Given *N* Forecasting Models, What To Do?

Fabrizio Culotta

1. Introduction

It is well known that the future is uncertain. Against this uncertainty, economic agents plan their economic activity accordingly. In this planning, producing forecasts of the quantity of interest is the traditional way of uncovering possible not-yet-realized trajectories. Feedback from estimated future dynamics will then influence actual planning and business activities. This is true also for private decision-makers, like firms and other types of organizations, but especially for public policy-makers since their activities produce effects at the whole country level.

The increasing availability of data, together with progress in computational techniques, have incentivized researchers to construct more sophisticated forecasting models and to increase the accuracy of their performances. Nowadays, available forecasting models range from classical econometric models, e.g. ARIMA, to non-parametric models, e.g. exponential smoothing, to machine-learning, e.g. trees and neural networks. It results in a plethora of single forecasting models available to both private and public decision-makers. Since the late '70s, a group of academic researchers proposed the idea of competition among different forecasting models (Makridakis et al., 1982). It emerged that statistically sophisticated models do not necessarily produce more accurate forecasts, whereas combinations of them outperform vis-$\acute{a}$-vis single models. Moreover, the ranking of forecasting models depends on the accuracy measure being as well as on the adopted forecast horizon. The success of the first so-called *M-competition* (M stands to Makridakis) allowed us to carry on the tradition of forecasting competitions (Hyndman, 2020) until today with the recent M4 and M5 competitions (Petropoulos and Makridakis, 2020; Makridakis et al., 2021). Given a set of time series at different frequencies, several models compete to produce the best forecast. Models? performances are then ranked based on some accuracy measures. Based on the idea of competition among different forecasting methods, this work compares their forecasting performances on a given time horizon. Unlike the tradition of Ms competitions, which are based on thousands of time series at different time frequencies, a single univariate time series is selected at the monthly frequency.

The motivation of this choice is to show that, in the simplest exercise of forecasting a single time series, the ex-ante choice of the model is likely to be misleading because a model ranking exists and it is specific to time (hence, frequency) and of measurement object of the single series. Indeed, when a set of forecasting models is available, a semi-automatic algorithm of model selection based on some performance measures would be a superior choice for the various decision-makers. In the case at hand, the choice of the monthly unemployment rate is dictated by the fact that it is the most common measure of the (mis-)functioning of the labour market and, as such, is of utmost importance for policymakers.

Forecasting models are finally ranked based on some accuracy measures. The main findings confirm that, given N forecasting models, combination techniques outperform single uncombined models in terms of accuracy and reduce the risk of adopting a single forecasting model.

Fabrizio Culotta, University of Genoa, Italy, fabrizio.culotta@edu.unige.it, 0000-0002-3910-3088

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Fabrizio Culotta, *Given N Forecasting Models, What To Do*, © Author(s), CC BY 4.0, DOI 10.36253/979-12-215-0106-3.55, in Enrico di Bella, Luigi Fabbris, Corrado Lagazio (edited by), *ASA 2022 Data-Driven Decision Making. Book of short papers*, pp. 317-322, 2023, published by Firenze University Press and Genova University Press, ISBN 979-12-215-0106-3, DOI 10.36253/979-12-215-0106-3

2. Forecasting Models

The comparative forecasting exercise presented in this work comprises a set of 23 different uncombined and combined models. The selected time series on which all models are trained is the deseasoned dynamics of the Italian unemployment rate over the years $2004 - 2019$ at the monthly frequency freely available from the ISTAT data warehouse (http://dati.istat.it/). The observational period is split between the training set, from January 2004 to June 2019, and the test set, from July to December 2019. The set of selected forecasting models contains some ARIMA-like models, some Exponential Smoothing models, to machine learning models. It also contains combinations of them based on some model averaging techniques. For sake of brevity, the succinct list is reported in table 1. All the computations are carried out with the statistical software $\mathcal{R}$ by using the most recent packages. Model specifications and other details can be provided upon request.

