International Journal of Forecasting 16 (2000) 451–476

www.elsevier.com/locate/ijforecast

The M3-Competition: results, conclusions and implications

Spyros Makridakis, Michèle Hibon*

INSEAD, Boulevard de Constance, 77305 Fontainebleau, France

Abstract

This paper describes the M3-Competition, the latest of the M-Competitions. It explains the reasons for conducting the competition and summarizes its results and conclusions. In addition, the paper compares such results/conclusions with those of the previous two M-Competitions as well as with those of other major empirical studies. Finally, the implications of these results and conclusions are considered, their consequences for both the theory and practice of forecasting are explored and directions for future research are contemplated.

Keywords: Comparative methods — time series: univariate; Forecasting competitions; M-Competition; Forecasting methods, Forecasting accuracy

1. Introduction

Forecasting accuracy is a critical factor for, among other things, reducing costs and providing better customer service. Yet the knowledge and experience available for improving such accuracy for specific situations is not always utilized. The consequence is actual and/or opportunity losses, sometimes of considerable magnitude. Empirical studies in the field of forecasting have compared the post-sample forecasting accuracy of various methods so that their performance can be determined in an objective, measurable manner. The M-Competitions are such empirical studies that have compared the performance of a large number of major time series methods using recognized experts who provide the forecasts for their method of expertise. Once the forecasts from each expert have been obtained they are evaluated and compared with those of the other experts as well as with some simple methods used as benchmarks. Forecasting competitions assure objectivity while also guaranteeing expert knowledge.

This paper summarizes the results of the latest of the Makridakis, or M-Competitions, the M3. It presents the conclusions that can be drawn from such results and compares them with those of the two previous M-Competitions, as well as with those of other major empirical studies. In addition, the implications of these results and conclusions are discussed and their

*Corresponding author. Tel.: +33-1-6072-4000; fax: +33-1-6074-5500.

E-mail addresses: spyros.makridakis@insead.fr (S. Makridakis), michele.hibon@insead.fr (M. Hibon).

0169-2070/00/$ – see front matter
PII: S0169-2070(00)00057-1

consequences for the theory and practice of forecasting are explored. The M-Competitions refer mainly to business and economic time series, although their results/conclusions may well be relevant to other disciplines. The paper ends with suggestions for future research and some concluding remarks. The M-Competitions refer mainly to business and economic time series, although their results/conclusions may well be relevant to other disciplines.

2. The history of accuracy studies and competitions

As far back as 1969, Reid (1969, 1975) and Newbold and Granger (1974) compared a large number of series to determine their post-sample forecasting accuracy. However, these early accuracy studies based their comparisons on a limited number of methods. Makridakis and Hibon (1979) was the first effort to compare a large number of major time series methods across multiple series. Altogether 111 time series were selected from a cross section of available data, covering a wide range of real-life situations (business firms, industry and macro data). The major conclusion of the Makridakis and Hibon study was that simple methods, such as exponential smoothing, outperformed sophisticated ones. Such a conclusion was in conflict with the accepted view (paradigm) of the time and was not received well by the great majority of commentators, mostly statisticians (see the commentary following the Makridakis & Hibon, 1979, study). To respond to the criticisms and to incorporate the suggestions of the various commentators for improvements, Makridakis continued the empirical comparisons of time series by launching the M-Competition (Makridakis et al., 1982).

In the M-Competition the number of series utilized was increased to 1001 and the number of methods to 15 (with another nine variations of these methods also included). Furthermore, more accuracy measures were employed while the data were subdivided into various categories (micro, macro, industry, etc.) in order to determine the reasons why some method(s) outperformed others. However, the most important innovation of the M-Competition (hence the name Competition) was that an expert was designated to run the 1001 series (or a subsample of 111 when the amount of work to implement a method was too much to use all 1001 series) in his/her area of expertise. Each expert provided his/her forecasts that were compared, in a post-sample fashion, with actual values not used in developing the forecasting model. These forecast errors were then used to compute the various reported accuracy measures (see Makridakis et al., 1982).

The results of the M-Competition were similar to those of the earlier Makridakis and Hibon study and can be summarized as follows:

(a) Statistically sophisticated or complex methods do not necessarily provide more accurate forecasts than simpler ones.

(b) The relative ranking of the performance of the various methods varies according to the accuracy measure being used.

(c) The accuracy when various methods are being combined outperforms, on average, the individual methods being combined and does very well in comparison to other methods.

(d) The accuracy of the various methods depends upon the length of the forecasting horizon involved.

Many researchers have replicated the conclusions of the M-Competition in four important ways. First, the calculations on which the study was based were re-verified and their appropriateness widely accepted. Second, new methods have been introduced and the results obtained have been found to agree with those of the M-Competition (Geurts & Kelly, 1986; Clemen, 1989; Fildes, Hibon, Makridakis & Meade, 1998). Third, many researchers (Hill &

Fildes, 1984; Lusk & Neves, 1984; Koehler & Murphree, 1988) have used the M-Competition data and have reached similar conclusions. Finally, additional studies using new data series have agreed with the above four conclusions (Armstrong & Collopy, 1992, 1993; Makridakis et al., 1993; Fildes et al., 1998) and have demonstrated, above any reasonable doubt, the validity of these four conclusions. Yet, there are still emotional objections to empirical accuracy studies (see Newbold, 1983) and criticisms for all types of empirical work (see Fildes & Makridakis, 1995, for a full discussion of such objections/criticisms and the implications for the field of forecasting).

The M-2 Competition (Makridakis et al., 1993) was a further attempt to provide an additional forum to study the accuracy of various forecasting methods and better understand the factors that affect forecasting accuracy. Again, as in the M-Competition, a call to participate in the M2-Competition was published in the *International Journal of Forecasting*, announcements were made during the International Symposium of Forecasting and a written invitation was extended to all known experts of the various time series methods. The M2-Competition was organized in collaboration with four companies and included six macro-economic series. It was designed and run on a *real-time* basis. This meant that the companies not only provided the participating experts with actual data, about the past and present, but they were also committed to answer their questions about such data, the factors that affected their business and the variables they were considering while forecasting the series that were given to the participants. The macro-economic data were from the USA, whose economic situation was known at the time to the participants. The competition was run for two years and the participating experts had to forecast for the next 15 months, as is the case when predictions in business firms are being made for next year's budget, sometime in September or October. The first year, in addition to the data, the participants were also provided with supplementary information about the industry and the company involved. As the competition was run on a real-time basis the actual state of the economy was known to the participating experts, who could also find, from published sources, additional information about the industry each company was operating, if they wished so.

A year later the actual values for the last 15 months were given to the participating experts so that they could check the accuracy of the forecasts they had made a year earlier. Furthermore, the experts were given additional information, concerning the forthcoming year, about the industry and the company. They could also write or call a contact person in each company if they desired helpful hints or clarifications about the industry/company and/or the data.

The results of the M2-Competition were practically identical to those of the M-Competition. Statistically sophisticated or complex methods did not provide more accurate forecasts than simpler ones. The relative ranking of the performance of the various methods varied according to the accuracy measure being used. The accuracy of combining various methods outperformed, on average, the individual methods used. And, the accuracy of the different methods depended upon the length of the forecasting horizon involved.

Although, the conclusions of the Makridakis and Hibon (1979) study could be questioned as they depended upon the forecasting skills of two individuals (Makridakis and Hibon), those of the M- and M2-Competitions were above such criticisms. In addition, every conceivable effort was being made to achieve as high a degree of objectivity as possible. Such efforts included finding knowledgeable participants to run each method expertly and to assure that their forecasting procedure was well documented so that

it could be replicated by other researchers and be available for later scrutiny. Such replication and scrutiny has indeed taken place. The data of the M- and M2-Competitions have been made available to more than 600 researchers who have studied every single aspect of the methods (for example, see Lusk & Neves, 1984) and the computations (Simmons, 1986). Moreover, new and different data sets (Grambsch & Stahel, 1990; Fildes, 1992; Armstrong & Collopy, 1993) further confirm the conclusions of the M-Competition and increase our confidence for generalizing them to new data sets and different situations.

The strong empirical evidence, however, has been ignored by theoretical statisticians (see Fildes & Makridakis, 1995) who have been hostile to empirical verifications (for example, see Newbold, 1983). Instead, they have concentrated their efforts in building more sophisticated models without regard to the ability of such models to more accurately predict real-life data. The M3-Competition is a final attempt by the authors to settle the accuracy issue of various time series methods. Its major aim has been to both replicate *and* extend the M- and the M2-Competitions. The extension involves the inclusion of more methods/researchers (in particular in the areas of neural networks and expert systems) and more series. The replication was intended to determine whether or not the major conclusions of the M-Competition would hold with the new, much enlarged, set of 3003 time series

3. Organizing and conducting the M3-Competition

The 3003 series of the M3-Competition were selected on a quota basis to include various types of time series data (micro, industry, macro, etc.) and different time intervals between successive observations (yearly, quarterly, etc.). In order to ensure that enough data were available to develop an adequate forecasting model it was decided to have a minimum number of observations for each type of data. This minimum was set as 14 observations for yearly series (the median length for the 645 yearly series is 19 observations), 16 for quarterly (the median length for the 756 quarterly series is 44 observations), 48 for monthly (the median length for the 1428 monthly series is 115 observations) and 60 for 'other' series (the median length for the 174 'other' series is 63 observations). Table 1 shows the classification of the 3003 series according to the two major groupings described above. All the time series data are strictly positive; a test has been done on all the forecasted values: in the case of a negative value, it was substituted by zero. This avoids any problem in the various MAPE measures.

Table 1
The classification of the 3003 time series used in the M3-Competition

Time interval between successive observations	Types of time series data						
	Micro	Industry	Macro	Finance	Demographic	Other	Total
Yearly	146	102	83	58	245	11	645
Quarterly	204	83	336	76	57		756
Monthly	474	334	312	145	111	52	1428
Other	4			29		141	174
Total	828	519	731	308	413	204	3003

As in the M-Competition, the participating experts were asked to make the following numbers of forecasts beyond the available data they had been given: six for yearly, eight for quarterly, 18 for monthly and eight for the category 'other'. Their forecasts were, subsequently, compared by the authors (the actual values referred to such forecasts were not available to the participating experts when they were making their forecasts and were not, therefore, used in developing their forecasting model). A presentation of the accuracy of such forecasts together with a discussion of the major findings is provided in the next section.

The M3-Competition was given a lot of publicity in the *International Journal of Forecasting*, during forecasting conferences, on the Internet and by mailing individualized letters to recognized experts in various time series forecasting methods. In doing so we sought to attract the maximum number of participants, in particular from the new areas of neural networks and expert systems where claims of superior forecasting performance were continuously being made. While announcing the M3-Competition we received many hundreds of requests for information and we sent the 3003 series to more than 100 potential participants. Moreover, many other researchers must have downloaded the 3003 series from the Internet site: http://www.insead.fr/facultyresearch/forecasting that contained the data. However, as the deadline was approaching the number of participants submitting forecasts could be counted on the fingers of two hands, despite multiple reminders and the extension of the deadline. What was most disappointing was the large number of experts in neural networks and expert systems who dropped out after they had received the M3-Competition data and had indicated their intention to participate in the M3-Competition.

