Representativeness: A New Criterion for Selecting Forecasts

Fotios Petropoulos and Enno Siemsen

PREVIEW Statistical criteria for selecting a best forecasting method from a group of candidates have been proposed, studied, and implemented widely in forecasting software. Very well known are information criteria, such as the AIC, which balance performance and complexity, and validation techniques, which examine forecasting performance in a holdout sample. So it's a breath of fresh air to have a distinctly new take on method selection, which is what Fotios and Enno are presenting here. They offer strong evidence that method selection can be improved by accounting for the *representativeness* of the forecasts.

KEY POINTS

- The forecasting community has not fully coalesced around the proper ways to select a forecasting method, and several approaches have been developed. None of the established model-selection rules make use of the forecasts that will eventually be used for decision making. Information criteria and validation/cross-validation approaches explicitly assume that whichever method produced the best forecasts previously will be the best model forward.
- Our new approach for selecting among forecast models is based on examination of the forecasts made for real-time future periods, the actuals for which are not yet knowable. It is based on *representativeness*, the degree to which these forecasts are a natural continuation of the observed data.
- We describe and illustrate a new criterion for method selection that considers the representative of the forecasts as well as the accuracy with which the methods fit the observed data. We call this the REP.
- Finally, we compare REP with two main existing criteria for method selection, the AIC and cross-validation (CV) using a large number and wide variety of time series from previous M competitions. We believe the results are highly promising and point to deeper exploration into the psychology of human input into forecast-method selection.

INTRODUCTION: THE CHALLENGE OF SELECTING A MODEL

It's the age-old dilemma: What is the best forecasting model for you? This has always been a challenging question for researchers and practitioners alike. Being able to identify the best model could lead to substantial performance benefits. Further, performance-ranking different forecasting models supports setting better weights for forecast combinations.

A fundamental problem, however, is that the model that has performed comparatively well given the observed (historical) data will not necessarily be the model that performs best in the future. The underlying patterns in observed data are always subject to change.

While forecast model selection is important, the forecasting community has not fully coalesced around the proper ways to accomplish this, and several approaches have been developed. One aspect these approaches have in common is that they focus wholly on the observed data – not considering their forecasts of the future in the selection decision.

We believe this is a significant oversight. In this paper, after discussion of existing approaches to model selection, we propose an enhancement that incorporates the *representativeness* of a model's forecasts as a new component in the model-selection decision. It does so by providing a reality check on the reasonableness of the generated forecasts. Testing on numerous actual time series shows this enhancement to improve forecast performance.

JUDGMENT IN MODEL SELECTION

Recent studies have shown that individuals are able to select models such that the average forecasting performance is as good as selection based on some statistical criteria. An examination of judgment in model selection was featured in the Summer 2019 issue of *Foresight* (Issue 54), beginning with Fotios's article on the application of judgement for model selection (Petropoulos, 2019). Several Commentaries pointed out that the commendable performance of judgmental model selection offers an attractive alternative to the reliance on automatic selection in forecasting software as well as a way forward for those with algorithmic aversion.

In addition, Fotios reported that combining forecasts, either by aggregating the forecasts of small groups (judgmental aggregation) or averaging the forecasts of models selected statistically with those selected judgmentally, could significantly outperform straight statistical approaches. Moreover, while he found that statistical approaches will select the best models more often (compared to humans), they will also select the worst model more frequently. In other words, humans are better able to avoid the worst outcomes.

The way we apply judgment to the task of model selection is fundamentally different from the way statistical approaches such as information criteria work. The key difference seems to lie in what information is being used. Statistical approaches work "backwards" in the sense of examining observed data values and measuring forecasting performance over past periods (both in terms of performance in tracking the in-sample fit or the out-of-sample performance on data held out from model fit).

We humans, in contrast, look "forward" and compare the pattern of forecasts produced by each forecasting model with our mental extrapolations of the data pattern. Brain imaging experiments by Weiwei Han and colleagues (2019) revealed that humans reject models for which the forecasts *look* unreasonable. In other words, we use a visualization of forecasts for which the actual data are not yet available to select models that produce a pattern that best matches past data: the forecasts that are *representative*.

