To combine or not to combine: selecting among forecasts and their combinations

Michèle Hibon*, Theodoros Evgeniou

INSEAD, Technology Management, Boulevard de Constance, 77305 Fontainebleau, France

Abstract

Much research shows that combining forecasts improves accuracy relative to individual forecasts. In this paper we present experiments, using the 3003 series of the M3-competition, that challenge this belief: on average across the series, the best individual forecasts, based on post-sample performance, perform as well as the best combinations. However, this finding lacks practical value since it requires that we identify the best individual forecast or combination using post sample data. So we propose a simple model-selection criterion to select among forecasts, and we show that, using this criterion, the accuracy of the selected combinations is significantly better and less variable than that of the selected individual forecasts. These results indicate that the advantage of combining forecasts is not that the best possible combinations perform better than the best possible individual forecasts, but that it is less risky in practice to combine forecasts than to select an individual forecasting method.

Keywords: Combined forecasts; M3-competition; Forecasting accuracy; Model selection

1. Introduction

Combining forecasts introduced by Bates and Granger (1969) is often considered as a successful alternative to using just an individual forecasting method. For example, in the machine learning and statistics communities, bootstrapping, bagging, stacking, and boosting are some well-documented approaches based on the idea of combining methods that lead to better predictive performance (Breiman, 1996, 1998; LeBlanc & Tibshirani, 1996; Shapire, Freund, Bartlett & Lee, 1998). Previous work on time series forecasting also argues that predictive performance increases through combining forecasts (Armstrong, 1989, 2001; Clemen, 1989; Makridakis & Winkler, 1983; Makridakis et al., 1982). Terui and Van Dijk (2002) have investigated combinations of forecasts for non-linear methods, with comparable results. However, there has also been work that questions whether one should always combine forecasts. For example, recently, Larrick and Soll (2003) show that under some conditions it is better not to combine forecasts. In this paper we investigate empirically the following two questions:

a) Do the best possible combinations have better predictive performance than the best possible individual forecasts?
b) When we do not know which is the best individual forecast or combination, is it riskier to select among individual forecasts than to select among their combinations?

Our empirical results indicate that managers and researchers should treat statements of the type “combining forecasts is better” with some caution: the best combinations are, on average across series, no better than the best individual forecasts. On the other hand when we do not know which individual forecasting method is the best, selecting among combinations in practice leads to a choice that has, on average, significantly better performance than that of a selected individual method. Therefore the advantage of combinations is not that the best are better than the best individual forecasts, on average, but that selecting among combinations is less risky than selecting among individual forecasts.

Our experiments use 14 methods and all 3003 time series from the M3-competition (Makridakis & Hibon, 2000), a large-scale experimental study along the lines of the work of Makridakis and Winkler (1983) that was done for the original M-competition. Throughout the paper, we use the terms “forecasts” and “forecasting methods” interchangeably, the latter being the methods used to generate the forecasts.

2. Setup of the study

2.1. Hypotheses tested

Although not needed for our final findings, we first replicate, for comparison, the experiments of Makridakis and Winkler (1983) for the M3-competition. In particular we first test the hypothesis, suggested by them:

**H1.** The best among all combinations of methods is no more accurate than the best individual method on average across series, when the same combination (individual method) is used for all series.

Moreover, we examine whether the relative advantage, in terms of best achieved accuracy, of combining methods decreases as the number of methods combined increases.

In Makridakis and Winkler (1983) the combinations (individual methods) are fixed over all series (the same one is used for all series). Given that, as we comment below, some of the methods we used (i.e. ForecastPro) are derived methods—e.g. methods derived from other methods using a selection rule—our test of H1 is not technically correct, although the effects of using derived methods were limited since, for example, the best individual method for all series turned out not to be a derived one.

In practice one uses a different combination (individual method) for each series. This leads to the more realistic version of hypothesis 1, namely:

**H2.** The best among all combinations of methods is no more accurate than the best individual method on average across series, when a different combination (individual method) is used for each series.

Moreover, as before, we examine whether the relative advantage, in terms of best achieved accuracy, of combining methods decreases as the number of methods combined increases.