In the next section the results for 24 methods, subdivided into six categories, are presented. Such methods include both all those utilized in the M-Competition plus seven new ones from the areas of neural networks, expert systems and decomposition. Table 2 lists the 24 methods included in the M3-Competition, with a brief description for each, and the various subcategories to which they belong.

4. The results of the M3-Competition

Five accuracy measures (symmetric MAPE, Average Ranking, Median symmetric APE, Percentage Better, and Median RAE) were used to analyze the performance of the various methods. For a short description of these accuracy measures see Appendix A, while for greater details see Makridakis et al. (1982) and Armstrong and Collopy (1992). Appendix B includes many tables with full results for each of some of these accuracy measures for all the 3003 series and for the different categories of data and the various time horizons. The Internet site: http://www.insead.fr/facultyresearch/forecasting contains the full details of these accuracy measures together with more extensive sets of tables and figures (corresponding to Appendices B and C). Although there is a great number of tables and too many numbers we believe that they provide researchers and practitioners with useful information to judge, for their specific situation, the relative accuracy of the various methods covered in the M3-Competition. In the remainder of this section we analyze and summarize the results of these tables and provide our own interpretation and conclusions concerning these results. The other papers included in this special issue present descriptions, by each of the participating experts, of the methods listed in Table 2 and their own interpretation of the results. In a future issue the *International Journal of Forecasting* will publish commentaries concerning the results and conclusions of the M3-Competition.

Table 2
The 24 methods included in the M3-Competition classified into six categories

Method	Competitors	Description
Naïve/simple		
1. Naïve2	M. Hibon	Deseasonalized Naïve (Random Walk)
2. Single	M. Hibon	Single Exponential Smoothing
Explicit trend models		
3. Holt	M. Hibon	Automatic Holt's Linear Exponential Smoothing (two parameter model)
4. Robust-Trend	N. Meade	Non-parametric version of Holt's linear model with median based estimate of trend
5. Winter	M. Hibon	Holt–Winter's linear and seasonal exponential smoothing (two or three parameter model)
6. Dampen	M. Hibon	Dampen Trend Exponential Smoothing
7. PP-autocast[a]	H. Levenbach	Damped Trend Exponential Smoothing
8. Theta-sm	V. Assimakopoulos	Successive smoothing plus a set of rules for dampening the trend
9. Comb S-H-D	M. Hibon	Combining three methods: Single/Holt/Dampen
Decomposition		
10. Theta	V. Assimakopoulos	Specific decomposition technique, projection and combination of the individual components
ARIMA/ARARMA model		
11. B–J automatic	M. Hibon	Box–Jenkins methodology of 'Business Forecast System'
12. Autobox1[a]	D. Reilly	Robust ARIMA univariate Box–Jenkins with/without Intervention Detection
13. Autobox2[a]		
14. Autobox3[a]		
15. AAM1	G. Melard,	Automatic ARIMA modelling with/without intervention analysis
16. AAM2	J.M. Pasteels	
17. ARARMA	N. Meade	Automated Parzen's methodology with Auto regressive filter
Expert system		
18. ForecastPro[a]	R. Goodrich, E. Stellwagen	Selects from among several methods: Exponential Smoothing/Box Jenkins/Poisson and negative binomial models/Croston's Method/Simple Moving Average
19. SmartFcs[a]	C. Smart	Automatic Forecasting Expert System which conducts a forecasting tournament among four exponential smoothing and two moving average methods
20. RBF	M. Adya, S. Armstrong, F. Collopy, M. Kennedy	Rule-based forecasting: using three methods — random walk, linear regression and Holt's, to estimate level and trend, involving corrections, simplification, automatic feature identification and re-calibration
21. Flores/Pearce1	B. Flores,	Expert system that chooses among four methods based on the characteristics of the data
22. Flores/Pearce2	S. Pearce	
23. ForecastX[a]	J. Galt	Runs tests for seasonality and outliers and selects from among several methods: Exponential Smoothing, Box–Jenkins and Croston's method
Neural networks		
24. Automat ANN	K. Ord, S. Balkin	Automated Artificial Neural Networks for forecasting purposes

[a] Commercially available forecasting packages. Professionals employed by those companies generated the forecasts utilized in this Competition.

Table 3
Comparison of various methods with Naïve2 as the benchmark

	Forecasting horizon(s)				
	1	Average: 1–4	Average: 1–6	Average: 1–12	Average: 1–18
Theta	2.1%	2.2%	2.1%	2.3%	2.5%
ForecastPro	1.9%	2.0%	1.9%	2.1%	2.3%
ForecastX	1.8%	1.8%	1.8%	2.0%	2.0%
Comb S-H-D	1.6%	1.5%	1.5%	1.8%	2.0%
Dampen	1.7%	1.6%	1.5%	1.8%	1.8%
RBF	0.6%	1.1%	1.3%	1.5%	1.7%
ARARMA	0.8%	0.8%	0.7%	0.7%	0.7%

4.1. The accuracy of various methods: comparisons to a benchmark

The absolute accuracy of the various methods is not as important as how well these methods perform relative to some benchmark. The simplest benchmark is Naïve2 (a random walk model that is applied to seasonally adjusted data by assuming that seasonality is known; see Appendix A for a brief description of Naïve2). Another easy benchmark is Dampen Trend Exponential Smoothing (Gardner & McKenzie, 1985). Table 3 lists the difference in the forecasting performance of the six most accurate forecasting methods, with a symmetric MAPE (sMAPE) below 14%, as well as ARARMA (the most sophisticated time series method) in relation to Naïve2. Table 3 shows sMAPE(Naïve2) − sMAPE(selected method), averaged across series, using the results in Appendix B (Table 6).

It is clear that the accuracy of practically all methods included in Table 3 is considerably better than that of Naïve2 which *only* captures the seasonality in the data. This is a very encouraging contribution which illustrates that the six methods listed in Table 3 can accurately predict other time series patterns, in addition to seasonality.

The comparisons of Table 4 are similar to those of Table 3 except, however, that Dampen Trend Exponential Smoothing is used as the benchmark (a negative sign signifies that the accuracy of the method listed is worse than that of Dampen). Table 4 shows sMAPE(Dampen) − sMAPE(selected method), averaged across series, using the results in Appendix B (Table 6).

Table 4
Comparison of various methods with Dampen as the benchmark

	Forecasting horizon(s)				
	1	Average: 1–4	Average: 1–6	Average: 1–12	Average: 1–18
Theta	0.4%	0.6%	0.5%	0.5%	0.6%
ForecastPro	0.2%	0.4%	0.4%	0.3%	0.4%
ForecastX	0.1%	0.2%	0.3%	0.2%	0.1%
Comb S-H-D	−0.1%	0.0%	0.0%	0.0%	0.1%
RBF	−1.1%	−0.5%	−0.2%	−0.3%	−0.1%
ARARMA	−0.9%	−0.8%	−0.9%	−1.1%	−1.1%

In Table 4 the differences in the forecasting performance (as far as symmetric MAPE is concerned) are small. The overall *extra* improvement (average of all 18 horizons) of the two most accurate methods is around 0.5% (half of one percent). As the actual overall symmetric MAPE of these methods is around 13%, this 0.5% represents an improvement, in symmetric MAPE, of Theta and ForecastPro of around 4%. The equivalent improvement of ForecastX is less than 1% while for RBF and ARARMA it is negative, meaning that Dampen is, on average, more accurate than these two methods. The accuracy of the remaining methods used in the M3-Competition, as far as the average symmetric MAPE is concerned, is worse than that of Dampen in most forecasting horizons (see Table 6).

If similar comparisons as those shown in Tables 3 and 4 are made with the remaining accuracy measures the results are, in most cases, similar to those shown in Tables 3 and 4. Although several forecasting methods outperform Dampen the differences involved are small and, in most cases, not statistically significant. This is particularly true for specific forecasting horizons and particular types of data.

Figs. 1 to 9 in Appendix C (C1–C25 on the web site) show, in graphical form, the differences of Dampen from the most important methods of the M3-Competition for three or four different forecasting horizons. Such figures confirm the good performance of Dampen while at the same time demonstrate that several methods consistently outperform Dampen. The forecast user will have to decide if the extra improvement in accuracy justifies the additional effort or cost that may be required when using time series methods other than Dampen.

4.2. The four conclusions of the M-Competition

This section examines the question of whether or not the four major conclusions of the M-Competition also apply to the 3003 data of the M3-Competition.

(1) *Statistically sophisticated or complex methods do not necessarily produce more accurate forecasts than simpler ones*. Tables 3 and 4 (see also the tables in Appendix B, the figures in Appendix C and the tables and figures on the INSEAD website) illustrate, beyond the slightest doubt, that statistically sophisticated methods *do not necessarily* outperform simple ones. This does not mean that some sophisticated methods do not do well or that it is always obvious how a method can be classified as simple or sophisticated. However, Tables 3 and 4 suggest that we cannot advance the statement that sophisticated time series methods outperform, on average, simple ones like Dampen trend.

(2) *The rankings of the performance of the various methods vary according to the accuracy measure being used*. Table 7 shows the method that gives the 'best' results (when more than one method is designated as 'best', their accuracy is the same within one decimal). Table 7 suggests that the 'best' method varies according to the accuracy measure being used and the type of data (micro, industry, macro, etc.) involved. Such differentiation becomes clearer in Tables 8 to 11 where the data are further subdivided into yearly, quarterly, monthly and 'other'.

(3) *The accuracy of the combination of various methods outperforms, on average, the specific methods being combined and does well in comparison with other methods*. In the various tables and figures, the method 'Comb S-H-D' is the simple arithmetic average of three methods: Single, Holt and Dampen Trend Exponential Smoothing. Table 5 shows the symmetric MAPE of Single, Holt and Dampen as well as that of their combination. Clearly, the combination is more accurate than the three individual methods being combined for practically all forecasting horizons, although its dif-

Table 5
Symmetric MAPE of Single, Holt and Dampen, and their combination

	Forecasting horizon(s)				
	1	Average: 1–4	Average: 1–6	Average: 1–12	Average: 1–18
Single	9.5%	11.7%	12.7%	13.1%	14.3%
Holt	9.0%	11.7%	12.9%	13.4%	14.6%
Dampen	8.8%	11.1%	12.0%	12.4%	13.6%
Comb S-H-D	8.9%	11.1%	12.0%	12.4%	13.5%

ference from Dampen is small (since Dampen does extremely well so does Comb S-H-D).

