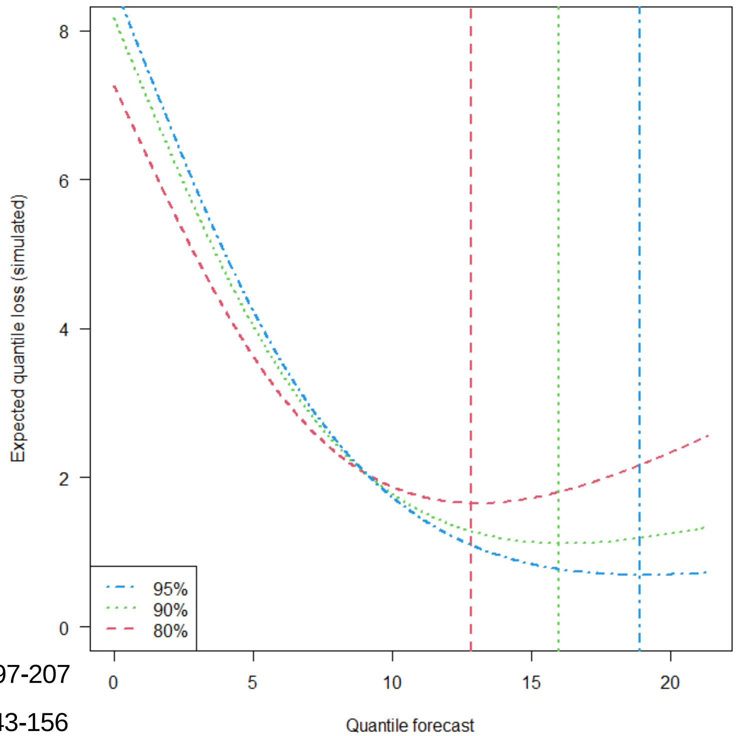
- Use quantile loss (Gneiting 2011), as in quantile regression (Koenker & Hallock, 2001):

$$L_\alpha(y, \hat{y}) \coloneqq \rho_\alpha(y - \hat{y})$$

$$\rho_\alpha(x) = \begin{cases} \alpha x, & x \geq 0 \\ (\alpha - 1)x, & x < 0 \end{cases}$$

- Note flatness at the minimum!





Gneiting, T. Quantiles as optimal point forecasts. *International Journal of Forecasting*, **2011**, *27*, 197-207

Koenker, R. & Hallock, K. F. Quantile Regression. *Journal of Economic Perspectives*, **2001**, *15*, 143-156

Point forecast error measures (PFEMs)

Takeaways

- Tailor your PFEM to the business problem
- Know which PFEM your software tries to minimize
- Tailor your point forecasts to the PFEM you want to minimize
- Do not evaluate the *same* point forecast using *different* PFEMs
- Ideally: forecast full predictive densities, then extract PFEM-optimal point forecasts
- Gently educate users and clients about potential pitfalls in inappropriate PFEMs

Agenda

Point (including quantile) forecast evaluation

Significance checking

Accuracy benchmarks

Is the difference between errors/scores statistically significant?

Comparing *two* forecasts

- Consider two sets of point/interval/density forecasts A and B with associated errors/scores $\left(S_i^A\right)$ and $\left(S_i^B\right)$ for $i = 1, \ldots, n$

- The scores will differ. How do we assess whether the better one is not only better by chance?

- The Diebold-Mariano test (Gneiting & Katzfuss, 2014, and elsewhere): under quite weak conditions,

$$z \coloneqq \sqrt{n}\,\frac{\frac{1}{n}\sum_{i=1}^{n}\left(S_i^A - S_i^B\right)}{\hat{\sigma}} \sim N(0,1), \text{where } \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(S_i^A - S_i^B\right)^2$$

