Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

International Journal of Forecasting 27 (2011) 635–660

international journal of forecasting

www.elsevier.com/locate/ijforecast

Advances in forecasting with neural networks? Empirical evidence from the NN3 competition on time series prediction

Sven F. Crone[a,*], Michèle Hibon[b], Konstantinos Nikolopoulos[c]

[a] Lancaster Centre for Forecasting, Lancaster University Management School, Lancaster, UK
[b] Decision Sciences, INSEAD, Fontainebleau, France
[c] Decision Sciences Research Centre, Manchester Business School, Manchester, UK

Available online 12 May 2011

Abstract

This paper reports the results of the NN3 competition, which is a replication of the M3 competition with an extension of the competition towards neural network (NN) and computational intelligence (CI) methods, in order to assess what progress has been made in the 10 years since the M3 competition. Two masked subsets of the M3 monthly industry data, containing 111 and 11 empirical time series respectively, were chosen, controlling for multiple data conditions of time series length (short/long), data patterns (seasonal/non-seasonal) and forecasting horizons (short/medium/long). The relative forecasting accuracy was assessed using the metrics from the M3, together with later extensions of scaled measures, and non-parametric statistical tests. The NN3 competition attracted 59 submissions from NN, CI and statistics, making it the largest CI competition on time series data. Its main findings include: (a) only one NN outperformed the damped trend using the sMAPE, but more contenders outperformed the AutomatANN of the M3; (b) ensembles of CI approaches performed very well, better than combinations of statistical methods; (c) a novel, complex statistical method outperformed all statistical and CI benchmarks; and (d) for the most difficult subset of short and seasonal series, a methodology employing echo state neural networks outperformed all others. The NN3 results highlight the ability of NN to handle complex data, including short and seasonal time series, beyond prior expectations, and thus identify multiple avenues for future research.

Keywords: Time series forecasting; Empirical evaluation; NN3 competition; Artificial neural networks; Computational intelligence

1. Introduction

Back in 1993, Chatfield wondered, "*Neural networks: forecasting breakthrough or passing fad*?"; and the question still remains largely unanswered today. On the one hand, if we consider only the number of publications relating to artificial neural networks (NN), the answer would seem to indicate that they were a breakthrough: motivated by their theoretical properties of non-parametric, data driven universal approximation of any linear or nonlinear function, the

[*] Corresponding author. Tel.: +44 1524 5 92991.
E-mail address: s.crone@lancaster.ac.uk (S.F. Crone).

0169-2070/$ - see front matter © 2011 International Institute of Forecasters. Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.ijforecast.2011.04.001

last two decades have witnessed over 5000 publications in academic journals and conference proceedings on forecasting with NNs across a wide range of disciplines (Crone & Preßmar, 2006). In two recent surveys on forecasting publications, Fildes et al. note that while the last 25 years have seen rapid developments in forecasting across a broad range of topics, computer intensive methods such as NNs have contributed the largest number of publications of any area in operational research (Fildes, Nikolopoulos, Crone, & Syntetos, 2008), and they form one of the top four areas of growth in forecasting journals (Fildes, 2006). Their growth in prominence appears to be easy to justify: the majority of publications indicate the competitive or even superior performance of NNs, from publications on single benchmark time series such as the popular airline passenger dataset (Faraway & Chatfield, 1998; Kolarik & Rudorfer, 1994; Tang & Fishwick, 1993), to representative subsets of established benchmarks from previous forecasting competitions (Foster, Collopy, & Ungar, 1992; Hill, O'Connor, & Remus, 1996; Sharda & Patil, 1992). In one of the few evaluative reviews, Adya and Collopy (1998) found eleven studies that met the criteria for a valid and reliable empirical evaluation, and NNs were more accurate in 8 of these (73%). However, their evaluation of the experimental design and the implementation of the NNs also raised concerns regarding the validity and reliability of the results in 37 of 48 studies (77%). For novel algorithms which are not evaluated following a rigorous experimental design, the results from an ex post evaluation (where the test data are known to the authors) may not be sufficiently reliable, but require an objective, unbiased ex ante evaluation in order to determine their true empirical accuracy under varying data conditions.

If, on the other hand, we considered only the empirical post-sample accuracies demonstrated by NNs, a different answer to Chatfield's question (1993) arises. In contrast to their optimistic publications, NNs have failed to provide objective evidence of their ex ante forecasting accuracy in large scale empirical evaluations in the form of forecasting competitions. The most renowned empirical investigation conducted to date — the M3 competition (Makridakis & Hibon, 2000) — indicated a comparatively poor performance from a single NN contestant. Thus, the performances of NNs for batch forecasting fell far short of their presumed potential.

At the same time, forecasting competitions conducted in computer science and machine learning (e.g., the Santa Fe competition, see Weigend & Gershenfeld, 1994, or the EUNITE competition, see Suykens & Vandewalle, 1998a) attracted a large number of NN and CI algorithms. Although these demonstrated the superior performance of NNs, the algorithms were often not evaluated against statistical methods, using only a single time series (and time origin), or a small set of heterogeneous time series. These setups ignored the evidence within the forecasting field as to how to design valid and reliable empirical evaluations (see for example Fildes, Hibon, Makridakis, & Meade, 1998), severely limiting the validity and reliability of their findings. As a consequence of the poor experimental designs, the forecasting community largely ignored these findings.

The discrepancy between NNs' superior theoretical capabilities, together with their promising accuracies in various publications on known datasets and some real world applications, and the lack of empirical accuracy in large scale ex ante evaluations, has raised serious concerns in the forecasting domain as to their adequacy for forecasting. As a consequence, Chatfield (as quoted by Armstrong, 2006) suspects a positive bias in NN publications, due to a "file-drawer problem" of negative results, leading Armstrong (2006) to conclude that too much research effort is being devoted to this method. However, to date, this skepticism is founded only on the performance of a single contestant in one large scale evaluation of automatic forecasting.