(4) *The performance of the various methods depends upon the length of the forecasting horizon*. Table 12 lists the best method, using symmetric MAPE, for short, medium and long forecasting horizons. This table indicates that the best method varies with the forecasting horizon, in particular when subcategories of the data are involved (the same conclusion can be seen from the other tables and figures of Appendix B as well as from the tables/figures on the INSEAD website). An exception is the forecasting performance of Theta, a new method used for the first time in the M3-Competition, which seems to perform consistently well across both forecasting horizons *and* accuracy measures (see Tables 7–11).

4.3. M3-Competition: implications for the theory and practice of forecasting

Better predictions remain the foundation of all science and the primary purpose of forecasting which must strive to achieve such an objective by all possible means. Pure theory and elaborate/sophisticated methods are of little practical value unless they can contribute to improving the accuracy of post-sample predictions. This study, the previous two M-Competitions and many other empirical studies have proven, beyond the slightest doubt, that elaborate theoretical constructs or more sophisticated methods do not necessarily improve post-sample forecasting accuracy, over simple methods, although they can better fit a statistical model to the available historical data. The authors of this paper believe that the time has come to accept this finding so that pragmatic ways can be found to improve, as much as possible, post-sample predictions. Such improvement can result in considerable benefits at the operational level of business firms, and other organizations (e.g., smaller inventories, superior scheduling, more effective allocation of resources, etc.), and can be exploited to provide better customer service. Each percentage improvement in post-sample forecasting accuracy can result in savings of many millions of dollars, less wasted resources, and/or better service. In order to improve forecasting accuracy, both research statisticians and practical forecasters must work together to advance the field of forecasting, with the single objective in mind of how to ameliorate its practical value and usefulness (Fildes & Makridakis, 1995).

4.4. Suggestions for further research

The reason for the anomalies between the theory and practice of forecasting is that real-life time series are not stationary while many of them also contain structural changes as fads, and fashions can change established patterns and affect existing relationships. Moreover, the randomness in such series is high as competitive

actions and reactions cannot be accurately predicted and as unforeseen events (e.g., extreme weather conditions) affecting the series involved can and do occur. Finally, many time series are influenced by strong cycles of varying duration and lengths whose turning points cannot be predicted, making them behave like a random walk. It is for these reasons that simple methods (e.g., Single exponential smoothing which does not extrapolate any trend) can outperform, in certain cases, statistically sophisticated ones that identify and extrapolate the trend (and other patterns) in the data.

Fildes and Makridakis (1995) have identified the following areas for research so that the accuracy of time series methods can be improved by taking into account the real-life behavior of data:

- Exploiting the robustness of simple methods that are less influenced than advanced ones by structural changes in the data.
- Modeling the trend in a more practical way by realizing that many series are random walks and that established trends in the data can and do change (a good example of such an approach is Dampen Trend Exponential Smoothing).
- As the forecasting performance of different methods is related to the forecasting horizon it would be possible to develop methods that combine the advantages of the methods that more accurately predict the short term *and* those that are more effective in forecasting the long term.
- As model fit is not a good indicator of the post-sample forecasting accuracy of the various methods it would be worthwhile to develop methods/models where the selection is done using out of sample criteria (see Chatfield, 1995).
- It may be possible that the post-sample accuracy of time series methods can be improved by incorporating multivariate information that will affect the future behavior of such series so that predictions can be improved.

5. Conclusions

This Competition has confirmed the original conclusions of M-Competition using a new and much enlarged set of data. In addition, it demonstrated, once more, that simple methods developed by practicing forecasters (e.g., Brown's Single and Gardner's Dampen Trend Exponential Smoothing) do as well, or in many cases better, than statistically sophisticated ones like ARIMA and ARARMA models. In addition, the M3-Competition has reached three additional conclusions that need further confirmation. First, a new method, Theta, seems to perform extremely well. Although this method seems simple to use (see article describing Theta for deciding the extent of simplicity/complexity of this method) and is not based on strong statistical theory, it performs remarkably well across different types of series, forecasting horizons and accuracy measures. Hopefully, new methods, similar to Theta, can be identified and brought to the attention of practicing forecasters. Second, ForecastPro, another new method not utilized in the M-Competition, also did well. In the spirit of Brown's attempts to obtain more accurate forecasts, this approach is empirically based and eclectic in nature. It identifies and uses the most appropriate method from a set of possible choices. Finally, this Competition, like Fildes et al. (1998), has shown that a specific method (i.e., Robust-Trend) can outperform all others when yearly data are involved. It may be possible that other methods can be found that can also outperform existing ones in specific situations and, therefore, be used exclusively for such situations only. Clearly, more research will be needed to establish the reason why, for instance, Robust-Trend is so

well suited for yearly data. Is it some inherent aspect of such a method or rather its robust estimation procedure? Similar questions will need to be answered, through additional research, for other methods. Is, for instance, the reason for Theta's excellent performance its way of deseasonalizing the data, its estimation procedure, or its ability to deal with extreme values? These and similar questions, if answered, can contribute to improving forecasting accuracy a great deal and make the field of forecasting more useful and relevant.

As with the previous two M-Competitions, the data for M3 are available to any researcher who wants to use them. This can be done by contacting Michèle Hibon at michele.hibon@insead.fr, or by downloading the M3-Competition data from the site: http://www.insead.fr/facultyresearch/forecasting. We hope that this new data set of the 3003 series will become the basis for more empirical research in the field of forecasting and that its impact on the science and practice of forecasting will prove to be even more significant than that of the M-Competition data. We strongly believe that more empirical research is needed to advance the field of forecasting and make it more practical and relevant for business and other organizations requiring predictions. Ignoring empirical findings is contrary to rational thinking and scientific inquiry.

We are convinced that those criticizing Competitions, and empirical studies in general, should stop doing so and instead concentrate their efforts on explaining the anomalies between theory and practice and on working to improve the accuracy of forecasting methods. Emotional criticisms are not appropriate for good science. Everyone in the field of forecasting ought to heed the advice of Kuhn (1962) that ''Discovery commences with the awareness of anomaly. ... It then continues with a more of less extended exploration of the area of anomaly. And it closes when the paradigm theory has been adjusted so that the anomalous has become the expected.'' Perhaps the time has come to follow the example of a recent conference on the 'Future of Economics' (see *The Economist*, March 4th, 2000, p. 90) and start debating, in a serious and scientific manner, the future of forecasting.

Appendix A

The five accuracy measures utilized in the M3-Competition

The five accuracy measures employed to describe the results of the M3-Competition are the following.

- *Symmetric mean absolute percentage error*
  The symmetric MAPE (sMAPE) is defined as

$$\sum \frac{|X - F|}{(X + F)/2} * 100$$

where X is the real value and F is the forecast.The symmetric MAPE is the average across all forecasts made for a given horizon.

By using the symmetric MAPE, we avoid the problem of large errors when the actual, X, values are close to zero and the large difference between the absolute percentage errors when X is greater than F and vice versa (e.g., the absolute, non-symmetric, percentage error when $X = 100$ and $F = 50$ is 50%, while when $X = 50$ and $F = 100$ it is 100%. On the other hand, the symmetric MAPE in both cases is 66.67%). In addition, the symmetric MAPE fluctuates between -200% and 200% while the non-symmetric measure does not have limits.

- *Average ranking*

For each series, the average rankings are computed by sorting, for each forecasting horizon, the symmetric absolute percentage error of each method from the smallest (taking

the value of 1) to the largest. Consequently, once the ranks for all series have been determined, the mean rank is calculated for each forecasting horizon, over all series. An overall average ranking is also calculated by averaging the ranks over six, eight or 18 forecasts, for each method.

• *Percentage better*

The percentage better measure counts and reports the percentage of time that a given method has a smaller forecasting error than another method. Each forecast made is given equal weight. Our comparisons in Appendix B and in Fig. 7 (C7 to C11 on the website) use Dampen as the benchmark to present the percentage of time that this method does better than the others.

• *Median symmetric APE (median symmetric absolute percentage error)*

The median symmetric absolute percentage error is found and reported for each method/forecasting horizon. Such a measure is not influenced by extreme values and is more robust than the average absolute percentage error. In the case of the M3-Competition the differences between symmetric MAPEs and Median symmetric APEs were much smaller than the corresponding values in the M-Competition as care has been taken so that the level of the series not be close to zero while, at the same time, using symmetric percentage errors which reduce their fluctuations.

• *Median RAE (relative absolute error)*

The RAE is the absolute error for the proposed model relative to the absolute error for the Naïve2 (no-change model). It ranges from 0 (a perfect forecast) to 1.0 (equal to the random walk), to greater than 1 (worse than the random walk). The RAE is similar to Theil's U2, except that it is a linear rather than a quadratic measure. It is designed to be easy to interpret and it lends itself easily to summarizing across horizons and across series as it controls for scale and for the difficulty of forecasting. The Median RAE (MdRAE) is recommended for comparing the accuracy of alternative models as it also controls for outliers (for information on the performance of this measure, see Armstrong & Collopy, 1992).

Defining Naïve2

The forecasts of Naïve2 are simply the last available data value X_t, assuming that seasonality is known. It is defined as follows:

$$F_{t+i} = X_t^*(S_j)$$

where X_t^* is the seasonally adjusted value of X_t, that is X_t/S_j, S_j is the seasonal index, computed using the classical decomposition method, for the j period (quarter or month), and $i = 1,2,\ldots,m$ (where $m = 6$ for yearly data, 8 for quarterly and 'other' and 18 for monthly).

In statistical terms Naïve2 is a random walk model applied to seasonally adjusted data. As such Naïve2 assumes that the trend in the data cannot be predicted and that the best forecast for the future is the most recent value, after the seasonality has been taken into consideration.

