Incorporating risk preferences in forecast selection

Nikolaos Kourentzes Ivan Svetunkov

Energy Finance Christmas Workshop (EFC25)
12 December 2025

State of play of model selection

- Model or method selection is typically based on some (single-dimension) summary statistic:
  - cross-validated MSE or other error metric.
  - information criteria, like the Akaike Information Criterion.
- Good summary statistics guard us against overfitting.
  - Information criteria explicitly penalize for model complexity.
  - Cross-validated errors implicitly do the same.
  - 1-step ahead cross-validated MSE is equivalent to AIC.
- Arguably, selection should match the supported decision horizon, something that many metrics ignore.
- Forecasting focused → preferences of decision makers & stake/holders?

Where do forecast errors come from?

$$\mathbb{E}[E(f_n) - E(f^\star)] = \mathbb{E}[E(f^\star_{\mathcal{F}}) - E(f^\star)] + \mathbb{E}[E(f_n) - E(f^\star_{\mathcal{F}})] + \mathbb{E}[E(\tilde{f}_n) - E(f_n)]$$
$$= \mathcal{E}_{\mathrm{app}} + \mathcal{E}_{\mathrm{est}} + \mathcal{E}_{\mathrm{opt}},$$

Error of your forecast

Error of DGP (irreducible error)

Error of best possible solution from pool of methods/models explored (F)

Error due to estimation (including variable selection) of your solution to the best possible within pool of models

Error because your optimization is not perfect (e.g., tolerance)

This is what we care about in this talk, can we get a feel of the size of this?

These are encapsulated in the predictive distribution

A probabilistic treatment of model statistics

- We'll keep it simple by discussing only the "model case", i.e., there is a likelihood. **You can replace the likelihood with cross-validated errors and generalize to any method.**

- A fairly general expression of likelihood for regression problems (state space formulation) looks like

error at observation t

$$\mathcal{L}(\boldsymbol{\theta}, \sigma^2|\mathbf{y}) = (2\pi\sigma^2)^{-n/2}\exp\left(-\frac{1}{2}\sum_{t=1}^{n}\frac{\varepsilon_t^2}{\sigma^2}\right)\left|\prod_{t=1}^{n} r(\mathbf{v}_{t-1})\right|^{-1}$$

Variance of population

blah blah...
I assumed it is normally distributed

Potentially after
transformations
(multiplicative errors)

- So, it is "just" a Sum of Squared (**sampled**) Errors

A probabilistic treatment of model statistics

- First, what is standard practice?
- We "simplify"

$$\ln(\mathcal{L}) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2}\sum_{t=1}^{n}\left(\frac{\varepsilon_t^2}{\sigma^2}\right) - \sum_{t=1}^{n}\ln|r(\mathbf{v}_{t-1})|,$$

These just shift the mean of the log-likelihood, and are often (erroneously!) ignored. All that is left is ½ of the normalised SSE (so it could all be replaced by cross-validated errors).

- And obtain a selection criteria

$$\text{AIC} = 2k - 2\ln(\mathcal{L}^*)$$

Number of model parameters (including σ)

Maximised likelihood

- The model with the lowest AIC is the model we want

A probabilistic treatment of model statistics

- We make an "intentional typo"

Constant

Constant

$$\ln(\mathcal{L})_t = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2}\sum_{t=1}^{n}\left(\frac{\varepsilon_t^2}{\sigma^2}\right) - \sum_{t=1}^{n}\ln|r(\mathbf{v}_{t-1})|,$$

- We now have a likelihood expression per observation, let's name it something imaginative… EFC25 meet point likelihood λ, point likelihood meet EFC25!

If this is observational, then it is distributed somehow

$$\lambda_t = \ln(L_t) = -\frac{1}{2}\left(\ln(2\pi\sigma^2) + \frac{\varepsilon_t^2}{\sigma^2}\right)$$