In order to explore the persistent gap between the theoretical capabilities and empirical accuracy of NNs, we conducted a forecasting competition in order to provide valid and reliable empirical evidence of the accuracy of NNs, as well as to evaluate and disseminate potential progress in modelling NNs and to determine the conditions under which different algorithms perform well. Our motivation for conducting yet another competition follows the same arguments as those of the original M-competition (see Makridakis et al., 1982): a full decade has passed since the start of the M3 competition, a decade which has seen the development of extended NN paradigms (e.g., recurrent Echo State NN, see Jaeger & Haas, 2004), theoretical advances in methodologies for specifying NNs (see, e.g., Crone & Kourentzes, 2010; Liao &

Fildes, 2005; Qi & Zhang, 2001), and the appearance of a range of novel computer intensive algorithms in CI for forecasting (including new algorithms, e.g. Support Vector Regression, see Smola & Schölkopf, 2004; and methodologies, e.g. method combination by boosting, see Freund & Schapire, 1997). In addition, there has been substantial progress in information technology, which may facilitate the application of existing algorithms and novel extensions to large scale forecasting competitions that were not feasible before due to the limited computational resources available. As new alternatives now exist, the choices made with regard to selecting and using appropriate forecasting methods need to be revisited.

To evaluate the progress in NNs, and to allow a comparison with the original M3 contestants over time, we utilised a subset of 111 monthly industry time series taken from the M3 dataset for which the original predictions were available. The dataset contains a balanced sample of seasonal and non-seasonal, short and long time series, in order to evaluate the conditions under which a given algorithm performs well. The competition was open to all NN and CI methods. To reduce potential biases, we also allowed novel statistical methodologies (e.g., that of Billah, King, Snyder, & Koehler, 2006) and newer software releases (e.g., the latest versions of Autobox, ForecastPro or R), which had been developed but had not yet been assessed in competitions, to participate as benchmarks. NN3 attracted 59 submissions, making it the largest competition in CI and forecasting to date. The results were evaluated using multiple error metrics, including the original symmetric mean absolute percent error (sMAPE), the mean absolute scaled error (MASE), as proposed by Hyndman and Koehler (2006), and two non-parametric tests proposed by Koning, Franses, Hibon, and Stekler (2005) in a follow-up analysis of the M3-data: analysis of the mean (ANOM) and multiple comparisons to the best method (MCB). In short, we attempted to consider all recommendations on how to conduct a valid and reliable empirical evaluation, while balancing the effort and resources of the contestants, in order to attract a more representative sample of algorithms. As the competition followed the original design of the M3, it was launched under the name *NN3 competition*. This paper summarises its findings, discusses the results of the experiments, and suggests directions for future research.

The rest of the paper is structured as follows: Section 2 discusses previous forecasting competitions in both forecasting and CI, their relevance for deriving empirical evidence, guidelines for their setup, and discrepancies in the findings of the forecasting and CI competitions, in order to justify us in conducting another one. As CI competitions have not followed consistent designs, the best practices derived from the experimental design of forecasting competitions are explored in more detail in order to disseminate them to a interdisciplinary readership. Sections 3 and 4 describe the setup and the results of the empirical evaluation, taking these best practices into consideration. Section 5 provides a brief discussion of the most important findings, followed by the conclusions and implications for future research.

2. Evidence from competitions in forecasting and computational intelligence

2.1. *Competitions in forecasting*

In the absence of the universal (theoretical or empirical) dominance of a single ‘best method’, competitions are an established means of providing objective evidence on the empirical ex ante accuracy of forecasting methods, and of guiding rational choices between algorithms and methodologies for a given set of data conditions. Forecasting competitions have received a substantial amount of attention and have initiated stimulating discussions within the academic forecasting community, opening up new areas of academic research (e.g. model selection and evaluation) and leading to improved practices on valid and reliable competitions and experimental designs (Ord, Hibon, & Makridakis, 2000). An overview and discussion of the impact of empirical evaluations is given by Fildes and Makridakis (1995) and Fildes and Ord (2002). In contrast, time series prediction competitions which have been conducted outside the forecasting community, including those in computer science, machine learning, engineering and CI, have pursued different experimental designs that have ignored best practices on how to conduct competitions, thus limiting both their validity and their reliability. In order to assess the empirical evidence provided in each field to date, and to contrast the lack of dissemination of algorithms, applications and best practices across the two domains, we briefly summarize the existing competitions in forecasting and CI, and provide an overview in Table 1.

In forecasting research, a series of competitions have been conducted that have received a substantial amount of attention. Drawing upon the criticisms of earlier competitions on time series data (Groff, 1973; Makridakis & Hibon, 1979; Newbold & Granger, 1974; Reid, unpublished, 1972), Makridakis et al. conducted a series of enlarged forecasting competitions where experts could submit the predictions of their preferred algorithms: the M-Competition (Makridakis et al., 1982) used two datasets of 1001 and 111 time series respectively, and taking into account suggestions made at a meeting of the Royal Statistical Society. A smaller subset of the data was offered in order to allow the participation of algorithms which required time and cost intensive manual tuning by experts (e.g., the ARIMA models required more than one hour per time series). The subsequent M2-competition (Makridakis et al., 1993) focussed on non-automatic, real time judgmental forecasts of 23 time series, and hence is less relevant for our quantitative competition design. None of the earlier competitions attracted any submissions of NNs or CI methods, as these algorithms did not emerge until the late 1980s; e.g., in the case of NNs, through the (re-)discovery of the back-propagation algorithm (Rumelhart, Hinton, & Williams, 1994). The competitions also did not receive submissions using some other CI methods such as CART (Breiman, 1984), fuzzy logic (Zadeh, 1965) or evolutionary computation (Fogel, 1994), although these algorithms had already been developed.