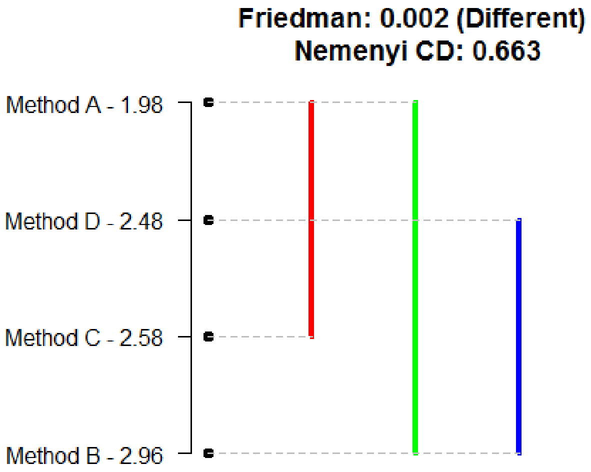
- Assess this using your favorite normality test (e.g., Shapiro-Wilk)

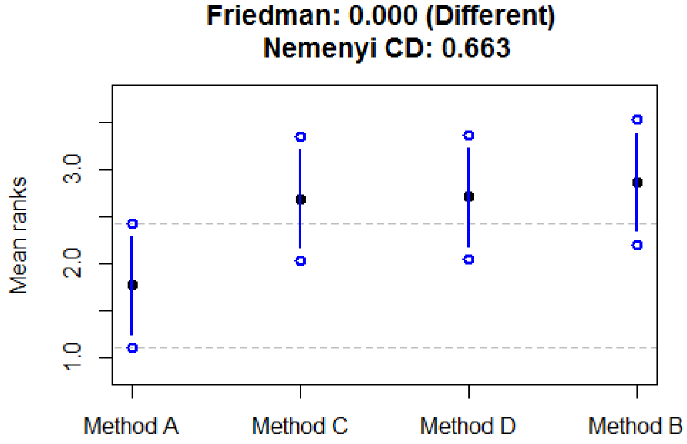
Gneiting, T. & Katzfuss, M. Probabilistic Forecasting. *Annual Review of Statistics and Its Application*, **2014**, *1*, 125-151

Is the difference between errors/scores statistically significant?

Comparing *multiple* forecasts

- We will typically have *multiple* forecasting methods (and multiple time series, and multiple evaluation time periods) – possible tests, all rank-based and usable for any PFEM:
  - – Multiple Comparisons with the Best (MCB) test (Koning et al., 2005)
  - – Same, against the average method (ANOM; also Koning et al., 2005)
  - – Friedman-Nemenyi test (Demšar, 2006)
- Empirical comparison by Hibon et al. (2012)
- Implemented in the tsutils package for R on CRAN
- All of this still only assesses *statistical significance*, not *business relevance*





Demšar, J. Statistical Comparisons of Classifiers over Multiple Data Sets. *Journal of Machine Learning Research, JMLR.org,* **2006**, *7*, 1-30

Hibon, M.; Crone, S. & Kourentzes, N. Statistical Significance of Forecasting Methods: An Empirical Evaluation of the Robustness and Interpretability of the MCB, ANOM and Nemenyi Test. International Symposium on Forecasting, **2012**

Koning, A. J.; Franses, P. H.; Hibon, M. & Stekler, H. O. The M3 competition: Statistical tests of the results. *International Journal of Forecasting,* **2005**, *21*, 397-409

