MDPI

Article

On the disagreement of forecasting model selection criteria

Evangelos Spiliotis [1]*, Fotios Petropoulos [2] and Vassilios Assimakopoulos [1]

[1] Forecasting and Strategy Unit, School of Electrical and Computer Engineering, National Technical University of Athens, Greece
[2] School of Management, University of Bath, UK
* Correspondence: spiliotis@fsu.gr

Abstract: Forecasters have been using various criteria to select the most appropriate model from a pool of candidate models. This includes measurements on the in-sample accuracy of the models, information criteria, and cross-validation, among others. Although the latter two options are generally preferred due to their ability to tackle overfitting, in univariate time series forecasting settings limited work has been conducted to confirm their superiority. In this study we compare such popular criteria for the case of the exponential smoothing family of models using a large data set of real series. Our results suggest that there is significant disagreement between the suggestions of the examined criteria and that, depending on the approach used, models of different complexity may be favored with possibly negative effects on forecasting accuracy. Moreover, we find that simple in-sample error measures can effectively select forecasting models, especially when focused on the recent observations of the series.

Keywords: Model Selection; Information Criteria; Time Series; Exponential Smoothing; M4 Competition

1. Introduction

The "no free lunch" theorem [1] suggests that *"...for any algorithm, any elevated performance over one class of problems is offset by performance over another class"*. This theorem also holds true in time series forecasting settings, meaning that no model can optimally forecast any series and, subsequently, that the most appropriate model should be identified per case to improve overall forecasting accuracy. Indeed, if model selection could be done perfectly, then the accuracy gains would be substantial [2]. Unfortunately, due to the uncertainty involved in the process [3], the task of model selection has been proven to be very challenging in practice, especially when performed for numerous series [4] or for data that involve anomalies, outliers, and shifts [5]. To that end, the forecasting literature has developed various criteria, rules, and methods to improve model selection and automate forecasting.

Early attempts on model selection involve the utilization of information criteria, such as the Akaike's information criterion [AIC; 6] and the Bayesian information criterion [BIC; 7], choosing the most appropriate model by comparing their ability to fit the historical observations with the number of parameters used for producing forecasts. Other attempts include rule-based selections [8], i.e., using a set of heuristics to define when a model should be preferred over another. These rules, that typically build on time series features (e.g., strength of trend and seasonality), can also be determined analytically by processing the forecast errors of various models across different series of diverse features [9]. The rise of machine learning has facilitated the development of such feature-based model selection algorithms, also called "meta-learners" [10]. Another promising alternative is to select models judgmentally, allowing that way the incorporation of human experience [11] and avoiding making unreasonable choices [12]. In any case, particular emphasis should be given on the pool of candidate models considered for making selections [13].

Citation: Spiliotis, E.; Petropoulos, F.; Assimakopoulos, V. On the disagreement of forecasting model selection criteria. *Forecasting* **2023**, *1*, 1–12. https://doi.org/

Received:
Revised:
Accepted:
Published:

Another direction for selecting forecasting models is based on cross-validation [14], specially designed to overcome issues related with overfitting and to enhance the generalization of the forecasting methods used. In time-series forecasting applications, cross-validation variants include blocked approaches that mitigate the problem of serial correlations [15,16] and techniques that omit data close to the period used for evaluating performance [17]. Therefore, along with information criteria, cross-validation is nowadays among the most popular approaches used for effective model selection [18].

Despite the development of several model selection approaches, researchers and practitioners have been relatively indecisive about the criteria that they should use in practice for identifying the most accurate forecasting models. Each criterion comes with particular advantages and limitations, rendering often their use subject to the judgment and the experience of the forecaster or even the settings of the software utilized for producing the forecasts. We argue that tuning model selection processes is critical for improving accuracy and that empirical evidence should be exploited to define which criterion should be considered for the data set and forecasting application at hand. To motivate our argument, we evaluate some of the most widely used model selection criteria on a large data set of real series, considering the exponential smoothing family of models, a standard method for time series forecasting. Our analysis focuses on both the precision of the criteria and the forecasting accuracy of the underlying selection approaches, providing evidence about the way the two measures are correlated. Also, we investigate the disagreement of the examined criteria and discuss its implications for forecasting practice.

The rest of the paper is organized as follows. Section 2 provides an introduction to the model selection criteria used in our study. Section 3 presents the exponential smoothing family of models, forming the pool of candidate models used in our experiments. Section 4 evaluates empirically the performance of the selection criteria, describing the experimental setup and discussing the results. Finally, Section 5 concludes the paper.

2. Model selection criteria

2.1. Criteria based on in-sample accuracy measurements

The simplest and fastest approach to select a forecasting model from a pool of candidate models is to compare their accuracy in the in-sample of the series. This is because in-sample accuracy can be directly measured when fitting the models, requiring no further computations. The most common measures used in this direction are the mean squared error (MSE) and the mean absolute error (MAE), defined as follows:

$$MSE = \frac{1}{n}\sum_{t=1}^{n}(y_t - f_{t|n})^2, \tag{1}$$

$$MAE = \frac{1}{n}\sum_{t=1}^{n} \mid y_t - f_{t|n} \mid, \tag{2}$$

where n is the sample size (number of in-sample observations), y_t is the observed value of the series being forecast at point t, and $f_t|_n$ is the forecast provided by the model given n observations for estimating its parameters. Smaller values of MAE and MSE suggest better fit.

