Evaluating forecasts and accuracy metrics

Nikolaos Kourentzes

Skövde Artificial Intelligence Lab

Skövde University, Sweden

sail.his.se

18 June 2024

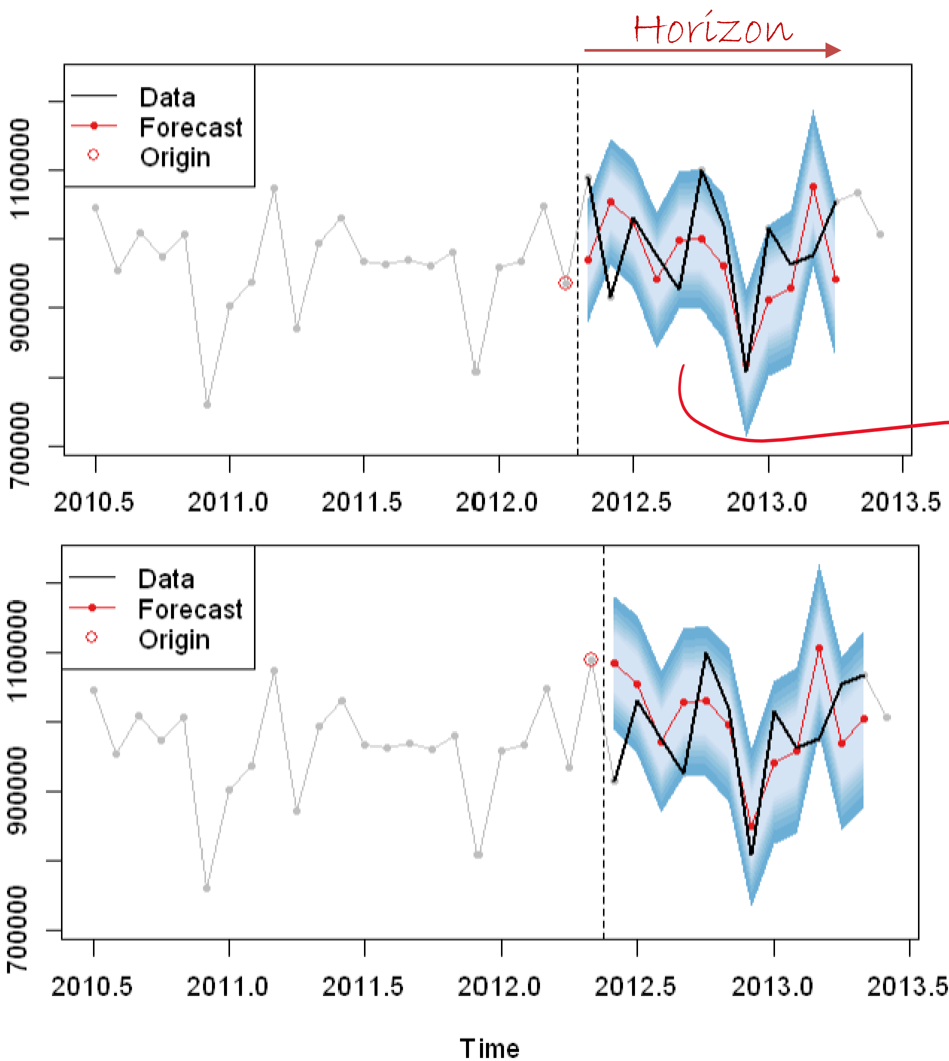
Some starting points

1. **Forecasting is not the end-target!** Forecasting error metrics are convenient and helpful, but likewise not the complete story.
    - The **supported decisions may transform the forecasts** in ways that make closely following the data only a part of what matters (e.g., in inventory we first check how much we have in stock before we order)
    - However, just **not evaluating forecasts would be an excuse**. Beyond any challenges, there is a **problem of attribution**. How do we find how much value is added by the forecasts, if they are to be transformed when used?
    - In practice many firms strive for accurate forecasts, but the supported decisions would not change by much by more accurate forecasts. This is due to simplistic decision-making heuristics.

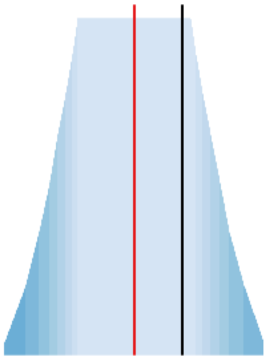
2. **There is no best error metric!** The application context drives the choice of the metric.
    - Nonetheless, there are some metrics we can let them rest in peace.

What is a forecast?

To design metrics, we need to know what we are measuring



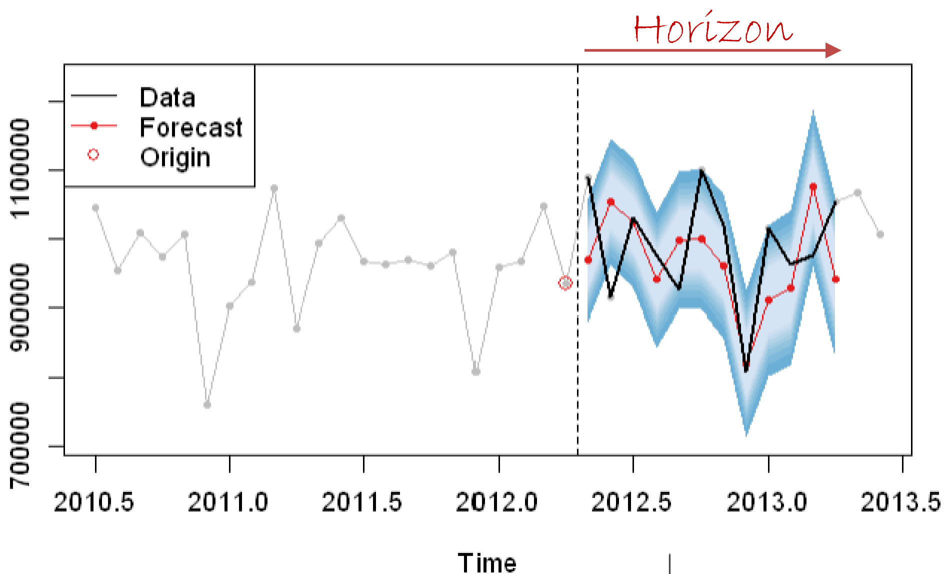
- Point (mean) forecast
- Origin → Target dates
- Distribution
- Horizon



- Distance
- Bias
- “Coverage”

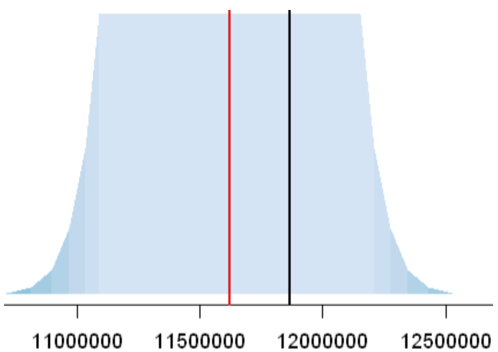
What is a forecast?