Hypotheses H1 and H2 consider the best possible combination or individual forecast chosen using the post-sample data—“best” in terms of post-sample accuracy. However in practice we cannot select a method with hindsight. A central issue in forecasting is the problem of model selection: how to select among a number of possible models without using post-sample data (Akaike, 1977; Vapnik, 1998). Model selection generally has the risk that it can lead to a choice that is much worse than the best possible choice in terms of post-sample accuracy—for example due to over-fitting. Therefore even though the best possible (with hindsight) individual method may be as good as the best possible combination, the selected individual method in practice (with no hindsight) may not be as good as the selected combination. We therefore test the hypothesis:

**H3.** The predictive performance of the selected combination is no better than that of the selected individual method.
We also investigate whether selecting among individual methods is more risky than selecting among combinations. We test the hypothesis:

**H4.** The risk, measured as the difference in post-sample performance between the selected and the best possible, for selecting among individual methods is no higher than the risk for selecting among combinations.

Hypotheses H3 and H4 are motivated by two types of previous theoretical results:

a) a key effect of combining methods is a decrease in the variance of the performance across the combinations relative to the variance across the individual methods, for various measures of variance (Breiman, 1998; Evgeniou, Pontil & Elisseeff, 2004);

b) the variance across a number of methods is related to the risk (as defined in hypothesis H4) of selecting among these methods (Bousquet & Elisseeff, 2002; Vapnik, 1998).

Clearly, testing H3 and H4 requires a model selection criterion. Here we use a simple criterion (defined in Section 4) based on the sliding simulation idea of Makridakis (1990). Therefore our results are specific to this model-selection criterion: more general conclusions would require consideration of other model selection criteria, or a general theoretical analysis that considers any model selection criterion with certain characteristics. The latter is an open theoretical problem.

2.2. Data and performance measures

Our database consists of the 3003 time series of the M3-competition (Makridakis & Hibon, 2000), which has 645 yearly series, 756 quarterly series, 1428 monthly series and 174 other series. The number of forecasting (post-sample) periods was six for yearly data, eight for quarterly, 18 for monthly, and eight for other.

Our experiments use all four types of data: yearly, quarterly, monthly, and other; the conclusions we report hold for all four types, but to avoid clutter, we present only the quarterly results.

We have selected the following 14 forecasting methods\(^1\) from the M3-competition to be used in the combinations:

Single, Holt, Dampen, Autobox2, Robust-Trend, Auto-regressive autoregressive moving average (ARARMA), Automatic neural network modeling (AutomatANN), Flores-Pearce 1, PP-Autocast, ForecastPro, SmartForecasts’ Automatic Forecasting System (SmartFcs), Theta, Rule-Based Forecasting (RBF), ForecastX engine (ForcX).

From the original 24 methods of the M3-competition we dropped the following methods: Naive2 and Winter (the worst methods), Comb S-H-D (it is already a combination of the individual methods Single, Holt and Dampen), Automatic ARIMA modeling with and without intervention (AAM1 and AAM2-no forecasts available for yearly and other series). We retained only one version of Autobox, Flores-Pearce, and Theta.

In this study, a combination of forecasts always consists of a simple (unweighted) average:

\[
\hat{X}_t = \frac{X^{(1)}_t + X^{(2)}_t + X^{(3)}_t + \ldots + X^{(n)}_t}{n}
\]

where \(t\) is the forecast lead time and \(n\) is the number of methods being combined. We consider all possible combinations of the 14 methods. For a given number \(n\) of methods combined we get \(\binom{14}{n}\) combinations, which is 91 for \(n=2\), 364 for \(n=3\), 1001 for \(n=4\), 2002 for \(n=5\), 3003 for \(n=6\), and 3432 for \(n=7\) (it is the same for \(n\) and for \(14-n\)).

We evaluate forecast accuracy using the symmetric MAPE (sMAPE) on the post-sample data, defined as:

\[
\sum \frac{|X - F|}{X + F} \times 100
\]

where \(X\) is the actual value and \(F\) is the forecast. Strengths and weaknesses of this accuracy measure

---

\(^1\) Some of these methods (for example ForecastPro and SmartFcs) are not “individual methods” but “derived methods” using other methods and a selection rule. This affects only hypothesis H1—which assumes a fixed method used for all series—as mentioned in Section 2.1.
have been reported and criticized before, for example by Goodwin and Lawton (1999) and Koehler (2001). Despite the criticisms of the sMAPE, we have used it in order to be consistent with the M3-Competition study (Makridakis & Hibon, 2000).

Fig. 1. “Over series”: For each possible combination we average the sMAPE over all series, and then we identify which combination gives the best (lowest average) sMAPE on the post-sample data among the available choices.

For the comparisons we consider two post-sample data performance measures:

a) LowMAPE: the best (lowest average) sMAPE on the post-sample data among the available choices.