Agenda

Point (including quantile) forecast evaluation

Significance checking

Accuracy benchmarks

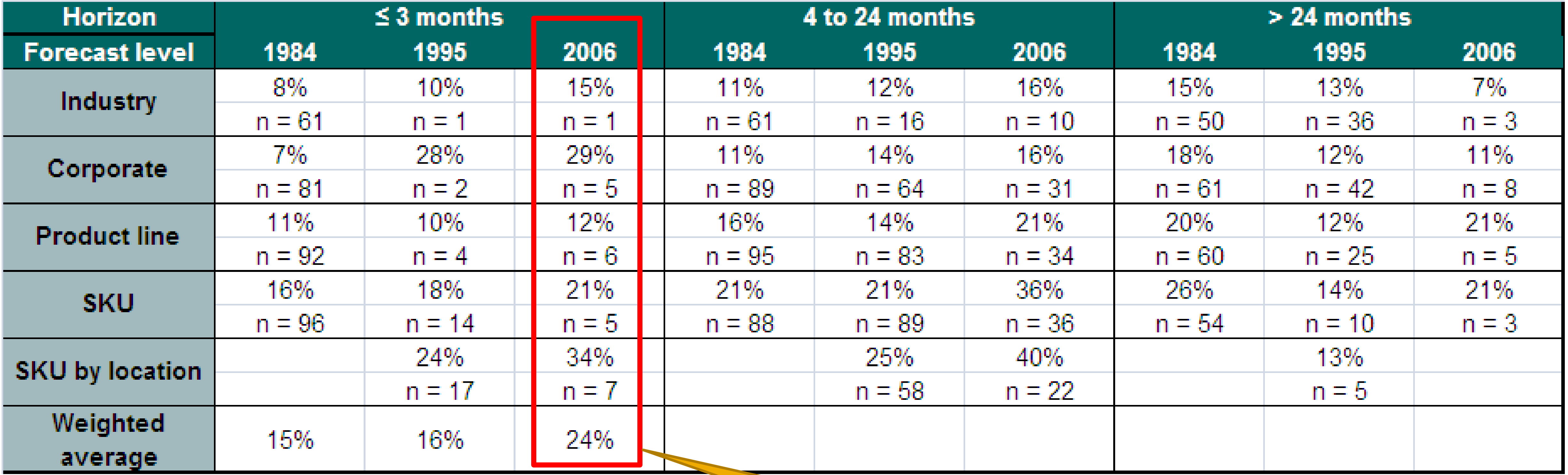
Accuracy benchmarks

- Often customers confront you with the following:
  - “We know that the industry average MAPE is 30%, so you need to achieve at least 25%.”
  - “This project needs to reach a MAPE of 20% to continue funding.”
  - “Can you guarantee that your software will yield 15% MAPE?”

- There are two possible reactions to that:
  - Smile and agree (…to get the deal)
  - Tell the customer that such an answer cannot be made without further investigation of the KPIs (… and risk losing the contract)

- Why is that so? Let’s look at one example of benchmarks!