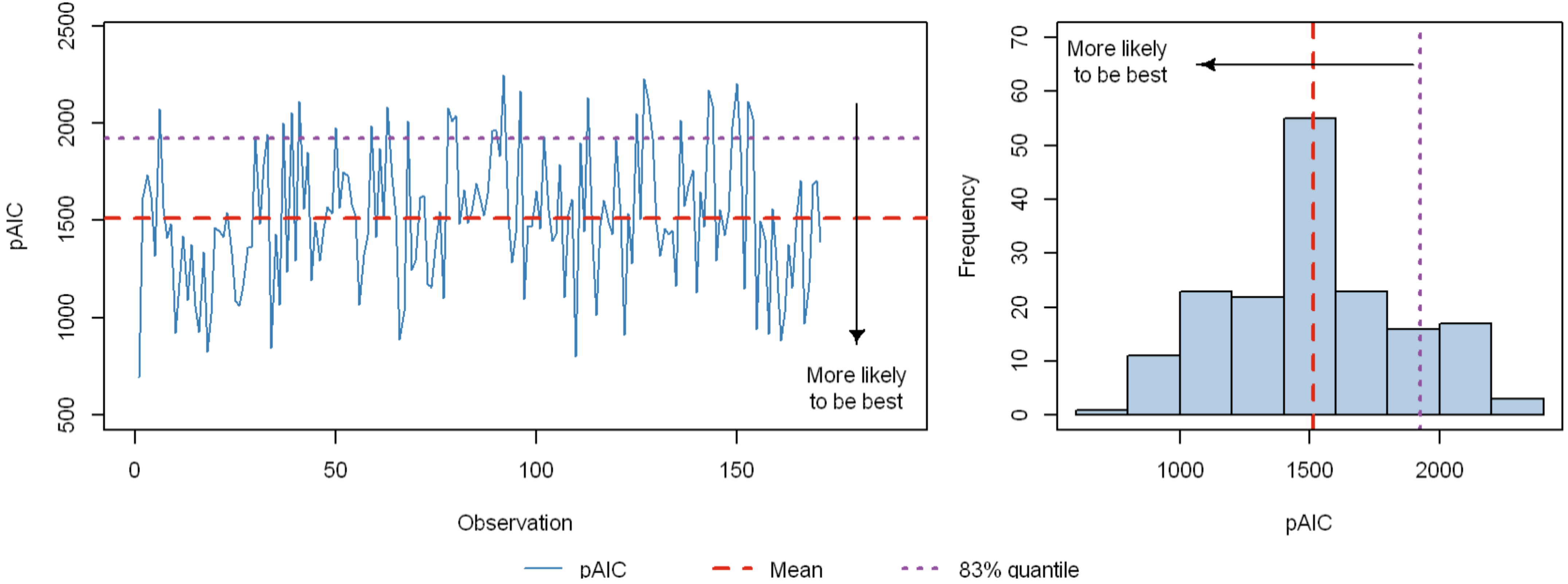
- And likewise, we can have a point AIC, instead of a summary AIC.

This is why we retain the constant above. A matter of scale

$$\text{pAIC}_t = 2k - 2n\lambda_t \longrightarrow \text{AIC} = \frac{1}{n}\sum_{t=1}^{n}\text{pAIC}_t$$

A probabilistic treatment of model statistics

The $pAIC_t$ of an example model fit



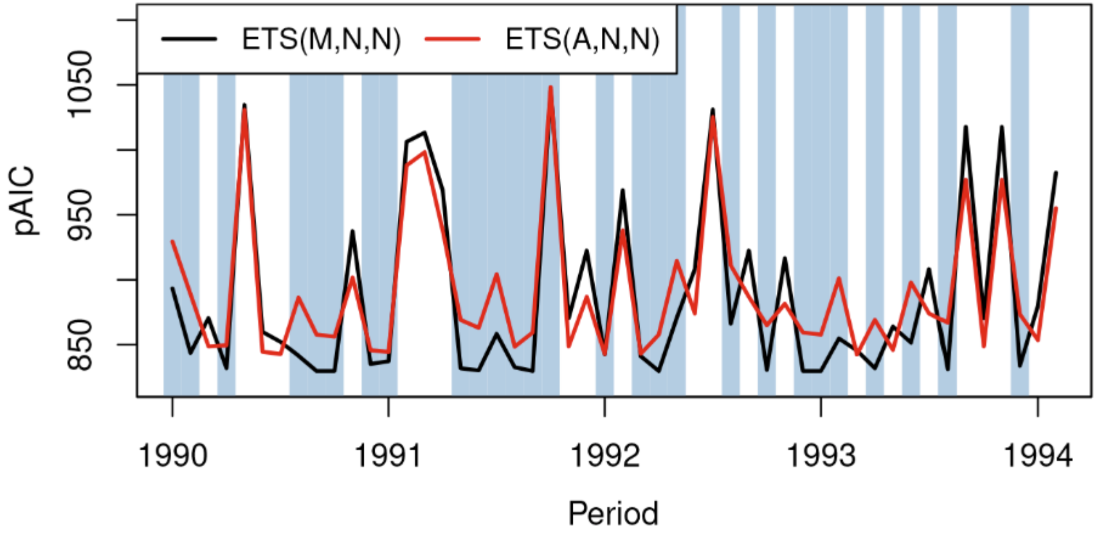
There are many ways to summarise parametrically or nonparametrically this distribution

→ diverse ranking of models

A probabilistic treatment of model statistics

Let us compare the pAIC of two models on a series. (Blue highlight when ETS(M,N,N) is more plausible.)

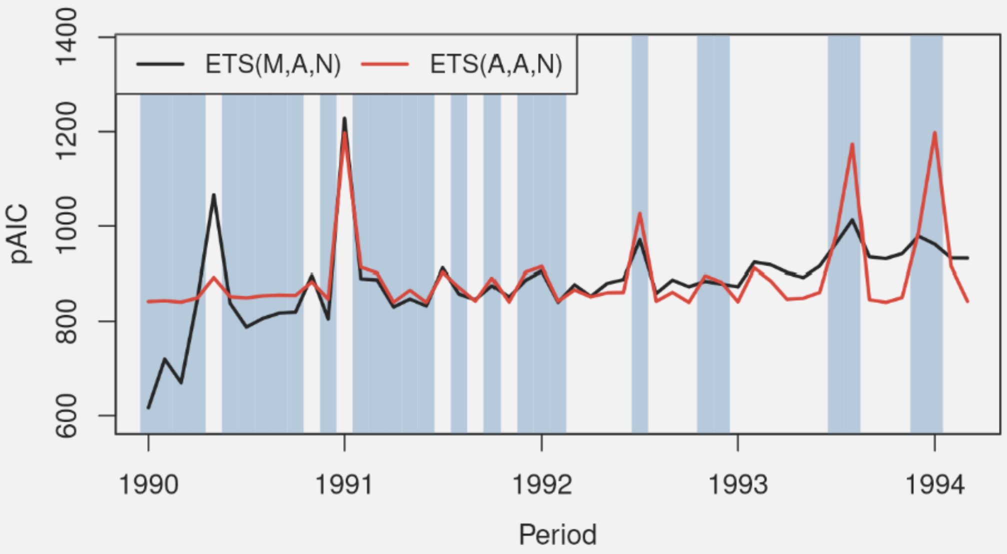
ETS(A,N,N), AIC: 892.4
ETS(M,N,N), AIC: 887.7