Although other measures (e.g., measures based on percentage errors, relative errors, relative measures, and measures based on scaled errors) can be used instead of MSE and MAE, given their limitations [19] and the fact that model selection is typically performed in a series-by-series fashion, the aforementioned scale-dependent measures can be considered sufficient for the described task.

In terms of statistical properties, MAE is an appropriate measure for evaluating the ability of a forecasting model to specify the median of the future values of a series [20], while MSE for measuring the ability of a model to specify its mean [21]. Moreover, since MSE builds on squared errors, it penalizes more large errors than small ones and is therefore

more sensitive to outliers than MAE [22]. Therefore, although MSE has long been the standard measure of choice for selecting and optimizing time series forecasting models, it becomes evident that there may be settings where MAE provides superior results.

Model selection criteria that build on in-sample measurements theoretically come with two major issues. First, since they focus just on how well the model fits the historical observations, they are prone to overfitting. In general, sophisticated models that consist of multiple parameters have the capacity to fit series better than simpler models, although the latter may result in more accurate post-sample forecasts. Second, since in-sample measurements evaluate accuracy across the complete sample of historical observations, they may favor models that do not necessarily perform well in the most recent past. Typically, the latest information available is of higher importance for producing accurate post-sample forecasts, meaning that models should put particular emphasis on the last part of the series, especially when its underlying patterns (e.g., seasonality and trend) have changed.

Although the first issue can be tackled by the criteria presented in the following two subsections, the second one can be mitigated by simply adjusting the time window on which the accuracy measures are computed. For the sake of simplicity, in this study we consider some variants of the MSE and MAE measures, to be called MSEh and MAEh, that capture the in-sample accuracy of the examined models in the last h observations of the series, as follows:

$$MSEh = \frac{1}{h} \sum_{t=n-h+1}^{n} (y_t - f_{t|n})^2, \tag{3}$$

$$MAEh = \frac{1}{h} \sum_{t=n-h+1}^{n} | y_t - f_{t|n} |, \tag{4}$$

where h is the forecasting horizon, i.e., the number of periods the model is tasked to forecast in the post-sample. Although h can in principle be selected based on the particular requirements of the forecasting task at hand (e.g., set equal to a full calendar year or the last observation), we argue that the forecasting horizon is a reasonable and practical alternative for determining the time window. Moreover, by selecting a sufficiently large evaluation window (e.g., greater than 1 observation), the results are expected to be more representative and less sensitive to potential extreme values.

2.2. Information criteria

Information criteria have become particularly popular for model selection as they are fast to compute but can also mitigate overfitting [23]. To do so, instead of selecting the model that best fits the series, as measured by an accuracy measure in the in-sample data, they make choices by penalizing the in-sample accuracy of the candidate models according to their complexity, as realized based on the number of parameters that have to be estimated to form each model.

Specifically, information criteria build on complexity-penalized maximum likelihood estimations (as described in Section 3 for the case of exponential smoothing models). The most notable variants of information criteria include the AIC, the AIC corrected for small sample sizes (AICc), and the BIC, defined as follows:

$$\text{AIC} = -2\ln(L) + 2k, \tag{5}$$

$$\text{AICc} = \text{AIC} + \frac{2k(k+1)}{n-k-1}, \tag{6}$$

$$\text{BIC} = \text{AIC} + k(\ln(n) - 2), \tag{7}$$

where L is the likelihood and k is the total number of parameters. In all cases, smaller values imply better performance. As seen, by definition, BIC assigns larger penalties to more sophisticated model than AIC, thus favoring simpler models of comparable likelihood. The same stands for AICc, provided that the sample size is relatively small (AICc approximates AIC as the sample size increases).

Although there have been theoretical arguments over the use of particular information criteria over others, Burnham and Anderson [24] demonstrate that AIC and AICc can be derived in the same Bayesian framework as BIC, just by using different prior probabilities. As a result, information criterion selection should be based on the assumptions made about reality and models, taking also into consideration empirical evidence about the forecasting performance of each criterion in the application at hand. Simulation studies suggest that AICc tends to outperform BIC and also recommend its use over AIC, even when sample sizes are relatively large [25,26]. This may justify the utilization of AICc in popular model selection software, such as the `auto.arima` and `ets` functions of the *forecast* package for R that allow the automatic selection of ARIMA and exponential smoothing models, respectively [27].

2.3. Criteria based on cross-validation

Cross-validation is another approach for selecting among forecasting models. The greatest advantage of this approach is that it focuses on the post-sample accuracy of the candidate models, thus selecting models that perform well in their actual task, regardless how well they managed to fit the historical data, while also making no assumptions about how model complexity should be measured or penalized. As a result, cross-validation has become very popular for model selection and optimization, especially in applications that involve sophisticated models (e.g., neural networks) where standard selection criteria are either challenging to compute or sensitive to overfitting. On the negative side, cross-validation is applicable only to series that have enough observations to allow the creation of hold-out samples and computationally expensive (the more the hold-out samples created, the greater the cost becomes).

