Beyond summary performance metrics for forecast selection and combination

Ivan Svetunkov & Nikos Kourentzes

Lancaster University, UK

18th July 2025

Marketing Analytics and Forecasting | Lancaster University Management School

A couple of moments for model selection and combination

```
\documentclass[preprint,review ,authoryear,12pt]{elsarticle}
%% Use the option review to obtain double line spacing
%% \documentclass[preprint,review,12pt]{elsarticle}

\usepackage{amssymb,multirow,amsmath,array,lscape,rotating,booktabs}
\usepackage{graphicx,epstopdf}

\usepackage{color}          % Used for revisions
\usepackage[normalem]{ulem} % Used for revisions
% For additions use {\color{red}...}
% For deletions use \sout{{\color{blue}...}}

% \usepackage[nomarkers]{endfloat} # Move all figures and tables at the end for JBR submission

% \usepackage{refcheck}  % Used to list unused references

\begin{document}
\begin{frontmatter}

\title{Beyond summary performance metrics for forecast selection and combination}

\author[label1]{Nikolaos Kourentzes\corref{cor1}}
\ead{nikolaos@kourentzes.com}
\cortext[cor1]{Correspondance: N Kourentzes, Department of Management Science, Lancaster University Management School, Lancaster, Lancashire, LA1 4YX, UK.}
\address[label1]{Lancaster University Management School \\ Department of Management Science, Lancaster, LA1 4YX, UK}
\author[label1]{Ivan Svetunkov}
% \author[label2]{Stephan Kolassa}
% \address[label2]{SAP Switzerland, Bahnstrasse 2, 8274 T{\"a}gerwilen, Switzerland}

\begin{abstract}
In doing forecast selection or combination we typically rely on some performance metric. For example, this could be the Akaike Information Criterion (AIC) or some cross-validated accuracy measure. From these we can either pick the top performer, or construct combination weights. There is ample empirical evidence demonstrating the appropriateness of such metrics, both in terms of resulting forecast accuracy and automation of the forecasting process. Yet, these performance metrics are summary statistics, and they do not reflect higher moments of the metrics. This poses similar issues to analysing only point forecasts to assess the risks associated with a prediction, instead of looking at the whole distribution of values. Looking at summary statistics does not reflect the uncertainty in the ranking of alternative forecasts, and therefore the uncertainty in selection and combination of forecasts. We propose a modification in the use of the AIC and an associated procedure for selecting a single forecast or constructing combination weights that aims to go beyond the use of summary statistics to characterise each forecast. We demonstrate that our approach does not require an arbitrary dichotomy between forecast selection, combination or pooling, and switches appropriately depending on the time series on hand and the pool of forecasts considered. The performance of the approach is evaluated empirically on a large number of real time series from various sources.  We argue that the proposed approach is useful for model selection and combination more generally, beyond the forecasting domain.
\end{abstract}
\begin{keyword}
Model selection \sep Forecast combination \sep Akaike Information Criterion \sep Uncertainty \sep Cross-validation
\end{keyword}

\end{frontmatter}

\section{The Plan}
```

Introduction

How many of you have used combinations for forecasting?

How do you combine forecasts?

There's lots of papers on different combination methods:

- Means, trimmed and winsorised (Jose and Winkler, 2008);
- Median (Agnew, 1985; Stock and Watson, 2004);
- Weighted combinations (Elliott, 2011; Elliott and Timmermann, 2016; Kolassa, 2011);
- Claeskens et al. (2016) explains why simple combinations are more robust than the weighted ones.

Introduction

How many models do you have in your pool? Do you need all of them?

Geweke and Amisano (2011) shows that using a subset of models helps.

In your pool of models:

- some will perform consistently poorly, and need to be removed;
- some will do consistently well. Keep them!

But how can we decide?

Kourentzes et al. (2018) develop a heuristic to reduce the pool of models.

Introduction

Are you sure that your combination is adequate?

Will it change tomorrow?