Appendix B

Table 6
Average symmetric MAPE: all data

Method	Forecasting horizon										Average of forecasting horizon						# obs
	1	2	3	4	5	6	8	12	15	18	1 to 4	1 to 6	1 to 8	1 to 12	1 to 15	1 to 18	
Naïve2	10.5	11.3	13.6	15.1	15.1	15.9	14.5	16	19.3	20.7	12.62	13.57	13.76	14.24	14.81	15.47	3003
Single	9.5	10.6	12.7	14.1	14.3	15	13.3	14.5	18.3	19.4	11.73	12.71	12.84	13.13	13.67	14.32	3003
Holt	9	10.4	12.8	14.5	15.1	15.8	13.9	14.8	18.8	20.2	11.67	12.93	13.11	13.42	13.95	14.6	3003
Dampen	8.8	10	12	13.5	13.7	14.3	12.5	13.9	17.5	18.9	11.05	12.04	12.14	12.44	12.96	13.63	3003
Winter	9.1	10.5	12.9	14.6	15.1	15.9	14	14.6	18.9	20.2	11.77	13.01	13.19	13.48	14.01	14.65	3003
Comb S-H-D	8.9	10	12	13.5	13.7	14.2	12.4	13.6	17.3	18.3	11.1	12.04	12.13	12.4	12.91	13.52	3003
B–J automatic	9.2	10.4	12.2	13.9	14	14.8	13	14.1	17.8	19.3	11.42	12.41	12.54	12.8	13.35	14.01	3003
Autobox1	9.8	11.1	13.1	15.1	16	16.8	14.2	15.4	19.1	20.4	12.3	13.67	13.78	14	14.56	15.23	3003
Autobox2	9.5	10.4	12.2	13.8	13.8	14.9	13.2	15.2	18.2	19.9	11.48	12.44	12.63	13.1	13.7	14.41	3003
Autobox3	9.7	11.2	12.9	14.6	15.8	16.5	14.4	16.1	19.2	21.2	12.08	13.43	13.64	14.01	14.57	15.33	3003
Robust-Trend	10.5	11.2	13.2	14.7	15	15.9	15.1	17.5	22.2	24.3	12.38	13.4	13.73	14.57	15.42	16.3	3003
ARARMA	9.7	10.9	12.6	14.2	14.6	15.6	13.9	15.2	18.5	20.3	11.83	12.92	13.12	13.54	14.09	14.74	3003
Automat ANN	9	10.4	11.8	13.8	13.8	15.5	13.4	14.6	17.3	19.6	11.23	12.38	12.58	12.96	13.48	14.11	3003
Flores/Pearce1	9.2	10.5	12.6	14.5	14.8	15.3	13.8	14.4	19.1	20.8	11.68	12.79	13.03	13.31	13.92	14.7	3003
Flores/Pearce2	10	11	12.8	14.1	14.1	14.7	12.9	14.4	18.2	19.9	11.96	12.77	12.81	13.04	13.61	14.29	3003
PP-autocast	9.1	10	12.1	13.5	13.8	14.7	13.1	14.3	17.7	19.6	11.2	12.21	12.4	12.8	13.34	14.01	3003
ForecastPro	8.6	9.6	11.4	12.9	13.3	14.3	12.6	13.2	16.4	18.3	10.64	11.69	11.86	12.14	12.6	13.19	3003
SmartFcs	9.2	10.3	12	13.5	14	15.1	13	14.9	18	19.4	11.23	12.34	12.49	12.94	13.48	14.13	3003
Theta-sm	9.8	11.3	12.6	13.6	14.3	15	12.7	14	16.2	18.3	11.81	12.76	12.77	13.04	13.4	13.88	3003
Theta	8.4	9.6	11.3	12.5	13.2	14	12	13.2	16.2	18.2	10.44	11.49	11.62	11.95	12.42	13.01	3003
RBF	9.9	10.5	12.4	13.4	13.2	14.2	12.8	14.1	17.3	17.8	11.56	12.28	12.42	12.77	13.25	13.75	3003
ForecastX	8.7	9.8	11.6	13.1	13.2	13.9	12.6	13.9	17.8	18.7	10.82	11.73	11.89	12.22	12.81	13.49	3003
AAM1	9.8	10.6	11.2	12.6	13	13.5	14.1	14.9	18	20.4	11.04	11.76	12.43	13.04	13.77	14.63	2184
AAM2	10	10.7	11.3	12.9	13.2	13.7	14.3	15.1	18.4	20.7	11.21	11.95	12.62	13.21	13.97	14.85	2184

Table 7
Methods which give the best results: all data

Accuracy measure	Micro (828)	Industry (519)	Macro (731)	Finance (308)	Demographic (413)	Other (204)
Symmetric MAPE	Theta ForecastPro	ForecastX/ ForecastPro	RBF/ARARMA Theta/ Robust-Trend	AAM1/ AAM2	ForecastX Dampen ForecastPro/RBF SmartFcs Comb S-H-D	Comb S-H-D ARARMA ForecastPro
Average RANKING	Theta	ForecastPro Theta/ ForecastX Comb S-H-D	Robust-Trend	AAM1/ AAM2	Robust-Trend ForecastX	Theta Autobox2/ ARARMA ForecastPro
Median APE	Theta ForecastPro	ForecastX Theta	Robust-Trend ARARMA	Autobox3 ForecastPro	RBF Robust-Trend	Theta Autobox2
Median RAE	Theta	Theta RBF/ Comb S-H-D	Robust-Trend ARARMA RBF	Robust-Trend ARARMA Theta AAM1/AAM2	RBF	ARARMA Theta Autobox2 Comb S-H-D

Table 8
Methods which give the best results: yearly data

Accuracy measure	Micro (146)	Industry (102)	Macro (83)	Finance (58)	Demographic (245)	Total (645)
Symmetric MAPE	Robust-Trend Flores/Pearce2 SmartFcs Autobox2	Theta Comb S-H-D Autobox2	Robust-Trend ARARMA	Autobox2 Single Naïve2	ForecastX RBF	RBF ForecastX Autobox2 Theta Robust-Trend
Average RANKING	Robust-Trend Theta/ Autobox2	Theta Comb S-H-D/ Robust-Trend RBF	Robust-Trend ARARMA RBF	Single Naïve2/ Autobox2 ForecastX/ ForecastPro	ForecastX ForecastPro/ PP-autocast	RBF/ForecastX Theta/ Robust-Trend/ Autobox2
Median APE	Robust-Trend SmartFcs	Robust-Trend	Robust-Trend ForecastPro	Single Naïve2 Autobox2	ForecastX ForecastPro RBF Theta/ Autobox2	RBF Flores/Pearce1 PP-autocast Dampen
Median RAE	Robust-Trend SmartFcs/ Theta/ Autobox2	Robust-Trend Theta-sm Theta	Robust-Trend ARARMA RBF		RBF Theta	RBF/ Theta/ Robust-Trend Comb S-H-D

Table 9
Methods which give the best results: quarterly data

Accuracy measure	Micro (204)	Industry (83)	Macro (336)	Finance (76)	Demographic (57)	Total (756)
Symmetric MAPE	Theta Comb S-H-D ForecastX	Comb S-H-D RBF ForecastX PP-autocast	Theta Comb S-H-D	Theta PP-autocast ForecastPro	Theta/ SmartFcs Dampen	Theta Comb S-H-D Dampen PP-autocast
Average RANKING	Theta Holt Comb S-H-D	Comb S-H-D PP-autocast ForecastX	Theta Comb S-H-D Dampen	Theta ARARMA Comb S-H-D	Theta/ Dampen ARARMA	Theta Comb S-H-D
Median APE	ForecastX Comb S-H-D Holt	ForecastX Comb S-H-D Theta Robust-Trend PP-autocast	Theta RBF Flores/Pearce1	Theta Winter SmartFcs	ARARMA Robust-Trend	Robust-Trend Theta Comb S-H-D ForecastX/ Dampen PP-autocast
Median RAE	Holt Theta Comb S-H-D/ Robust-Trend	Comb S-H-D/ Theta/ Robust-Trend Holt	Theta/ Comb S-H-D	Theta/ Winter	Theta ARARMA Comb S-H-D	Theta Comb S-H-D Robust-Trend

Table 10
Methods which give the best results: monthly data

Accuracy Measure	Micro (474)	Industry (334)	Macro (312)	Finance (145)	Demographic (111)	Other (52)	Total (1428)
Symmetric MAPE	Theta ForecastPro	ForecastPro ForecastX B–J automatic	ARARMA RBF	AAM1 AAM2	ForecastX SmartFcs Single ForecastPro	Comb S-H-D B–J automatic AAM1	Theta ForecastPro
Average RANKING	Theta ForecastPro	ForecastPro ForecastX Theta B–J automatic Comb S-H-D	Robust-Trend Holt Winter ARARMA/ AAM1	AAM1 AAM2	Robust-Trend	Theta AAM1/ AAM2 ARARMA/ Comb S-H-D	Theta ForecastPro Comb S-H-D
Median APE	Theta ForecastPro	ForecastPro B–J automatic ForecastX Theta	Robust-Trend Holt AAM1	AAM1/ AAM2 Autobox3 Autobox1	Robust-Trend ARARMA/ RBF	ARARMA AAM2	ForecastPro Theta
Median RAE	Theta Theta-sm ForecastPro/ Automat ANN		AAM1/ Robust-Trend Holt ARARMA	AAM1/ AAM2	Robust-Trend ARARMA	ARARMA AAM2 AAM1 Theta	

Table 11
Methods which give the best results: other data

Accuracy measure	Micro	Industry	Macro	Finance (29)	Demographic	Other (141)	Total (174)
Symmetric MAPE						Theta Autobox2 Comb S-H-D/ Robust-Trend ARARMA	ARARMA Theta/ Autobox2
Average RANKING				PP-autocast Dampen		ForecastX/ Autobox2 Robust-Trend Theta	ForecastX/ Autobox2 Theta ForecastPro/ Robust-Trend
Median APE				Automat ANN		ForecastX Autobox2	ForecastX/ Autobox2 Theta/ ForecastPro/ Robust-Trend
Median RAE							

Table 12
Methods which give the best results: symmetric MAPE — monthly data

Average step horizons	Types of time series data						
	Micro (474)	Industry (334)	Macro (312)	Finance (145)	Demographic (111)	Other (52)	Total (1428)
Short 1–3	SmartFcs Theta ForecastPro Automat ANN	ForecastPro ForecastX Dampen Comb S-H-D Theta	Most of the methods	Autobox2/ Automat ANN ForecastX	Most of the methods	Most of the methods	Theta ForecastPro SmartFcs Automat ANN ForecastX
Medium 4–12	Theta ForecastPro	ForecastPro ForecastX	Most of the methods	AAM1/ AAM2	Most of the methods	Comb S-H-D B–J automatic	ForecastPro Theta ForecastX
Long 13–18	Theta ForecastPro	Theta ForecastX/RBF ForecastPro Dampen	Robust-Trend RBF ARARMA AAM1	AAM1/ AAM2	Single Naïve2/ SmartFcs ForecastX/ Dampen ForecastPro	AAM1 ARARMA RBF/ Comb S-H-D	Theta ForecastPro RBF
Overall 1–18	Theta	ForecastPro ForecastX	ARARMA RBF	AAM1/ AAM2	ForecastX SmartFcs Single ForecastPro	Comb S-H-D B–J automatic AAM1	Theta ForecastPro