Motivated by this realization, we devised an algorithmic approach to forecast selection based on the concept of *representativeness*. A full description and analysis of the algorithm is in our new article in the journal *Management Science* (Petropoulos and Siemsen, 2022). Here we will describe and illustrate how model selection works based on this new criterion of representativeness. We begin with an overview of model-selection criteria.

MODEL-SELECTION CRITERIA

Most statistical criteria for model selection are based, directly or indirectly, on forecast-error metrics, with smaller errors over a designated period being preferable.

Information Criteria

Information criteria are based on how well the model fits the in-sample data but contains a penalty for the size/complexity of the model. The penalty is designed to avoid overfitting – introduction of additional model complexity that adds little new value.

Information criteria, such as the Akaike's Information Criterion (AIC) and Bayesian Information Criterion (BIC), are in wide use. Implementations for them are readily provided within open-source forecasting packages, such as the *forecast* and *smooth* packages for R statistical software. They are also reported in many commercial packages and used in some of these for model selection. Some versions build in corrections for small sample size. For instance, the functions ets() and auto.arima() of R's *forecast* package rely, by default, on the AIC values corrected for small sample sizes (AICc).

The in-sample fit of each model is measured either by the mean squared error (MSE) or the likelihood function (which measures how likely we would see the observed data, given the model). The penalty is based on the number of model parameters to be estimated. The premise of an information criterion follows Occam's Razor in that, among all solutions with the same performance, the simplest one should be selected. In comparing the values of an information criterion across models, lower values indicate superior performance.

One caveat is that information criteria values are not always directly comparable across different models, especially if we are comparing models from different families, such as an exponential smoothing vs. an ARIMA model. The difficulty arises from how the likelihood function is computed and how initial values have been specified; for instance, differencing is often required for ARIMA models but not exponential smoothing. In general, then, information criteria should not be used when the data are transformed in different ways, or when different orders of integration have been applied.

In addition, information criteria may erroneously select a simpler model when the sample size is small and the sample data are highly variable. That said, the selection of a misspecified model –

either one that is simpler or one that is more complex – will not necessarily degrade forecast performance, a point we return to later in this article.

Baki Billah and colleagues (2006) compared different statistical criteria for choosing among exponential-smoothing models and reported that information criteria, particularly the AIC, performed best in their simulations. Stephan Kolassa (2011) showed how the values of information criteria may be used to calculate weights for combining the forecasts of different models.

Validation and Cross-Validation

Validation and *cross-validation* approaches split the available data into training and validation sets, using the former to fit a model and the latter to measure the out-of-sample performance of the forecasts. The model with the best validation performance is then taken forward to produce forecasts for future periods.

The principle of out-of-sample evaluation is that forecasting performance should be measured on data that have not been used in the training of the models.

In the validation approach, only one set of forecasts is produced for a time series, and these are tested against the data in the validation set of that series. This approach is simple and relatively fast; however, it is still slower than selection based on information criteria, as the selected model needs to be fitted twice: once for the validation step and once to produce forecasts for future periods. In addition, the forecast-error metrics that are calculated blend multiple forecast horizons from a fixed origin, so no insights emerge as to how forecast accuracy changes as the horizon of the forecast lengthens.

A better alternative to validation is cross-validation, in which model fitting and evaluation is repeated from multiple forecast origins. While there are several forms in use, for time-series data the rolling origin evaluation is most appropriate. Here, a model is fit with the training data and forecasts made for multiple horizons in the validation set. The window of the training data is expanded to include the first time period in the validation set. The model is reestimated and forecasts made from the new origin. The process continues until all data in the validation set have been incorporated into the training set. The procedure can be time-consuming and requires the availability of long time series to support evaluations over multiple origins.

Additionally, for cross-validation in general, one needs to decide between expanding or rolling training windows, overlapping or nonoverlapping validation sets, and how often validation forecasts should be produced. For additional reading on time-series validation and cross-validation, the reader is referred to Tashman (2000) and Bergmeir and Benítez (2012).