Fig. 2. “Per series”: For each series, we first calculate the sMAPE of each possible combination and find the lowest and highest values. We then average these values (each potentially coming from a different combination) over all series.
b) HighMAPE: the worst (highest average) sMAPE on the post-sample data among the available choices.

There are two distinct ways of computing each of these measures. The first corresponds to hypothesis H1, which we test, as mentioned, for comparison with the study done by Makridakis and Winkler (1983), and the second to hypothesis H2:

H1: “Over series” case: For each method (or combination) we find the average sMAPE over all series (and all forecast horizons)2 and then we identify the method (or combination) that gives the lowest and highest average sMAPE. We refer to this as the best (worst) “over series” method (or combination) (Fig. 1).

H2: “Per series” case: For each series we first calculate the sMAPE of each possible method (or combination), and find the lowest and highest sMAPEs. We then average each of these two sets of values over all series. We emphasize that there can be a different best or worst method (or combination) for each series. We refer to this as the best (worst) “per series” method (or combination) (Fig. 2).

3. Combinations versus individual methods

3.1. The “over series” case

In Makridakis and Winkler (1983) only the “over series” case was considered. We replicate that work for the case of the M3-competition and show the results for the quarterly data in Table 1 and Fig. 1. As in Makridakis and Winkler (1983), the two curves represent, for each number n of methods combined, the LowMAPE and HighMAPE in the “over series” case. In Fig. 1 the LowMAPE and HighMAPE for the best individual method are those for n = 1. Clearly the best possible among all combinations is the one corresponding to the minimum of the LowMAPE curve, which is a combination of size n = 7 in this case of quarterly data (the same is true for the monthly and yearly series).

As expected, the curves approach each other and meet at n = 14 where there is only one combination. Consistent with the results by Makridakis and Winkler (1983), the HighMAPE decreases rapidly as we combine more methods. The marginal impact of including an additional method decreases as the number of methods increases.

We are mainly interested in the best possible performance attained, so the LowMAPE curve is the one that is relevant. This curve declines when n increases from 1 to 7 and then eventually slightly increases (shown also in Table 1), as also found in Makridakis and Winkler (1983) for the M-competition data.

However, the LowMAPE curve in Fig. 1 is misleading. We performed significance tests to check if the differences between the sMAPE of the best combination (LowMAPE) for each n and the sMAPE of the best individual method (LowMAPE) are significant:

H_0: Differences between the two means = 0;  
Ha: Differences between the two means ≠ 0 (combinations are different from individual methods).

For all experiments reported below, we performed a t-test to check if the difference between two population means (i.e. the sMAPE of the best combination and the sMAPE of the best individual method) is significant, as it was done by Flores and Pearce (2000) for the M3-competition. Mariano and Diebold (1995) provide a general framework for such tests.

Table 1 shows the different values of the sMAPE and the corresponding P-value of the tests for n = 5, 7 (the best possible combination), and 14. We conclude that there are no significant differences

<table>
<thead>
<tr>
<th></th>
<th>sMAPE</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Best among all individual methods</td>
<td>8.96</td>
<td>–</td>
</tr>
<tr>
<td>Best among all combinations of 5</td>
<td>8.73</td>
<td>0.68</td>
</tr>
<tr>
<td>Best among all combinations of 7</td>
<td>8.72</td>
<td>0.67</td>
</tr>
<tr>
<td>Combination of all 14 methods</td>
<td>8.81</td>
<td>0.80</td>
</tr>
</tbody>
</table>

P-values of comparison with best individual method—first row. There is no significance difference between the four cases.

---

2 The calculations of the sMAPEs can be done either for each post-sample horizon or on average across the horizons. We have calculated them both ways. Since the results were similar, as also observed in Makridakis and Winkler (1983), we report only the case where the sMAPE is calculated across the post-sample horizons.
between the sMAPE values, therefore we cannot reject hypothesis H1:

a) the best possible combination “over series” is similar to the best possible individual method, on average, across the series;
b) the size of the combinations does not make a significant difference in terms of the best attainable performance.

3.2. The “per series” case

The results for the “per series” case are given in Table 2 and Fig. 2. The LowMAPE and HighMAPE for the best individual methods are again those for $n=1$. As before, the best possible combination is similar to the best possible individual method, as the $P$-value of the second row in Table 2 shows. Moreover, we note that when we select for each series among both the individual methods and all their combinations (third row in Table 2) performance increases slightly, but not significantly.