In time-series forecasting settings, where data is non-stationary and have serial dependencies, cross-validation is typically implemented using the rolling origin evaluation approach [28]. According to this approach, a period of historical data $N \leq n - h$ is first used for fitting a forecasting model. Then, the model is used to produce h-step-ahead forecasts and its accuracy is assessed based on the actual values of the series in the corresponding period using a measure of choice (e.g., MSE or MAE). Subsequently, the forecast origin is shifted by T periods, the model is re-fitted using the new in-sample data ($N + T$), and new forecasts are produced, contributing another assessment. This process is repeated till there are no data left for testing and the overall performance of the model is determined based on its average accuracy over the conducted evaluations.

For the sake of simplicity, and to accelerate computations, in this study we consider a fixed origin evaluation, i.e., a single evaluation set that consists of the last h observations of the original in-sample data ($N = n - h$). Also, in accordance to the in-sample selection measures used in our study, we use MSE and MAE for assessing forecasting post-sample accuracy. The examined cross-validation approaches, to be called MSEv and MAEv, are defined as follows:

$$MSEv = \frac{1}{h} \sum_{t=n-h+1}^{n} (y_t - f_{t|n-h})^2, \tag{8}$$

$$MAEv = \frac{1}{h} \sum_{t=n-h+1}^{n} | \, y_t - f_{t|n-h} \, | \, . \tag{9}$$

Note that MSEv/MAEv are computed on the same sample as MSEh/MAEh. The only difference is that the forecasts used with the MSEv/MAEv criteria are computed using $n - h$

Table 1. Exponential smoothing family of models. Applicable models are highlighted in bold face.

Additive Error				Multiplicative Error			
	Seasonality				Seasonality		
Trend	N	A	M	Trend	N	A	M
N	**ANN**	**ANA**	ANM	N	**MNN**	**MNA**	**MNM**
A	**AAN**	**AAA**	AAM	A	**MAN**	**MAA**	**MAM**
Ad	**AAdN**	**AAdA**	AAdM	Ad	**MAdN**	**MAdA**	**MAdM**
M	AMN	AMA	AMM	M	MMN	MMA	MMM
Md	AMdN	AMdA	AMdM	Md	MMdN	MMdA	MMdM

observations, while the forecasts used with the MSEh/MAEh criteria using the complete in-sample (n observations).

3. Forecasting models

Exponential smoothing, originally introduced by Brown [29], is considered the workhorse of time series forecasting, being among the oldest and simplest, yet most effective and widely used methods for univariate predictions [for an encyclopedic review on exponential smoothing, please refer to section 2.3.1 of 30]. The key advantage of the method is that it is fast to compute [31], easy to implement in software [32], and results in competitive accuracy compared to more sophisticated methods in various applications, including financial, economic, demographic, and demand data, among others, as demonstrated empirically by recent forecasting competitions [33,34]. In addition, the forecasts are produced based on intuitive models that are closely connected with key time series features (e.g., trend and seasonality), thus being easy to communicate with managers or be adjusted based on judgment [12].

The key idea behind exponential smoothing is that more recent observations are more valuable to forecasting. As a result, the method produces forecasts by putting exponentially more weight on the most recent past of the series. The degree on which the weight is decreased as we move further to the past is determined by the smoothing parameters of the model. Moreover, according to its state-space expression [23], typically referred as ETS, exponential smoothing can be realized as a combination of three components, namely the error (E), trend (T), and seasonal (S) components. The error component can be either additive (A) or multiplicative (M), while the trend and seasonal components can be none (N), additive (A), or multiplicative (M). In addition, additive and multiplicative trends can be damped (d), if needed. Consequently, the ETS framework involves a total of 30 exponential smoothing models (or model forms) that can be acronymized using the respective symbols of the three components, as shown in Table 1. From these models, some may result to infinite forecast variances for long forecast horizons [35], while others involve a multiplicative trend that is not recommended to use in practice [36]. Therefore, the `ets` function of the *forecast* package for R, used for implementing exponential smoothing in our study, limits the candidate models to 15 for seasonal data and 6 for non-seasonal data.

As an example, the simplest form of exponential smoothing, that accounts just for level variations (simple exponential smoothing or ANN) can be expressed as:

$$\begin{aligned} \hat{y}_{t+h} &= l_t, \\ l_t &= \alpha y_t + (1-\alpha) l_{t-1}, \end{aligned}$$

where l_t is the state of the level component at period t and α is the smoothing parameter of that level. Note that calculating the state of the component at $t = 1$ (l_1) requires the estimation of some initial state values (l_0). Effectively, this is another parameter of the model and its estimation is part of the fitting process.

Table 2. Categorization of ETS models based on their complexity, i.e., number of estimated components.