Times-Series Features

Apart from information criteria and validation approaches, model selection has been based on *time-series features*, such as trend, seasonality, autocorrelation, cycle, randomness, series length and variability, and interdemand interval for intermittent data (Petropoulos and colleagues, 2014). Since there is no universal method that is best for every time series, the goal is to identify the best forecasting method for the particular features of "my data."

Given a pool of models and a set of reference series, the features are used to train *meta-learning algorithms.* Here each series in the reference set is split into a training and a test set (Talagala and colleagues, 2021). The training set of each series is used to calculate values for the time-series features as well as to produce forecasts for the test set, and this is done for each of the models in the pool.

A meta-learning model is trained to select between models by comparing how performance of the forecasts of the various models in the pool is related to the values of the features. The reference series can be publicly available data sets, such as the M4 competition data, or even synthetically generated series that possess the desired features (Kang and colleagues, 2020). The R packages *tsfeatures* and *gratis* can be used to calculate the values of time-series features and generate synthetic series based on them.

A variation of this approach is the rule-based forecasting RBF system from Armstrong, Collopy, and Adya. These are sets of "if/then" rules distilled from experience of forecasting experts to select and combine among a set of simple time-series models (Adya and colleagues, 2001).

SELECTION BY REPRESENTATIVENESS

These various selection criteria all consider past data as well as the forecasts that correspond to observations up until the latest available time period. Information criteria use in-sample comparisons between actual and predicted – also called *fitted* – values. Validation approaches use out-of-sample forecasts; however, these forecasts still occur during past time periods (the validation sets correspond to appropriate holdouts).

None of these approaches, however, make use of the forecasts for *future* periods--that is, forecasts for which the actual observations remain unknown. More explicitly, none of the established model-selection rules make use of the forecasts that will eventually be used for decision making. Information criteria and validation/cross-validation approaches explicitly assume that whichever method produced the best forecasts previously will be the best model forward.

This principle strikes us as, at the very least, naive. Our new approach for selecting among forecast models is based on examination of the forecasts made for real-time future periods, the actuals for which are not yet knowable. The evaluation of such forecasts, given that the actual data are not yet available, is based on the degree to which these forecasts are a natural continuation of the past observed data. We call this *representativeness.* For example, a flat/level

forecast would not be representative of data with a strong trend, nor would a trend-only model be representative of data that exhibit strong seasonal behavior.

REP Defined

We define the *representativeness gap* as the lack of representativeness of a forecast compared to past actuals, and use this concept in conjunction with the performance of the in-sample forecasts (fitted values). This approach effectively replaces the complexity penalty applied by information criteria by the representativeness gap. In essence, we propose that selecting between forecasting models should be a balance between in-sample fit and representativeness. Conceptually, this new criterion, REP, can be expressed as:

REP = performance gap + representativeness gap

Similar to information criteria, model forecasts with lower REP values should be preferred to those with higher. If two sets of forecasts offer the same in-sample performance, the one with better representativeness – a lower representativeness gap – should be selected. This is not the same as selecting the least complex of two equally performing models.

Representativeness (and its gap) can be measured in different ways, all of which are based on comparisons of the forecasts for the future with the available data up to the present. Such comparisons are asynchronous, in the sense that the forecasts and the data to be compared refer to different time periods. As such, some scaling and transformations may be required to place the “present” and the “future” on the same level.

For series with multiplicative (trend or seasonal) patterns, logarithmic transformations can be needed. If the data are periodic – with patterns that repeat at regular intervals such as monthly seasonality – the comparisons must be aligned, so that the respective periods (e.g., the same month) of the past data and the forecasts are compared.

Measurement of representativeness can not only be made for point forecasts but prediction intervals as well. Other choices include the error metrics used, the length of the forecast horizon (i.e., length of the data window for the comparisons), the use of a single window or multiple windows of past data, and the use of equal or unequal weights to average representativeness across multiple windows of data.