Accuracy benchmarks

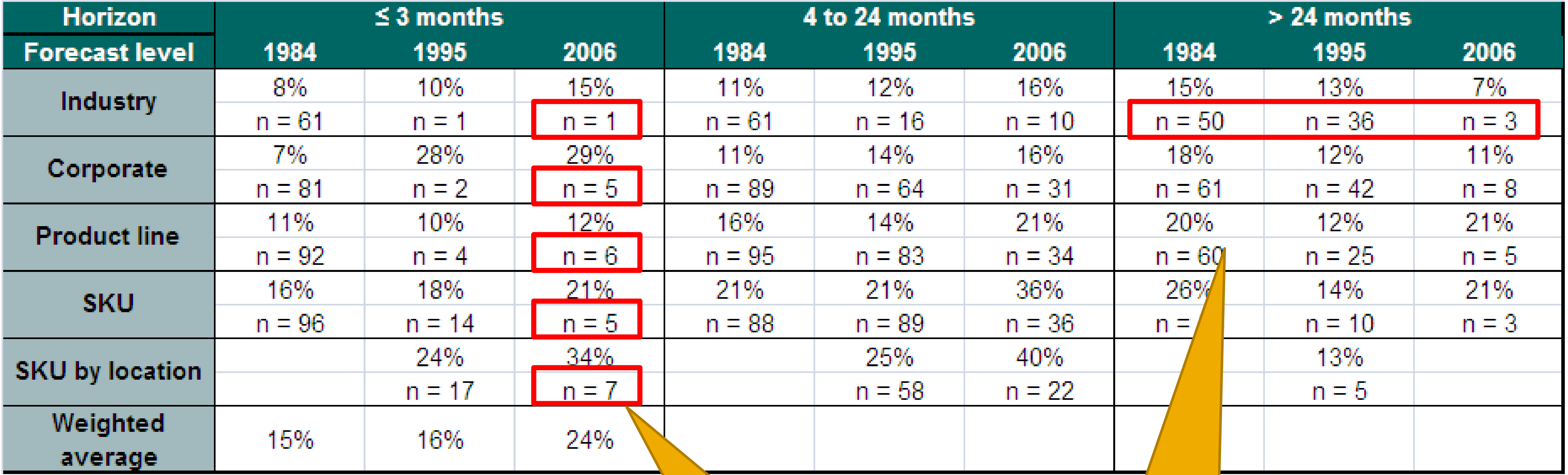


Horizon	≤ 3 months			4 to 24 months			> 24 months		
Forecast level	1984	1995	2006	1984	1995	2006	1984	1995	2006
Industry	8% n = 61	10% n = 1	15% n = 1	11% n = 61	12% n = 16	16% n = 10	15% n = 50	13% n = 36	7% n = 3
Corporate	7% n = 81	28% n = 2	29% n = 5	11% n = 89	14% n = 64	16% n = 31	18% n = 61	12% n = 42	11% n = 8
Product line	11% n = 92	10% n = 4	12% n = 6	16% n = 95	14% n = 83	21% n = 34	20% n = 60	12% n = 25	21% n = 5
SKU	16% n = 96	18% n = 14	21% n = 5	21% n = 88	21% n = 89	36% n = 36	26% n = 54	14% n = 10	21% n = 3
SKU by location		24% n = 17	34% n = 7		25% n = 58	40% n = 22		13% n = 5	
Weighted average	15%	16%	24%						

Can we use this column as meaningful benchmarks?

McCarthy, T. M.; Davis, D. F.; Golicic, S. L. & Mentzer, J. T. The Evolution of Sales Forecasting Management: A 20-Year Longitudinal Study of Forecasting Practices. *Journal of Forecasting,* **2006**, *25*, 303-324

Accuracy benchmarks



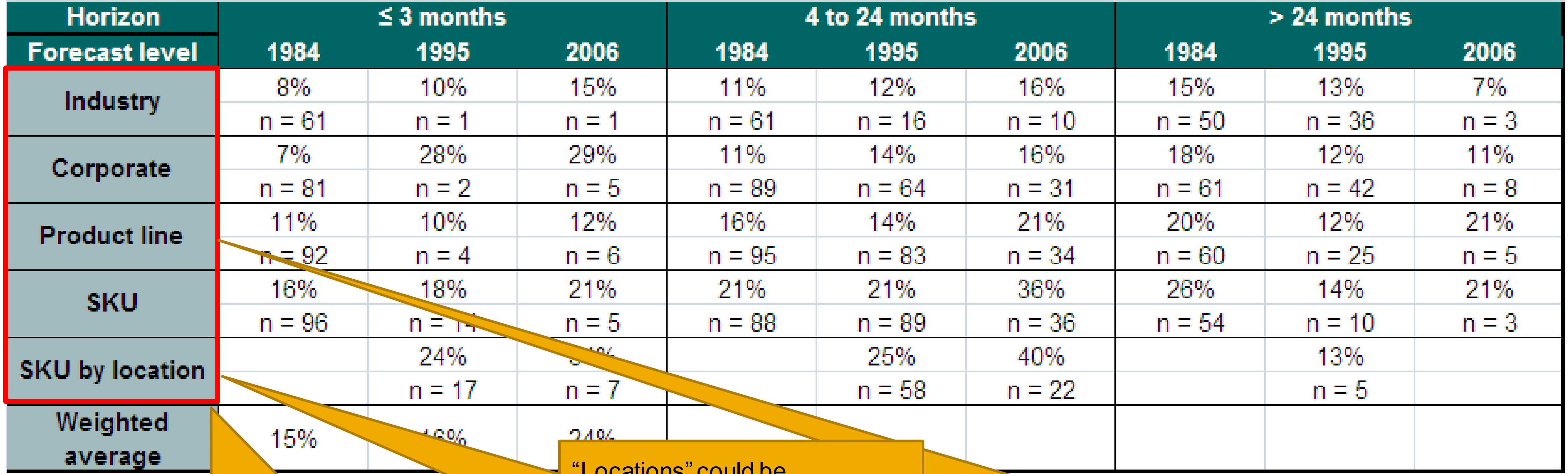
Horizon	≤ 3 months			4 to 24 months			> 24 months		
Forecast level	**1984**	**1995**	**2006**	**1984**	**1995**	**2006**	**1984**	**1995**	**2006**
Industry	8%	10%	15%	11%	12%	16%	15%	13%	7%
	n = 61	n = 1	n = 1	n = 61	n = 16	n = 10	n = 50	n = 36	n = 3
Corporate	7%	28%	29%	11%	14%	16%	18%	12%	11%
	n = 81	n = 2	n = 5	n = 89	n = 64	n = 31	n = 61	n = 42	n = 8
Product line	11%	10%	12%	16%	14%	21%	20%	12%	21%
	n = 92	n = 4	n = 6	n = 95	n = 83	n = 34	n = 60	n = 25	n = 5
SKU	16%	18%	21%	21%	21%	36%	26%	14%	21%
	n = 96	n = 14	n = 5	n = 88	n = 89	n = 36	n = [illegible]	n = 10	n = 3
SKU by location		24%	34%		25%	40%		13%	
		n = 17	n = 7		n = 58	n = 22		n = 5	
Weighted average	15%	16%	24%						