Complexity	Models
Low	ANN, MNN
Moderate	AAN, MAN, ANA, MNA, MNM
Significant	AAdN, MAdN, AAA, MAA, MAM
High	AAdA, MAdA, MAdM

As more components are added to the model, more equations and parameters are considered for producing the forecasts. This results in more generic models that can effectively account for more complicated time series patterns, such as trend [37] and seasonality [38]. However, as discussed earlier, more sophisticated models involve more parameters and, as a result, higher parameter uncertainty [3], possibly rendering their use less accurate compared to other, simpler model forms. By default, `ets` uses likelihood for estimating the parameters of the models and AICc for selecting the most appropriate model form. Depending on the form of the error component, likelihood is defined as follows:

$$L_A = -\frac{n}{2}\ln\left(\sum_{t=1}^{n}(y_t - f_{t|n})^2\right), \tag{10}$$

$$L_M = -\frac{n}{2}\ln\left(\sum_{t=1}^{n}\left(\frac{y_t - f_{t|n}}{f_{t|n}}\right)^2\right) - \sum_{t=1}^{n}\ln(\mid f_{t|n} \mid), \tag{11}$$

where L_A and L_M correspond to the likelihood of the models that involve an additive and a multiplicative error component, respectively.

In terms of parameters k used, all exponential smoothing models involve a minimum of three parameters, namely α, l_0, and σ, which corresponds to the standard deviation of the residuals. Then, models with a trend component will involve two additional parameters (for smoothing and initializing trend), models with a damped trend component will involve three additional parameters (for initializing, smoothing, and damping trend), while models with a seasonal component will involve $s + 1$ additional parameters (s for initializing and one for smoothing seasonality), where s is equal to the seasonal periods of the data (e.g., 12 for monthly and 4 for quarterly series). For the sake of simplicity, and to allow comparisons between series of different seasonal periods, in this study we categorize ETS models based on their complexity in four categories, namely "Low", "Moderate", "Significant", and "High" complexity, as presented in Table 2. As seen, models of low complexity involve only estimations about the level of the series, models of moderate complexity involve estimations either about the trend or the seasonality of the series, models of significant complexity involve estimations either about a damped trend or seasonality, while models of high complexity are both damped and seasonal.

4. Empirical evaluation

4.1. Experimental setup

To empirically evaluate the performance of the model selection criteria described in Section 2, we consider a subset of the M4 competition data [33]. The M4 data set is frequently used for benchmarking by the forecasting community as it originally involves 100,000 series from various domains (micro, macro, industry, finance, demographic, and other) and frequencies (yearly, quarterly, monthly, weekly, daily, and hourly) that are very diverse and representative of several real-life applications [39]. Moreover, since it is publicly available, it facilitates replication of results and allows comparisons with past and future studies [40]. A detailed description of the data set and the structural characteristics of its series, including their length, forecastability, trend, seasonality, linearity, stability, skewness,

kurtosis, non-linear autoregressive structure, and self-similarity is provided by Spiliotis *et al.* [39].

Specifically, we consider the yearly, quarterly, and monthly series of the M4 data set (95,000 series), but exclude those that are too short in length to perform cross-validation. Therefore, the data set used in our study involves a total of 91,444 series (20,077 yearly, 23,760 quarterly, and 47,607 monthly).

For each series, we fit all the exponential smoothing models that are recommended to use with the *forecast* package, i.e., a total of 15 models for seasonal and 6 models for non-seasonal series, respectively, as presented in Table 1. Then, we employ the examined model selection criteria and, based on their recommendations, we track the forecasting accuracy of each approach, as measured by the mean absolute scaled error (MASE), defined as follows:

$$\text{MASE} = \frac{1}{h} \frac{\sum_{t=n+1}^{n+h} |y_t - f_{t|n}|}{\frac{1}{n-s}\sum_{i=s+1}^{n} |y_i - y_{i-s}|}. \tag{12}$$

MASE, originally proposed by Hyndman and Koehler [41], is equal to the mean absolute error scaled by the in-sample one-step-ahead mean absolute error of seasonal Naive. MASE was preferred over other accuracy measures as it has been officially used in M4 for ranking the original submissions and it has better statistical properties (it is independent of the data scale, becomes infinite or undefined only when all the errors of the Naive method are equal to zero, has a defined mean and finite variance, and penalizes equally positive and negative forecast errors [19]).

In addition to MASE, we also track the number of series where the criteria have successfully identified the most accurate model in terms of out-of-sample MASE accuracy. To that end, it is possible to evaluate both the precision of the criteria (proportion of selections that were actually the "best") and the effect of such a precision score on the post-sample accuracy (absolute accuracy, as measured by MASE). This distinction is important since higher precision may not necessarily lead to better accuracy, meaning that a criterion of lower precision that somehow avoids selecting the worst alternatives could perform better overall than an approach that selects more frequently the best option but also often chooses some of the worst possible forecasting models [42].

Finally, in order to add depth in our analysis, and given that the M4 considered the symmetric mean absolute percentage error (sMAPE) in addition to MASE to evaluate the forecasting accuracy of the submitted methods, we also utilize sMAPE, which is defined as follows:

$$\text{sMAPE} = \frac{2}{h} \sum_{t=n+1}^{n+h} \frac{|y_t - f_{t|n}|}{|y_t| + |f_{t|n}|} * 100\%. \tag{13}$$

4.2. Results and discussion

Table 3 summarizes the forecasting accuracy of the examined criteria, both per data frequency and in total. As expected, criteria that build on cross-validation (MSEv and MAEv) result in better forecasts overall. However, this result is mostly driven by the superior performance of these criteria in the yearly data. If we focus on the seasonal series, we find that criteria that build on in-sample accuracy measurements (MAE and MSE) lead to more accurate forecasts. In fact, we can see that the average rank of the models selected by MAE is the lowest among the criteria considered for all the examined data frequencies. Also, interestingly enough, MAEh and MSEh are consistently less accurate approaches than MAE and MSE, respectively, indicating that focusing on the last part of the series does not guarantee better model selection performance.