To design metrics, we need to know what we are measuring

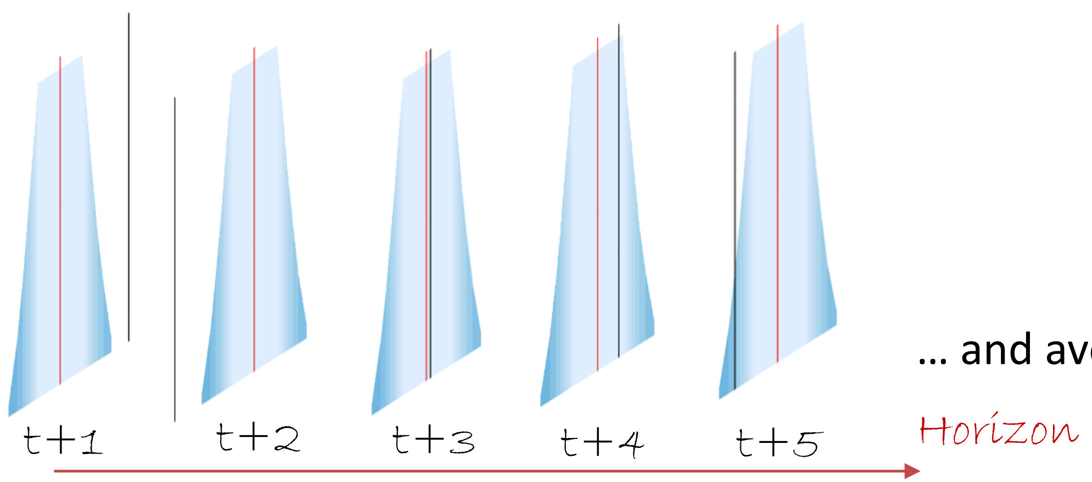


or

Sum across horizon
e.g., inventory



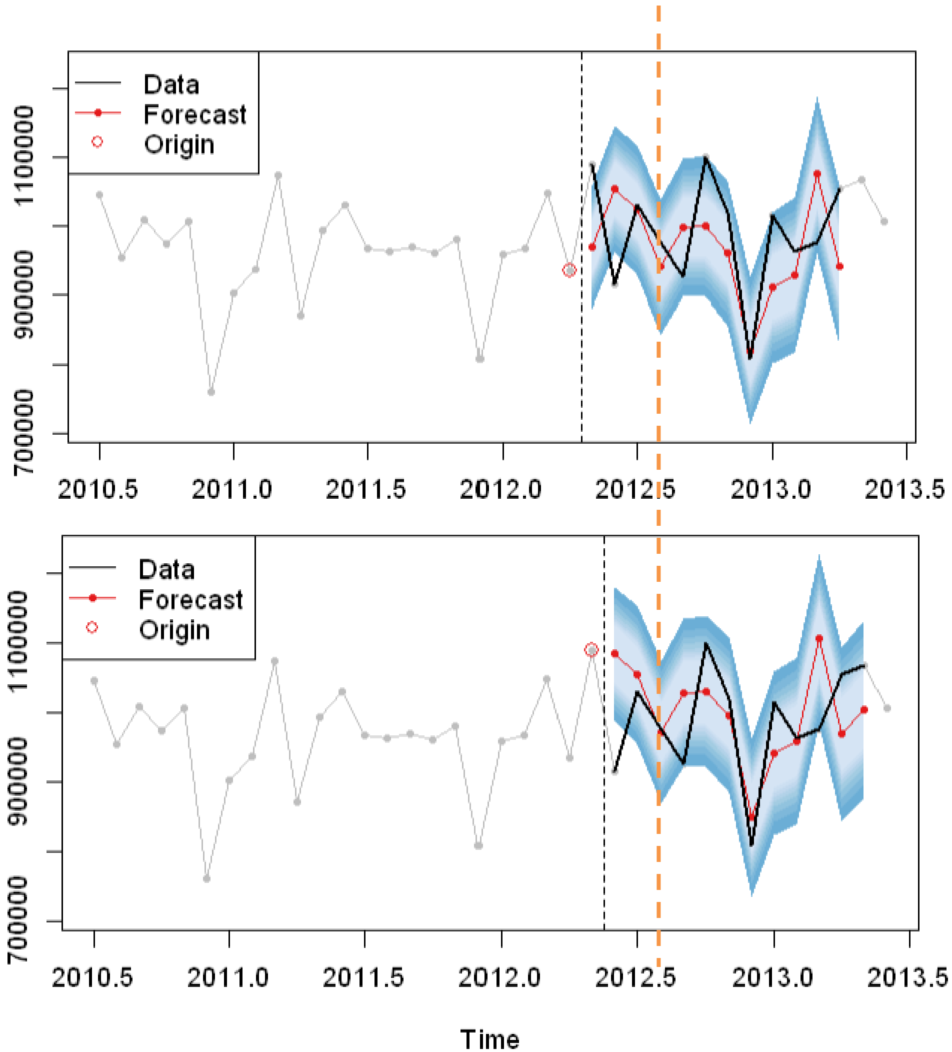
Per horizon

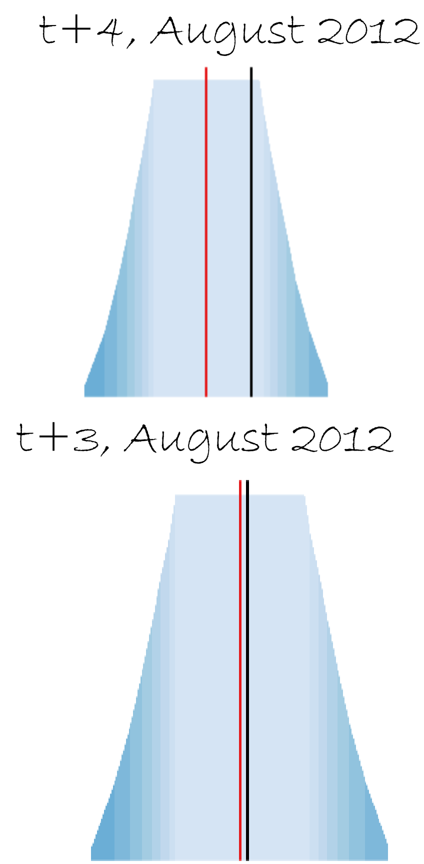


… and average

What is a forecast?

To design metrics, we need to know what we are measuring





Now the focus is on a specific period and not horizon.

What about the measuring stick?

Given some collections of errors e_i

- Use **as is** → track the **bias**, if you are over or under-forecasting
- Take the **square** → track how well you are getting the **mean** of the target
- Take the **absolute** → track how well you are getting the **median** of the target

At school you may have done some physics. **Things have units**!

- 98 ice creams (sold) – 100 ice creams (predicted) = -2 ice creams (I will eat them)
- e^2 = (ice creams) $^{\mathbf{2}}$ (take the square root if this matters)

What about the measuring stick?

What about the distribution?

- What do you want to measure from the distribution?
  - Quantile? Pinball loss (or similar)
  - Interval? Mean Interval Score (or similar – coverage errs towards wider!)
  - The whole thing? You got options! (see Gneiting & Raftery, 2007)

- Are costs symmetric?
  - Going over or under the quantile, is it equally costly?
    - We can talk about bias in quantiles as well.
    - Similarly, we have quantiles (absolute) and expectiles (quadratic).
    - Same building blocks to get your metrics.

And what about summarising?

If we mix errors from different distributions (series, horizons, aggregation, etc.) we need to use **scale independent errors**

- Suppose we sell affogato (coffee and ice cream)

We have:

- 50 ice cream scoops RMSE
- 10 coffee pours RMSE
- The direct average would be 30 half-ice cream scoop half-coffee chimeric measurements. This does not work.

We can normalize in various ways:

- Divide by the mean (it also carries units and the scale, but assumes stationarity)
- Divide by the standard deviation (or equivalent - also carries units! And normalises scale, think of z-score) – I see scaled errors, like MASE, RMSSE in this category.
- Relative errors, per period, or on summary errors.
- Percentage errors  (a.k.a. the easy review comment)

Percentage errors

The aim is to remove scale and units. We do that by dividing each error by its respective observation.

- Easy to calculate (careful of zeros)
- Easy to misunderstand (negative and positive errors are not weighted equally, **never use for bias**)

What about MAPE – should we stop?
(Mean Absolute Percentage Error)

For academic work, yes. There is no reason, use other normalization approaches.

For practice it is more complicated

- Is it intuitive? Yes, but misunderstood
- Is it connected to the decision? Probably not
- Is it putting a heavy focus on point predictions? Yes
- Is it as critical as climate change? No

How to summarise?

That should be easy!

- Arithmetic average
- Geometric mean – use when summarise ratios (or simply, is 0 or 1 the “neutral number”?)

How to take these averages?

- What is your objective? What do you want to measure?
- Careful about mixing sample sizes (e.g., fewer long-term forecasts)
- Careful about mixing difficulty of forecasts (e.g., short and long term forecasts)

Should we use weighted averages? (e.g., wMAPE)

- Weigh statistically (normalization?)
- Based on the application → we will return to this

How many errors?

Observations are comprised by **structure** and **noise**.

For simplicity let us assume additive errors:

$$\underbrace{y_t = \mu_t}_{\text{Structure}} + \underbrace{\varepsilon_t}_{\text{Noise}}$$

$$e_t = y_t - \hat{y}_t$$
$$e_t = (\mu_t + \varepsilon_t) - \hat{y}_t$$
$$e_t = (\mu_t - \hat{y}_t) + \varepsilon_t$$

Which part is bigger, given that the true ε_t is unknown?

We expect ε_t to over periods to cancel out, so as long as we summarise over enough periods we should reduce the contribution of ε_t and measure reliably $\mu_t - \hat{y}_t$.

- Not all averages do this (over what are you averaging?)!
- Also consider what it means to average across horizons.

Thoughts on some common errors

Mean Error – ME and MsE

- Do not forget **bias**! It can be quite important on some applications. Bias can be calculated on mean, quantiles, etc.
- Difficult to make scale independent. Two favourites:
  - Relative bias to a benchmark or current process
  - ME/scaling factor; my go to scaling factor is: mean($|y_t - y_{t-1}|$) – think of MASE.

AMsE – Absolute Mean scaled Error

- Calculate your MsE and then take the absolute. We calculate the **magnitude of bias – helpful for summarizing across series, so that bias does not vanish.**

Thoughts on some common errors

RMSSE, MASE

- My go to scale independent errors
- We scale by $s^2 = \text{mean}((y_t - y_{t-1})^2)$ and $s = \text{mean}(|y_t - y_{t-1}|)$ to match the loss order of the numerator.
- For me, these belong to the same family of MSE/σ^2 and MAE/σ.
  - σ assumes stationarity of the series. s does not. In principle you should model select, but since this is the measuring stick, err towards s.
  - This makes the horizon question of the denominator mute, and the interpretation simply becomes "normalized errors".

(see Athanasopoulos & Kourentzes, 2023)

I use this scaling liberally to make things scale independent – just be careful to match the loss order of the numerator!

  - e.g., Pinball/s

Thoughts on some common errors

Percentage errors

- I never use these – but I do not reject papers that do
- I try to avoid doing consultancy on it – so far so good!
- If I find it in use, I do not attack it, just put a large sticker over it that it does not mean what people think!
- sMAPE (symmetric MAPE) is immediate rejection and pills for my pressure
- MPE (Mean Percentage Error) is major revision on a sunny day
- wMAPE, unless your weighting scheme is very carefully thought out goes in the sMAPE bucket for me. You can manipulate the reported values.

It is 2024 – if it is an abstract academic work about "insert your model here" then report on the quality of probabilistic forecasts!