Very few respondents…

… and the number of respondents decreases over time

McCarthy, T. M.; Davis, D. F.; Golicic, S. L. & Mentzer, J. T. The Evolution of Sales Forecasting Management: A 20-Year Longitudinal Study of Forecasting Practices. *Journal of Forecasting,* **2006**, *25*, 303-324

Accuracy benchmarks



Horizon	≤ 3 months			4 to 24 months			> 24 months		
Forecast level	1984	1995	2006	1984	1995	2006	1984	1995	2006
Industry	8% n = 61	10% n = 1	15% n = 1	11% n = 61	12% n = 16	16% n = 10	15% n = 50	13% n = 36	7% n = 3
Corporate	7% n = 81	28% n = 2	29% n = 5	11% n = 89	14% n = 64	16% n = 31	18% n = 61	12% n = 42	11% n = 8
Product line	11% n = 92	10% n = 4	12% n = 6	16% n = 95	14% n = 83	21% n = 34	20% n = 60	12% n = 25	21% n = 5
SKU	16% n = 96	18% n = 14	21% n = 5	21% n = 88	21% n = 89	36% n = 36	26% n = 54	14% n = 10	21% n = 3
SKU by location		24% n = 17	[illegible] n = 7		25% n = 58	40% n = 22		13% n = 5	
Weighted average	15%	16%	24%						

Do we all mean the same thing here?

“Locations” could be

- Retail stores
- Distribution centers (DCs)
- Entire geographies

What about other channels?

Same problem with “Product lines”

McCarthy, T. M.; Davis, D. F.; Golicic, S. L. & Mentzer, J. T. The Evolution of Sales Forecasting Management: A 20-Year Longitudinal Study of Forecasting Practices. *Journal of Forecasting,* **2006**, *25*, 303-324

Accuracy benchmarks

Takeaways

- Put not your trust in published benchmarks of forecast accuracy (Kolassa, 2008)…
  - … for they suffer from a number of problems…
  - … most importantly a lack of comparability

- Better: use *internal* benchmarks
  - Forecasting methods currently used
  - Very simple benchmarks like the overall mean (which may be surprisingly hard to beat; Andy T, 2014)

- Best: look for best practices in forecasting!
  - Applying best practices will not ensure better accuracy – but it will help
  - If your practices & processes are already optimal, don't worry about external benchmarks

Andy T. Is it unusual for the MEAN to outperform ARIMA? https://stats.stackexchange.com/q/124955/1352, **2014**

Kolassa, S. Can we obtain valid benchmarks from published surveys of forecast accuracy? *Foresight: The International Journal of Applied Forecasting,* **2008**, 6-14

Questions & Answers

Accuracy budgets

- With forecast accuracy, every year we always heard about so what will be your budget for next year forecast accuracy? What will be your approach to this question? Is looking at the product coeff variation to calculate the budget/target fc accuracy make sense?

- There is a limit to achievable accuracy (try forecasting a coin toss with higher than 50% accuracy!)
- I would try not to commit to a specific accuracy target (output side), rather explain which avenues I would pursue to improve accuracy (input side):
  - Analyze bad forecasts
  - Improve data cleansing
  - Possibly try other models (better data usually beats better models: https://stats.stackexchange.com/a/355395/1352)
- In parallel: try to educate people responsible for my targets

Recommended metrics

- Regarding accuracy benchmarks in the Business Context: what metric would you propose

- RMSE if my goal is unbiased expectation forecasts; hinge loss if it is quantile forecasts
- In each case, “benchmarking” should be done internally, not externally: compare the method under investigation to a simple benchmark, or to the method currently in use

Time dynamics of accuracy

- For non-stationary time series, I've found it useful to produce an estimate of the forecasting error as a function of time. This is because the accuracy of the model may change over time. Is this something that clients find useful?