Table 13
Average symmetric MAPE: yearly data

Method	Forecasting horizon						Average		# obs
	1	2	3	4	5	6	1 to 4	1 to 6	
Naïve2	8.5	13.2	17.8	19.9	23	24.9	14.85	17.88	645
Single	8.5	13.3	17.6	19.8	22.8	24.8	14.82	17.82	645
Holt	8.3	13.7	19	22	25.2	27.3	15.77	19.27	645
Dampen	8	12.4	17	19.3	22.3	24	14.19	17.18	645
Winter	8.3	13.7	19	22	25.2	27.3	15.77	19.27	645
Comb S-H-D	7.9	12.4	16.9	19.3	22.2	23.7	14.11	17.07	645
B–J automatic	8.6	13	17.5	20	22.8	24.5	14.78	17.73	645
Autobox1	10.1	15.2	20.8	24.1	28.1	31.2	17.57	21.59	645
Autobox2	8	12.2	16.2	18.2	21.2	23.3	13.65	16.52	645
Autobox3	10.7	15.1	20	22.5	25.7	28.1	17.09	20.36	645
Robust-Trend	7.6	11.8	16.6	19	22.1	23.5	13.75	16.78	645
ARARMA	9	13.4	17.9	20.4	23.8	25.7	15.17	18.36	645
Automat ANN	9.2	13.2	17.5	20.3	23.2	25.4	15.04	18.13	645
Flores/Pearce1	8.4	12.5	16.9	19.1	22.2	24.2	14.22	17.21	645
Flores/Pearce2	10.3	13.6	17.6	19.7	21.9	23.9	15.31	17.84	645
PP-autocast	8	12.3	16.9	19.1	22.1	23.9	14.08	17.05	645
ForecastPro	8.3	12.2	16.8	19.3	22.2	24.1	14.15	17.14	645
SmartFcs	9.5	13	17.5	19.9	22.1	24.1	14.95	17.68	645
Theta-sm	8	12.6	17.5	20.2	23.4	25.4	14.6	17.87	645
Theta	8	12.2	16.7	19.2	21.7	23.6	14.02	16.9	645
RBF	8.2	12.1	16.4	18.3	20.8	22.7	13.75	16.42	645
ForecastX	8.6	12.4	16.1	18.2	21	22.7	13.8	16.48	645

Table 14
Average symmetric MAPE: quarterly data

Method	Forecasting horizon							Average			# obs
	1	2	3	4	5	6	8	1 to 4	1 to 6	1 to 8	
Naïve2	5.4	7.4	8.1	9.2	10.4	12.4	13.7	7.55	8.82	9.95	756
Single	5.3	7.2	7.8	9.2	10.2	12	13.4	7.38	8.63	9.72	756
Holt	5	6.9	8.3	10.4	11.5	13.1	15.6	7.67	9.21	10.67	756
Dampen	5.1	6.8	7.7	9.1	9.7	11.3	12.8	7.18	8.29	9.33	756
Winter	5	7.1	8.3	10.2	11.4	13.2	15.3	7.65	9.21	10.61	756
Comb S-H-D	5	6.7	7.5	8.9	9.7	11.2	12.8	7.03	8.16	9.22	756
B–J automatic	5.5	7.4	8.4	9.9	10.9	12.5	14.2	7.79	9.1	10.26	756
Autobox1	5.4	7.3	8.7	10.4	11.6	13.7	15.7	7.95	9.52	10.96	756
Autobox2	5.7	7.5	8.1	9.6	10.4	12.1	13.4	7.73	8.89	9.9	756
Autobox3	5.5	7.5	8.8	10.7	11.8	13.4	15.4	8.1	9.6	10.93	756
Robust-Trend	5.7	7.7	8.2	8.9	10.5	12.2	12.7	7.63	8.86	9.79	756
ARARMA	5.7	7.7	8.6	9.8	10.6	12.2	13.5	7.96	9.09	10.12	756
Automat ANN	5.5	7.6	8.3	9.8	10.9	12.5	14.1	7.8	9.1	10.2	756
Flores/Pearce1	5.3	7	8	9.7	10.6	12.2	13.8	7.48	8.78	9.95	756
Flores/Pearce2	6.7	8.5	9	10	10.8	12.2	13.5	8.57	9.54	10.43	756
PP-autocast	4.8	6.6	7.8	9.3	9.9	11.3	13	7.12	8.28	9.36	756
ForecastPro	4.9	6.8	7.9	9.6	10.5	11.9	13.9	7.28	8.57	9.77	756
SmartFcs	5.9	7.7	8.6	10	10.7	12.2	13.5	8.02	9.16	10.15	756
Theta-sm	7.7	8.9	9.1	9.7	10.2	11.3	12.1	8.86	9.49	10.07	756
Theta	5	6.7	7.4	8.8	9.4	10.9	12	7	8.04	8.96	756
RBF	5.7	7.4	8.3	9.3	9.9	11.4	12.6	7.69	8.67	9.57	756
ForecastX	4.8	6.7	7.7	9.2	10	11.6	13.6	7.12	8.35	9.54	756
AAM1	5.5	7.3	8.4	9.7	10.9	12.5	13.8	7.71	9.05	10.16	756
AAM2	5.5	7.3	8.4	9.9	11.1	12.7	14	7.75	9.13	10.26	756

Table 15
Average symmetric MAPE: monthly data

Method	Forecasting horizon										Average of forecasting horizons						# obs
	1	2	3	4	5	6	8	12	15	18	1 to 4	1 to 6	1 to 8	1 to 12	1 to 15	1 to 18	
Naïve2	15	13.5	15.7	17	14.9	14.7	15.6	16	19.3	20.7	15.3	15.13	15.29	15.57	16.18	16.91	1428
Single	13	12.1	14	15.1	13.5	13.1	13.8	14.5	18.3	19.4	13.53	13.44	13.6	13.83	14.51	15.32	1428
Holt	12.2	11.6	13.4	14.6	13.6	13.3	13.7	14.8	18.8	20.2	12.95	13.11	13.33	13.77	14.51	15.36	1428
Dampen	11.9	11.4	13	14.2	12.9	12.6	13	13.9	17.5	18.9	12.63	12.67	12.85	13.1	13.77	14.59	1428
Winter	12.5	11.7	13.7	14.7	13.6	13.4	14.1	14.6	18.9	20.2	13.17	13.28	13.52	13.88	14.62	15.44	1428
Comb S-H-D	12.3	11.5	13.2	14.3	12.9	12.5	13	13.6	17.3	18.3	12.83	12.79	12.92	13.11	13.75	14.48	1428
B–J automatic	12.3	11.7	12.8	14.3	12.7	12.6	13	14.1	17.8	19.3	12.78	12.74	12.89	13.21	13.96	14.81	1428
Autobox1	13	12.2	13	14.8	14.1	13.4	14.3	15.4	19.1	20.4	13.27	13.42	13.71	14.1	14.93	15.83	1428
Autobox2	13.1	12.1	13.5	15.3	13.3	13.8	13.9	15.2	18.2	19.9	13.51	13.52	13.76	14.16	14.86	15.69	1428
Autobox3	12.3	12.3	13	14.4	14.6	14.2	14.8	16.1	19.2	21.2	12.99	13.47	13.89	14.43	15.2	16.18	1428
Robust-Trend	15.3	13.8	15.5	17	15.3	15.6	17.4	17.5	22.2	24.3	15.39	15.42	15.89	16.58	17.47	18.4	1428
ARARMA	13.1	12.4	13.4	14.9	13.7	14.2	15	15.2	18.5	20.3	13.42	13.59	14	14.41	15.08	15.84	1428
Automat ANN	11.6	11.6	12	14.1	12.2	13.9	13.8	14.6	17.3	19.6	12.31	12.55	12.92	13.42	14.13	14.93	1428
Flores/Pearce1	12.4	12.3	14.2	16.1	14.6	14	14.6	14.4	19.1	20.8	13.74	13.93	14.22	14.29	15.02	15.96	1428
Flores/Pearce2	12.6	12.1	13.7	14.7	13.2	12.9	13.4	14.4	18.2	19.9	13.26	13.21	13.33	13.53	14.31	15.17	1428
PP-autocast	12.7	11.7	13.3	14.3	13.2	13.4	14	14.3	17.7	19.6	13.02	13.11	13.37	13.72	14.36	15.15	1428
ForecastPro	11.5	10.7	11.7	12.9	11.8	12.3	12.6	13.2	16.4	18.3	11.72	11.82	12.06	12.46	13.09	13.86	1428
SmartFcs	11.6	11.2	12.2	13.6	13.1	13.7	13.5	14.9	18	19.4	12.16	12.58	12.9	13.51	14.22	15.03	1428
Theta-sm	12.6	12.9	13.2	13.7	13.4	13.3	13.7	14	16.2	18.3	13.1	13.2	13.44	13.65	14.09	14.66	1428
Theta	11.2	10.7	11.8	12.4	12.2	12.4	12.7	13.2	16.2	18.2	11.54	11.8	12.13	12.5	13.11	13.85	1428
RBF	13.7	12.3	13.7	14.3	12.3	12.8	13.5	14.1	17.3	17.8	13.49	13.18	13.4	13.67	14.21	14.77	1428
ForecastX	11.6	11.2	12.6	14	12.4	12.2	12.8	13.9	17.8	18.7	12.32	12.31	12.46	12.83	13.6	14.45	1428
AAM1	12	12.3	12.7	14.1	14	14	14.3	14.9	18	20.4	12.8	13.2	13.63	14.05	14.78	15.69	1428
AAM2	12.3	12.4	12.9	14.4	14.3	14.2	14.5	15.1	18.4	20.7	13.03	13.45	13.87	14.25	15.01	15.93	1428

Table 16
Average symmetric MAPE: other data

Method	Forecasting horizon							Average			# obs
	1	2	3	4	5	6	8	1 to 4	1 to 6	1 to 8	
Naïve2	2.2	3.6	5.4	6.3	7.8	7.6	9.2	4.38	5.49	6.3	174
Single	2.1	3.6	5.4	6.3	7.8	7.6	9.2	4.36	5.48	6.29	174
Holt	1.9	2.9	3.9	4.7	5.8	5.6	7.2	3.32	4.13	4.81	174
Dampen	1.8	2.7	3.9	4.7	5.8	5.4	6.6	3.28	4.06	4.61	174
Winter	1.9	2.9	3.9	4.7	5.8	5.6	7.2	3.32	4.13	4.81	174
Comb S-H-D	1.8	2.8	4.1	4.7	5.8	5.3	6.2	3.36	4.09	4.56	174
B–J automatic	1.8	3	4.5	4.9	6.1	6.1	7.5	3.52	4.38	5.06	174
Autobox1	2.4	3.3	4.4	4.9	5.8	5.4	6.9	3.76	4.38	4.93	174
Autobox2	1.6	2.9	4	4.3	5.3	5.1	6.4	3.19	3.86	4.41	174
Autobox3	1.9	3.2	4.1	4.4	5.5	5.5	7	3.39	4.09	4.71	174
Robust-Trend	1.9	2.8	3.9	4.7	5.7	5.4	6.4	3.32	4.07	4.58	174
ARARMA	1.7	2.7	4	4.4	5.5	5.1	6	3.17	3.87	4.38	174
Automat ANN	1.7	2.9	4	4.5	5.7	5.7	7.4	3.26	4.07	4.8	174
Flores/Pearce1	2.1	3.2	4.3	5.2	6.2	5.8	7.3	3.71	4.47	5.09	174
Flores/Pearce2	2.3	2.9	4.3	5.1	6.2	5.7	6.5	3.67	4.43	4.89	174
PP-autocast	1.8	2.7	4	4.7	5.8	5.4	6.6	3.29	4.07	4.62	174
ForecastPro	1.9	3	4	4.4	5.4	5.4	6.7	3.31	4	4.6	174
SmartFcs	2.5	3.3	4.3	4.7	5.8	5.5	6.7	3.68	4.33	4.86	174
Theta-sm	2.3	3.2	4.3	4.8	6	5.6	6.9	3.66	4.37	4.93	174
Theta	1.8	2.7	3.8	4.5	5.6	5.2	6.1	3.2	3.93	4.41	174
RBF	2.7	3.8	5.2	5.8	6.9	6.3	7.3	4.38	5.12	5.6	174
ForecastX	2.1	3.1	4.1	4.4	5.6	5.4	6.5	3.42	4.1	4.64	174