FAMILY	**Label**	**Model**	Reference	**R package**
ARIMA	ARIMA	ARIMA	Hyndman and Khandakar (2008)	forecast
	ARFIMA	Fractionally-differenced ARIMA	Peiris and Perera (1988)	forecast
	GARMA	Gegenbauer-ARIMA	Dissanayake et al. (2016)	garma
	SSARIMA	State-space ARIMA	Svetunkov and Boylan (2020)	smooth
Exponential Smoothing	ES	Exponential Smoothing	Brown (1956)	ets
	HOLT	Linear Exponential Smoothing	Holt and Modigliani (1960)	forecast
	THETA	Exponential Smoothing with drift	Assimakopoulos and Nikolopoulos (2000)	forecast
	CES	Complex Exponential Smoothing	Svetunkov and Kourentzes (2018)	smooth
	GUM	State-space Exponential Smoothing	Svetunkov and Kourentzes (2018)	smooth
Machine Learning	ARML	Bagged AR	Bergmeir et al. (2016)	caretForecast
	BAG	Bagged Exponential Smoothing		forecast
	NN	Fast-forward Neural Network		forecast
Hybrid	ADAM	Augmented Dynamic Adaptive Model	Hyndman and Khandakar (2008)	smooth
	BATS	GUM with ARMA errors	De Livera et al. (2011)	forecast
	ATA	Combination of ES and ARIMA	Yapar et al. (2017)	ATAforecasting
	SPL	Cubic Spline	Chambers and Hastie (2017)	forecast
Combinations	COMB1	Combination of ETS,SSARIMA,GUM and CES		smooth
	COMB2	Combination of ARIMA,ETS,THETA,NN and BATS		forecastHybrid
	COMB3	Combination of ARML and SPL with simple weights		ForecastCombinations
	COMB4_BG	COMB3 with Bates-Granger weights		ForecastComb
	COMB4_InW	COMB3 with Inverse Rank approach		ForecastComb
	COMB4_Me	COMB3 with Dynamic weighting scheme		ForecastComb
	COMB5	Combination of all models except COMBs		ForecastCombinations

Table 1: *Selection of forecasting models.*

Once all forecasting models have been estimated, it is interesting to compare statistics of model fitting in terms of moments of the corresponding error distribution. At this aim, table 2 below provides rank values (column RANK) for each forecasting model based on a total score (SCORE). The latter statistics is computed as the sum of the single scores reported in terms of mean (RANK_MEAN), standard deviation (RANK_SD), skewness (RANK_SKEWNESS), and kurtosis (RANK_KURTOSIS).

FAMILY	MODEL	RANK_MEAN	RANK_SD	RANK_SKEWNESS	RANK_KURTOSIS	SCORE	RANK
ARIMA	ARIMA	2	14	23	10	49	13
	ARFIMA	20	15	20	7	62	19
	GARMA	15	9	11	6	41	11
	SSARIMA	21	18	19	18	76	23
Exponential Smoothing	ES	14	8	9	3	34	5
	HOLT	12	7	10	4	33	4
	THETA	22	22	15	16	75	21
	CES	5	20	17	19	61	18
	GUM	23	21	16	15	75	21
Machine Learning	ARML	18	11	1	23	53	14
	BAG	19	12	13	9	53	14
	NN	1	23	5	8	37	8
Hybrid	ADAM	17	17	22	14	70	20
	BATS	13	10	12	5	40	10
	ATA	3	19	14	12	48	12
	SPL	6	6	8	1	21	1
Combinations	COMB1	4	16	21	13	54	16
	COMB2	16	13	18	11	58	17
	COMB3	9	3	2	21	35	7
	COMB4_BG	7	1	7	17	32	3
	COMB4_InvW	8	2	4	20	34	5
	COMB4_MED	10	4	3	22	39	9
	COMB5	11	5	6	2	24	2

Table 2: *Ranking of forecasting models in terms of model fitting*.

What emerges from table 2 is that, in terms of model fitting, the best-performing forecasting model is SPL followed by COMB5, COMB4_BG, COMB4_InW, and so on. In detail, the error distribution of the NN model is associated with the lowest mean error, COMB4_BG with the lowest dispersion. Whereas ARML and SPL are characterized by the lowest skewness and kurtosis, respectively. Despite model fitting being an important quality feature of forecasting models, it is not the definitive dimension to consider when a decision-maker needs to adopt a single forecasting model. As shown in the next section, the accuracy of forecasting performances may deliver different conclusions.

3. Results

Figure 1 shows the forecasts produced by each model on the test set over a time horizon of six months. It is possible to observe that ARML model fails in capturing the dynamics of actual data despite its model fitting performances being characterized by the lowest skewness. On the contrary, the COMB2 forecasts closely mimic the dynamics of the Italian unemployment rate despite its model fitting performance are not the best in any moments of the error distribution.