When it comes to information criteria, we observe that balancing complexity with in-sample accuracy is critical. Specifically, we find that AIC, which considers a relatively small penalty for complexity, does not improve accuracy over L, being slightly worse

Table 3. Forecasting accuracy (MASE) of the examined criteria used for model selection. The average rank of the selected models is also displayed. The results are presented per data frequency and for the complete data set. The bold numbers highlight the most accurate criterion per case.

Criterion	MASE				Average Rank			
	Yearly	Quarterly	Monthly	Total	Yearly	Quarterly	Monthly	Total
MSE	3.471	1.151	0.923	1.542	3.200	6.865	6.675	5.962
MAE	3.441	**1.141**	**0.921**	1.531	**3.177**	**6.721**	**6.543**	**5.850**
MSEh	3.485	1.162	0.925	1.548	3.240	6.962	6.663	5.989
MAEh	3.459	1.163	0.924	1.543	3.219	6.959	6.656	5.980
L	3.432	1.159	0.934	1.541	3.180	6.906	6.728	5.995
AIC	3.436	1.158	0.936	1.543	3.246	6.900	6.739	6.014
AICc	3.407	1.158	0.939	1.538	3.256	6.924	6.795	6.051
BIC	3.426	1.162	0.948	1.548	3.265	7.057	6.971	6.180
MSEv	**3.349**	1.175	0.940	**1.530**	3.178	7.194	6.887	6.152
MAEv	3.367	1.175	0.940	1.534	3.187	7.190	6.881	6.150

Table 4. Percentage of time series where the examined criteria have selected a model of low, moderate, significant, and high complexity in terms of estimated components. The last column displays the percentage of time series where the most accurate model truly falls in the respective categories. The figures are computed based on the complete data set (91,444 series).

Complexity	MSE	MAE	MSEh	MAEh	L	AIC	AICc	BIC	MSEv	MAEv	Actual
Low	1.81	8.15	7.04	8.92	1.96	23.77	27.99	41.78	12.31	12.44	12.12
Moderate	22.59	25.85	36.13	35.93	24.82	39.48	39.18	36.98	40.06	39.94	37.14
Significant	33.11	32.92	37.2	36.38	32.93	23.27	21.16	15.51	36.17	35.92	38.68
High	42.49	33.08	19.63	18.77	40.3	13.48	11.67	5.74	11.45	11.7	12.07

overall. The same stands for BIC, probably because it applies a relatively large penalty to more sophisticated models and, therefore, tends to select models of insufficient learning capacity. This is evident in the seasonal series, where both L and AIC outperform BIC. Thus, our results confirm the superiority of AICc among the information criteria examined, supporting its utilization in popular forecasting software.

Note that, as presented in Appendix A, similar conclusions can be drawn when sMAPE is used instead of MASE for measuring forecasting accuracy.

To validate our previous claims, we proceed by computing the percentage of series where the examined criteria have selected a model of low, moderate, significant, and high complexity in terms of estimated components, as defined in Table 2. The results are presented in Table 4 along with the proportion of series where the most accurate model truly falls in the respective categories. The results confirm that L and the criteria that build on in-sample accuracy measurements are indeed more likely to select more sophisticated models, especially when they are based on squared errors (MSE, MSEh, and L). Using the most recent historical data (MSEh and MAEh) or considering some penalty for complexity (AIC, AICc, and BIC) can mitigate this issue. However, it is evident that some of the former approaches are more effective in properly balancing in-sample accuracy with complexity. For example, we find that although all information criteria tend to use low complexity models more often than actually required, BIC is clearly more extreme, selecting high complexity models in just 6% of the cases and low complexity models in 42% of the series. Overall, the criteria that build on cross-validation and, to some degree AICc, are proven to be the most balanced approaches, following more precisely the actual tendencies of the data set.

Table 5 presents the precision of the examined criteria used for model selection, both per data frequency and in total. In contrast to the results of Table 3, we observe that, on average, the MSEh and MAEh approaches manage to identify more frequently the most accurate model. This finding supports our claim that precision does not always guarantee accuracy and demonstrates the value added by less precise, yet more robust

Table 5. Percentage of time series where the examined criteria used for model selection have successfully identified the most accurate alternative. The results are presented per data frequency and for the complete data set. The bold numbers highlight the most successful criterion per case.