(if you read this slide in the future, yes, it is that many years ago and we still talk about MAPE)

General framing for applications

Forecasts are not the end target; how close can you get to the decision?

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

Why are you measuring errors?

- A fun activity for the family!
- To estimate parameters
- To model/method select
- To calibrate a process, e.g., estimate empirical quantiles
- To control a process, e.g., manage forecasts by exception

Can change the appropriate metric

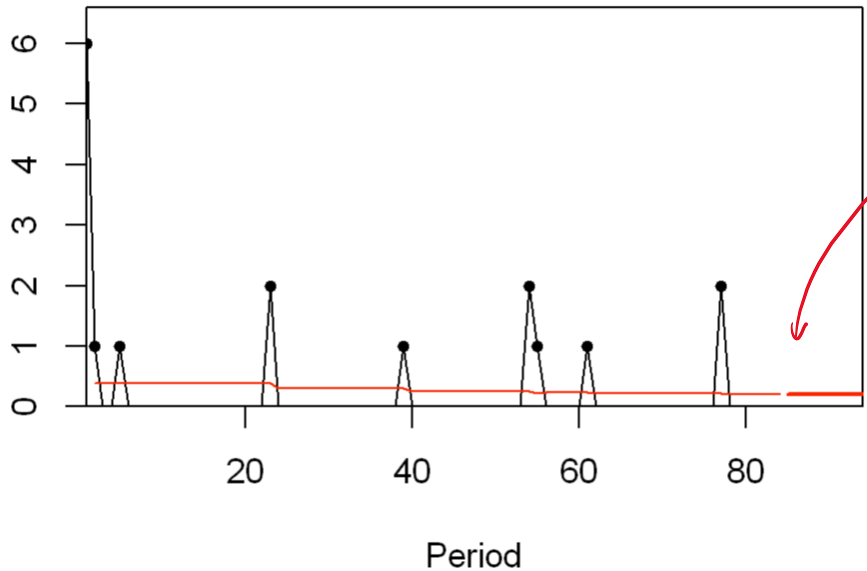
Who are the stakeholders?

Are they responsible for a single decision (use of forecasts) or multiple?

Thoughts on some “common” applications

Intermittent demand

- It has zeros! Metrics that fail on zeros will fail here. These are most errors that normalize per period.
- Absolute errors track the median of the target distribution. For intermittent demand this may well be zero. What is the value of a zero forecast?



What are the units of this forecast? If Croston’s and the like it is not ice creams ☹

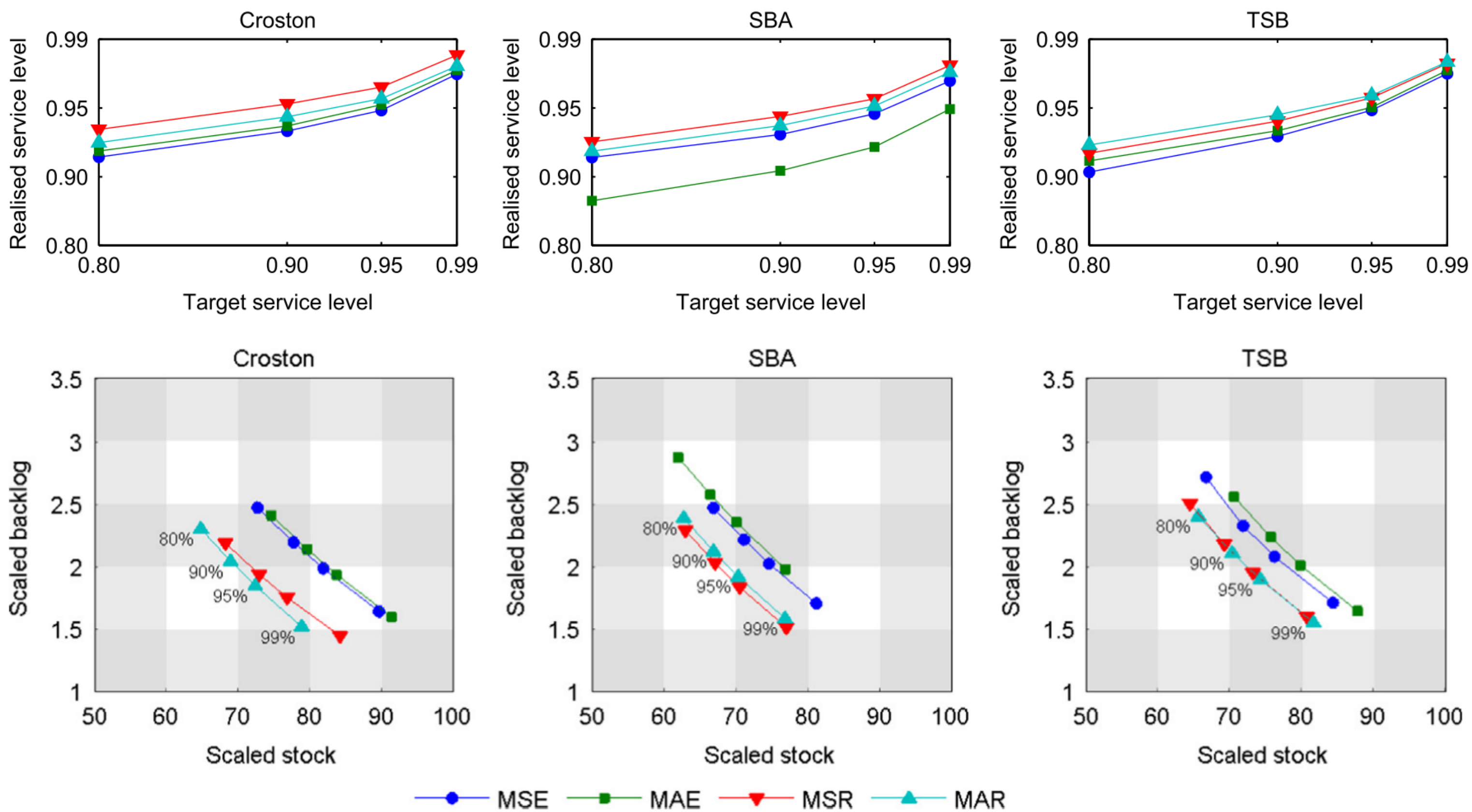
- Is it the accuracy per period (hard) that we care about, or the quantile over the lead-time demand? (easy-ish)

Thoughts on some "common" applications

(see Kourentzes, 2014)

Intermittent demand

- Better yet, if you are doing forecasts for inventory simulate the effect

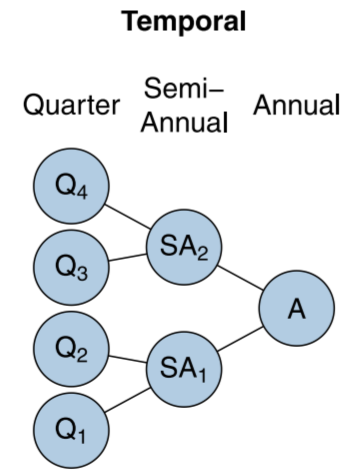


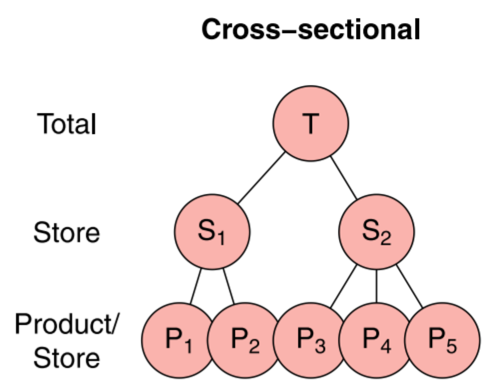
Don't tell anyone, but that paper uses MASE to evaluate intermittent demand forecasts…

Thoughts on some “common” applications

Hierarchical forecasting

- Most levels are “statistical devices”
  - Should we measure accuracy? An average error across levels seems to be popular. Careful of usefulness and scaling issues!
- Objective: different error metric per level/planning objective?