Out of the 756 quarterly series, an individual method is the best for 324 (43%) of the series, a combination of size 2 for 233, of size 3 for 144, and there are only 21 series for which a combination of size larger than 7 is the best. This shows also that the best combinations use only a few (<7) methods most of the time. This is in agreement with Armstrong (2001), who suggests a combination of at least five methods, but notes that adding more methods leads to diminishing rates of improvement.

Fig. 2 also shows that the best attainable performance decreases (i.e. the LowMAPE increases) as we combine more and more methods. The $P$-values for the comparison between the best individual methods and the best combinations of size $n=5$, 7, 14 are given in Table 2. There is no significant difference between the sMAPE of the best individual method and the sMAPE of the best combination of five methods, but the difference is significantly worse when we combine more than seven methods. It may be that as we combine more and more methods, we introduce poorer methods, which worsen performance—in agreement with Armstrong (2001) and the observation above that the best combinations use only a few methods.

In conclusion, we cannot reject hypothesis H2. When we use the best for each series:

a) The best possible combination is on average similar to the best possible individual method;
b) The performance decreases as we combine more and more forecasts.

Fig. 2 also shows that there is a significantly higher spread (the HighMAPE curve minus the LowMAPE curve) across the performance of individual methods (this is the spread for $n=1$) than across the performance of any combinations, which is an expected effect of averaging. That the spread (similarly for the variance) decreases as the number of methods combined increases, was also shown in Makridakis and Winkler (1983). This result suggests that, when we do not know which method or combination is best, it may be riskier to choose an individual method than it is to choose a combination, with the risk decreasing with $n$. We define risk as: $\text{Risk} = \text{difference of the post-sample (periods 4–8 in the next section) accuracy of the chosen method from the best possible post-sample (periods 4–8 again) accuracy}$. We note this as “chosen-best”.

Note that we could define the risk using the variance of the performances, but in practice we care about how much worse the chosen method (or combination) is from the best possible one, hence we consider the risk as defined above.

Table 2
Average sMAPE of the best individual method (combination) for periods 1–8 when this is not fixed for all the series—best “per series”

<table>
<thead>
<tr>
<th></th>
<th>sMAPE</th>
<th>$P$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Best among all</td>
<td>5.55</td>
<td>–</td>
</tr>
<tr>
<td>individual methods</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Best among all</td>
<td>5.48</td>
<td>0.85</td>
</tr>
<tr>
<td>combinations</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Best among all</td>
<td>5.22</td>
<td>0.38</td>
</tr>
<tr>
<td>individual methods</td>
<td></td>
<td></td>
</tr>
<tr>
<td>and all combinations</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Best among all</td>
<td>6.24</td>
<td>0.10</td>
</tr>
<tr>
<td>combinations of 5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Best among all</td>
<td>6.75</td>
<td>0</td>
</tr>
<tr>
<td>combinations of 7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Combination of all 14</td>
<td>8.81</td>
<td>0</td>
</tr>
<tr>
<td>methods</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$P$-values of comparison with best individual method—first row. We also report in bold the “best” or “not significantly different from the best” performances. For the third row, out of the 756 quarterly series an individual method was the best for 324 (43%) of the series ($P$-values $= 0$ corresponds to $P$-value $< 0.01$).
4. Risk of model selection: hypotheses H3 and H4

Until now we have used post-sample data to find the LowMAPE combination (or individual method) for each series. The real issue is whether it is riskier to choose a combination or an individual method when the forecaster must select a method (or combination) based on a fit-set (validation set) of data rather than the test-set. For that we need to specify a model-selection rule.

We use a simple model selection criterion based on the sliding simulation idea of Makridakis (1990), also described by Tashman (2000). The reason we use this model selection criterion, although very simple, is that practically we cannot use any criterion that needs within-sample information about the forecasting methods since the forecasts of the model fitting of the individual methods are not available for the M3-competition. The only information available is the post-sample forecasts. Generally the observations we make in this section may not be conclusive since if one were to use a different model selection criterion the results may be different—although we conjecture that the conclusions will be qualitatively the same.

4.1. A simple practical model selection criterion

As in Makridakis (1990) and Tashman (2000), our model selection criterion uses data that have not been used for estimating the parameters of the methods—that initially serve the role of “validation data” for selection. In particular the evaluation of a combination (or individual method) uses the first two post-sample periods for yearly data, three for quarterly data, six for monthly data, and three for other data. Since we know the actual values for these post-sample “validation data” periods we can calculate the average sMAPE over these periods. We then choose the combination (individual method) with the smallest average sMAPE over these first post-sample periods. We can then calculate the sMAPE of the chosen combination (or individual method) for the remaining post-sample periods: 3–6 for yearly data, 4–8 for quarterly and other data, and 7–18 for monthly data. This effectively results in the use of one third of the post-sample data for “fitting” and the remaining two-thirds for “testing”.