Criterion	Yearly	Quarterly	Monthly	Total
MSE	19.95	7.62	7.77	10.40
MAE	20.70	9.26	8.94	11.61
MSEh	22.11	**11.68**	**11.76**	**14.01**
MAEh	**22.33**	11.47	11.70	13.97
L	19.86	8.17	8.09	10.69
AIC	20.46	9.25	8.46	11.30
AICc	20.44	9.28	8.36	11.25
BIC	20.39	9.02	7.86	10.91
MSEv	20.86	11.27	11.37	13.43
MAEv	20.76	11.14	11.38	13.38

model selection approaches, such as cross-validation and information criteria. Moreover, our findings are in agreement with those reported by [43], demonstrating that forecasting models should be able both to fit the historical data well and to result in "representative" forecasts in the sense that the predictions should mimic the most recent patterns of the series. Consequently, future research could focus on the further development of information criteria, expanding their formulas to account simultaneously for in-sample accuracy, forecast representativeness, and model complexity.

As a final step to our analysis, we count the percentage of series where the model being selected based on a particular criterion is the same to that being selected according to another criterion. Our results are visualized in Figure 1. We find that information criteria, and especially AIC and AICc, tend to agree significantly with each other, resulting however in very different suggestions to other criteria. This is particularly true for the criteria that build on cross-validation, displaying an agreement with the rest of the model selection approaches of less than 16%, on average. Moreover, we observe that although changing the measure used for evaluating accuracy (MAE versus MSE) can affect the model being selected, the impact of this choice is relatively smaller compared to the model selection approach itself. Thus, we conclude that the disagreement of existing forecasting model selection criteria can be substantial and that the selection of the most appropriate approach should be primarily based on the data set and application at hand, being also supported by empirical investigation.

5. Conclusions

We empirically evaluated the forecasting performance of popular model selection criteria using more than 90,000 real time series and considering fifteen models from the exponential smoothing family. We found that criteria that build on cross-validation can result in better forecasts overall but observed cases (seasonal data) where simple in-sample accuracy measurements can produce significantly more accurate results. We also noticed that information criteria offered a fair balance between the approaches that build either on in-sample or post-sample accuracy measurements, but identified notable discrepancies among their choices, driven by the different penalties they consider to avoid the use of unnecessarily sophisticated models. Moreover, we concluded that the measure used for assessing forecasting accuracy (e.g., absolute versus squared errors) has a lower impact on forecasting performance compared to the criteria used for model selection per se.

A key finding of our study is that, when it comes to model selection, robustness is probably more important than precision. In other words, in order for a selection criterion to result in accurate forecasts, it is more crucial to systematically avoid choosing the worst forecasting models than selecting more frequently the most accurate model. In this respect, it should not be surprising if two criteria of significant disagreement and different precision scores resulted in similar forecasting performance. An interesting direction to improve the

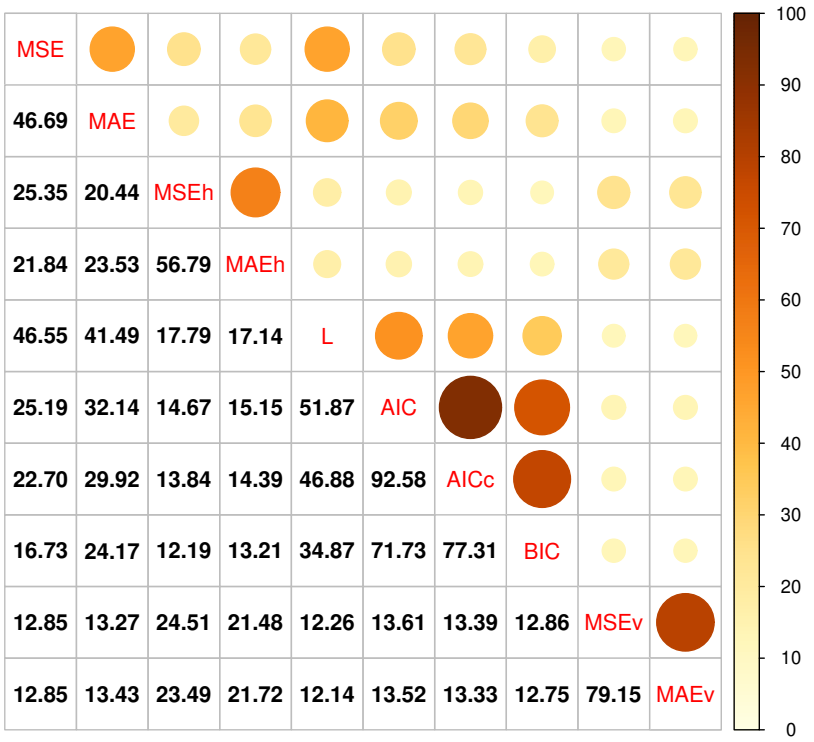

Figure 1. Percentage of time series where the model being selected based on a particular criterion is the same to that being selected according to another criterion. The percentages are computed in a pairwise fashion by considering the complete data set (91,444 series).

robustness of model selection approaches would be to introduce criteria that concurrently balance in-sample accuracy, forecast representativeness, and model complexity. According to our results, the first component allows the use of sufficiently sophisticated models, the second improves the precision of the selection process, while the third offers a "safety net" to overfitting.