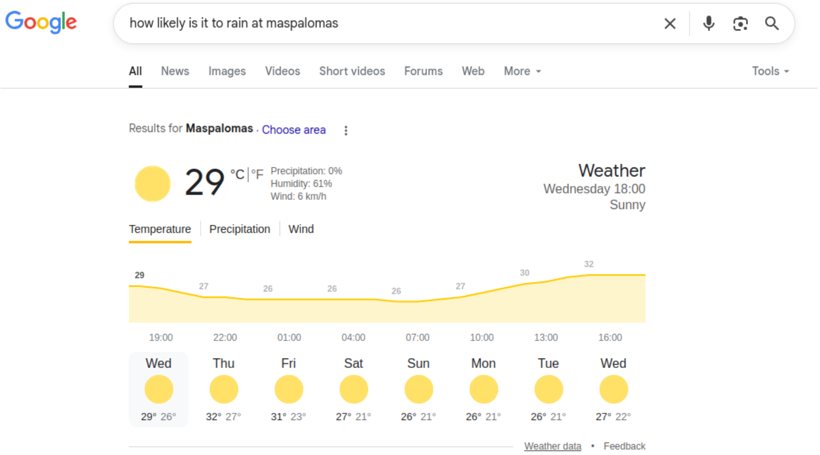
Yes! Everything is random! You cannot be sure of anything...

Yet all the combination approaches rely on some summary statistics from a sample.

Vehtari et al. (2017) use the leave-one-out CV errors to get the standard errors for forecasts.

This way we can compare model performance...

Point likelihood



Point likelihood

Take a look at AIC:

$$\mathrm{AIC}_m = 2k_m - 2\ell_m, \tag{1}$$

ℓ_m is the log-likelihood value, k_m is the number of estimated parameters of a model m in the pool.

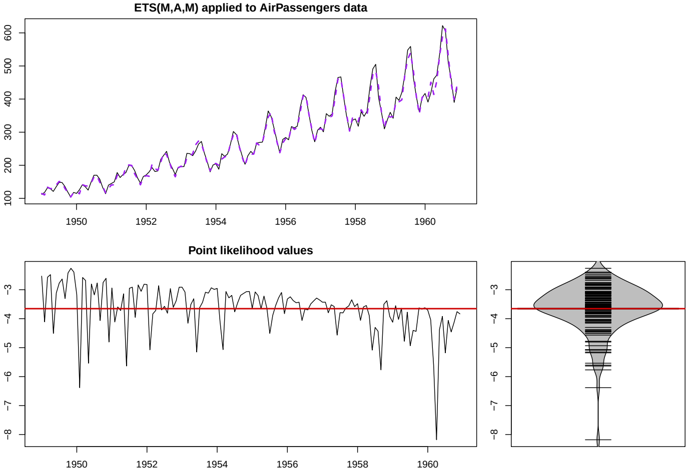
This relies on the summary statistics, log-likelihood ℓ_m

It shows how likely it is to have such a model given the data.

This is a sum of point log-likelihoods:

$$\ell_m = \sum_{t=1}^{T} \ell_{m,t} \tag{2}$$

Point likelihood



Point IC

So, we can use point log-likelihoods to compare models.

The only issue is the bias in the likelihood estimate (Akaike, 1974).

We propose a point AIC:

$$\mathrm{pAIC}_{m,t} = 2k_m - 2T\ell_{m,t}. \tag{3}$$

T is needed to bring the point likelihood value to the overall level.

Other criteria can be modified like that as well.

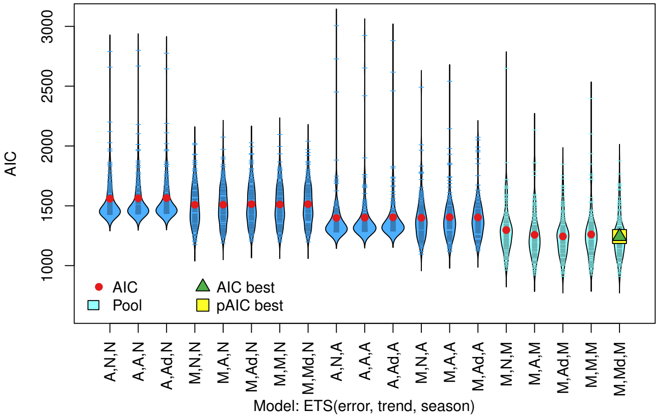
Pooling with point IC

How can we use that?

1. Apply models to the data;
2. Extract pAIC values for each of them;
3. Use Nemenyi test (Demšar, 2006) to group models;
4. Combine their forecasts.

Pooling with point IC

ETS on air passengers data!