The choice of the number of post-sample “fitting” periods is arbitrary. We tried three for yearly data and four for quarterly and other data without any differences in the results. Finally, a possible caveat of the criterion we use is that it leads to choices that are good for short-term forecasting (since it uses periods 1–3 as validation data), but are not necessarily good for long-term forecasting (periods 4–8). However this is “equally unfair” for both individual methods and their combinations, although there are indications that combinations are advantageous for long-term forecasting (Armstrong, 2001).

4.2. Hypothesis H3

The results for H3 are summarized in Table 3. For each row we report the sMAPE for the “testing periods”, namely periods 4–8 for the quarterly data,

<table>
<thead>
<tr>
<th></th>
<th>sMAPE</th>
<th>P-value</th>
<th>Percentage of series for which an individual method was chosen (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Chosen among all individual methods</td>
<td>8.56</td>
<td>(6.11)</td>
<td></td>
</tr>
<tr>
<td>2. Chosen among all combinations</td>
<td>7.10</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>3. Chosen among all individual methods and all combinations</td>
<td>6.75</td>
<td>(5.59)</td>
<td>28</td>
</tr>
<tr>
<td>4. Chosen among all combinations of 5</td>
<td>8.76</td>
<td>0.73</td>
<td></td>
</tr>
<tr>
<td>5. Chosen among all combinations of 7</td>
<td>9.08</td>
<td>0.38</td>
<td></td>
</tr>
<tr>
<td>6. Combination of all 14 methods</td>
<td>10.34</td>
<td>0.01</td>
<td></td>
</tr>
<tr>
<td>7. Chosen among all individual methods and all combinations of 5</td>
<td>7.62</td>
<td>0</td>
<td>52</td>
</tr>
<tr>
<td>8. Chosen among all individual methods and all combinations of 7</td>
<td>7.61</td>
<td>(6.05)</td>
<td>53</td>
</tr>
</tbody>
</table>

The corresponding P-values of comparing rows 2–8 with the first row are shown. We also report in bold the “best” or “not significantly different from the best” performances. (P-values = 0 corresponds to P-value < 0.01).
of the chosen combination (or individual method) on average across the series. In parentheses we also report the best possible performance attained for periods 4–8—replicating the experiment for H2, this time for periods 4–8 instead of 1–8. We note that hypothesis H2 still can not be rejected for the post-sample periods 4–8: there is no significant difference between the best performances for rows 1, 2, 3, 4, 7, and 8 of Table 3, but there is a significant difference between row 1 and rows 5, 6—to avoid cluttering we do not show the $P$-values for these comparisons.

We report the $P$-values for comparing the performance of the chosen individual method (row 1) and the performance of the choice for each of the rows 2–8. From Table 3 we conclude that hypothesis H3 can be rejected.

The combination chosen among all combinations (second row of Table 3) is significantly better than the chosen individual method (first row of Table 3).

We also note the following two secondary observations:

a) When we choose among both individual methods and all their combinations (third row) performance increases, although not significantly, relative to choosing only among combinations (the corresponding $P$-value is 0.50). An individual method is chosen for 28% of the series in this case.

b) When we choose among all individual methods and all combinations of a specific size $n$, the performance is significantly better than when we select only among individual methods or only among combinations of the same size $n$. In this case the performance is similar to that attained when we choose among all combinations and individual methods, and an individual method is chosen for about 50% of the series.

These results indicate that in practice the question of “whether it is better to combine or to use an individual method” should be rephrased as “how to select the best among a number of individual methods and their combinations”, which is the standard model selection problem in statistics. We conjecture that the success of such a selection depends on the variance/spread of the performances of the individual methods.

This is in agreement with existing theory (Breiman, 1998; Bousquet & Elisseeff, 2002; Evgeniou et al., 2004) which shows bounds on the predictive performance of a selected model which are of the form “future error of chosen = best attainable error + variance” with high probability, where appropriate (and various) measures of variance are defined (Breiman, 1998; Bousquet & Elisseeff, 2002; Evgeniou et al., 2004; Vapnik, 1998).