In 1998, the popular M3-Competition evaluated the accuracies of 24 algorithms on 3003 univariate empirical time series of historical data (Makridakis & Hibon, 2000), the largest dataset ever to be used in such a competition. The time series were selected from various domains of micro- and macroeconomic, industrial, financial and demographic activity, and from different time frequencies (yearly, quarterly and monthly data), in order to cover a wide range of time series structures and different data conditions. All of the methods were implemented by academic experts and commercial software providers, leading to the most representative ex ante evaluation of forecasting methods to date.

Across all time series, two methods generally outperformed all other methods: the software expert system ForecastPro using automatic model selection and the parameterisation of exponential smoothing (ES) and ARIMA models (Goodrich, 2000), and Theta, a decomposition approach combining exponential smoothing and regressing around a damped trend line (Assimakopoulos & Nikolopoulos, 2000). Further statistical analysis by Koning et al. (2005) has provided statistical evidence for a group of four methods with higher accuracies, which also includes rule based forecasting (Adya, Armstrong, Collopy, & Kennedy, 2000) and Comb S-H-D, an equally weighted combination of the Brown's single, Holt's linear trend and Gardner's damped trend ES methods (computed by Hibon) in the top performers.

Despite the initial interest shown by various CI researchers, only one group ended up submitting results to the competition using a NN methodology (Balkin & Ord, 2000). However, their fully automated methodology AutomatANN performed only moderately well relative to the majority of the twenty statistical approaches, and was not ranked among the top performers (Makridakis & Hibon, 2000, Table 15). The limited participation of CI approaches has been attributed to the high computational costs of building and parameterising these methods for each time series, but also to the absence of methodologies that would allow automation beyond manual tuning by a human expert. However, the poor performance was neither expected nor explained sufficiently.

The conclusions which had been drawn from previous M-competitions (Makridakis et al., 1982, 1993) were confirmed in the M3-competition (see Makridakis & Hibon, 2000), verified through follow-up studies (see, e.g., Fildes, 1992), and extended to provide additional insights (Fildes et al., 1998):

(H1) the characteristics of the data series are an important factor in determining the relative performances of different methods;
(H2) the accuracy of a method depends upon the length of the forecasting horizon;
(H3) the relative performance rankings of methods vary with the accuracy measure;
(H4) the sampling variability of the performance measures renders comparisons which are based on

single time series unreliable: comparisons based on multiple time origins are recommended;

(H5) combinations of predictions tend to be quite accurate, and often outperform the individual methods; and

(H6) sophisticated methods do not necessarily provide more accurate forecasts than simpler ones.

Consequently, valid competitions have developed a rigorous design, including the use of a representative number of time series (and, where possible, a rolling origin design), the use of multiple robust error metrics, a comparison with established (statistical) benchmark algorithms, and the analysis of the data conditions under which a method performs well (Tashman, 2000), in order to obtain valid and reliable results. Conclusion H6 seems to be particularly relevant, as NNs and other computer intensive methods — just like sophisticated statistical algorithms such as ARIMA before them — do not guarantee an enhanced forecasting performance as a result of their proven capabilities or theoretical features; instead, they require an objective evaluation against simpler benchmarks. No competitions on a similar scale have been conducted since the M3 (including the MH competition on transportation data of varying time frequencies, conducted in 2007 by Hibon, Young and Scaglione, and the tourism forecasting competition conducted by Athanasopoulos, Hyndman, Song, and Wu (2011). This leaves the M3 as the most recent large scale evaluation in the forecasting domain, and explains the impact and prominence of the disappointing results of NN in empirical forecasting, based upon the one entry of the only CI-contestant AutomatANN (Balkin & Ord, 2000), which are yet unchallenged.

Conversely, the findings of the M3 cannot be considered as representative of the wide class of NN paradigms, which have evolved over time. Despite a myriad of published NN methodologies, only one methodology was evaluated, limiting the representativeness of the results for the class of NNs (which encompasses a variety of feed-forward and recurrent architectures), and for CI as a whole. Also, the M3 attracted no interest from the computer science, engineering and machine learning communities, where CI and other artificial intelligence approaches had been advanced for years, introducing a sample selection bias of algorithms (an omission possibly caused by disseminating the call for papers only through the International Institute of Forecasters (IIF), i.e. the IJF and the International Symposium on Forecasting (ISF), which may also have limited the dissemination of the results across disciplines). Consequently, the poor performance of a single NN approach in the M3 cannot be considered as being representative of the whole class of algorithms. Furthermore, almost a decade has passed since M3, meaning that the results may no longer reflect the capabilities of today's NNs. There is evidence of substantial theoretical progress in NNs, in forecasting both single time series (see for example de Menezes & Nikolaev, 2006, Preminger & Franck, 2007 and Terasvirta, van Dijk, & Medeiros, 2005) and representative sets of empirical time series (see, e.g., Liao & Fildes, 2005, Zhang & Qi, 2005), where new methodologies for fully automated applications of NN are developed. These have not yet been evaluated in an objective empirical competition. Lastly, the computational power today is far superior to that which was available in 1997 when automated NNs were first run for the M3 competition, which may enable a much wider participation, given the expanded community which now applies computationally intensive methods regularly. Thus, the results of the M3 may no longer be representative. However, in the absence of more recent forecasting competitions, its critical findings with regard to NNs remain unchallenged.