Our findings are relevant to forecasting research and practice. Over the years, some model selection approaches have become that standard that forecasters often ignore the alternatives available and overlook the improvements that more appropriate criteria could offer. This is especially true in large scale forecasting applications, such as in the retail, energy, and financial industries, where the amount of series to be forecast is that great that using automated and off-the-shelf forecasting software has become a necessity. We argue that tuning the model selection options provided by such software is critical, yet practical if based on empirical assessments. Moreover, in some cases, they may even be proven more efficient computationally wise with no loss in forecasting accuracy. Therefore, future work could expand the findings of our study by examining the performance of model selection criteria in data sets that are more focused on particular forecasting applications, also extending their examination for different families of models commonly used for batch, automatic forecasting, such as ARIMA, regression trees, and neural networks.

Author Contributions: Conceptualization, E.S., F.P. and V.A.; methodology, E.S., F.P. and V.A.; software, E.S.; validation, E.S. and F.P.; formal analysis, E.S. and F.P.; investigation, E.S., F.P. and V.A.; resources, E.S. and V.A.; data curation, E.S.; writing—original draft preparation, E.S., F.P. and V.A.; writing—review and editing, E.S., F.P. and V.A.; visualization, E.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data used in this research are available at the GitHub repository of the M4 competition https://github.com/Mcompetitions/M4-methods.

Table A1. Forecasting accuracy (sMAPE) of the examined criteria used for model selection. The average rank of the selected models is also displayed. The results are presented per data frequency and for the complete data set. The bold numbers highlight the most accurate criterion per case.

Criterion	MASE				Average Rank			
	Yearly	Quarterly	Monthly	Total	Yearly	Quarterly	Monthly	Total
MSE	15.065	10.212	13.176	12.821	3.190	6.868	6.692	5.969
MAE	15.022	**10.050**	**13.013**	**12.684**	**3.165**	**6.723**	**6.553**	**5.853**
MSEh	15.183	10.306	13.084	12.823	3.230	6.964	6.671	5.992
MAEh	15.258	10.256	13.025	12.796	3.206	6.957	6.665	5.981
L	15.307	10.276	13.173	12.889	3.168	6.901	6.728	5.991
AIC	15.039	10.211	13.194	12.824	3.235	6.896	6.734	6.008
AICc	14.784	10.200	13.272	12.805	3.245	6.919	6.789	6.045
BIC	14.802	10.143	13.359	12.840	3.256	7.059	6.963	6.174
MSEv	**14.463**	10.401	13.331	12.818	3.168	7.190	6.893	6.152
MAEv	14.543	10.399	13.309	12.824	3.177	7.188	6.886	6.150

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Forecasting accuracy according to sMAPE

Table A1 summarizes the forecasting accuracy of the examined criteria in terms of sMAPE in a similar fashion to Table 3.

References

1. Wolpert, D.; Macready, W. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation* **1997**, *1*, 67–82.
2. Fildes, R. Beyond forecasting competitions. *International Journal of Forecasting* **2001**, *17*, 556–560.
3. Petropoulos, F.; Hyndman, R.J.; Bergmeir, C. Exploring the sources of uncertainty: Why does bagging for time series forecasting work? *European Journal of Operational Research* **2018**, *268*, 545–554.
4. Fildes, R.; Petropoulos, F. Simple versus complex selection rules for forecasting many time series. *Journal of Business Research* **2015**, *68*, 1692–1701.
5. Doornik, J.A.; Castle, J.L.; Hendry, D.F. Short-term forecasting of the coronavirus pandemic. *International Journal of Forecasting* **2022**, *38*, 453–466.
6. Akaike, H., Information Theory and an Extension of the Maximum Likelihood Principle. In *Selected Papers of Hirotugu Akaike*; Springer New York: New York, NY, 1998; pp. 199–213.
7. Schwarz, G. Estimating the Dimension of a Model. *The Annals of Statistics* **1978**, *6*, 461–464.
8. Collopy, F.; Armstrong, J.S. Rule-Based Forecasting: Development and Validation of an Expert Systems Approach to Combining Time Series Extrapolations. *Management Science* **1992**, *38*, 1394–1414.
9. Petropoulos, F.; Makridakis, S.; Assimakopoulos, V.; Nikolopoulos, K. 'Horses for Courses' in demand forecasting. *European Journal of Operational Research* **2014**, *237*, 152–163.
10. Montero-Manso, P.; Athanasopoulos, G.; Hyndman, R.J.; Talagala, T.S. FFORMA: Feature-based forecast model averaging. *International Journal of Forecasting* **2020**, *36*, 86–92.
11. Han, W.; Wang, X.; Petropoulos, F.; Wang, J. Brain imaging and forecasting: Insights from judgmental model selection. *Omega* **2019**, *87*, 1–9.
12. Petropoulos, F.; Kourentzes, N.; Nikolopoulos, K.; Siemsen, E. Judgmental selection of forecasting models. *Journal of Operations Management* **2018**, *60*, 34–46.
13. Kourentzes, N.; Barrow, D.; Petropoulos, F. Another look at forecast selection and combination: Evidence from forecast pooling. *International Journal of Production Economics* **2019**, *209*, 226–235.
14. Stone, M. Cross-Validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society. Series B (Methodological)* **1974**, *36*, 111–147.
15. Bergmeir, C.; Benítez, J.M. On the use of cross-validation for time series predictor evaluation. *Information Sciences* **2012**, *191*, 192–213.
16. Racine, J. Consistent cross-validatory model-selection for dependent data: hv-block cross-validation. *Journal of Econometrics* **2000**, *99*, 39–61.
17. Burman, P.; Chow, E.; Nolan, D. A Cross-Validatory Method for Dependent Data. *Biometrika* **1994**, *81*, 351–358.
18. Bergmeir, C.; Hyndman, R.J.; Koo, B. A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics & Data Analysis* **2018**, *120*, 70–83.
19. Koutsandreas, D.; Spiliotis, E.; Petropoulos, F.; Assimakopoulos, V. On the selection of forecasting accuracy measures. *Journal of the Operational Research Society* **2022**, *73*, 937–954.