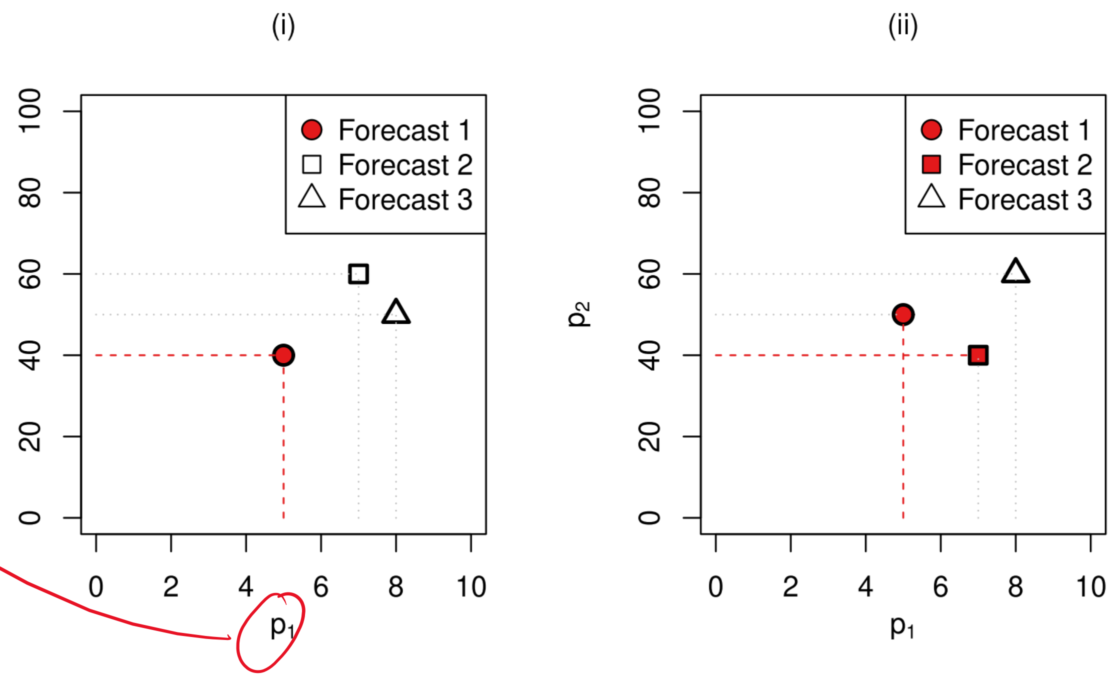




T/Q_j	T/SA_j	T/A
S_i/Q_j	S_i/SA_j	S_i/A
P_i/Q_j	P_i/SA_j	P_i/A

A potential way forward: multi-objective evaluation!

Different performance metrics across objectives/levels



i. Dominant forecast, irrespective of objective/level

ii. Partially dominant, need to weight performance metric

- Weighted averages are linearisations of (ii) but assume a common metric

- p_1 and p_2 do not have to be forecasting metrics.

(see Athanasopoulos & Kourentzes, 2023)

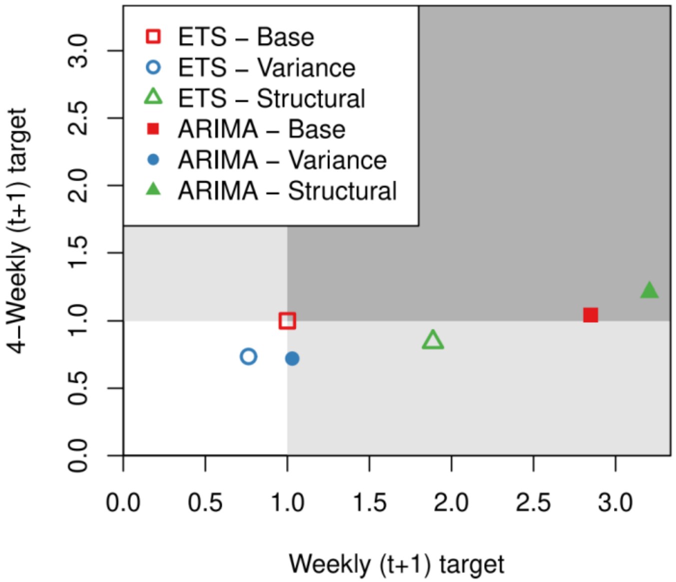
A potential way forward: multi-objective evaluation!

Example hierarchies of decisions in organisations.

Decision	Frequency	Time series	Output
	Call centre (Koole & Li, 2021)		
Budget planning	Quarterly	Monthly+	Budget
Capacity planning	Monthly	Weekly+	Training and hiring plans
Operational planning	Weekly	Weekly	Outsourced call volume
Scheduling	Weekly	Daily+	Agent schedules per type
Scheduling	Hourly	Intra-daily	Adaptations to schedules
	Tech manufacturer*		
Financial planning	Yearly	Quarterly+	High-level financial goals
Annual operations plan	Yearly	Monthly+	Resource allocation
Production planning	Monthly	Monthly	Aggregate demand planning
Master production plan	Weekly	Weekly	Detailed demand planning
Material planning	Weekly	Weekly	Supply requirements

* sourced from interviews; '+' series can be recorded at a higher aggregation level.

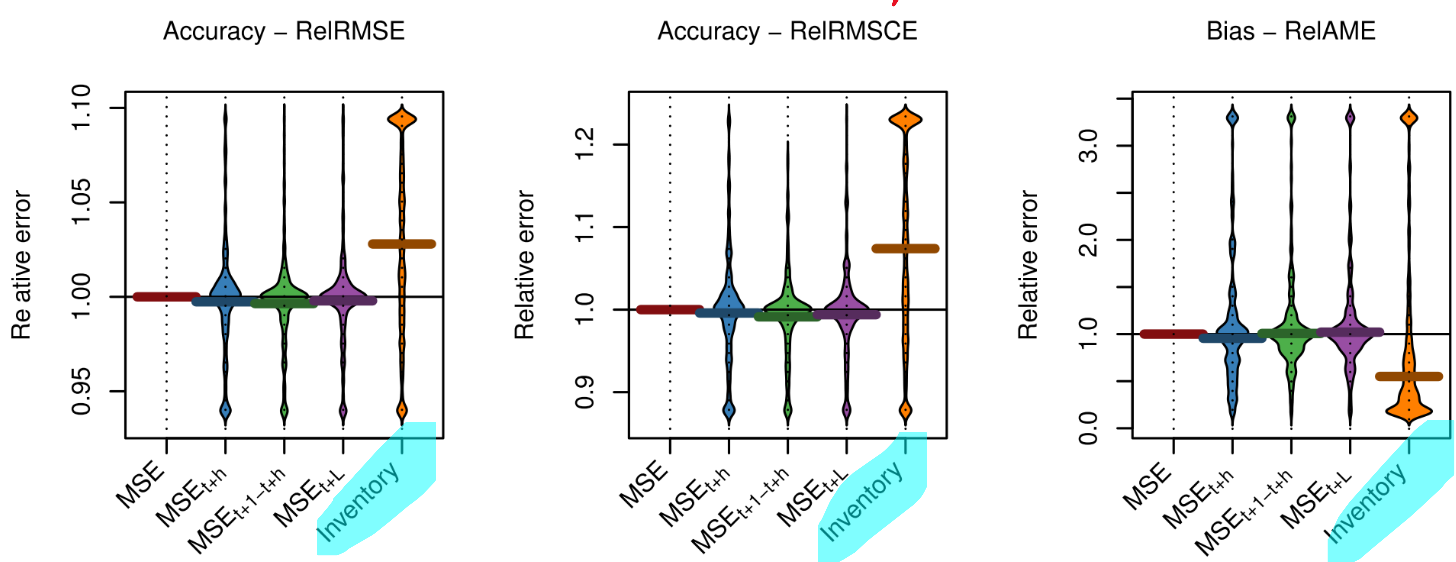
Even if we cannot find a dominant forecast, we can find dominant approaches ("variance" in this case)

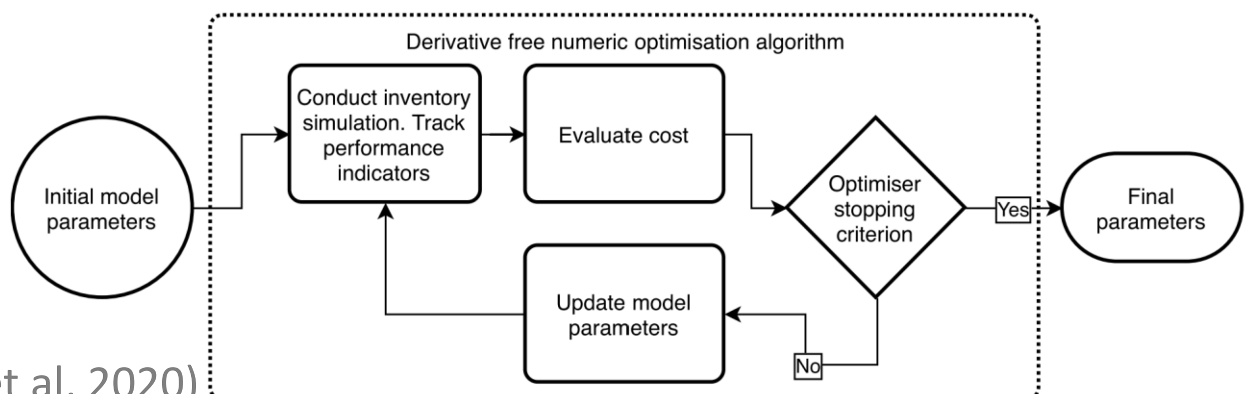


Does accuracy tell the full story?