ETS(M,N,N) tends to be less plausible when both models exhibit worse pAIC. Works well when "safe", **relatively less plausible on the difficult cases**.

How much risk are we willing to take?

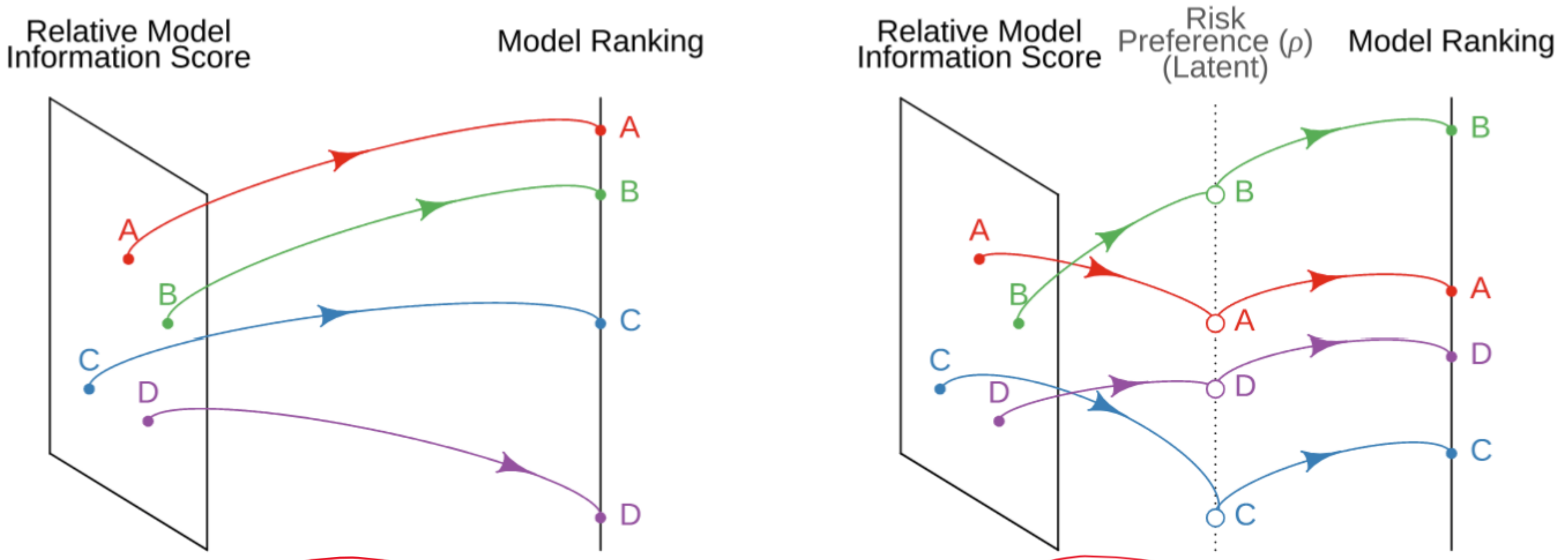
Also, consider the stationarity of your model selection statistic! ETS(M,A,N) has lower AIC, but its pAIC is non-stationary (and increasing!)



Risk preferences in model selection

Eventually, we are interested in the probability that the model we choose is the most plausible:

- **risk-averse**, corresponding to **upper quantiles** of the model statistic;
- **risk-neutral**, corresponding to the **median,**
- **risk-tolerant**, corresponding to **lower quantiles**.