20. Schwertman, N.C.; Gilks, A.J.; Cameron, J. A Simple Noncalculus Proof That the Median Minimizes the Sum of the Absolute Deviations. *The American Statistician* **1990**, *44*, 38–39.
21. Kolassa, S. Evaluating predictive count data distributions in retail sales forecasting. *International Journal of Forecasting* **2016**, *32*, 788–803.
22. Armstrong, J.S. Standards and practices for forecasting. In *Principles of Forecasting*; Springer, 2001; pp. 679–732.
23. Hyndman, R.J.; Koehler, A.B.; Snyder, R.D.; Grose, S. A state space framework for automatic forecasting using exponential smoothing methods. *International Journal of Forecasting* **2002**, *18*, 439–454.
24. Burnham, K.P.; Anderson, D.R. Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociological Methods & Research* **2004**, *33*, 261–304.
25. Billah, B.; King, M.L.; Snyder, R.D.; Koehler, A.B. Exponential smoothing model selection for forecasting. *International Journal of Forecasting* **2006**, *22*, 239–247.
26. Kolassa, S. Combining exponential smoothing forecasts using Akaike weights. *International Journal of Forecasting* **2011**, *27*, 238–251.
27. Hyndman, R.J.; Khandakar, Y. Automatic Time Series Forecasting: The forecast Package for R. *Journal of Statistical Software* **2008**, *27*, 1–22.
28. Tashman, L.J. Out-of-sample tests of forecasting accuracy: an analysis and review. *International Journal of Forecasting* **2000**, *16*, 437–450.
29. Brown, R.G. *Exponential smoothing for predicting demand*; Little: Cambridge, Massachusetts, 1956.
30. Petropoulos, F.; Apiletti, D.; Assimakopoulos, V.; Babai, M.Z.; Barrow, D.K.; Ben Taieb, S.; Bergmeir, C.; Bessa, R.J.; Bijak, J.; Boylan, J.E.; et al. Forecasting: theory and practice. *International Journal of Forecasting* **2022**, *38*, 705–871.
31. Makridakis, S.; Spiliotis, E.; Assimakopoulos, V. Statistical and Machine Learning forecasting methods: Concerns and ways forward. *PLOS ONE* **2018**, *13*, 1–26.
32. Fildes, R.; Goodwin, P.; Lawrence, M.; Nikolopoulos, K. Effective forecasting and judgmental adjustments: an empirical evaluation and strategies for improvement in supply-chain planning. *International Journal of Forecasting* **2009**, *25*, 3–23.
33. Makridakis, S.; Spiliotis, E.; Assimakopoulos, V. The M4 Competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting* **2020**, *36*, 54–74.
34. Makridakis, S.; Spiliotis, E.; Assimakopoulos, V. M5 accuracy competition: Results, findings, and conclusions. *International Journal of Forecasting* **2022**.
35. Hyndman, R.J.; Koehler, A.B.; Ord, J.K.; Snyder, R.D. *Forecasting with Exponential Smoothing: The State Space Approach*; Springer Verlag: Berlin, 2008.
36. Petropoulos, F.; Grushka-Cockayne, Y.; Siemsen, E.; Spiliotis, E. Wielding Occam's razor: Fast and frugal retail forecasting. *arXiv:2102.13209* **2022**.
37. Gardner, E.S. Exponential smoothing: The state of the art—Part II. *International Journal of Forecasting* **2006**, *22*, 637–666.
38. Winters, P.R. Forecasting sales by exponentially weighted moving averages. *Management Science* **1960**, *6*, 324–342.
39. Spiliotis, E.; Kouloumos, A.; Assimakopoulos, V.; Makridakis, S. Are forecasting competitions data representative of the reality? *International Journal of Forecasting* **2020**, *36*, 37–53.
40. Makridakis, S.; Assimakopoulos, V.; Spiliotis, E. Objectivity, reproducibility and replicability in forecasting research. *International Journal of Forecasting* **2018**, *34*, 835–838.
41. Hyndman, R.J.; Koehler, A.B. Another look at measures of forecast accuracy. *International Journal of Forecasting* **2006**, *22*, 679–688.
42. Petropoulos, F.; Spiliotis, E.; Panagiotelis, A. Model combinations through revised base rates. *International Journal of Forecasting* **2022**.
43. Petropoulos, F.; Siemsen, E. Forecast Selection and Representativeness. *Management Science* **2022**.