Calculating the REP for Point Forecasts

The calculation of REP follows a number of steps, each of which can be adapted to the specifics of the data at hand. **Figure 1** provides an illustrative example. The historical data are plotted in black in the top panel. These represent monthly sales of a toy product over a period of six years. An upward trend and a seasonal pattern are evident.

We produce forecasts with two methods: Holt's linear trend exponential smoothing (that ignored the seasonal pattern) and the Holt-Winters exponential smoothing with multiplicative seasonality. The forecasts for the next year (next 12 months) are depicted by the red and blue lines in the first panel.

1. The first step is to produce forecasts using all available historical data, without splitting a series into training and validation sets.

2. Specify the window of forecasts over which representativeness is to be measured. It simplifies things to match the forecast horizon with the periodicity of the data, so for monthly data – a periodicity of 12 – we set the forecast horizon to be 12 months ahead.

3. Split the sample data into buckets such that (i) each bucket is at least as long as the forecasts' window, and (ii) the first period of each bucket corresponds to the respective first period of the forecasts. In the toy example, we split the in-sample data into five windows/buckets, each a length of 12 months to match the forecasting horizon and each bucket beginning with the same month of the year. This ensures that seasonal patterns will be aligned when comparisons are performed.

4. The fourth step is to perform any needed scaling and transformations. For each bucket in our illustrative data, we aligned the scales and did a Box-Cox transformation to stabilize the variance and transform multiplicative patterns into additive ones. This is to be done for the forecasts as well.

5. For each individual bucket, the representativeness gap is measured in terms of the closeness of the past data points to the point forecasts. Most simply, we sum the absolute distances between a (scaled and transformed) bucket of past data and the forecasts. For our toy example, the result of this step is five sums of absolute differences.

6. The final step is to take an average of the resulting bucket sums from the prior step. This could be an unweighted average, but we recommend a weighted mean to give more emphasis to the more recent years. In particular, we recommend that the weight for each bucket is one half of its more recent adjacent bucket; that is, the weights are decreasing by 50% as we move back from the most recent bucket.

In Figure 1, the middle frame highlights the representativeness gap for the trended-only forecast – the red line in the top frame. The large representativeness gap reflects mainly the seasonal departure of the data from the trend line. The bottom frame shows the forecasts from a trended/seasonal (Holt-Winters) model – the blue forecasts. It is clear that the representativeness gap is much lower now that we've also accounted for seasonality.

Once the representativeness gap has been calculated, then the value of the REP criterion is calculated as the sum of the representativeness gap and the in-sample performance gap. For

the toy data, the in-sample performance gap is the mean of absolute differences between the actual and fitted values over the five years.