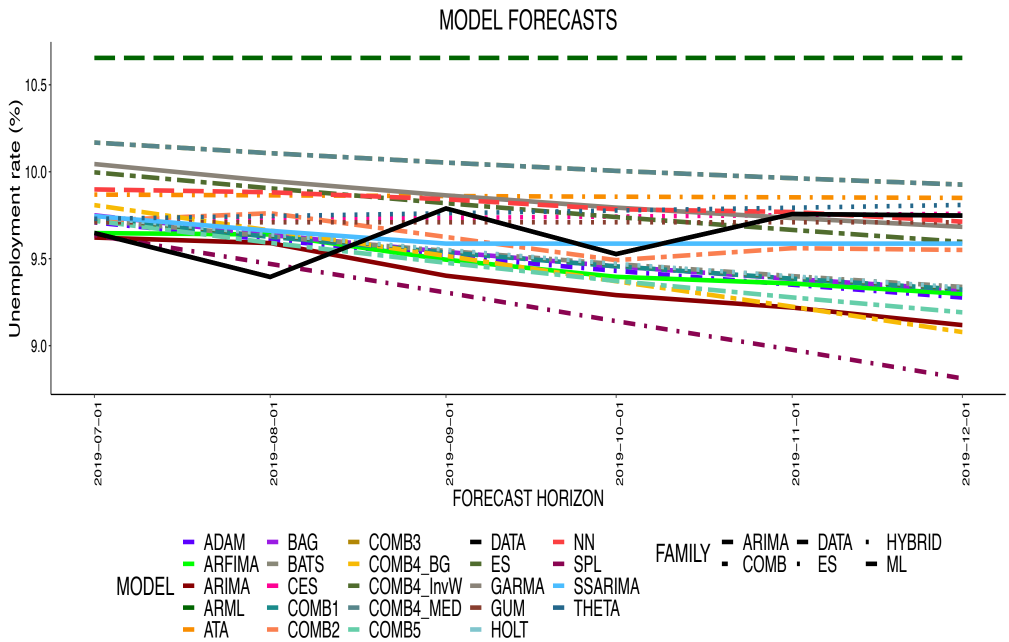

Figure 1: *Forecasts of Italian unemployment rate. ARIMA models (solid line): ARFIMA, ARIMA, GARMA, SSARMA. Combinations (COMB, two-dashed line): COMB1, COMB2, COMB3, COMB4_BG, COMB4_InvW, COMB4_MED. Exponentional Smoothing (ES, dotted line): CES, ES, GUM, HOLT, THETA. Hybrid models (dot-dashed line): ADAM, ATA, BATS, SPL. Machine Learning models (ML, long-dashedline): ARML, BAG, NN.*

These considerations confirm that model fitting, despite being an important aspect to consider for the selection of forecasting models, does not necessarily ensure that forecast performances are aligned with model fitting performances. Instead, the use of various ensembling techniques delivers satisfactory results compared to those of single uncombined models. On this point, note also from figure 1 that the actual dynamics of the unemployment rate is contained within the full set of forecasts. This means that a suitable model combination can be obtained by ensembling appropriately some of the models under scrutiny.

Finally, table 3 provides the values of various accuracy measures used in the various forecasting competitions: ME (mean error), MAE (mean absolute error), MPE (mean percentage error), MSE (mean squared error), MAPE (mean absolute percentage error), RMSSE (root mean squared scaled error), RAME (relative absolute mean error), RMAE (root mean absolute error) and RRMSE (relative root mean squared error).

FAMILY	MODEL	ME	MAE	MPE	MSE	MAPE	RMSSE	RAME	RMAE	RRMSE	SCORE	RANK
ARIMA	ARIMA	19	18	19	18	18	18	19	18	18	165	18
	ARFIMA	14	15	13	15	15	15	14	15	15	131	16
	GARMA	15	10	15	16	10	16	15	10	16	123	14
	SSARIMA	1	4	1	3	4	3	1	4	3	24	3
Exponential Smoothing	ES	9	12	8	11	12	11	9	12	11	95	11
	HOLT	6	11	6	10	11	10	6	11	10	81	9
	THETA	7	3	10	4	3	4	7	3	4	45	5
	CES	4	1	4	2	1	2	4	1	2	21	2
	GUM	3	2	3	1	2	1	3	2	1	18	1
Machine Learning	ARML	23	23	23	23	23	23	23	23	23	207	23
	BAG	10	14	9	13	14	13	10	14	13	110	13
	NN	13	6	14	6	6	6	13	6	6	76	7
Hybrid	ADAM	12	16	12	14	16	14	12	16	14	126	15
	BATS	8	9	7	9	9	9	8	9	9	77	8
	ATA	17	7	17	7	7	7	17	7	7	93	10
	SPL	22	22	22	22	22	22	22	22	22	198	22
Combinations	COMB1	11	13	11	12	13	12	11	13	12	108	12
	COMB2	2	5	2	5	5	5	2	5	5	36	4
	COMB3	20	19	20	19	19	19	20	19	19	178	19
	COMB4_BG	18	21	18	21	21	21	18	21	21	180	21
	COMB4_InvW	5	8	5	8	8	8	5	8	8	63	6
	COMB4_MED	20	19	20	19	19	19	20	19	19	178	19
	COMB5	16	17	16	17	17	17	16	17	17	150	17