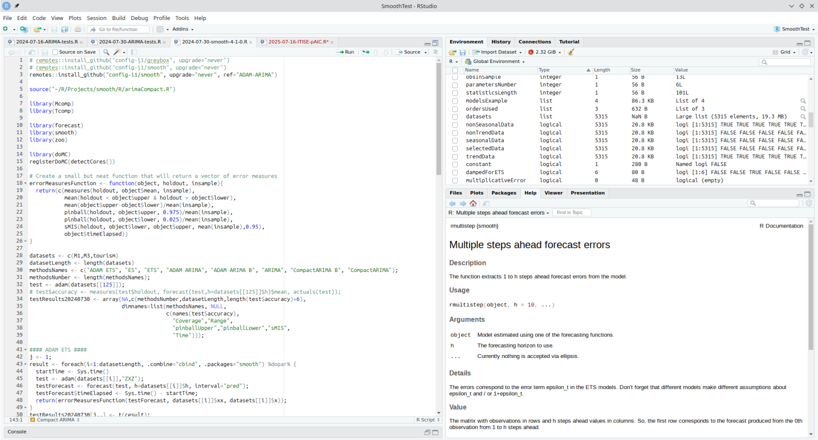


Pooling with point IC

You can do that with CV errors as well (e.g. rolling origin).

Nemenyi can be substituted by other tests.

Real data experiment



Real data experiment

We used M1 (Makridakis et al., 1982), M3 (Makridakis and Hibon, 2000), and Tourism competition data (Athanasopoulos et al., 2011).

A mixture of monthly, quarterly, annual data.

Horizons of 18, 8, and 6.

We measure RMSSE (Athanasopoulos and Kourentzes, 2023), sCE, and Computational Time.

`greybox` (Svetunkov, 2025a) and `smooth` (Svetunkov, 2025b) packages from R (R Core Team, 2024).

Real data experiment

We use ETS to select/combine forecasts (`adam()` function from `smooth`):

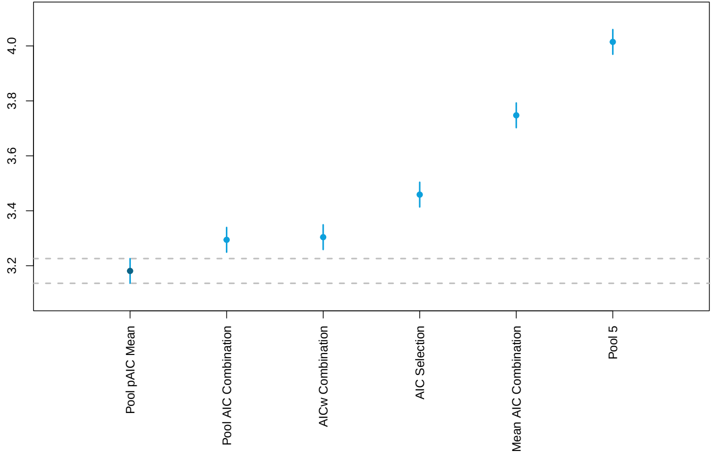
1. **AIC Selection** – apply all models, select the best based on AIC;
2. **AICw Combination** – weighted combination from Kolassa (2011);
3. **Mean Combination** – simple mean of the combination;
4. **Pool pAIC Mean** – form a pool based on Nemenyi (`rmcb()` from the `greybox` package), take the mean;
5. **Pool AIC Combination** - same as (4), but with AIC weights;
6. **Pool 5** - mean of the pool of the best 5 models.

Real data experiment

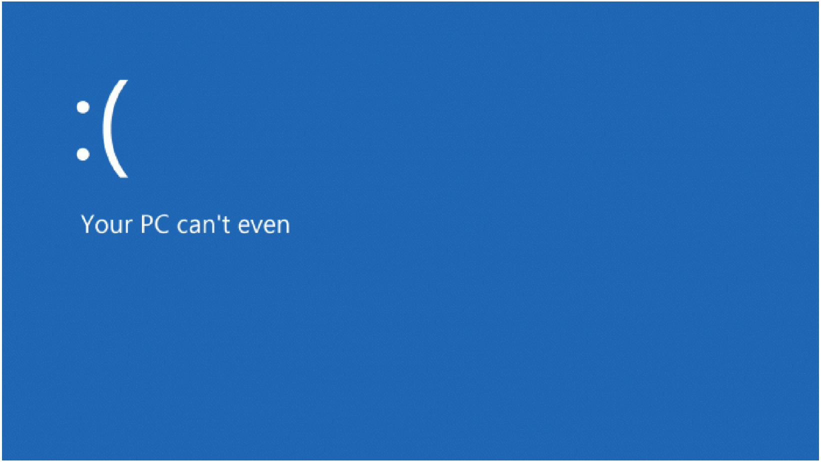
Method	min	Q1	median	Q3	max	mean	sCE	Time
AIC Selection	**0.015**	0.667	1.171	2.278	51.616	1.922	0.452	**0.382**
AICw Combination	0.021	0.658	**1.167**	2.286	51.255	1.904	0.458	0.738
Mean AIC Combination	0.034	0.741	1.258	2.428	372.899	2.116	**0.338**	0.990
Pool pAIC Mean	0.027	**0.653**	1.169	**2.238**	**50.598**	**1.877**	0.450	0.766
Pool AIC Combination	0.021	0.658	1.168	2.283	51.242	1.903	0.455	0.700
Pool 5	0.082	0.786	1.289	2.408	50.973	2.003	0.532	0.397