2.2. Competitions in CI

Competitions for determining the predictive accuracy of algorithms have been equally popular outside the forecasting domain, and many have been more recent than the M3. Regular data mining competitions have been conducted, albeit they have focussed on classification tasks, including the annual competitions at the KDD conference, which attracted over 1000 contestants in 2008, and the recently closed Netflix competition (www.netflixprice.com) for predicting movie choices, which attracted over 44,000 submissions (by awarding US$1 million in prize-money). As in forecasting, competitions for classification using CI generally follow a rigorous experimental design, adhere to established best practices for valid and reliable results, and often address sophisticated modelling

Table 1
Competition designs in forecasting and computational intelligence.

Competition name	Data properties				# of algorithms		Conditions evaluated			
	# of series	# of observ.	Data type		Statistics	NN& CI	Multiple metrics[k]	Multiple horizons[k]	Multiple types[k]	Time frequency[l]
			Univariate	Multivariate						
Empirical accuracy[a]	111	?	X		22	0	X	X	X	Y, Q, M
M1[b]	1001	15–150	X		24	0	X	X	X	Y, Q, M
M3[c]	3003	20–144	X		24	1	X	X	X	Y, Q, M
MH/Transport[d]	278	19–1502	X		3–10	1	X	X	X	Y, Q, M W, D, H
Santa Fe[e]	6	1000–300,000	X (2)	X (4)	0	14	–	–	–	Synthetic
KULeuven[f]	1	2000	X		0	17	–	–	–	Synthetic
2001 EUNITE[g]	1	35,040		X	1	24	–	–	–	30 m
ANNEXG[h]	1	1460		X	0	12	–	–	–	360 m
BI cup 2003[d]	1	365		X	0	10	–	–	–	D
CATS 2005[i]	1	4905	X		0	25	–	–	–	Synthetic
Predictive uncertainty[j]	4	380–21,000	X (1)	X (3)	0	20	–	–	–	Synthetic D, 3D
BI cup 2006[d]	1	1325	X (1)		0	?	–	–	–	15 m
NN3 (in this issue)	111	68–144	X		17	63	X	X	–	M

[a] Makridakis and Hibon (1979).
[b] Makridakis et al. (1982).
[c] Makridakis and Hibon (2000).
[d] Unpublished.
[e] Weigend and Gershenfeld (1994).
[f] Suykens and Vandewalle (1998a,b, 2000).
[g] Unpublished.
[h] Dawson et al. (2005).
[i] Lendasse et al. (2007).
[j] Cawley et al. (2007).
[k] X indicates the use of multiple error metrics, multiple forecasting horizons and data types, – indicates their absence.
[l] Y = yearly data; Q = quarterly data; M = monthly data; W = weekly data; D = daily data; H = hourly data; m = minutes.? indicates undisclosed information.

questions, e.g. the value of domain knowledge over agnostic prediction (Guyon, Saffari, Dror, & Cawley, 2008), or the extent to which the (in-)sample accuracy can be generalised to the out-of-sample performance accuracy (Cawley, Janacek, Haylock, & Dorling, 2007).

In contrast, only few competitions in the CI-domain have been dedicated to time series data, as shown in the exhaustive overview in Table 1, although some CI competitions on forecasting may have eluded our attention as these have often been on a small scale. A discussion of all CI competitions and their contributions is beyond the scope of this paper, but we will outline the most influential, in order to exemplify differences in the experimental design. The time series prediction and analysis competition organised by Weigend and Gershenfeld (1994) under the auspices of the Santa Fe Institute was the first dedicated CI competition to evaluate the forecasting capabilities of NNs using a variety of nonlinear time series datasets. The datasets were highly heterogeneous and required both univariate and multivariate time series prediction, including a physics experiment recording the oscillations and structural breaks of a NH5-Laser, tick-by-tick currency exchange rates, astrophysical data of light fluctuations from a white star, physiological data from a patient with sleep apnoea, and music from Bach's last (unfinished) Fuge. Given the heterogeneity of the data conditions, most of the participants predicted only one of the time series from a single origin (instead of — at least — all of the series), and no statistical benchmarks were evaluated. As a consequence, the comparative work undertaken in the competition remains rudimentary and does not provide sufficient evidence to enable us to draw conclusions as to the accuracy of any of the nonlinear algorithms (Makridakis, 1994). The lack of rigor seems particular disappointing, considering that the authors were aware of the design and findings of the M-competitions, and given that the late Clive Granger served on the competition's advisory board.

The largest CI competition on time series to date was organised by Suykens and Vandewalle in 2001 (unpublished) for the European Network on Intelligent Technologies for Smart Adaptive Systems (EUNITE, www.eunite.org —no longer online), which attracted 24 submissions from 16 contestants only a subset of the 56 that had registered to compete, similar to M3. It evaluated the accuracy of predicting a time series of the maximum electrical load using two years of half-hourly electricity load data, and additional explanatory variables of past temperatures and holidays (all provided by the Eastern Slovakian Electricity Corporation). Forecasts were made up to 31 days into the future from a single time origin. The best contestant used support vector regression (Chen, Chang, & Lin, 2004) to outperform the CI contestants and one 'statistical' contender using regression on decomposed time series components. Although all of the algorithms were published in a monograph (Sincák, Strackeljan, Kolcun, Novotný, & Szathmáry, 2002), it has received limited attention outside the electrical load literature.