- Definitely yes!
- For instance, in retail, it makes sense to compare forecast accuracy separately in promotions and for regular sales
- Alternatively, the error may be influenced by seasonality
- One main driver is that time series fluctuations typically happen *both* in the expectation and the variance – so higher means go with higher variances, which in turn increases error

Dealing with inflated accuracy expectations

- As a data scientist internally at a company, how do you deal with "marketing" from vendors ("improved accuracy up to 50%!") - convincing leadership that there are so many caveats around those improved accuracy claims?

- This is hard, because everyone wants to believe that the silver bullet has arrived
- As the data scientist:
  - Look into the claims
  - Point out caveats and weaknesses in the claims
  - But make sure to also point out that you are ready and willing to cooperate, not obstructing on principle
  - Establish your own credentials so people will trust your expertise (https://datascience.stackexchange.com/a/2406/2853)
- In my experience, executives usually didn’t rise by swallowing marketing claims whole – perhaps remind them of similar claims they have more experience with (e.g., on project durations, or ROI of a project…)
- In the end, we all need to collect our own experience with over-promises and under-deliveries

The MAPE and positive data

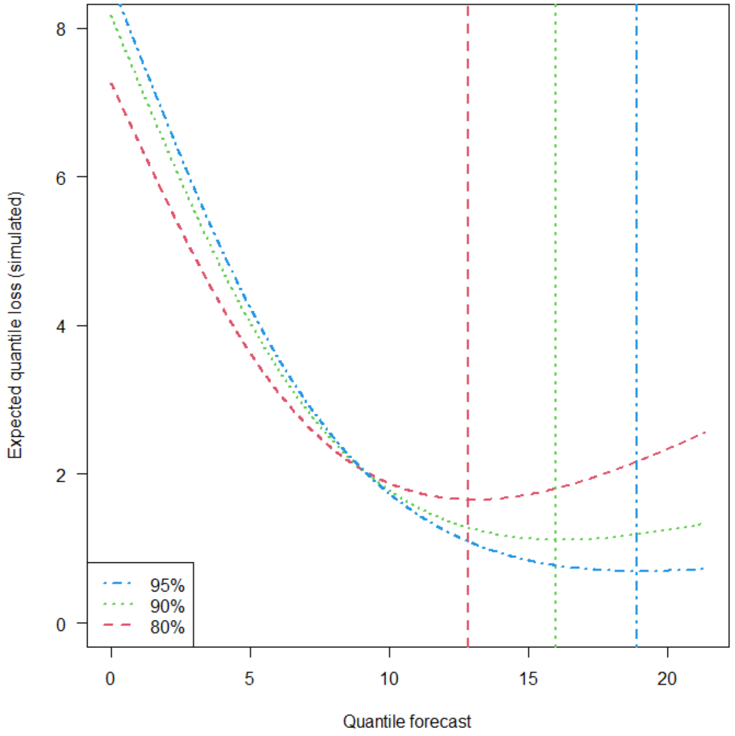
- Why MAPE only makes sense for strictly positive data?

- If just one of the actuals is zero, then we are dividing by zero, which is mathematically undefined
- Calculating $\frac{|\hat{y}-y|}{y}$ for $y < 0$ gives a negative APE – how to interpret?
- In principle, $\left|\frac{\hat{y}-y}{y}\right|$ is positive even if $y < 0$, so this can be calculated – but what does it tell us?
- Actually, other problems with the MAPE are IMO much more serious: https://stats.stackexchange.com/q/299712/1352

Quantile forecasts and small data

- Pinball assumes a lot of data. What to do on small samples?