Real data experiment

The p−value from the significance test is 0.000.
95% confidence intervals constructed.



Conclusions



Conclusions

- All modern combination approaches rely on summary statistics;
- We consider a distribution of point likelihoods;
- This allows forming pools of models, kicking out the bad ones (superforecasters?);
- Their combination is more robust.

Thank you for your attention!

Ivan Svetunkov

i.svetunkov@lancaster.ac.uk

Marketing Analytics and Forecasting

Lancaster University Management School

References I

Carson E Agnew. Bayesian consensus forecasts of macroeconomic variables. *Journal of Forecasting*, 4(4):363–376, 1985.

Hirotugu Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 1974.

George Athanasopoulos and Nikolaos Kourentzes. On the evaluation of hierarchical forecasts. *International Journal of Forecasting*, 39(04):1502–1511, 2023. ISSN 01692070. doi: 10.1016/j.ijforecast.2022.08.003.

George Athanasopoulos, Rob J Hyndman, Haiyan Song, and Doris C Wu. The tourism forecasting competition. *International Journal of Forecasting*, 27(3):822–844, 2011. ISSN 01692070. doi: 10.1016/j.ijforecast.2010.04.009.

References II

Gerda Claeskens, Jan R Magnus, Andrey L Vasnev, and Wendun Wang. The forecast combination puzzle: A simple theoretical explanation. *International Journal of Forecasting*, 32(3): 754–762, 2016.

Janez Demšar. Statistical Comparisons of Classifiers over Multiple Data Sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

Graham Elliott. Averaging and the optimal combination of forecasts. *University of California, San Diego*, 2011.

Graham Elliott and Allan Timmermann. *Economic Forecasting*. Princeton University Press, 1 edition, 2016.

John Geweke and Gianni Amisano. Optimal prediction pools. *Journal of Econometrics*, 164(1):130–141, 2011.

References III

Victor Richmond R Jose and Robert L Winkler. Simple robust averages of forecasts: Some empirical results. *International Journal of Forecasting*, 24(1):163–169, 2008.

Stephan Kolassa. Combining exponential smoothing forecasts using akaike weights. *International Journal of Forecasting*, 27 (2):238–251, 2011.

Nikolaos Kourentzes, Devon Barrow, and Fotios Petropoulos. Another look at forecast selection and combination: evidence from forecast pooling. *International Journal of Production Economics*, 2018.

Spyros Makridakis and Michele Hibon. The m3-competition: results, conclusions and implications. *International journal of forecasting*, 16(4):451–476, 2000.

References IV

Spyros Makridakis, A P Andersen, R Carbone, Robert Fildes, Michèle Hibon, R Lewandowski, J Newton, Emanuel Parzen, and Robert L Winkler. The accuracy of extrapolation (time series) methods: Results of a forecasting competition. *Journal of Forecasting*, 1(2):111–153, 1982. ISSN 02776693. doi: 10.1002/for.3980010202.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2024. URL https://www.R-project.org/.

James H. Stock and Mark W. Watson. Combination forecasts of output growth in a seven-country data set. *Journal of Forecasting*, 23(6):405–430, 2004. ISSN 02776693. doi: 10.1002/for.928.

References V

Ivan Svetunkov. *greybox: Toolbox for Model Building and Forecasting*, 2025a. URL https://CRAN.R-project.org/package=greybox. R package version 2.0.5.

Ivan Svetunkov. *smooth: Forecasting Using State Space Models*, 2025b. URL https://CRAN.R-project.org/package=smooth. R package version 4.3.0.

Aki Vehtari, Andrew Gelman, and Jonah Gabry. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5):1413–1432, September 2017. ISSN 1573-1375. doi: 10.1007/s11222-016-9696-4.