Various smaller competitions have also been run at conferences on computational intelligence, including the Competition on Artificial Time Series (CATS) for imputing missing values in synthetic data (Lendasse, Oja, Simula, & Verleysen, 2007), held at the 2004 IEEE International Joint Conference on Neural Networks (IJCNN); the Predictive Uncertainty Competition on environmental data at the 2006 IJCNN (Cawley et al., 2007); the (unpublished) 2003 and 2006 Business Intelligence Cups on predicting time series of sugar and retail sales, organised by Richard Weber at the IEEE Latin-American Summer School on Computational Intelligence (EVIC); the 2001 ANNEXG competition on river stage forecasting (Dawson et al., 2005), held at the 2002 BHS National Hydrology Symposium (the 2005 re-run attracted no competitors); and the KULeuven competition on synthetic data by Suykens and Vandewalle (1998a,b) held at the International Workshop on Advanced Black-Box Techniques for Nonlinear Modeling in 1998 (for the winner, see McNames, Suykens, & Vandewalle, 1999). Table 1 provides a structured summary of prior time series competitions, in both forecasting and CI, and points out differences in experimental designs between the two domains, to assess their contributions.

2.3. *Differences in competition design*

Few similarities emerge, but one stands out: each domain favours and evaluates almost exclusively its own preferred family of algorithms: forecasting competitions evaluate only statistical methods (and expert systems which configure these), with the exception of the single NN contender in the M3 and one in the unpublished MH-competition, while CI-competitions, in turn, have failed to evaluate statistical algorithms.

More noticeably, differences and discrepancies in the design of all CI-competitions become evident, which seriously impair their contribution. As a concession to the resources required to run a competition, both the forecasting and CI competitions each employed only one hold-out set, and hence a single time series origin. However, while all competitions in the forecasting domain have used representative sample sizes of hundreds or even thousands of time series in order to derive robust results, CI competitions have mostly evaluated accuracies on a single time series only. The few competitions which evaluated multiple time series, such as the Santa Fe and predictive uncertainty competitions, did so for distinct domains, with only one series per category, again limiting any generalisation of their findings. Had the same algorithm been used across multiple similar series, datasets or competitions, it would have allowed somewhat more reliable and insightful results to be obtained. Instead, the same authors applied different methodologies for each dataset, even within a given competition, thus leading to distinctly different models and preventing any comparisons. Also, none of the CI competitions compare the results with established benchmark methods, whether naïve methods (i.e. a random walk), simple statistical benchmarks which are used in the application domain (e.g., ES methods), or non-statistical methods in the same family of algorithms (e.g., a simple NN with default parameters to compete against a more sophisticated architecture). We therefore conclude that the recommendations on the design of empirical evaluations developed in forecasting have been ignored by the CI community. Makridakis and Hibon's (2000) original criticism holds: just like theoretical statisticians before them, NN researchers have concentrated their efforts on building more sophisticated models, with no regard to either the assessment of their accuracy or objective empirical verifications, successfully ignoring the strong empirical evidence of the M-competitions and the ground rules they have laid out on how to assess forecasting competitions. This substantially limits the validity and reliability of the evidence from the CI competitions to date, and therefore they cannot challenge the authority of the earlier M3-competition, where the class of NN failed to show any improvement in accuracy.

With the competitions in both domains being limited in their coverage of algorithms, the results of the M3 competition not being representative of CI, and more recent CI competitions being unreliable, the gap between the theoretical capabilities and empirical accuracies of NNs remains unexplored. In order to evaluate the potential progress in the development of NN and CI approaches, a new competition seemed the most suitable way to provide valid and reliable empirical evidence on their accuracies and the conditions under which different algorithms perform well, and to disseminate information about the potential progress in modelling NNs. For the sake of consistency, it seemed natural to use the original setup of the M3-competition and a homogeneous subset of the M3 data in the form of a replication, which will be discussed in detail in the next section.

In reviewing Table 1, we also note an important omission in the data conditions the two domains have explored. Previous forecasting competitions have focussed exclusively on low time series frequencies of yearly, quarterly or at most monthly data in a univariate context. Although this is an adequate reflection of the theme of operational forecasting set out by Makridakis' series of M-competitions, it does not allow us to generalise these findings to previously unexamined data conditions. In particular, it provides no insights for the quite different data conditions of high-frequency datasets of weekly, daily, hourly or shorter time intervals on which NNs have generally been evaluated in CI research. It appears that Armstrong's (2006) criticism of NNs is based not only on the limited evidence of a single contestant in the M3, but in itself remains is limited due to a substantial omission of the empirical data conditions, for which — following his arguments — no evidence exists. As the omitted data properties are representative of those on which NNs are regularly employed in practice (e.g., electrical load forecasting, Hippert, Pedreira, & Souza, 2001), this yields a possible explanation to the simultaneous skepticism and euphoria on NNs in forecasting and CI respectively. Hopefully, it will provide the motivation for the gap to be closed by conducting competitions for novel data conditions, including those involving high frequency data.

3. Design and organisation of the NN3 competition

3.1. Objectives

Following the rationale provided above, we sought to explore the current forecasting performances of NN and CI methods. The M-competitions focussed explicitly on a particular set of data conditions, which Makridakis proposed in the context of forecasting for operations. To assess our progress relative to M3, we will keep this tradition and restrict our competition to the operations context of monthly industry data, although there are other data conditions which might show quite different results for NN.

The NN3 competition was designed both to (partly) replicate and to extend the M3 competition. As a replication, the NN3 will utilise the data, experimental setup and original forecast submissions from M3, and evaluate the working hypotheses of earlier competitions (see Section 2) to challenge or confirm the prior findings. In addition, the NN3 represents an extension towards more methods/researchers from the areas of NN and CI, in order to assess advances in accuracy and to overcome the limitations of M3's representativeness. Previous forecasting competitions have led to an established 'research methodology' for a systematic, valid and reliable design of future competitions, which we have attempted to follow here. We will briefly review these design choices, the datasets and conditions, accuracy metrics, methods and benchmarks, and the process by which NN3 was conducted, in order to allow the verification of the experimental design and the dissemination of this knowledge to the CI community, and to facilitate replication studies.