Table 17
Methods which give the best results: symmetric MAPE

Time interval between observations	Types of time series data						
	Micro (828)	Industry (519)	Macro (731)	Finance (308)	Demographic (413)	Other (204)	Total (3003)
Yearly (645)	Robust-Trend Flores/Pearce2 SmartFcs Autobox2	Theta Comb S-H-D Autobox2	Robust-Trend ARARMA	Autobox2 Single Naïve2	ForecastX RBF		RBF ForecastX Autobox2 Theta Robust-Trend
Quarterly (756)	Theta Comb S-H-D ForecastX	Comb S-H-D RBF ForecastX PP-autocast	Theta Comb S-H-D	Theta PP-autocast ForecastPro	Theta/ SmartFcs Dampen		Theta Comb S-H-D Dampen PP-autocast
Monthly (1428)	Theta ForecastPro	ForecastPro ForecastX	ARARMA RBF	AAM1/ AAM2	ForecastX SmartFcs Single ForecastPro	Comb S-H-D B–J automatic AAM1	Theta ForecastPro
Other (174)				Dampen/ PP-autocast Automat ANN ForecastPro		Theta Autobox2 Robust-Trend Comb S-H-D	ARARMA Theta/ Autobox2
Total (3003)	Theta ForecastPro	ForecastPro/ ForecastX Theta	RBF/ ARARMA Theta/ Robust-Trend	AAM1 AAM2	ForecastX		Theta ForecastPro

Table 18
Methods which give the best results: average RANKING

Time interval between observations	Types of time series data						
	Micro (828)	Industry (519)	Macro (731)	Finance (308)	Demographic (413)	Other (204)	Total (3003)
Yearly (645)	Robust-Trend Autobox2 Theta	Theta Robust-Trend Comb S-H-D RBF	Robust-Trend ARARMA	Single Naïve2/ Autobox2 ForecastPro/ ForecastX	ForecastX PP-autocast ForecastPro		RBF/ ForecastX Theta/ Robust-Trend Autobox2
Quarterly (756)	Theta Holt Comb S-H-D	Comb S-H-D PP-autocast ForecastX	Theta Comb S-H-D Dampen	Theta ARARMA Comb S-H-D	Theta/ Dampen ARARMA		Theta Comb S-H-D
Monthly (1428)	Theta ForecastPro	ForecastPro ForecastX Theta Comb S-H-D	Robust-Trend Holt Winter ARARMA AAM1	AAM1/ AAM2	Robust-Trend	Theta Comb S-H-D ARARMA AAM1/ AAM2	Theta ForecastPro Comb S-H-D
Other (174)				PP-autocast Dampen		ForecastX/ Autobox2 Robust-Trend Theta	Autobox2 ForecastX Theta

Table 19
Methods which give the best results: median APE

Time interval between successive observations	Types of time series data						
	Micro (828)	Industry (519)	Macro (731)	Finance (308)	Demographic (413)	Other (204)	Total (3003)
Yearly (645)	Robust-Trend SmartFcs	Robust-Trend	Robust-Trend ForecastPro	Single Naïve2 Autobox2	ForecastX ForecastPro RBF Theta Autobox2		RBF Flores/Pearce1 PP-autocast
Quarterly (756)	ForecastX Comb S-H-D Holt	ForecastX Comb S-H-D Theta Robust-Trend PP-autocast	Theta RBF Flores/Pearce1	Theta Winter SmartFcs	ARARMA Robust-Trend		Robust-Trend Theta Comb S-H-D ForecastX
Monthly (1428)	Theta ForecastPro	ForecastPro B–J automatic ForecastX Theta	Robust-Trend Holt AAM1	AAM1/ AAM2 Autobox3 Autobox1	Robust-Trend ARARMA/ RBF	ARARMA AAM2	ForecastPro Theta Holt Comb S-H-D
Other (174)				Automat ANN		ForecastX Autobox2	ForecastX Autobox2 Theta ForecastPro

Table 20
Methods which give the best results: median RAE

Time interval between successive observations	Types of time series data						
	Micro (828)	Industry (519)	Macro (731)	Finance (308)	Demographic (413)	Other (204)	Total (3003)
Yearly (645)	Robust-Trend SmartFcs/ Theta/ Autobox2	Robust-Trend Theta-sm Theta	Robust-Trend ARARMA RBF		RBF Theta		
Quarterly (756)	Holt Theta Comb S-H-D/ Robust-Trend	Comb S-H-D/ Theta/ Robust-Trend Holt	Theta/ Comb S-H-D	Theta/ Winter	Theta ARARMA Comb S-H-D		
Monthly (1428)	Theta Theta-sm ForecastPro/ Automat ANN		AAM1/ Robust-Trend Holt ARARMA	AAM1/ AAM2	Robust-Trend ARARMA	ARARMA AAM2 AAM1 Theta	
Other (174)							

Table 21
Methods which give the best results: seasonal/non-seasonal data

	Types of time series data						
	Micro (828)	Industry (519)	Macro (731)	Finance (308)	Demographic (413)	Other (204)	Total (3003)
Seasonal (862)	ForecastPro Theta Dampen Comb S-H-D SmartFcs ForecastX			AAM1/ AAM2 ForecastPro ForecastX			ForecastPro Theta/ ForecastX/ Dampen Comb S-H-D
Non-Seasonal (2141)	Theta			AAM1/ AAM2			Theta ForecastPro ForecastX/ Comb S-H-D

Appendix C

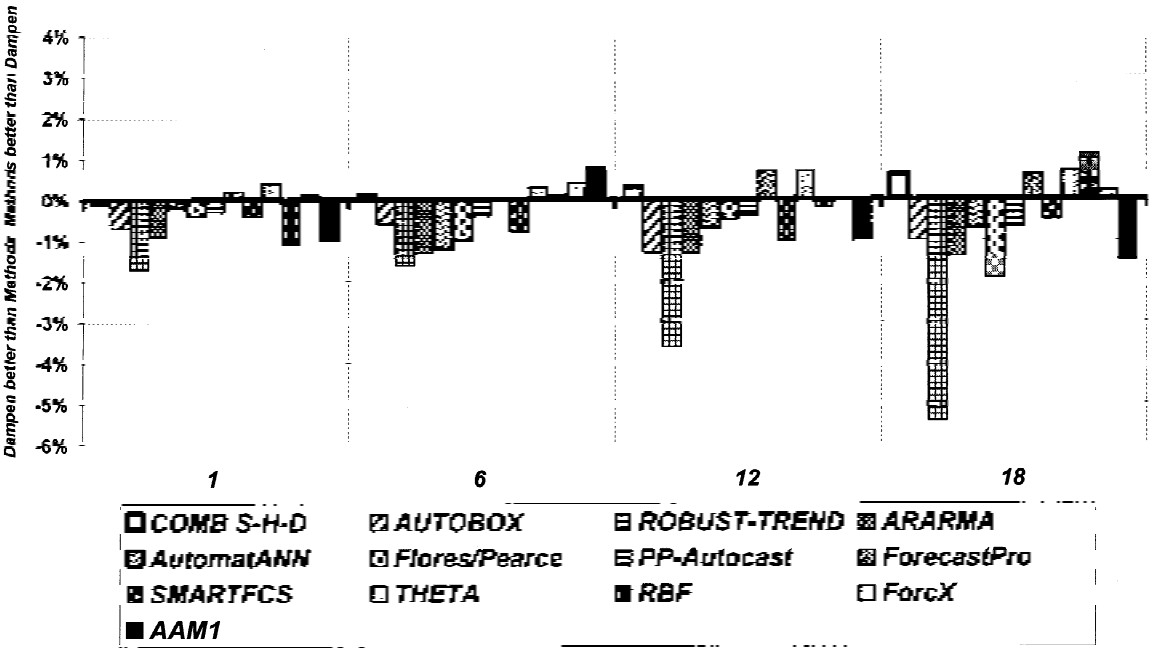

Fig. 1. Average symmetric MAPE (Dampen-Method): all data.

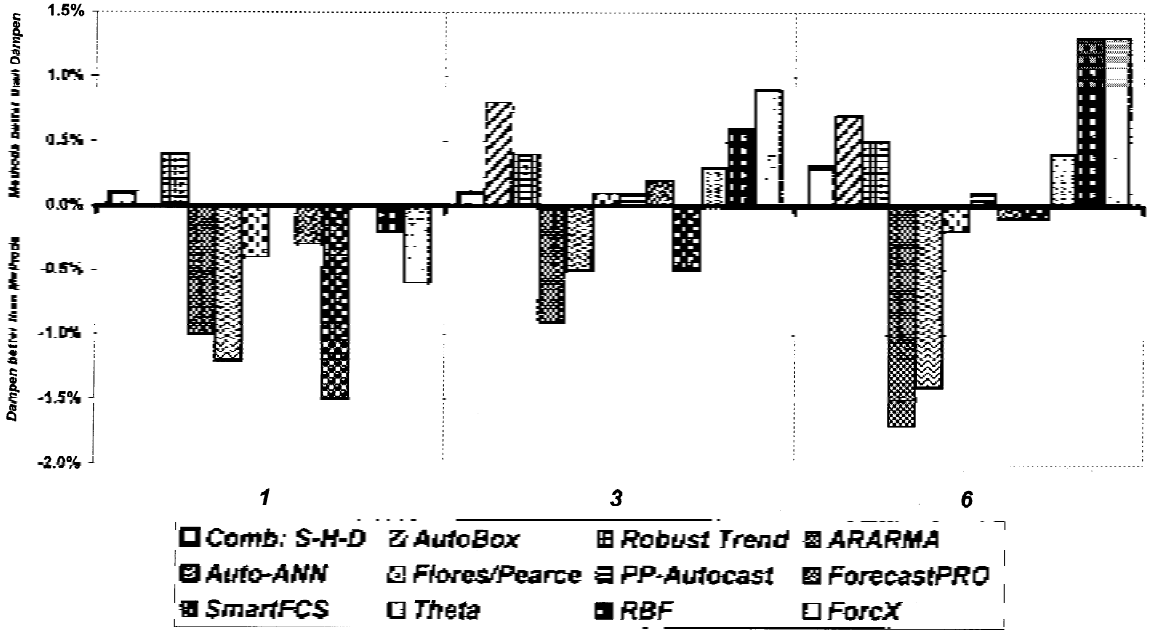

Fig. 2. Average symmetric MAPE (Dampen-Method): yearly data.