Forecasts on firm data

C for cumulative over the lead-time

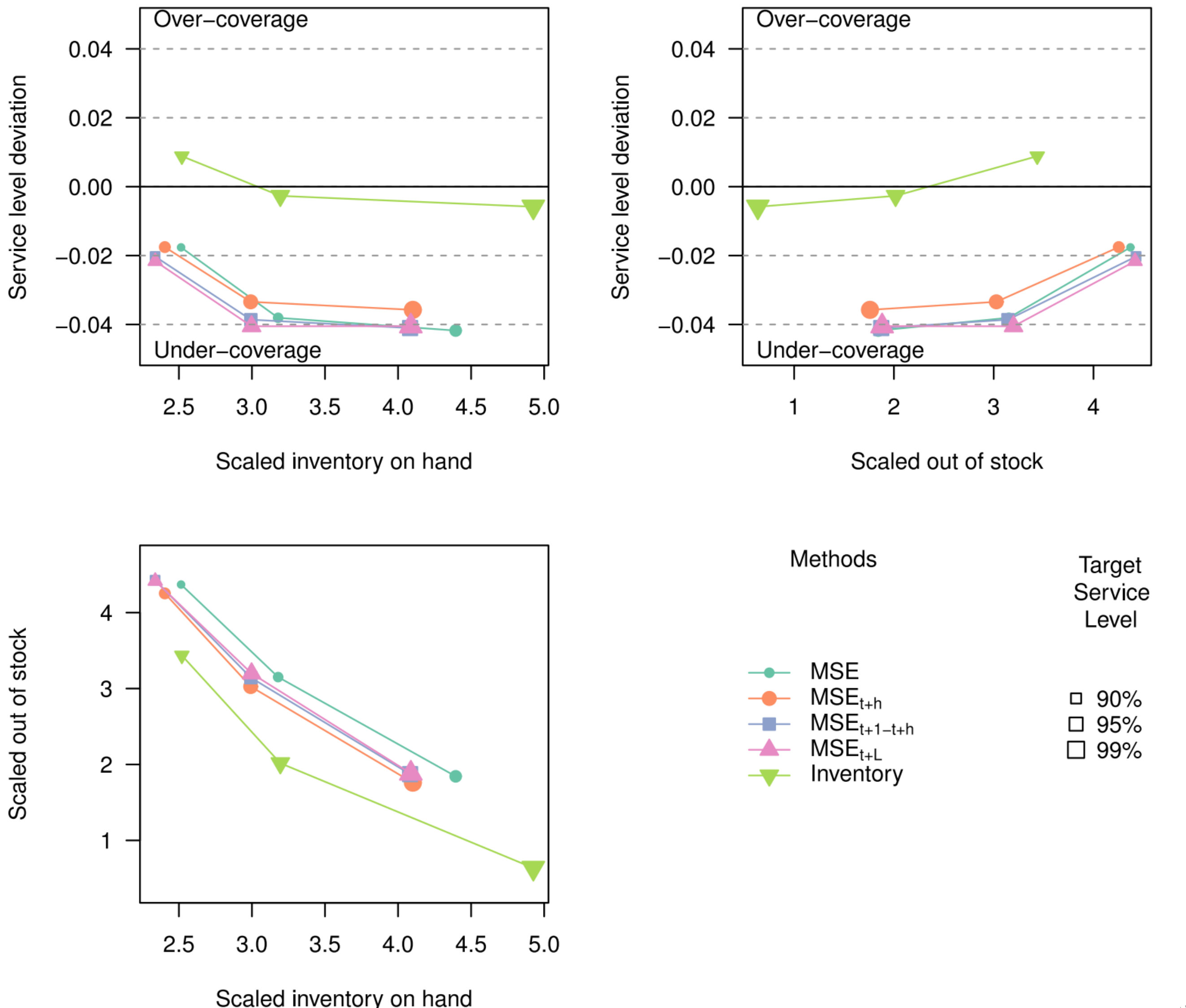




(see Kourentzes et al, 2020)

Does accuracy tell the full story?

Inventory performance on firm data

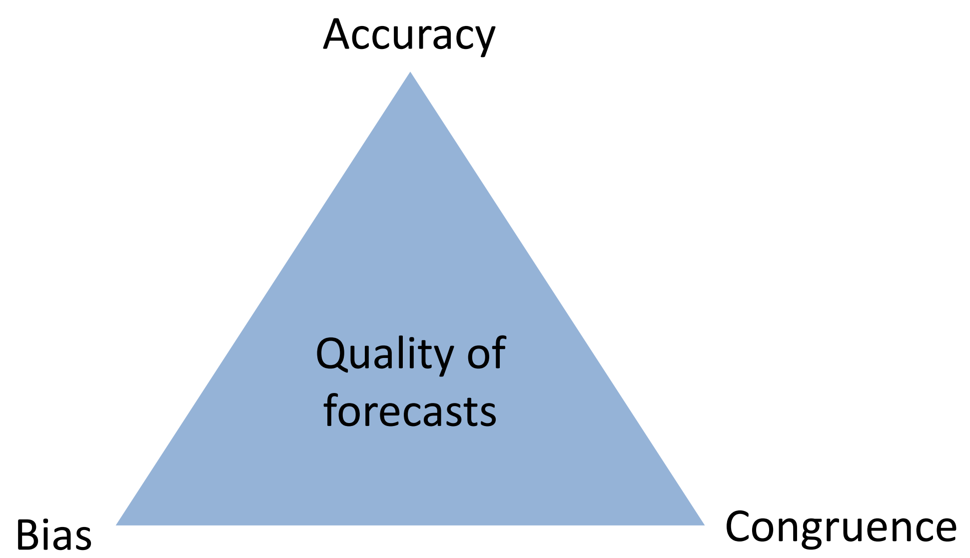


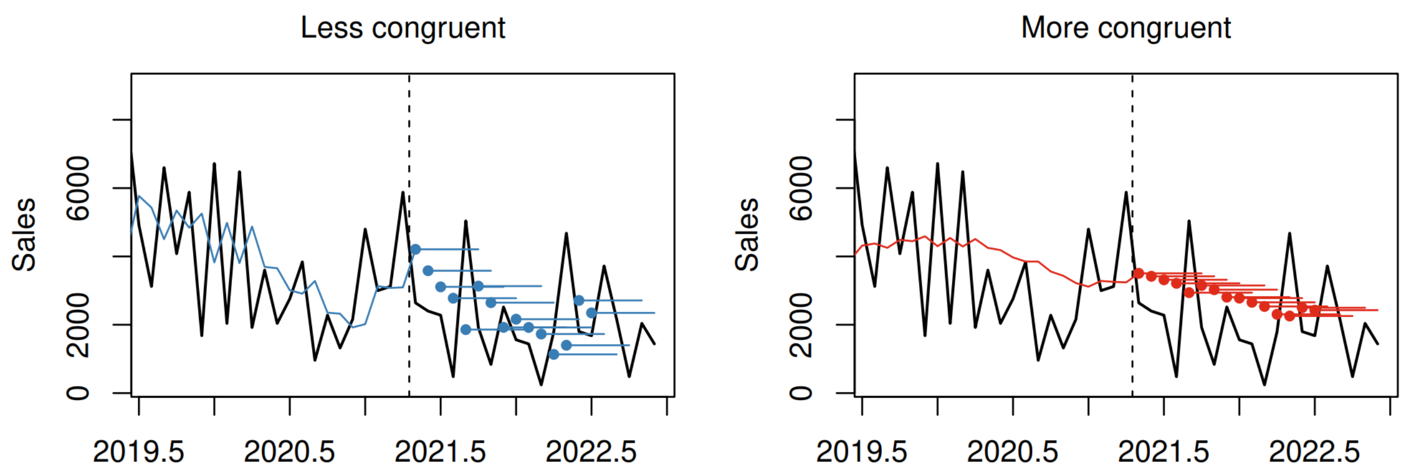
Should the metric much model loss?

If we are interested in the performance of spherical ETS in vacuum, sure!

- There are some cases we can design loss functions that much the use of the forecasts → evaluate on that as a metric!
- Often, it is difficult to get a clean loss function that matches the use case.
  - Matching loss & metric can lead to “overfit” solutions
  - Diversifying can lead to “regularized” solutions
    - And that is why climate change ranks higher than MAPE for me!
- Do we stop measuring forecast accuracy in some form?
  - I don’t think so:
    1. Issue of attribution, how do you measure the value of forecasts, how do you provide feedback to models/humans.
    2. Process optimization, there are steps in the forecasting process that add value, and they need a clean signal to improve – the complete decision obfuscates the signal.
    3. An argument suggests to not forecast!? It obfuscates the issue for me.

Some “out there” ideas





Congruence

Firms attempt to decrease the jitteriness of the forecasts in various ways

Forecasting model review interval	Responses	Are forecasts adjusted to be more 'stable' over time?	Responses
Every time	71.43%	No	19.05%
Longer review intervals	28.57%	Yes, in an ad-hoc manner	33.33%
		Yes, rule-based changes	47.62%

The argument is that it reduces supply chain stress and improves trust among supply chain members.