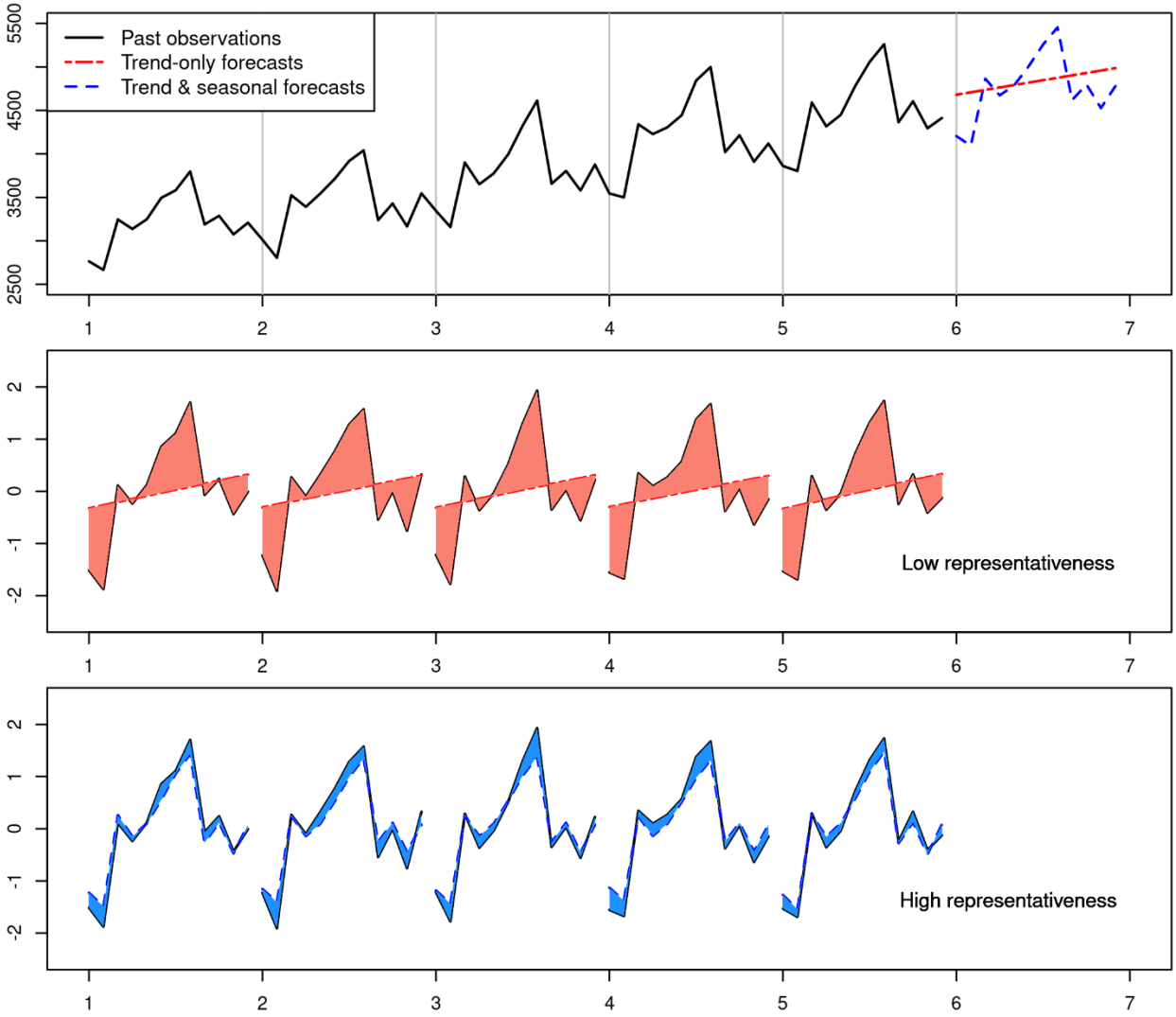

Figure 1. A toy example on measuring representativeness (and its gap); adapted from Petropoulos and Siemsen (2022).

PERFORMANCE COMPARISON OF SELECTION APPROACHES

To evaluate how our REP criterion fares against information criteria and cross-validation, we compared the three criteria on a large collection of real data from the M, M3, and M4 forecasting competitions. We used these three criteria to select and to combine models from the exponential-smoothing family. For the latter, we calculated weights based on how the individual models performed on the particular selection criterion.

Table 1 offers a selective summary of the empirical results. (These were more fully reported in Petropoulos and Siemsen, 2022). For Table 1, we present forecast accuracy results for the yearly and monthly data frequencies. The metric reported is the mean absolute scaled error (MASE), originally proposed by Rob Hyndman (see Hyndman, 2006) as a scale-free metric suitable for measuring accuracy across multiple time series (including intermittent series).

Table 1. Summary accuracy results; adapted from Petropoulos and Siemsen (2022).

		Yearly data	Monthly data
Selection	AICc	3.405	0.941
	CV	3.307	0.921
	REP	3.125	0.918
Combination	AICc	3.351	0.933
	CV	3.300	0.916
	REP	3.101	0.906

For point forecasts, the REP selection rule outperforms both the information criterion and cross-validation, and does so for both the yearly and monthly data. For the yearly frequency, REP was more accurate than AICc (Akaike's Information Criterion corrected for small sample sizes) and cross-validation (CV) by 8.2% and 5.5% respectively. Differences in performance were statistically significant in most of the cases, especially for yearly, quarterly, and monthly data.