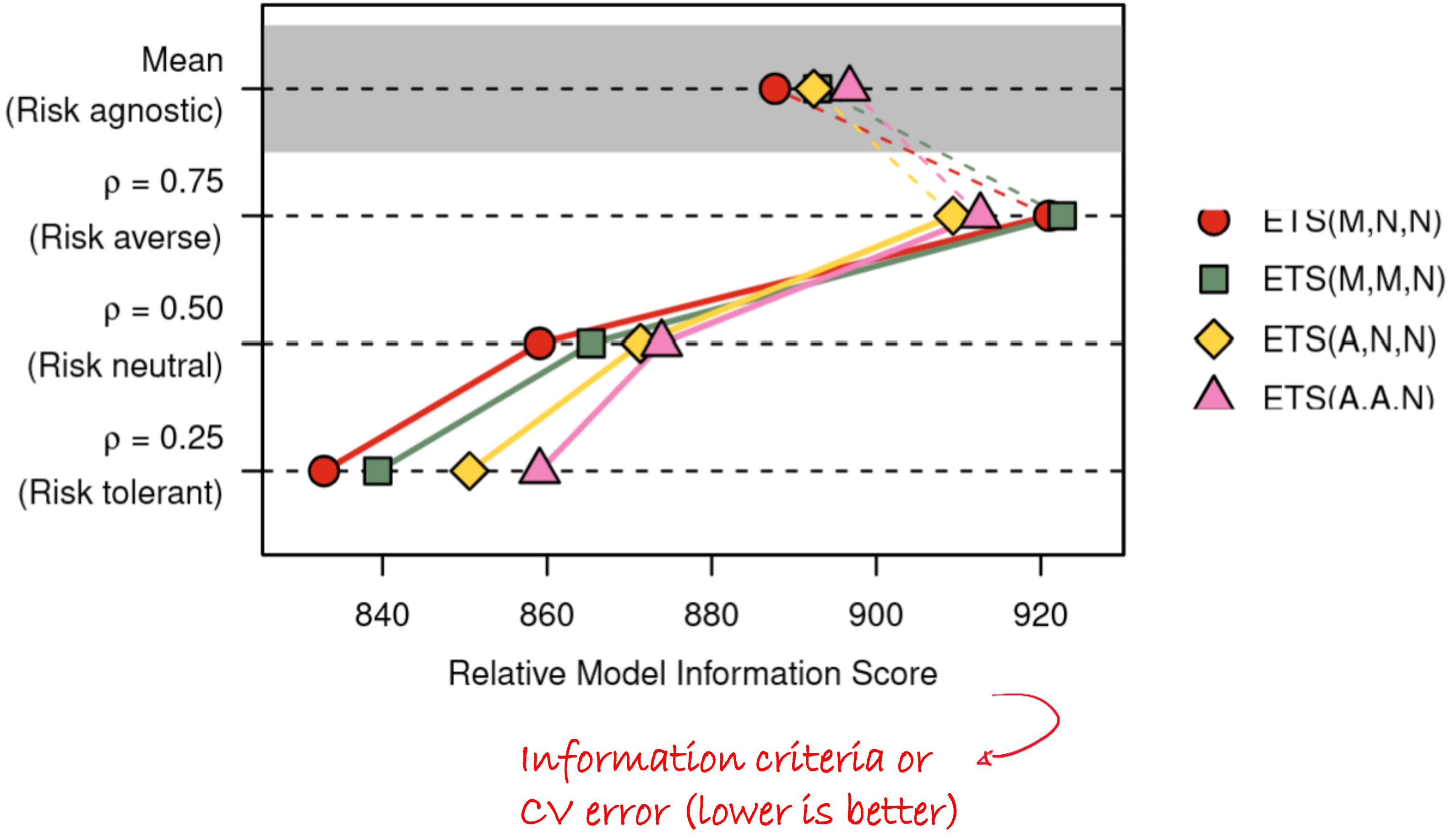


Model selection "model" $f: S_t \mapsto \mathbb{R}^1$ Latent risk preference $g: S_t \overset{\rho}{\mapsto} \mathbb{R}^1_\rho$ $f: \mathbb{R}^1_\rho \mapsto \mathbb{R}^1$