Congruence

$$\begin{array}{ccc}
 & \begin{array}{ccccc} h=1 & h=2 & \cdots & h=l-1 & h=l \end{array} & \\
\mathbf{e} = & \begin{pmatrix}
e_{1+l|l} & e_{1+l|l-1} & \cdots & e_{1+l|1+l-(l-1)} & e_{1+l|1} \\
e_{2+l|1+l} & e_{2+l|l} & \cdots & e_{2+l|2+l-(l-1)} & e_{1+l|2} \\
\vdots & \ddots & & & \vdots \\
 & & e_{t|t-h} & & \\
\vdots & & & \ddots & \vdots \\
e_{n-1|n-2} & e_{n-1|n-3} & \cdots & e_{n-1|n-1-(l-1)} & e_{n-1|n-1-l} \\
e_{n|n-1} & e_{n|n-2} & \cdots & e_{n|n-(l-1)} & e_{n|n-l}
\end{pmatrix} & \begin{matrix} t=l+1 \\ t=l+2 \\ \\ \vdots \\ \\ t=n-1 \\ t=n \end{matrix}
\end{array}$$

$$\begin{array}{cc}
\begin{pmatrix}
e_{1+l|l} & e_{1+l|l-1} & \cdots & e_{1+l|1+l-(l-1)} & e_{1+l|1} \\
e_{2+l|1+l} & e_{2+l|l} & \cdots & e_{2+l|2+l-(l-1)} & e_{1+l|2} \\
\vdots & \ddots & & & \vdots \\
 & & e_{t|t-h} & & \\
\vdots & & & \ddots & \vdots \\
e_{n-1|n-2} & e_{n-1|n-3} & \cdots & e_{n-1|n-1-(l-1)} & e_{n-1|n-1-l} \\
e_{n|n-1} & e_{n|n-2} & \cdots & e_{n|n-(l-1)} & e_{n|n-l}
\end{pmatrix} & \begin{matrix} \to \ \tau^2_{l+1} \\ \to \ \tau^2_{l+2} \\ \\ \vdots \\ \\ \to \ \tau^2_{n-1} \\ \to \ \tau^2_{n} \end{matrix} \\
\begin{array}{ccccc} \downarrow & \downarrow & & \downarrow & \downarrow \\ \sigma^2_1 & \sigma^2_2 & \cdots & \sigma^2_{l-1} & \sigma^2_l \end{array} &
\end{array}$$

Congruence

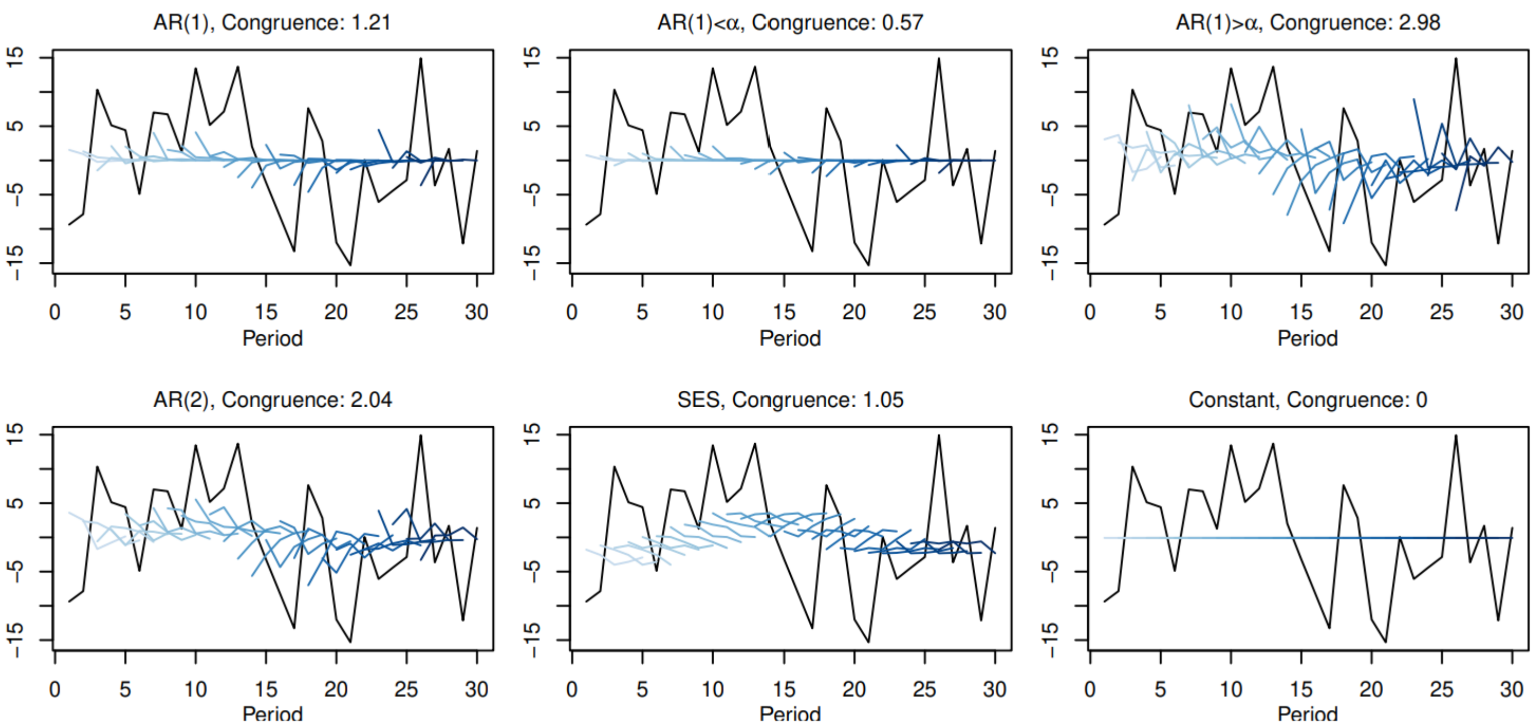
MSE

Congruence

Examples of under and over-congruent forecasts

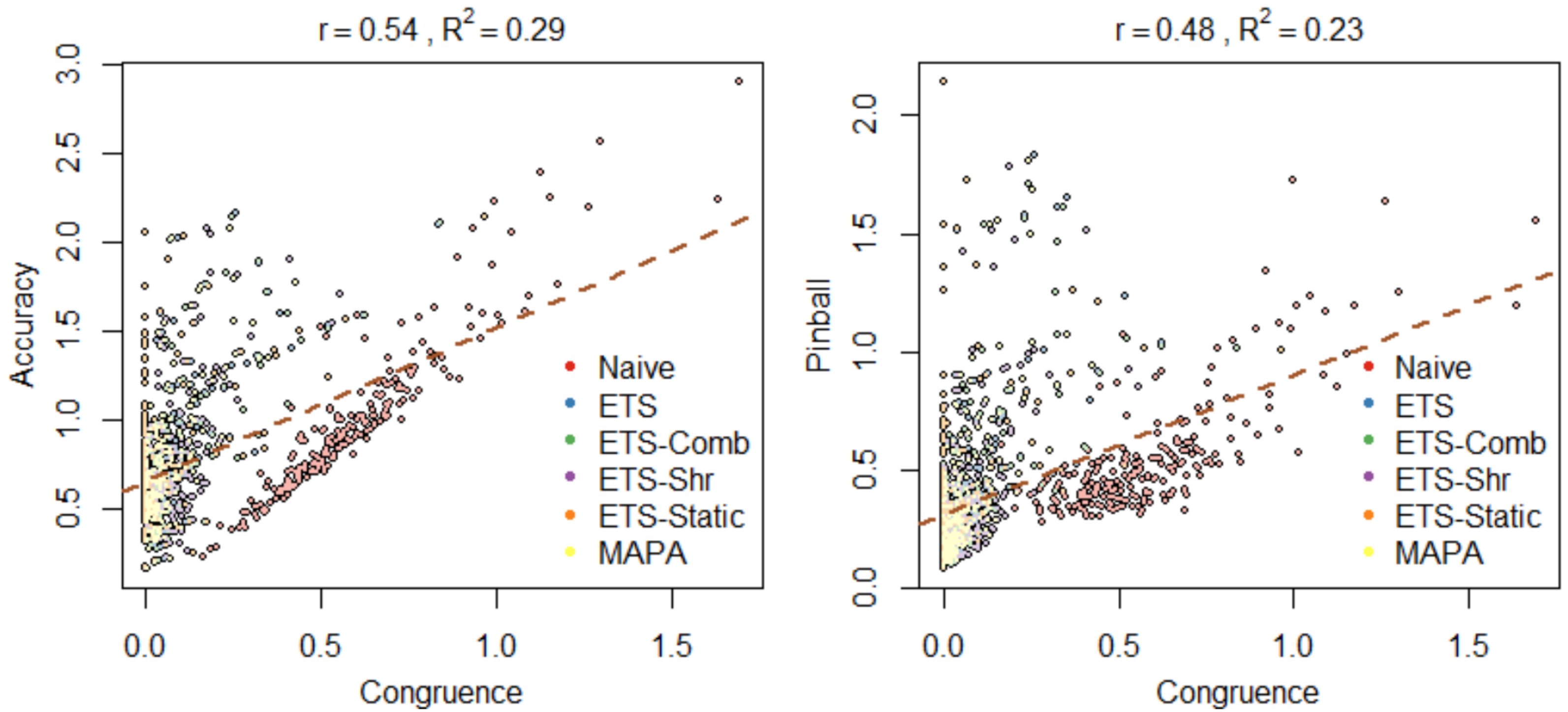
Metric	AR(1)	AR(1)$< \alpha$	AR(1)$> \alpha$	AR(2)	SES	Constant
Congruence τ	1.21	0.57	2.98	2.04	1.05	0
$\sqrt{\text{MSE}_{\text{Total}}}$	10.41	10.44	10.65	10.56	10.92	10.52