Table 3: *Ranking of forecasting models in terms of accuracy measures.*

As expected, the overall rank of forecasting models in terms of accuracy measures differs from the ranking in terms of model fitting presented in table 2. Now, the best-performing forecasting model is GUM, followed by CES and SSARIMA. Among all model combinations, only COMB2 and COMB4_InvW lie in a good position, being the fourth and the sixth best performing models respectively. Forecasting models SPL and ARML occupy the next-to-last and last positions, respectively.

4. Conclusions

Results confirm that it does not exist yet a single superior universal model. On the contrary, the ranking of different forecasting models is specific to the adopted training set. For example, when the time series of interest switches to the employment rates instead of unemployment rates, the rank of model performances changes. Secondly, results confirm that performances of machine learning and neural network models offer satisfactory alternatives to the traditional econometric models like ARIMA or the non-parametric Exponential Smoothing. Finally, the results stress the importance of model ensemble techniques as a solution to model uncertainty as well as a tool to improve forecast accuracy (Shaub, 2020).

Overall, the flexibility provided by a rich set of forecasting models, and the possibility to combine them, together represent an advantage for decision-makers often constrained to adopt solely pure, uncombined, forecasting models.

References

Assimakopoulos, V. and Nikolopoulos, K. (2000). The theta model: a decomposition approach to forecasting. *International Journal of Forecasting*, 16(4):521–530.

Bergmeir, C., Hyndman, R. J., and Benítez, J. M. (2016). Bagging exponential smoothing methods using stl decomposition and box-cox transformation. *International Journal of Forecasting*, 32(2):303–312.

Brown, R. G. (1956). Exponential smoothing for predicting demand. *NBER*.

Chambers, J. M. and Hastie, T. J. (2017). Statistical models. In *Statistical models in S*, pages 13–44. Routledge.

De Livera, A. M., Hyndman, R. J., and Snyder, R. D. (2011). Forecasting time series with complex seasonal patterns using exponential smoothing. *Journal of the American Statistical Association*, 106(496):1513–1527.

Dissanayake, G., Peiris, M. S., and Proietti, T. (2016). State space modeling of gegenbauer processes with long memory. *Computational Statistics & Data Analysis*, 100:115–130.

Holt, M. and Modigliani, J. (1960). Planning production, inventories, and workforce.

Hyndman, R. J. (2020). A brief history of forecasting competitions. *International Journal of Forecasting*, 36(1):7–14.

Hyndman, R. J. and Khandakar, Y. (2008). Automatic time series forecasting: the forecast package for r. *Journal of Statistical Software*, 27:1–22.

Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, E., and Winkler, R. (1982). The accuracy of extrapolation (time series) methods: results of a forecasting competition. *Journal of Forecasting*, 1(2):111–153.

Makridakis, S., Spiliotis, E., Assimakopoulos, V., Chen, Z., Gaba, A., Tsetlin, I., and Winkler, R. L. (2021). The m5 uncertainty competition: results, findings and conclusions. *International Journal of Forecasting*.

Peiris, M. and Perera, B. (1988). On prediction with fractionally differenced arima models. *Journal of Time Series Analysis*, 9(3):215–220.

Petropoulos, F. and Makridakis, S. (2020). The m4 competition: bigger. stronger. better. *International Journal of Forecasting*, 36(1):3–6.

Shaub, D. (2020). Fast and accurate yearly time series forecasting with forecast combinations. *International Journal of Forecasting*, 36(1):116–120.

Svetunkov, I. and Boylan, J. E. (2020). State-space arima for supply-chain forecasting. *International Journal of Production Research*, 58(3):818–827.

Svetunkov, I. and Kourentzes, N. (2018). Complex exponential smoothing for seasonal time series.

Yapar, G., Yavuz, İ., and Selamlar, H. T. (2017). Why and how does exponential smoothing fail? an in depth comparison of ata-simple and simple exponential smoothing. *Turkish Journal of Forecasting*, 1(1):30–39.