$z: f \circ g: S_t \overset{\rho}{\mapsto} \mathbb{R}^1$

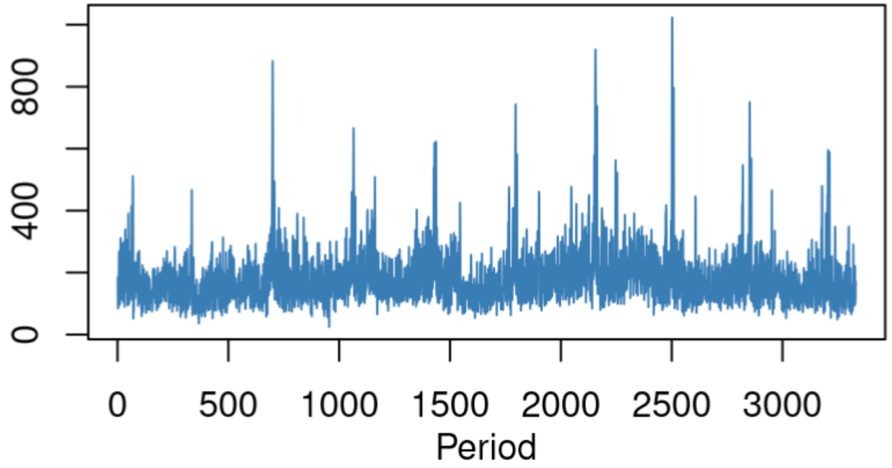
Risk preferences in model selection

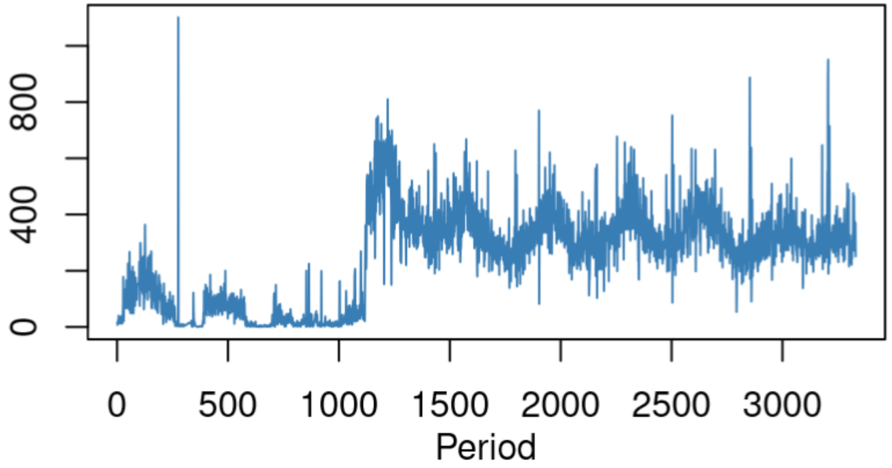
An example on a single time series



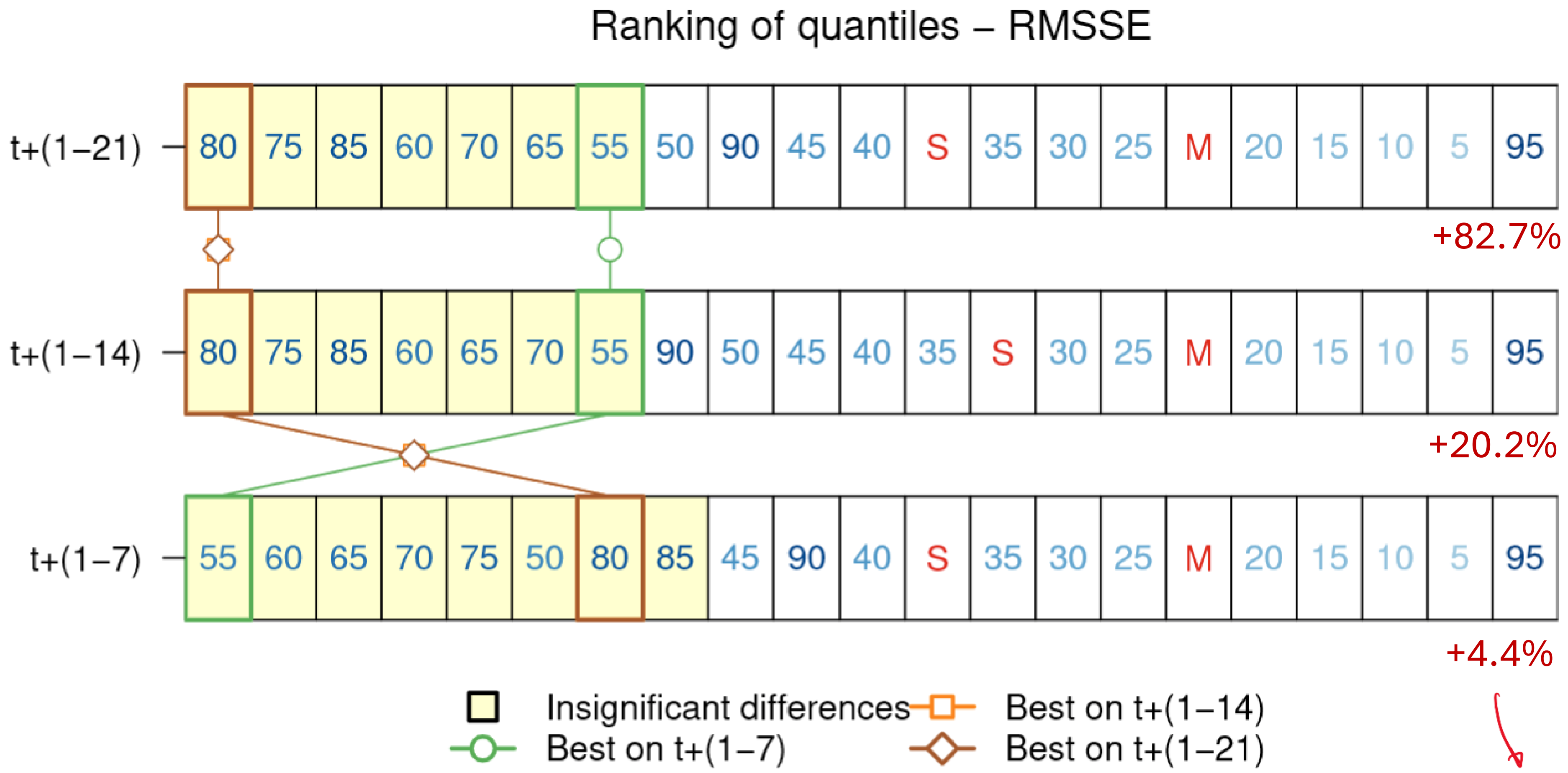
What is the impact on forecast accuracy?

- Forecast 883 items
- Daily data, 1021 to 3360 daily sales data
- Test on last 36 days, rolling origin forecasts with horizons of 7, 14, and 21 days.
- Forecast with ETS, perform model selection by mean, median, lower (5%-45%) quantile, upper quantiles (55%-95%), and sum of quantiles pAIC. Mean pAIC is same as standard AIC.