Data generating process



Congruence

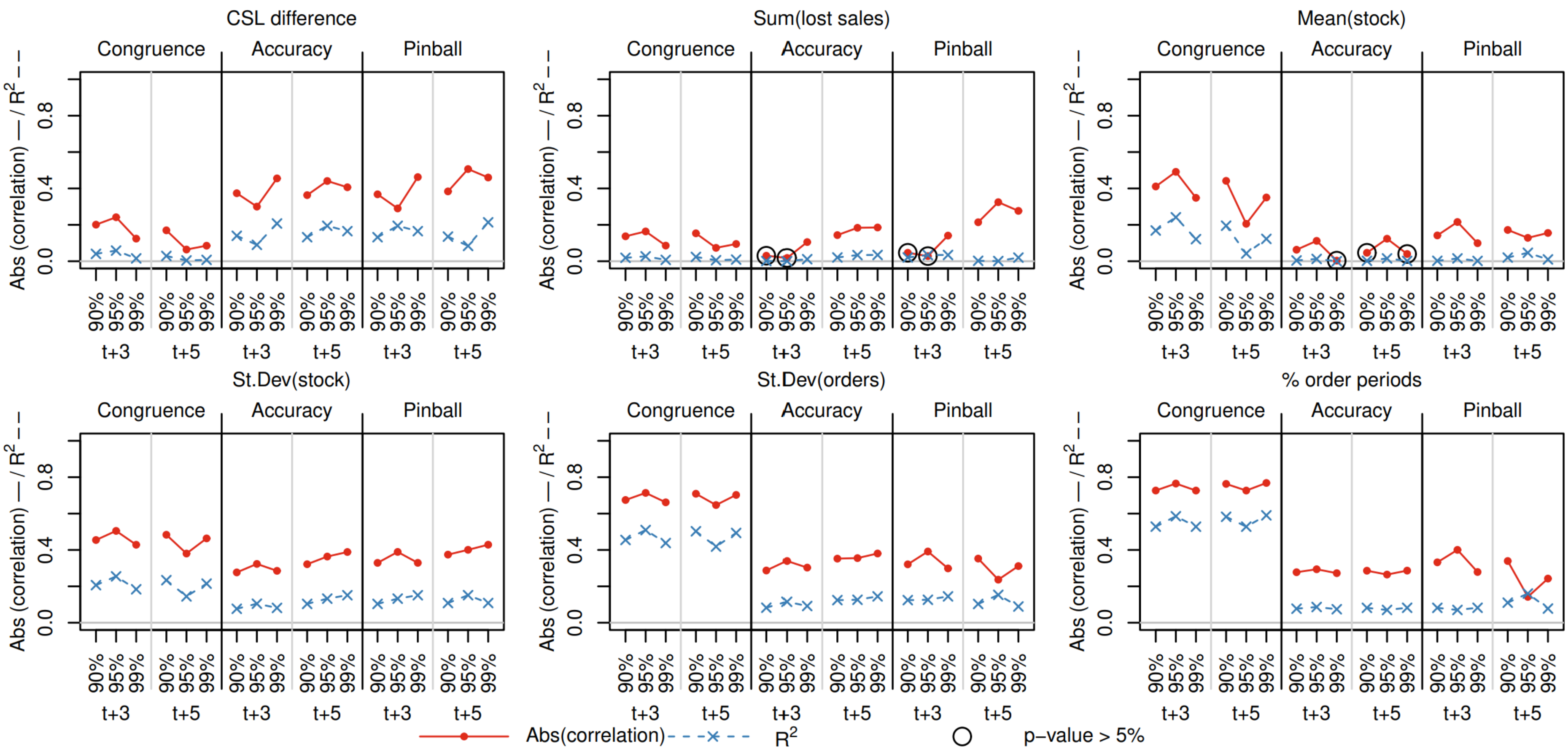
On real firm data



Although it is based on the same raw information, it exhibits limited correlation with accuracy (of mean and quantile)

Congruence

What does it do? (Let's go to the inventory outcome)



Congruence strongly connects with analyst decisions (how much/how often to order)

Congruence

Can we use this as a loss function or to model/method select?

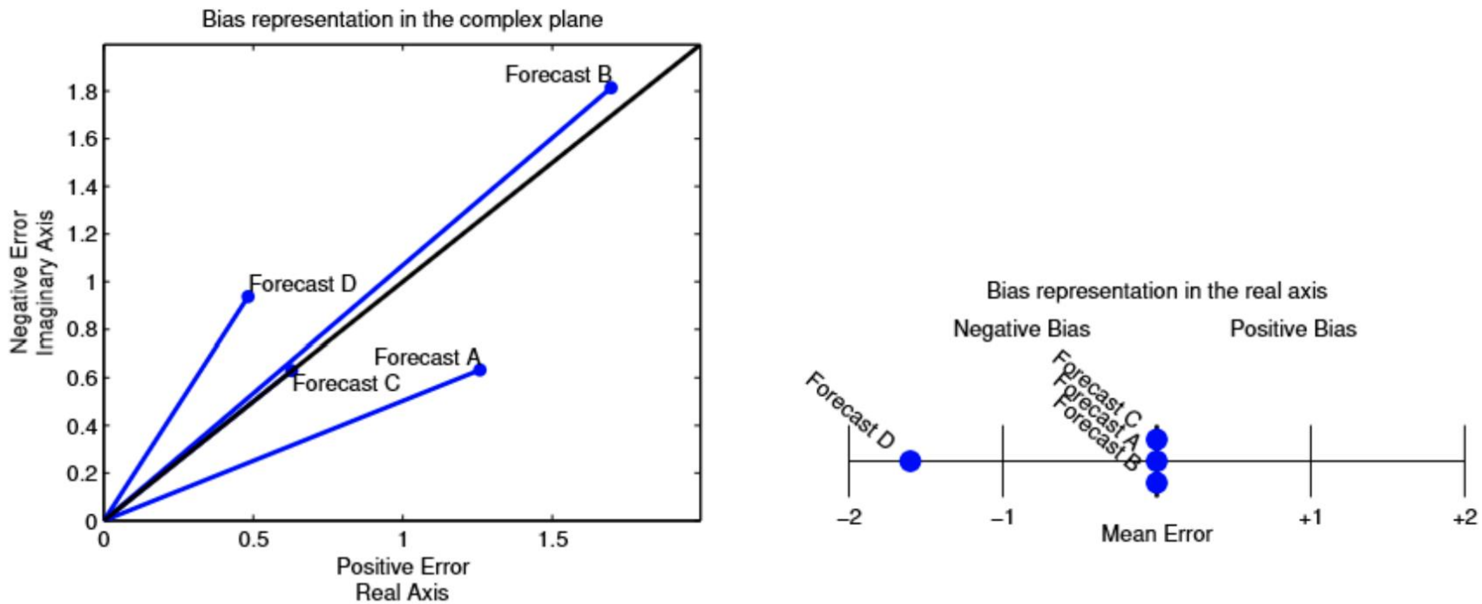
- We do not know yet (come to ISF2024 to hear more about this)
- My current understanding:
  - Low accuracy forecasts can improve in both congruence and accuracy
  - High accuracy forecasts can improve in congruence
  - High congruence can be harmful for accuracy
  - Over-congruence is desirable, under-congruence is not.
- Trivially connects to the Bullwhip Effect, so it may be useful to model select directly for other purposes.

Watch this space!

(see Pritularga & Kourentzes, 2024)

Some “far out” ideas

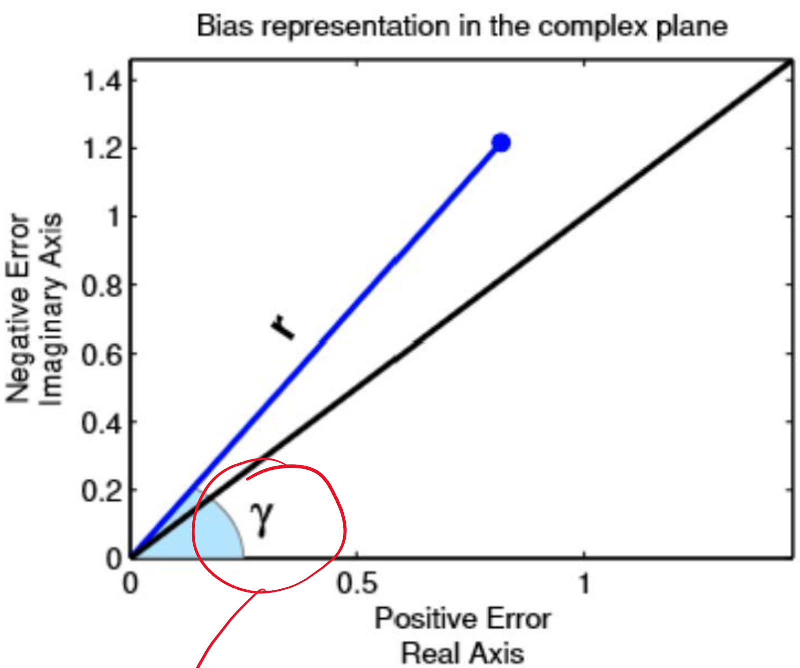
What happens if instead of sticking to $|e|$ or e^2 we do $\sqrt{e}$?