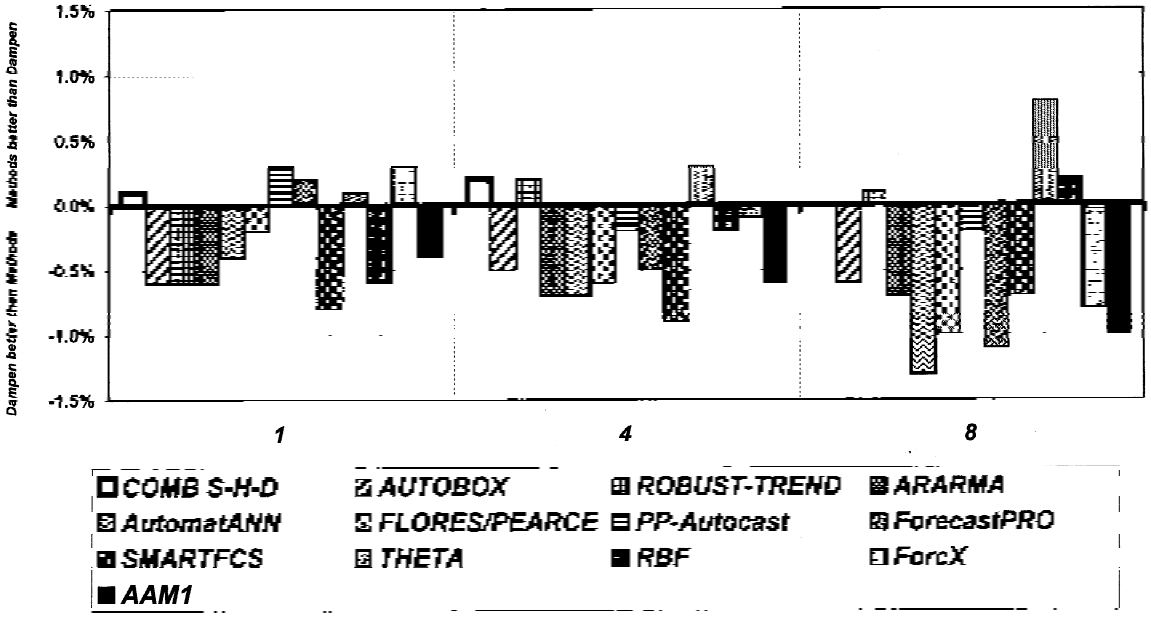

Fig. 3. Average symmetric MAPE (Dampen-Method): quarterly data.

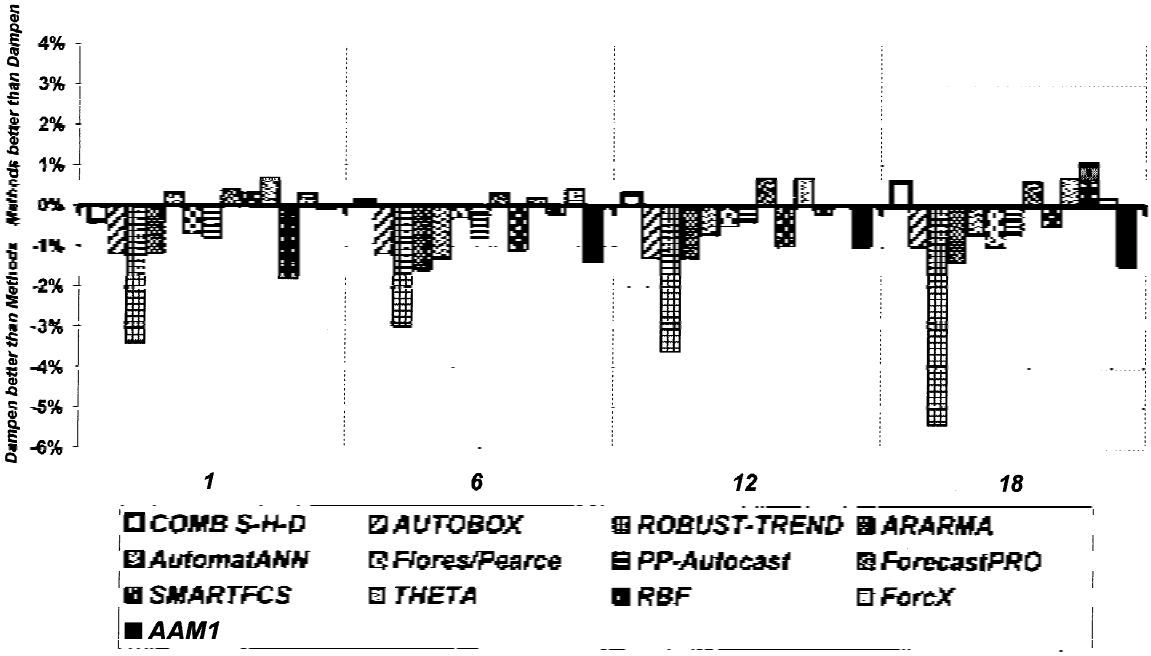

Fig. 4. Average symmetric MAPE (Dampen-Method): monthly data.

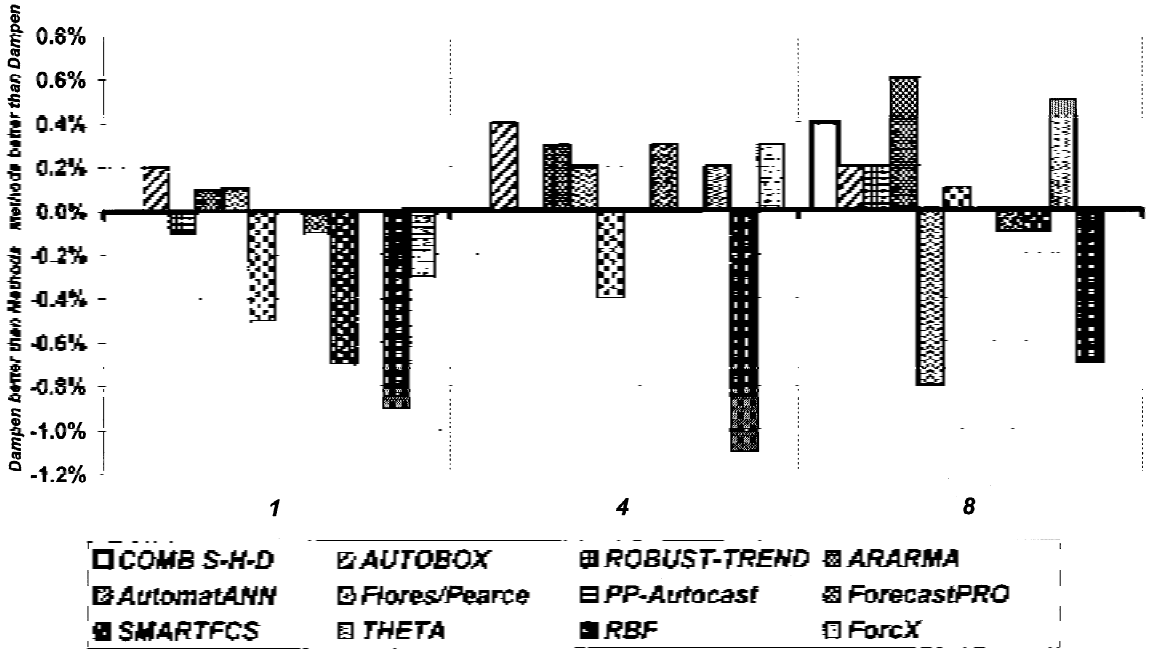

Fig. 5. Average symmetric MAPE (Dampen-Method): other data.

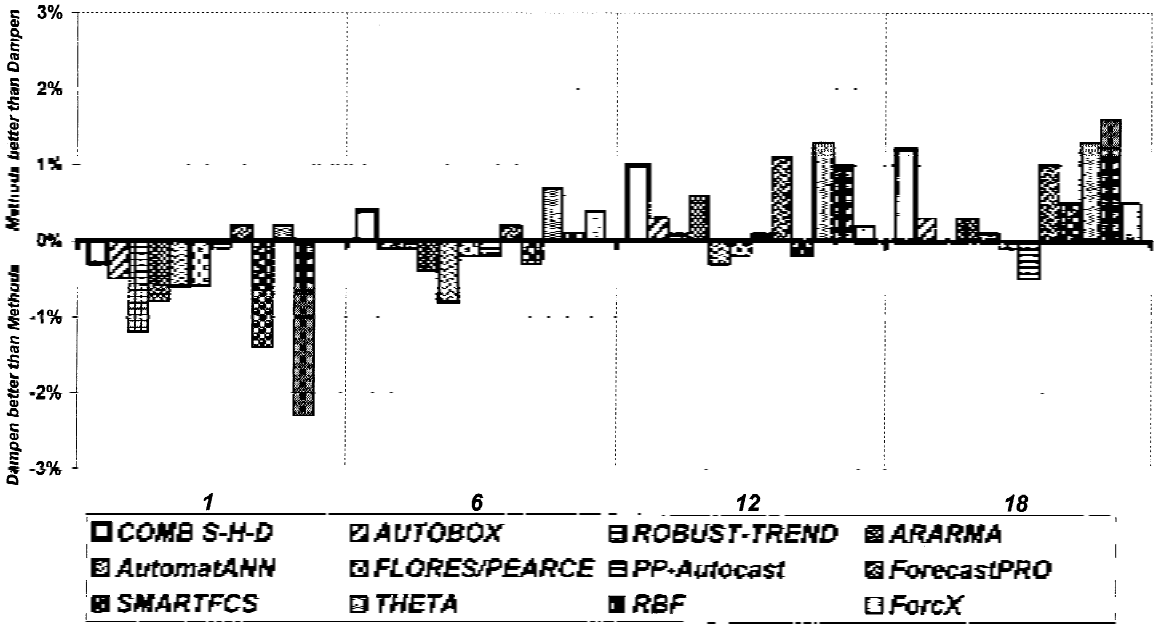

Fig. 6. Relative ranking (Dampen-Method): all data.