3.2. Datasets, working hypotheses and data conditions

The M3 dataset yielded substantial insights, but proved challenging for CI methods: the sample of 3003 time series was large, given the computational resources available in the 1990s, and the heterogeneity of the time series frequencies and data domains required multiple candidate methodologies (and human intervention at many stages), which limited automation and may have prevented many experts from participating with computationally intensive NN methods. In order to attract a representative number of contestants and algorithms to NN3, we sought to limit both the number of time series used and the heterogeneity of the data conditions (and thus the resulting insights), yet not enough that we could not derive reliable results. A set of 111 time series was selected randomly from the M3 monthly industry time series, representative of the M-competition's original focus of forecasting for operations (and in line with the size of the reduced M3 dataset for manual tuning). Time series of a single frequency were chosen in order to limit the competition's complexity to a single methodology for monthly data. We also hoped that using a sample would further mask the origin of the NN3 competition data, and thus prevent biases in the results through prior knowledge.

Four working hypotheses (WH) were considered in the evaluation. To determine the degree of automation or manual tuning required, and to address prevailing concerns on the computational demands of predicting a large number of time series with NNs, we allowed participants to chose between two (disguised) datasets of different sizes. The contestants were asked to predict either a reduced dataset of 11 time series or the complete set of 111 series (which included the reduced set) as accurately as possible. As a fully automated methodology could be applied to large datasets just as easily as smaller sets, more submissions for the reduced dataset would indicate the limitations of the automation through need for manual of extremely computational intensive approaches, and indicate the need for further research into methodologies (WH1).

A second working hypothesis (WH2) seeks to assess the relative accuracies of NNs and statistical approaches for longer forecasting horizons, where statistical algorithms have outperformed NNs in past studies (Hill et al., 1996). Each contestant is required to produce multiple forecasts y_{t+h} of $h = (1, \ldots, 18)$ steps into the future, which are later analysed for short (1–3 months), medium (3–12 months) and long (13–18 months) forecasting horizons in order to assess the differences in the results (see also H2).

Two further working hypotheses address the data conditions under which different methods perform well (see also H1). First, following the widespread

Table 2
NN3 datasets with data conditions of time series length and seasonality.

	Complete dataset				
	Short	Long	Reduced dataset		
			Normal	Difficult	Sum
Non-seasonal	25 (NS)	25 (NL)	4 (NN)	3 (ND)	57
Seasonal	25 (SS)	25 (SL)	4 (SN)	–	54
Sum	50	50	8	3	111

belief that NNs are data hungry and require long time series (WH3), balanced stratified samples were taken by time series length n, resulting in 50 long ($n > 100$) and 50 short ($n < 50$) time series. Second, in order to evaluate recent publications which conclude that NNs cannot forecast seasonal time series (WH4) see, e.g., (Curry, 2007; Nelson, Hill, Remus, & O'Connor, 1999; Zhang & Qi, 2005), stratified samples were taken to reflect the time series patterns of 50 seasonal and 50 non-seasonal time series (as per the original M3 classification). Series with structural breaks in the test set were manually identified and excluded.

The sample sizes were guided by the objective of deriving (statistically) valid and reliable results for each data condition from as small a dataset as possible, which created a lower bound of 25 time series in each cell (i.e., short-seasonal, long-seasonal, short-non-seasonal, and long-non-seasonal), resulting in 100 series as a core for the complete set. The reduced dataset contained 11 time series which we classified as difficult to forecast, of which four were seasonal and the remaining seven were non-seasonal (including outliers and structural breaks), and which served to ascertain whether or not non-automated methodologies are capable of forecasting across different data conditions. Table 2 summarises the time series conditions of both datasets.

The conditions within the reduced dataset were not intended to be statistically explored, due to the limited number of time series (3 ND + 4 NN + 4 SN), which could not provide reliable results. Nonetheless, the findings from the reduced dataset would be at least as valid as those from previous CI competitions using only a single time series to provide new insights.

3.3. Evaluation and error metrics

In order to evaluate the performances of the NN3 submissions and ensure consistency with the results of the M3-competition, we employed three of the metrics used in the M3 competition, namely sMAPE, MdRAE and AR (Makridakis & Hibon, 2000):

$$\text{sMAPE}_s = \frac{1}{n}\sum_{t=1}^{n}\frac{|X_t - F_t|}{(X_t - F_t)/2} \cdot 100, \qquad (1)$$

$$\text{MdRAE}_s = \text{median}(|r_t|), \quad \text{with } r_t = \frac{X_t - F_t}{X_t - F_t^*}, \qquad (2)$$

with X_t being the actual value in period t, F_t the forecast made for period t, n the number of observations forecasted by the respective forecasting method, and F_t^* the forecast made by the reference method Naïve2 (a random walk applied to seasonally adjusted data) for a given forecasting horizon h. AR is estimated by taking the ranks of sAPE for each forecasting horizon, over all series s. The errors are then averaged across all s series of a set, $s = (1, \ldots, S)$.

We also estimated two non-parametric tests proposed by Koning et al. (2005) in a follow-up analysis: ANOM and MCB, both using AR as the criterion. Finally, for the sake of consistency with the current literature, we have calculated the MASE, as proposed by Hyndman and Koehler (2006). In order to ensure a consistent computation of errors, we collaborated with Hibon, one of the original investigators of the M3 competition, and she computed all metrics as in the original competition.