Results on RMsSE (mean forecast accuracy) – Nemenyi test

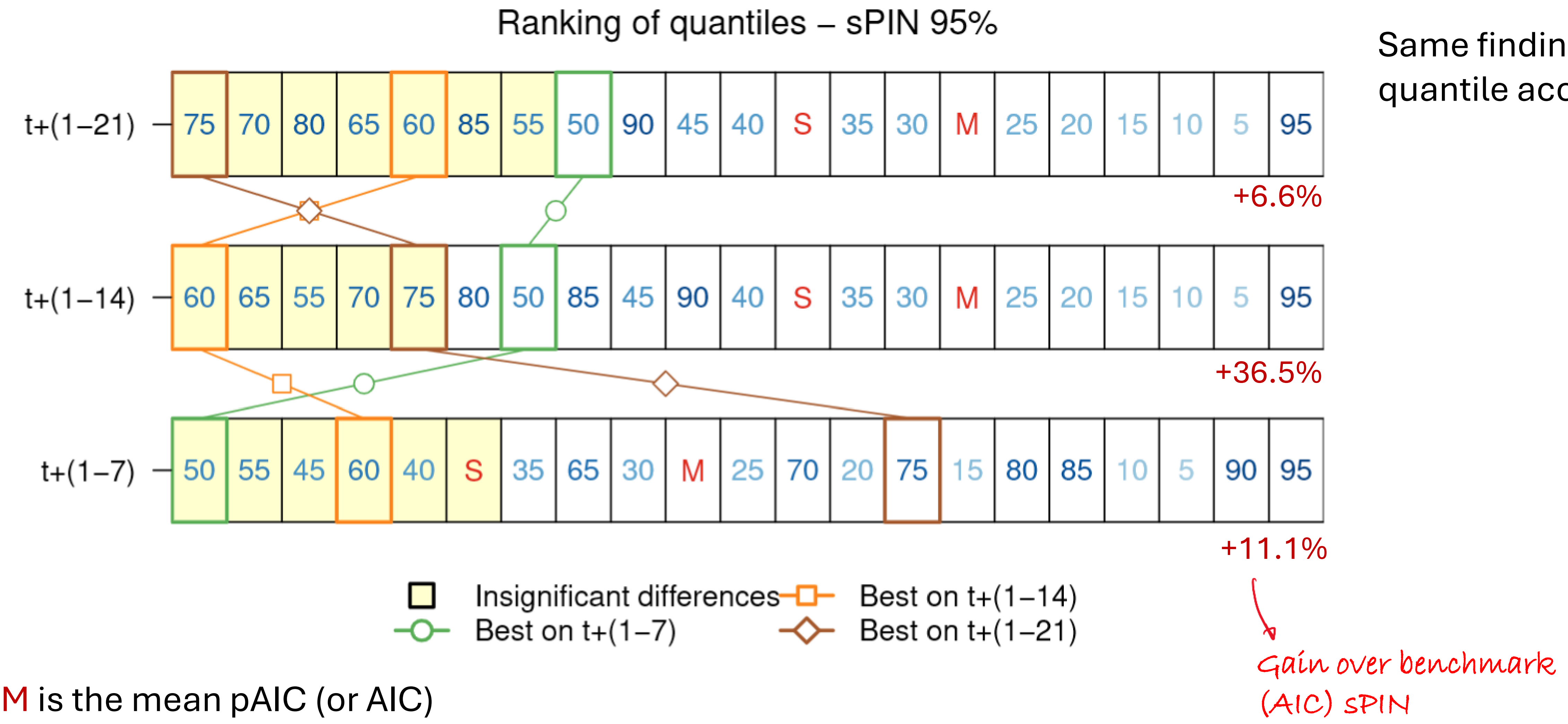


Gain over benchmark (AIC) RMSSE

- **Higher risk aversion performs better on longer horizons**
- **AIC (mean) ranks fairly low**
- **Extreme quantiles bring estimation issues**
- **Risk tolerant solutions do not perform well**