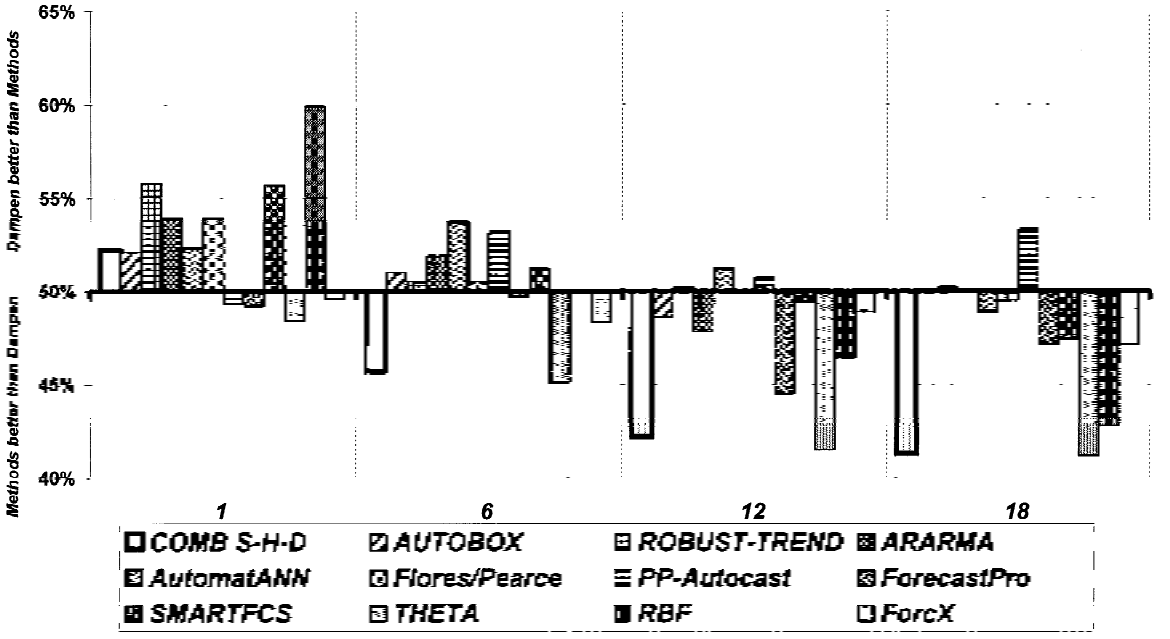

Fig. 7. Percentage of time Dampen is better than other methods: all data.

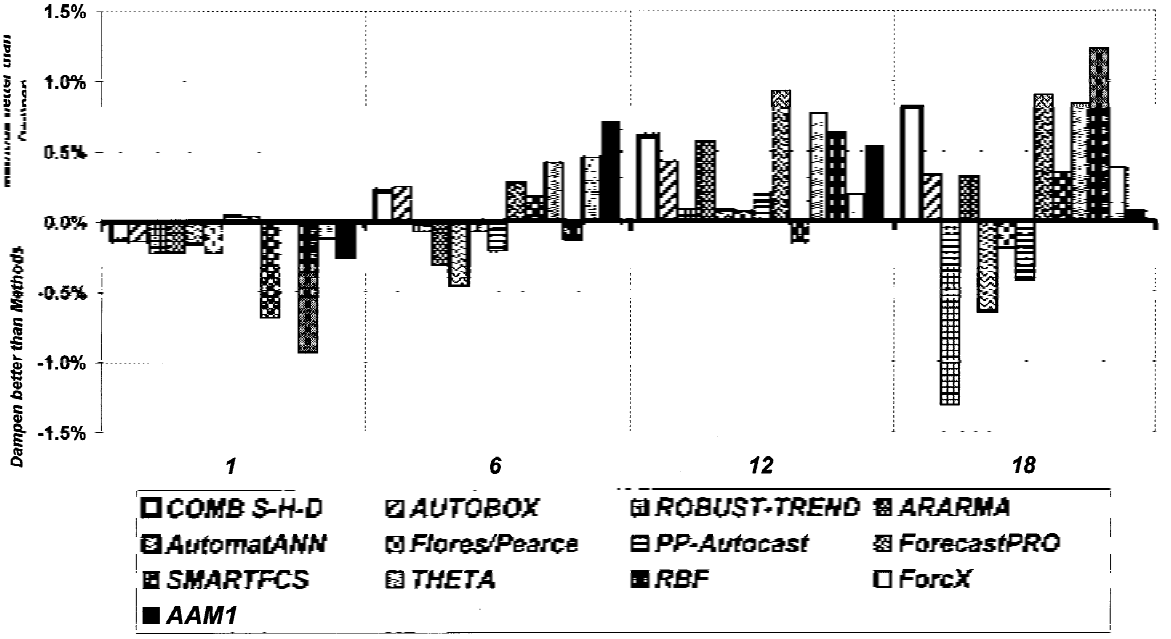

Fig. 8. Median APE (Dampen-Method): all data.

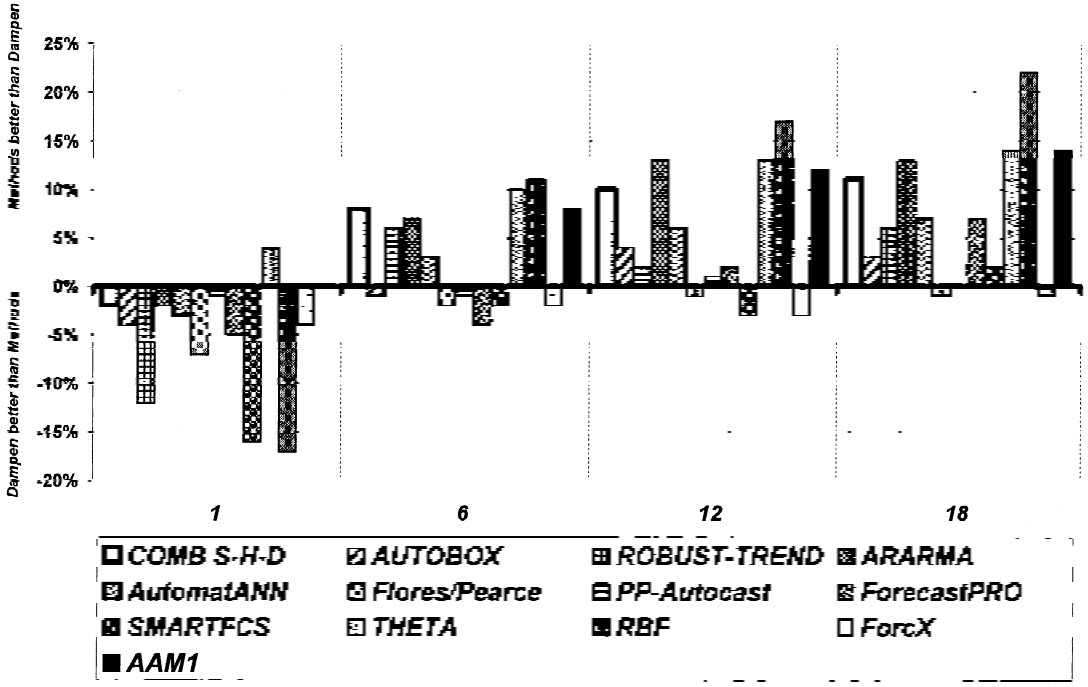

Fig. 9. Median RAE (Dampen-Method): all data.

References

Armstrong, J.S., & Collopy, F. (1992). Error measures for generalizing about forecasting methods: empirical comparisons (with discussion). *International Journal of Forecasting, 8*, 69–80, 99–111.

Armstrong, J. S., & Collopy, F. (1993). Causal forces: structuring knowledge for time series extrapolation. *Journal of Forecasting 12*, 103–115.

Chatfield, C. (1995). Positive or negative? *International Journal of Forecasting 11*, 501–502.

Clemen, R. (1989). Combining forecasts: a review and annotated bibliography with discussion. *International Journal of Forecasting 5*, 559–608.

Fildes, R. (1992). The evaluation of extrapolative forecasting methods (with discussion). *International Journal of Forecasting 8*, 81–111.

Fildes, R., & Makridakis, S. (1995). The impact of empirical accuracy studies on time series analysis and forecasting. *International Statistical Review 63*, 289–308.

Fildes, R., Hibon, M., Makridakis, S., & Meade, N. (1998). Generalising about univariate forecasting methods: further empirical evidence. *International Journal of Forecasting 14*, 339–358.

Gardner, E. S., & McKenzie, E. (1985). Forecasting trends in time series. *Management Science 31*, 1237–1246.

Geurts, M. D., & Kelly, J. P. (1986). Forecasting demand for special services. *International Journal of Forecasting 2*, 261–272.

Grambsch, P., & Stahel, W. A. (1990). Forecasting demand for special telephone services. *International Journal of Forecasting 6*, 53–64.

Hill, G., & Fildes, R. (1984). The accuracy of extrapolation methods: an automatic Box–Jenkins package SIFT. *Journal of Forecasting 3*, 319–323.

Koehler, A. B., & Murphree, E. S. (1988). A comparison of results from state space forecasting with forecasts from the Makridakis competition. *International Journal ofForecasting 4*, 45–55.

Kuhn, T. S. (1962). *The structure of scientific revolution*, University of Chicago Press, Chicago.

Lusk, E. J., & Neves, J. S. (1984). A comparative ARIMA analysis of the 111 series of the Makridakis competition. *Journal of Forecasting 3*, 329–332.

Makridakis, S., & Hibon, M. (1979). Accuracy of forecasting: an empirical investigation (with discussion). *Journal of the Royal Statistical Society A 142*, 97–145.

Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, E., &

Winkler, R. (1982). The accuracy of extrapolation (time series) methods: results of a forecasting competition. *Journal of Forecasting 1*, 111–153.

Makridakis, S., Chatfield, C., Hibon, M., Lawrence, M., Mills, T., Ord, K., & Simmons, L. F. (1993). The M-2 Competition: a real-time judgmentally based forecasting study. *International Journal of Forecasting 9*, 5–23.

Newbold, P. (1983). The competition to end all competitions. *Journal of Forecasting 2*, 276–279.

Newbold, P., & Granger, C. W. J. (1974). Experience with forecasting univariate time series and the combination of forecasts (with discussion). *Journal of Royal Statistical Society A 137*, 131–165.

Reid, D.J. (1969). *A comparative study of time series prediction techniques on economic data*. PhD Thesis, Department of Mathematics, University of Nottingham.

Reid, D. J. (1975). A review of short term projection techniques. In: Gordon, H. D. (Ed.), *Practical aspects of forecasting*, Operational Research Society, London, pp. 8–25.

Simmons, L. F. (1986). M-Competition — A closer look at Naïve2 and median APE: a note. *International Journal of Forecasting 4*, 457–460.

Biographies: Spyros MAKRIDAKIS is a Research Professor at INSEAD, Fontainebleau, France. His Ph.D is from New York University. He has consulted worldwide in the area of forecasting and has held teaching positions with several European and American institutions; as a research fellow at IIM in Berlin, an ICAME fellow at Stanford University and a Visiting Scholar at MIT and Harvard. Professor Makridakis is the author of approximately 100 articles and papers in various publications and has also authored or coauthored 18 books including *Forecasting, Planning and Strategy for the 21st Century* (Free Press, 1990).

Michèle HIBON is a Senior Research Fellow at INSEAD. She graduated from the University of Paris, holds a degree in Science and a diploma in advanced studies in physics. For the last several years she has been working on various studies dealing with forecasting accuracy of time series methods (particularly M-, M2- and M3-Competitions) and more recently in the area of Group Decision Support Systems.