4.3. Hypothesis H4

The results for H4 are summarized in Table 4: we report the difference between the performance of the best possible and the performance of the chosen shown in Table 3 (i.e. $2.45 = 8.56 - 6.11$, in the first row). We report the $P$-values comparing the difference “chosen -best” for individual methods (row 1) and the difference “chosen-best” for each of the rows 2–8. From Table 4 we conclude that hypothesis H4 can be rejected:

The difference “chosen-best” is significantly smaller when we choose among combinations (second row) than when we choose among individual methods (first row).

Moreover, the difference “chosen-best” decreases as $n$ increases.

<table>
<thead>
<tr>
<th>Table 4</th>
<th>Difference (“chosen-best”) between average sMAPE of the chosen method (or combination) for periods 4–8, and the best possible sMAPE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>“Chosen-best” sMAPE difference</td>
</tr>
<tr>
<td>1. Chosen among all individual methods</td>
<td>2.45</td>
</tr>
<tr>
<td>2. Chosen among all combinations</td>
<td>1.09</td>
</tr>
<tr>
<td>3. Chosen among all individual methods and all combinations</td>
<td>1.16</td>
</tr>
<tr>
<td>4. Chosen among all combinations of 5</td>
<td>1.69</td>
</tr>
<tr>
<td>5. Chosen among all combinations of 7</td>
<td>1.35</td>
</tr>
<tr>
<td>7. Chosen among all individual methods and all combinations of 5</td>
<td>1.82</td>
</tr>
<tr>
<td>8. Chosen among all individual methods and all combinations of 7</td>
<td>1.56</td>
</tr>
</tbody>
</table>

The corresponding $P$-values of comparing rows 2–8 with the first row are shown ($P$-values = 0 corresponds to $P$-value < 0.01).
5. Conclusions: to combine or not to combine? Is this the real question?

There has been a lot of work arguing that combining forecasts is better than using individual methods. The results of our experiments using data from the 3003 time series of the M3-competition challenge this belief. In particular they show that:

- If one always uses the same method or same combination for all time series, then the best individual methods and combinations perform similarly (Hypothesis H1) so that there is no inherent advantage to combining. There is, however, an advantage in terms of what is the worst possible performance attained: as also shown by Makridakis and Winkler (1983), the worst performance among the individual methods (what we call HighMAPE) is significantly worse than the worst performance among the combinations.

- If one uses a different method or combination for each time series, then the best possible individual method is again, on average, similar to the best possible combination (Hypothesis H2). However, the performance of combinations drops significantly as we combine more and more methods. We emphasize that this differs from the conclusions of Makridakis and Winkler (1983) since in that work only the “over series” (Hypothesis H1) case was considered, and not the “per-series” (Hypothesis H2) one.

- On the other hand, choosing an individual method out of a set of available methods is more risky than choosing a combination (Hypothesis H4). Therefore, in practice when one chooses among methods and their combinations, overall the chosen individual method (chosen by the selection method used here) may have significantly worse performance than the chosen combination (Hypothesis H3). This is the most important finding in this paper.

Studying how to select between methods and their combinations is an important direction of research: clearly it may not always be best to combine forecasts. For example, the spread (variance) of the individual methods may provide some insights on how to select among methods and their combinations. Characteristics other than the spread and variance, such as the bracketing rate used by Larrick and Soll (2003), may be also useful for this decision. We plan to explore this question using the data of the M3-competition in the future.

A limitation of this study is on how to choose among methods or combinations in an optimal way. Although the model selection procedure is necessitated by the data available, it is not a very powerful method. As we mentioned, the criterion used for selection may influence the conclusions of this work. Thus, an open question for research is to find a combination of forecasts that is robust with respect to the choice of the criterion function.

Acknowledgements

The authors would like to thank Robert Winkler, Spyros Makridakis and Keith Ord for their helpful comments and suggestions, and the two anonymous referees for their constructive comments.

References


**Biographies:** Michèle HIBON is a Senior Research Fellow and Lecturer at INSEAD. She graduated from the University of Paris, holds a degree in Science and a diploma in advanced studies in Physics. Her current research interests are in the areas of forecasting accuracy of time series methods including M-, M2-, and M3-competitions. She has worked in the area of Group Decision Support Systems.

Theodoros EVGENIOU is an Assistant Professor of Information Systems in the Technology Management department of INSEAD. He holds two B.S., an M.Eng., and a Ph.D. in Computer Science from MIT, where he has been affiliated with the Artificial Intelligence Laboratory and the Center for Biological and Computational Learning. His current research interests are in the areas of developing and implementing information based strategies, as well as data mining and business intelligence, both developing technologies in these fields and studying their strategic impact on business.