M is the mean pAIC (or AIC)
S is the sum of pAIC quantiles

Results on scaled Pinball (quantile accuracy) – Nemenyi test

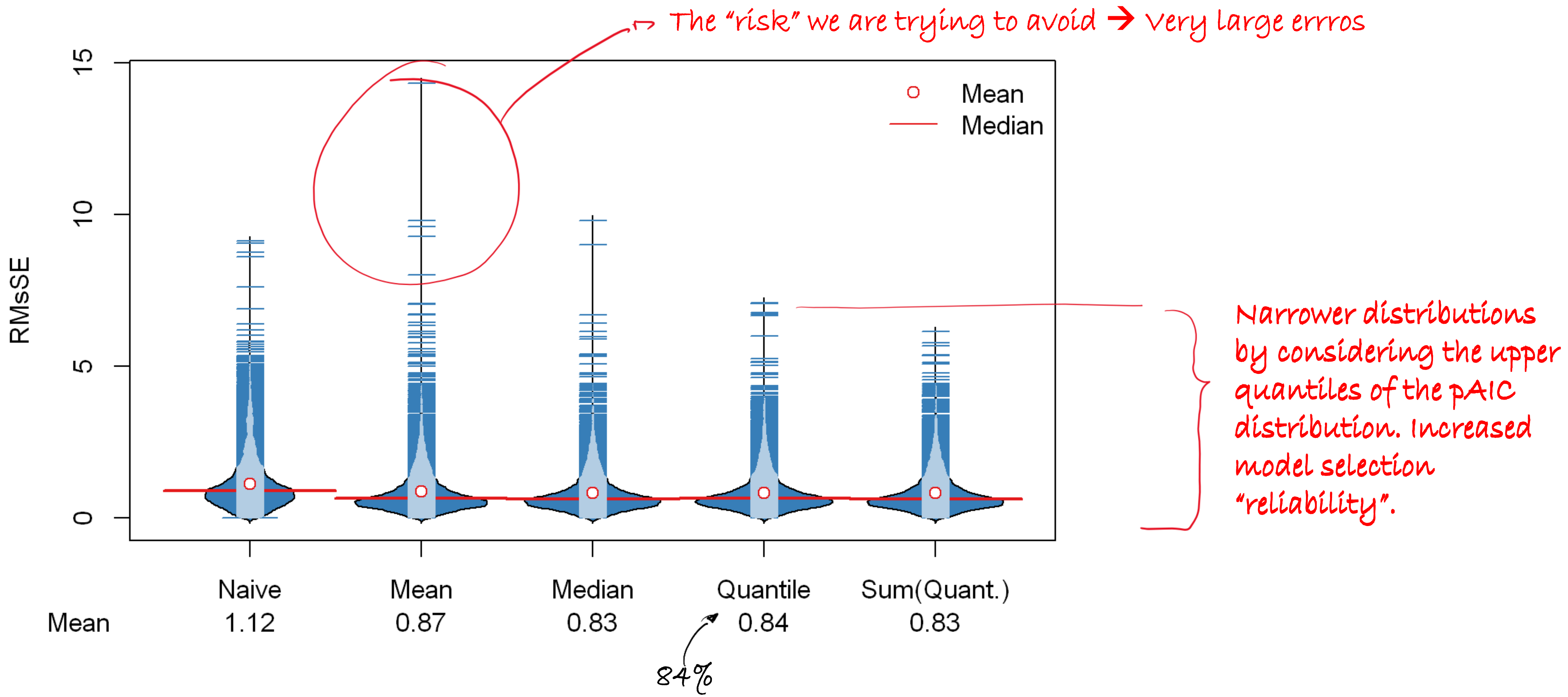


Same findings for quantile accuracy

M is the mean pAIC (or AIC)

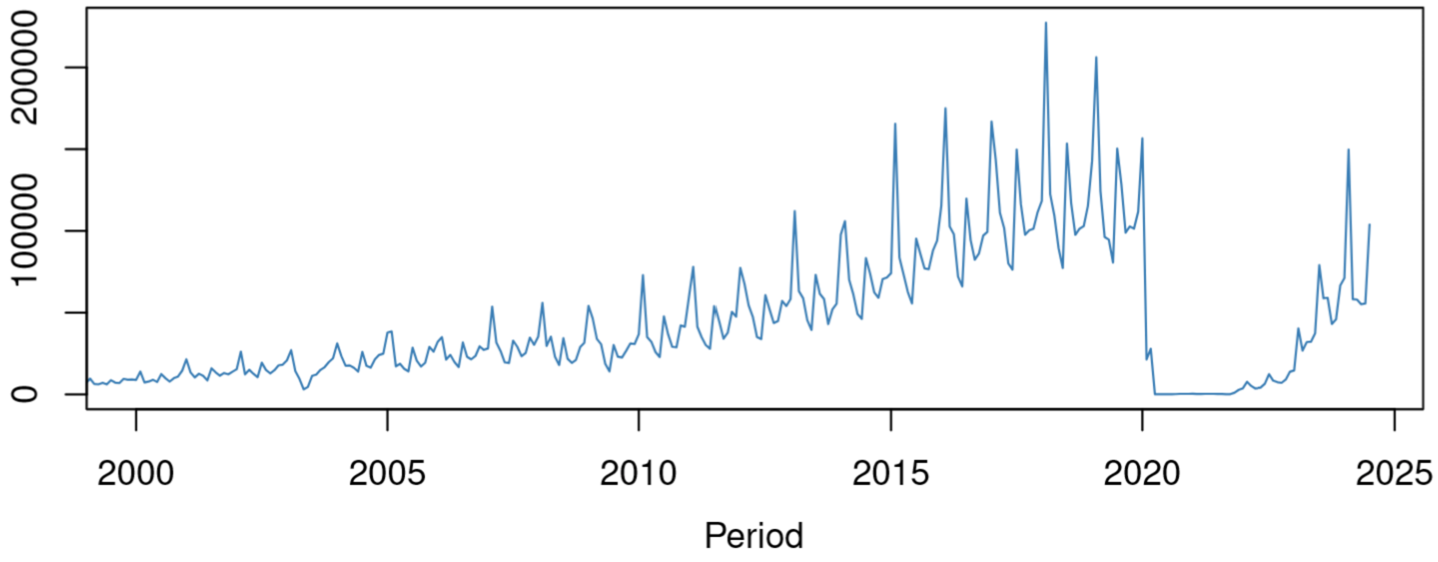
S is the sum of pAIC quantiles

Risk averse model selection – subset of 111 items (a category)