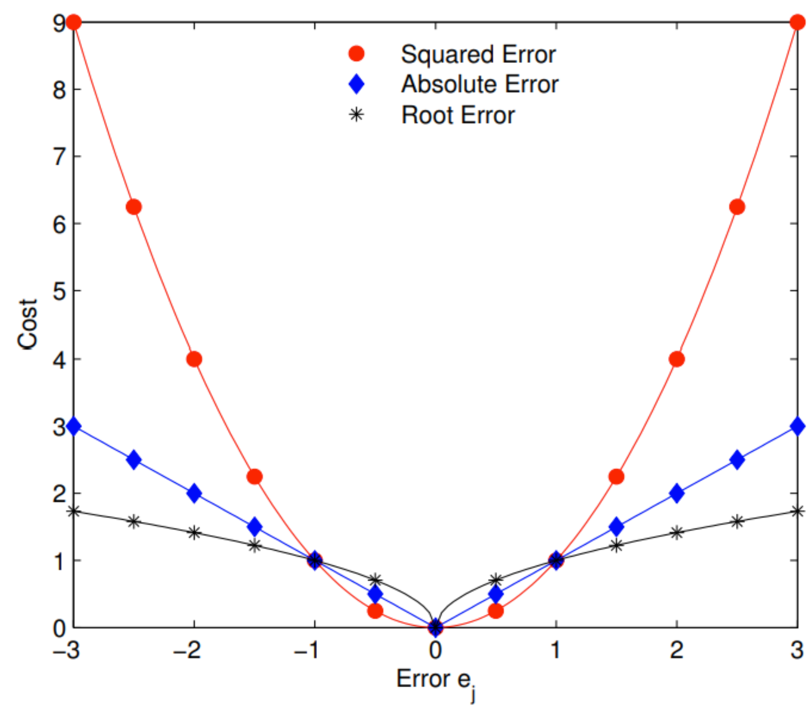
(a) MRE plotted on the complex plane. (b) ME plotted on the real number axis.

Some “far out” ideas

What happens if instead of sticking to $|e|$ or e^2 we do $\sqrt{e}$?



Normalise (-1 to 1) this
to the “bias coefficient”

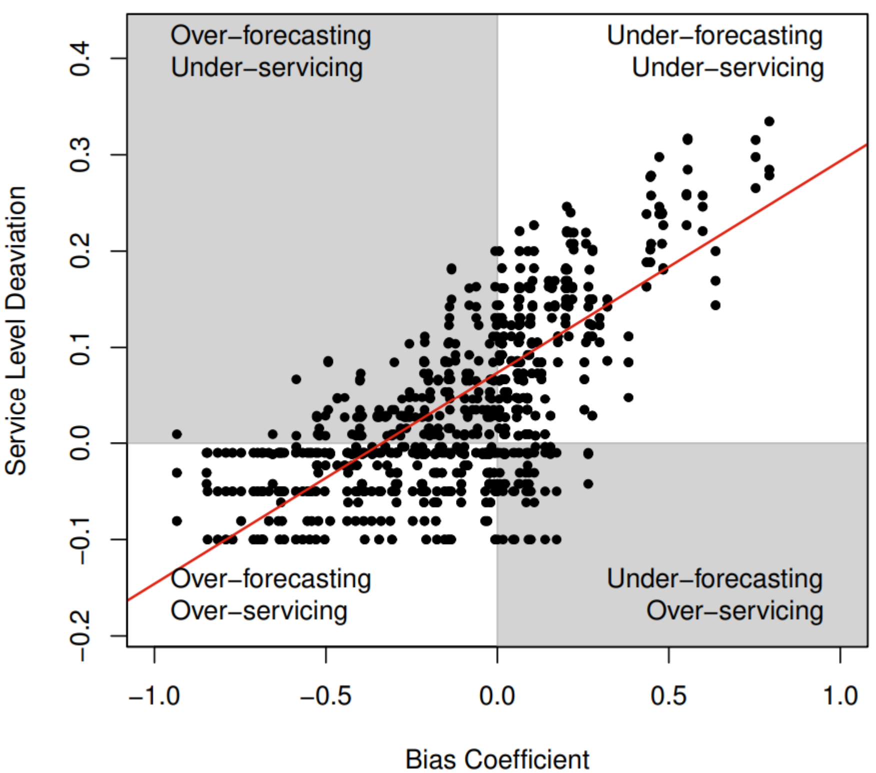


Some "far out" ideas

Do the inventory simulation and measure how much of its variance is explained

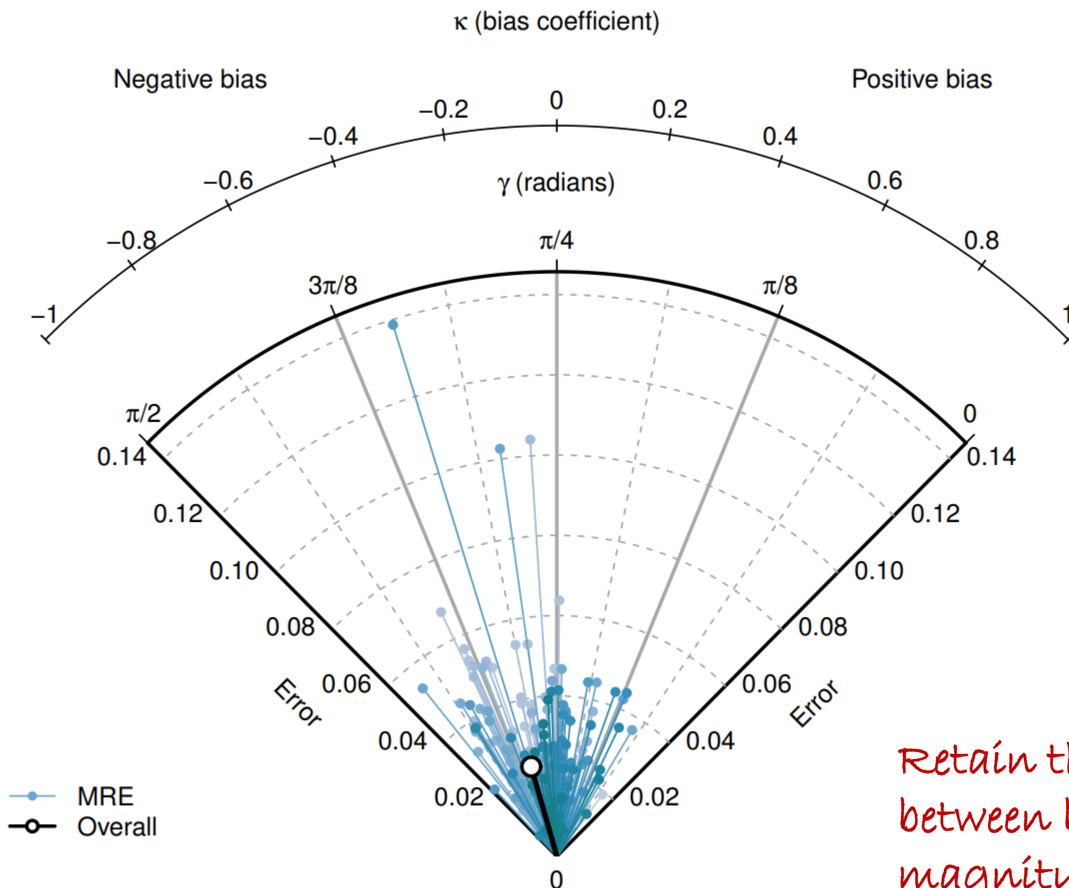
Metric	Scenario 1	Scenario 2
	Bias	
ME/mean	0.312 (+)	0.479 (+)
MPE	-	-
MDB	0.329 (+)	0.556 (+)
Bias Coef. κ	**0.374** (+)	**0.572** (+)
	Accuracy	
RMSE/mean	0.018 (-)	0.102 (-)
MAE/mean	0.028 (-)	0.105 (-)
MAPE	-	-
sMAPE	0.244 (-)	0.299 (-)
RMSSE	0.024 (+)	0.161 (+)
MASE	0.036 (+)	0.182 (+)
MAAPE	0.295 (-)	0.383 (-)
MMRE ($\lvert z \rvert$)	0.048 (-)	0.128 (-)

Sign of α_1 coefficient in brackets.



Some “far out” ideas

And it comes with nifty plots!



Retain the connection between bias and magnitude of error

(see Kourentzes et al., 2021)

Conclusions

Some starting points

1. **Forecasting is not the end-target!** Forecasting error metrics are convenient and helpful, but likewise not the complete story.
    - The **supported decisions may transform the forecasts** in ways that make closely following the data only a part of what matters (e.g., in inventory we first check how much we have in stock before we order)
    - However, just **not evaluating forecasts would be an excuse**. Beyond any challenges, there is a **problem of attribution**. How do we find how much value is added by the forecasts, if they are to be transformed when used?
    - In practice many firms strive for accurate forecasts, but the supported decisions would not change by much by more accurate forecasts. This is due to simplistic decision-making heuristics.

2. **There is no best error metric!** The application context drives the choice of the metric.
    - Nonetheless, there are some metrics we can let them rest in peace.

References

- Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, *102*(477), 359-378.
- Athanasopoulos, G., & Kourentzes, N. (2023). On the evaluation of hierarchical forecasts. *International Journal of Forecasting*, *39*(4), 1502-1511.
- Kourentzes, N. (2014). On intermittent demand model optimisation and selection. *International Journal of Production Economics*, *156*, 180-190.
- Kourentzes, N., Trapero, J. R., & Barrow, D. K. (2020). Optimising forecasting models for inventory planning. *International Journal of Production Economics*, *225*, 107597.
- Pritularga, K., & Kourentzes, N. (2024). Forecast congruence: a quantity to align forecasts and inventory decisions. *Available at SSRN*.
- Kourentzes, N., Svetunkov, I., & Trapero, J. R. (2021). Connecting forecasting and inventory performance: a complex task. *Available at SSRN 3878176*.

Thank you for your attention!
Questions?

Nikolaos Kourentzes

email: nikolaos@kourentzes.com

twitter @nkourentz

Blog: http://nikolaos.kourentzes.com

HÖGSKOLAN I SKÖVDE