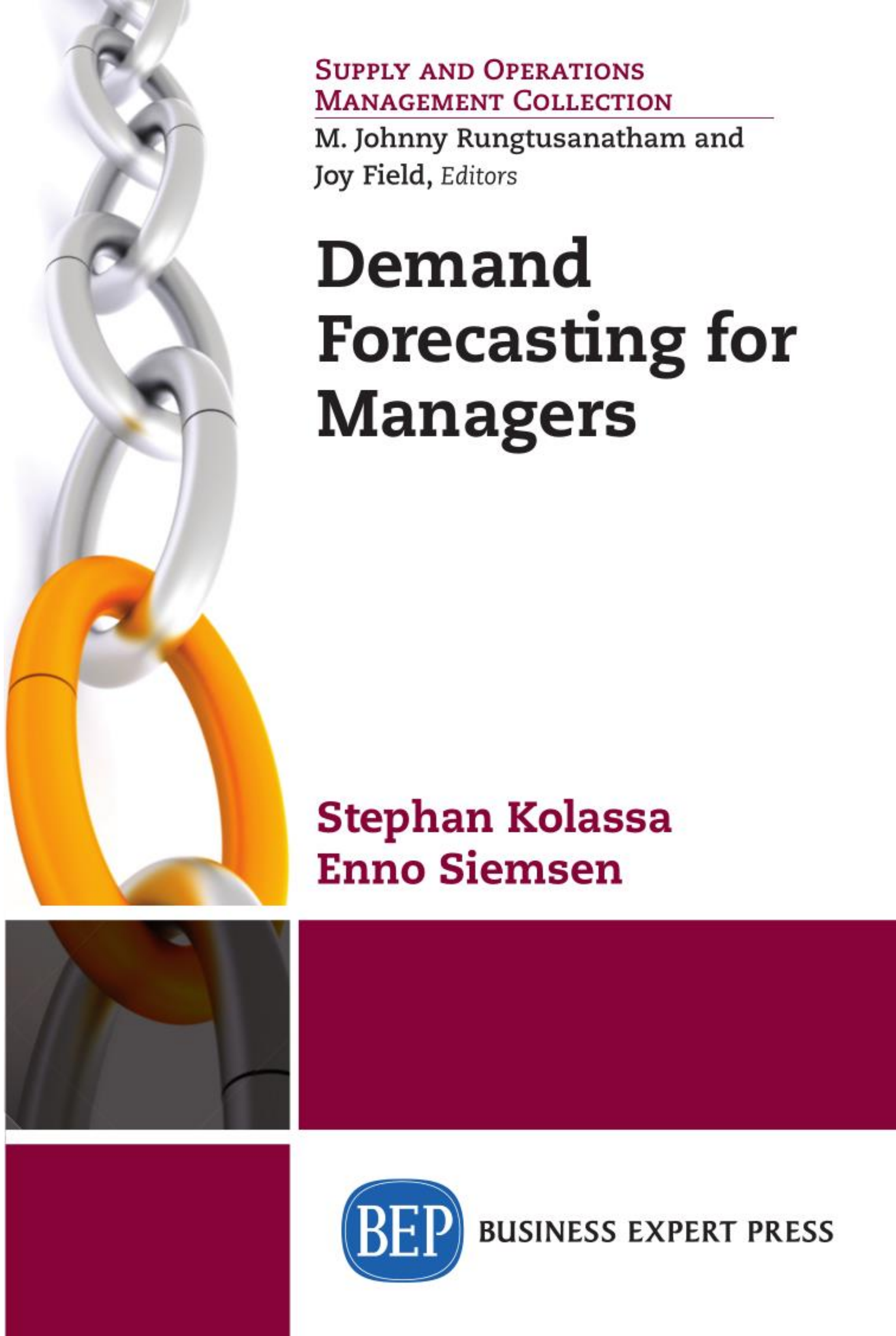
- Actually, I wouldn’t say it’s the *pinball loss* that assumes a lot of data – it’s *tail quantile forecasts* that assume a lot of data!
  - If you are shooting for a 95% quantile forecast, then only 5% of your observations should exceed it
  - But anything *below* your quantile forecast doesn’t really give you a lot of information!
  - This is actually the flatness of the pinball loss curve at its minimum!
- Try collecting more data
  - Longer holdout horizons
  - More time series
  - Lower granularity
- There is little else we can do – we can’t conjure data out of thin air!



Thank you.

Contact information:

Stephan Kolassa
Data Science Expert
Stephan.Kolassa@sap.com



SUPPLY AND OPERATIONS
MANAGEMENT COLLECTION
M. Johnny Rungtusanatham and
Joy Field, *Editors*

Demand Forecasting for Managers

Stephan Kolassa
Enno Siemsen

BEP BUSINESS EXPERT PRESS

Lancaster University
Management School

Follow us

www.sap.com/contactsap

THE BEST RUN SAP