In addition, we found that REP selects the best (among a class of exponential smoothing) models more often than information criteria, while more frequently avoiding the worst of these models. The good performance of REP was evident not only when it was used to select/combine models within the exponential-smoothing family, but also within the ARIMA family of models, or even between models of different families.

We note too that combinations of models based on REP outperformed AICc and CV in terms of estimating uncertainty as well as accuracy in the point forecasts, using the mean scaled interval score and a 95% confidence level.

We also performed sensitivity analyses to examine the performance of REP under different conditions. Two principal findings:

- First, we considered the case of using only the representativeness gap in measuring REP, leaving out the in-sample performance. We observed that even excluding the performance gap leaves the REP approach superior to selection based on the AICc . This is very important for cases in which the available forecasts are not accompanied by in-sample fits, as it is usual for purely judgmental forecasts.
- Second, we analyzed the performance of REP for different forecast horizons and found that REP performs strongly across all horizons (short, medium, and long) – and in fact, its advantage relative to the information and cross-validation criteria grows at longer horizons.

CONCLUSION

As a means of selecting between forecast models, *representativeness* – with the strong empirical performance of REP – has earned its way into the criteria for model selection. Similar to information criteria, REP consists of two parts: how well the model fits the historical data and a penalty. Contrary to information criteria, the penalty is not based on model complexity but rather on the representativeness of the forecasts, the degree to which the resulting forecasts are perceived as a natural extension of the historical data.

Our study offers more evidence that the infusion of human judgment into algorithms, such as the manner in which we use visualizations of the forecasts, can improve the performance of both existing algorithms and judgment alone. That said, there is still much to distill from how humans approach forecasting and how we can translate such insights to further improve our forecasting algorithms.

References

Adya, M., Collopy, F., Armstrong, J.S. & Kennedy, M. (2001/4). Automatic Identification of Time Series Features for Rule-Based Forecasting, *International Journal of Forecasting*, *17*(2), 143–157.

Bergmeir, C. & Benítez, J.M. (2012). On the Use of Cross-Validation for Time Series Predictor Evaluation, *Information Sciences*, *191*, 192–213.

Billah, B., King, M.L., Snyder, R. & Koehler, A.B. (2006/4). Exponential Smoothing Model Selection for Forecasting, *International Journal of Forecasting*, *22*(2), 239–247.

Han, W., Wang, X., Petropoulos, F., & Wang, J. (2019). Brain Imaging and Forecasting: Insights from Judgmental Model Selection. *Omega*, *87*, 1–9.

Hyndman, R.J. (2006). Another Look at Measures of Forecast Accuracy for Intermittent Demand, *Foresight*, *4*, 43-46.

Kang, Y., Hyndman, R.J., Li, F. GRATIS: (2020). Generating Time Series with Diverse and Controllable Characteristics, *Statistical Analysis and Data Mining*, *13*, 354– 376.

Kolassa, S. (2011). Combining Exponential Smoothing Forecasts Using Akaike Weights, *International Journal of Forecasting*, *27*(2), 238–251.

Petropoulos, F. (2019). Judgmental Model Selection, *Foresight*, *54*, 4–10.

Petropoulos, F., Makridakis, S., Assimakopoulos, V. & Nikolopoulos, K. (2014). “Horses for Courses” in Demand Forecasting, *European Journal of Operational Research*, *237*, 152–163.

Petropoulos, F. & Siemsen, E. (2022). Forecast Selection and Representativeness, *Management Science*, forthcoming.

Talagala, T.S., Li, F. & Kang, Y. (2021). FFORMPP: Feature-Based Forecast Model Performance Prediction, *International Journal of Forecasting*.

Tashman, L.J. (2000). Out-of-Sample Tests of Forecasting Accuracy: An Analysis and Review.

International Journal of Forecasting, *16*(4), 437–450.

Bios

Fotios Petropoulos is Professor of Management Science, University of Bath, and Foresight Editor for Forecasting Support Systems.
f.petropoulos@bath.ac.uk

Enno Siemsen is Professor of Operations and Information Management and Associate Dean at the University of Wisconsin School of Business.
esiemsen@wisc.edu