It was announced beforehand that the average sMAPE would be the metric used to determine the "winner", in order to allow those CI methods which are capable of using alternative loss functions (i.e. non-squared costs of errors) to align their approaches with the final criterion (see, e.g., the discussion by Zellner, 1986, following the M3). Despite the shortcomings of the sMAPE (Goodwin & Lawton, 1999), it was chosen both because it served as the primary criterion in the M3 competition and to make the NN3 results accessible to practitioners, whose predominant error metric is the MAPE. As the NN3 time series contained no zero, negative or small actual values X_t, and all submitted forecasts F_t were positive, we anticipate only limited biases.

This permits us to use Armstrong's (1985) version of sMAPE(1), as in the M3 competition, for reasons of comparison (see Hyndman & Koehler, 2006, for a more robust version of the sMAPE).

3.4. Methods and benchmarks

The competition invited contributions from all areas of machine learning, data mining and CI, including all NN paradigms and architectures, support vector regression, fuzzy logic, evolutionary and genetic algorithms, and hybrid methods utilising any kind of CI. In an attempt not to bias the results towards novel NN-methods, we also allowed novel statistical methodologies and newer software releases to be evaluated as benchmarks, further extending the representativeness of the NN3.

We personally invited both experts in statistical forecasting methods and commercial software vendors, in order to ensure the participation of the latest releases of the methods which had performed well in the original M3-competition, but with limited success. We are grateful for submissions from Eric Stellwagen of Business Forecasting Systems, applying the latest version of the expert system ForecastPro (B03); from Dave Reilly of Autobox, applying the latest version of the expert system for ARIMA and transfer function modelling (B05); and from Tucker McElroy, who submitted predictions from the Census X12 method (B6).

In order to assess the progress in NN modelling since the M3, the NN3 submissions needed to be compared to the original M3 submission of AutomatANN (Balkin & Ord, 2000, B00). Given the identical experimental setup and data taken from M3, our collaboration with one of the original conductors of the M3 competition allowed us to retrieve the 111 original predictions submitted to M3 and compare them directly with those of the NN3 contestants. Further to AutomatANN, five statistical benchmarks used in the M3 were recalled, including the Naïve-1 method (B04), three variants of Brown's single ES (B14), Holt's linear trend ES (B15) and Gardner's damped trend ES (B16), and their combination to Comb S-H-D (B17). Predictions for Theta (B7) were recomputed by the organisers, using a setup identical to that of the M3 competition.

In addition, we computed various CI benchmarks to provide additional levels of comparison for the entries, including a naïve support vector regression (SVR) approach (Crone & Pietsch, 2007, B01) and a naïve multilayer perceptron (MLP) model (B02), both of which replicate novice model building mistakes as a lower bound of errors for CI-methods. A novel NN extension of the successful Theta method, named Theta-AI (B08) by Nikolopoulos and Bougioukos, which determined optimal nonlinear weights for the Theta-lines, was withdrawn in order not to bias the results, as it was based on the Theta method, which is known a priori to perform well on the NN3 data.

3.5. Process of organising the competition

The competition design and feasibility were pretested in a small scale trial competition (held at the 2005 ISF, San Antonio, USA) using two time series, which facilitated feedback from 9 contestants and external experts, including a panel of IIF judges for a grant to fund NN3. The NN3 competition was first announced at the ISF 2006 in Santander, Spain, and was open for eight months from October 2007 to May 2008. Each contestant was required to submit predictions and a full description of their methodology, both of which have been published on the competition website[1] in order to facilitate replication. Following submission, each methodology was classified, to distinguish between CI contenders which were eligible to "win" the competition (identified by consecutive IDs C01–C59, given in the order of entry) and submissions that would serve as benchmarks: CI benchmarks (B00–B02), statistical benchmarks including forecasting packages (B03–B08), novel statistical methods submitted as benchmarks (B09–B13), and the original ES variants of M3 (B14–B17). The contestants had the option to withhold their identity prior to disclosing the final results, in order to limit any negative publicity for software vendors and participants. Some contestants did request to withhold their identity, and therefore their results are included in the tables with only their original submission IDs, to ensure consistency with previously disclosed results.

In order to limit any sample selection biases in the participation through the timing, location and audience of the conferences where the competition was

[1] www.neural-forecasting-competition.com/NN3/.

promoted, multiple special sessions were advertised and conducted at conferences throughout 2007, and across the domains of forecasting, CI, electrical engineering, data mining and machine learning. These included the 2007 ISF'07, New York, USA; the 2007 IEEE IJCNN, Orlando, USA; and the 2007 International Conference in Data Mining (DMIN'07) in Las Vegas, USA. The call for papers was disseminated via various email-lists, websites, online communities and newsletters across disciplines.

4. Results of the NN3 competition

4.1. Results on the complete dataset

The competition attracted 46 contestants who used NN and CI methods and 17 benchmark methods, making it the largest empirical evaluation in the areas of NN, CI and forecasting to date.

Table 3 presents the names of the NN3 contestants, a consecutive ID (assigned during the competition), and a summary of the algorithm that provided forecasts of the 111 series of the complete dataset. A discussion of all of the submissions is not feasible here, so we will limit our discussion to the methods which have stood out in some or all of the data conditions we analysed. A detailed description of each of the methodologies, including the 24 contenders who only provided forecasts for the 11 series of the reduced dataset, is available on the NN3 competition website, www.neural-forecasting-competition.com, for a detailed review and analysis.