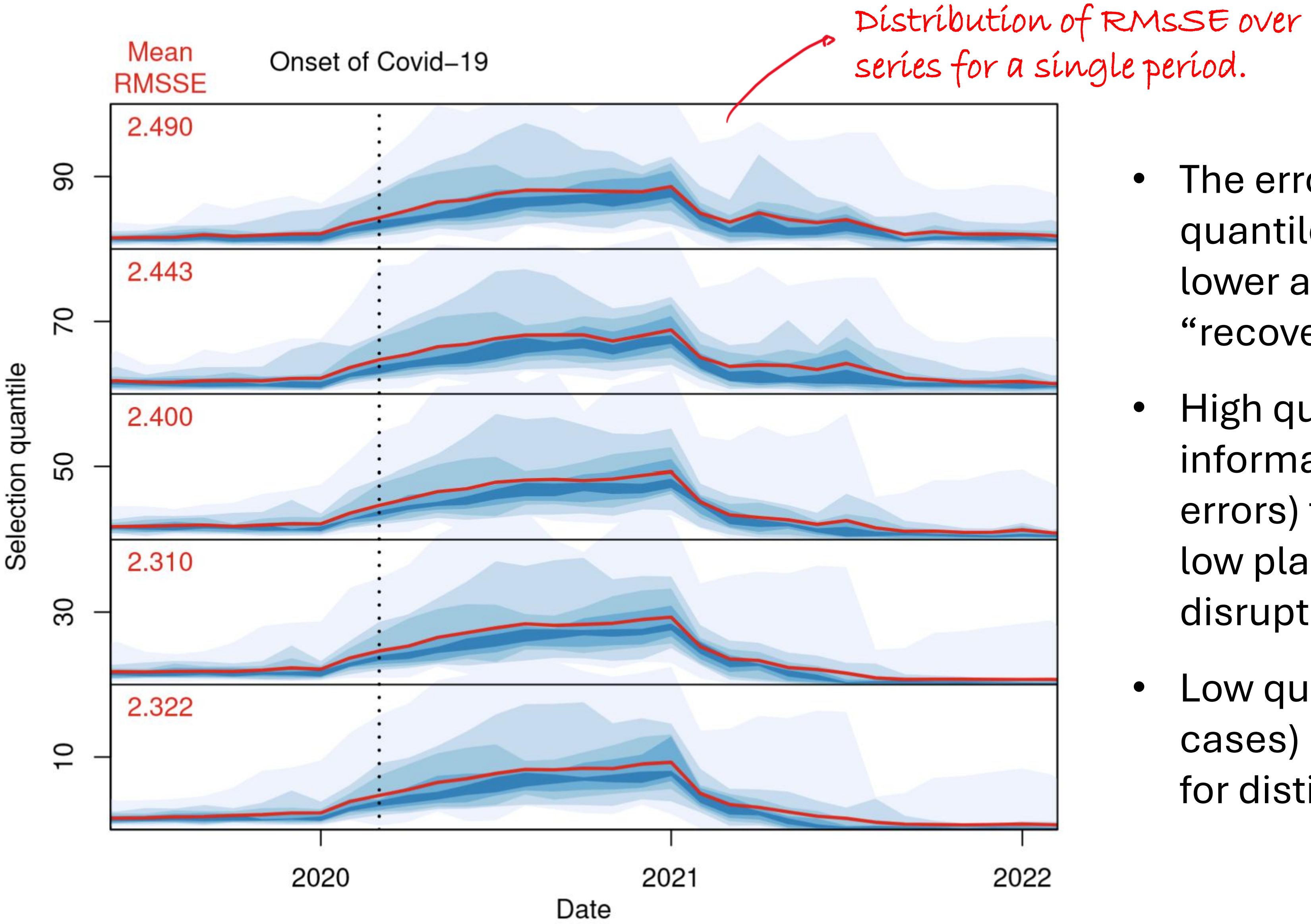
When does risk tolerance make sense?

Below are tourism flows during before, during, and after Covid-19 (between Australia and mainland China).



We produce rolling forecasts of 12 periods (1-year ahead) over 20 series of tourism flows.
Record the distribution of RMSSE over time/series.

When does risk tolerance make sense?



- The error distribution of lower quantiles (risk tolerant selection) is lower and less dispersed. It also “recovers” faster.
- High quantiles of the relative model information score (AIC or CV errors) focus on observations with low plausibility (hard ones). Under disruptions, plausibility plummets.
- Low quantiles (very plausible cases) become more informative for distinguishing between models.

Risk preferences and model selection

- Risk averse choices → overall better forecasting performance than literature standard. Risk tolerance useful under disruptions. Empirically, the exact choice of quantile does not matter (estimation issues can be a problem).
- Generally applicable → select the appropriate relative model information score.
  → do not focus on the summary statistic but consider the whole distribution;
- Combine instead of selecting forecasts: risk averse model pooling! (paper coming up!)
- Embed risk preferences of stakeholders/process onto models that give probabilistic forecasts (there are two separate uncertainties: model and forecast, the forecast variance ignores the risk of model misspecification, but it is conditioned on it!)
- Open question:
  Use the same quantile trickery to estimate model parameters. Does this make sense? Effectively we make models "blind" to specific errors during estimation. There may be benefits, but there are plenty of pitfalls as well!

Questions?

Nikolaos Kourentzes

nikolaos@kourentzes.com

Paper is under review – contact me for a copy!