Table 4 shows the results on the complete dataset as average sMAPE, MdRAE, MASE and AR values across 111 time series and 18 forecasting horizons. The relative ranks by error measure are given both across all methods and for the CI contestants alone (NN C).

Has progress been made, both within CI and in comparison to statistical methods? All 46 contenders submitted predictions for the reduced set of 11 time series, but only 22 contenders predicted all 111 time series in the complete set. The fact that under half of the contestants (47%) are able to predict more than 11 series provides evidence that the need for manual tuning and human intervention still dominates most methodologies. This reflects our experience, in both academia and practice, and is supported by the lack of commercial CI software for automatic time series forecasting (see also working hypothesis WH1). Nonetheless, the ability of 22 contestants to predict a large number of time series using CI indicates unsurprising progress in the development of methodologies that facilitate automation and/or in increased computational resources.

With regard to accuracy, the top 10 algorithms indicate some progress in accuracy, but not quite enough to confirm a breakthrough for NNs in the view of Chatfield (1993). Unsurprisingly, the top contenders for the M3 monthly data are also ranked highly for this subset: Theta (B07), ForecastPro (B03), Autobox (B05) and the ES variants DES (B16), Comb S-H-D (B17), SES (B14) and HES (B15). However, some new innovators have also joined the best performers. These algorithms will be introduced briefly here, as they have not been published elsewhere (see also the NN3 competition website).

Had the competition not been tailored to CI, Wildi's new statistical benchmark method (B09) would have won the competition, across all error metrics and against the tough competition of the 'winners' of the monthly M3 data. The prototype methodology extends the traditional adaptive state space approach, discounts errors exponentially by their distance to the forecast origin, estimates multiple-step-ahead out-of-sample errors (instead of 1-step-ahead in-sample errors) using a winsorised squared error loss function, and employs forecast combinations by building h separate models for each forecasting horizon $h = (1, 2, \ldots, 18)$, with their hyperparameters optimised for each h, and combining the 18 predictions using the median. A monograph on the algorithm is under preparation.

More in line with the competition's theme, the method of Illies, Jäger, Kosuchinas, Rincon, Sakenas and Vaskevcius (C27) ranked 3rd across all methods and provided the best results of all CI contenders. The methodology employs echo state networks (ESN), a novel paradigm of recurrent NNs with sparse, random connections in a so-called 'reservoir' of hidden neurons arranged in multiple layers. The time series were categorised into 6 clusters by time series length, thus ignoring the different data domains and properties, and pooling time series in different clusters (despite the unrelated natures of most of the series, a fact which was not known to the contestants). Each time series was first decomposed into its time

Table 3
NN3 participant IDs, names and method descriptions for the complete dataset of 111 series.

Code	Classification	Name	Description
C03	Contender: NN/CI	Flores, Anaya, Ramirez, Morales	Automated linear modeling of time series with self adaptive genetic algorithms
C11	Contender: NN/CI	Perfilieva, Novak, Pavliska, Dvorak, Stepnicka	Combination of two techniques: fuzzy transform and perception-based logical deduction
C13	Contender: NN/CI	D'yakonov	Simple kNN-method for time series prediction
C15	Contender: NN/CI	Isa	Growing fuzzy inference neural network
C17	Contender: NN/CI	Chang	K-nearest-neighbor and support-vector regression
C20	Contender: NN/CI	Kurogi, Koyama, Tanaka, Sanuki	Using first-order difference of time series and bagging of competitive associative nets
C24	Contender: NN/CI	Abou-Nasr	Recurrent neural networks
C26	Contender: NN/CI	de Vos	Multi-resolution time series forecasting using wavelet decomposition
C27	Contender: NN/CI	Illies, Jäger, Kosuchinas, Rincon, Sakenas, Vaskevcius	Stepping forward through echoes of the past: forecasting with echo state networks
C28	Contender: NN/CI	Eruhimov, Martyanov, Tuv	Windowed wavelet decomposition and gradient boosted trees
C30	Contender: NN/CI	Pucheta, Patino, Kuchen	Neural network-based prediction using long and short term dependence in the learning process
C31	Contender: NN/CI	Theodosiou, Swamy	A hybrid approach: structural decomposition, generalised regression neural networks and the Theta model
C36	Contender: NN/CI	Sorjamaa, Lendasse	A non-linear approach (self-organized maps) combined with a linear one (empirical orthogonal functions)
C37	Contender: NN/CI	Duclos-Gosselin	Fully-recurrent neural network learned with M.A.P. (Bayesian), Levenberg and genetic algorithms
C38	Contender: NN/CI	Adeodato, Vasconcelos, Arnaud, Chunha, Monteiro	Multilayer perceptron networks
C44	Contender: NN/CI	Yan	Multiple-model fusion for robust time series forecasting
C46	Contender: NN/CI	Chen, Yao	Ensemble regression trees
C49	Contender: NN/CI	Schliebs, Platel, Kasabov	Quantum inspired feature selection and neural network models
C50	Contender: NN/CI	Kamel, Atiya, Gayar, El-Shishiny	A combined neural network/Gaussian process regression time series forecasting system
C51	Contender: NN/CI	Papadaki, Amaxopolous	Dynamic architecture for artificial neural networks
C57	Contender: NN/CI	Corzo, Hong	Global neural network ensembles with M5 prime model trees
C59	Contender: NN/CI	Beliakov & Troiano	Time series forecasting using Lipschitz optimal interpolation
B09	Contender: Statistics	Wildi	An adaptive robustified multi-step-ahead out-of-sample forecasting combination approach
B10	Contender: Statistics	Beadle	Composite forecasting strategy using seasonal schemata
B11	Contender: Statistics	Lewicke	Paracaster software by parabolic systems fitting equations consisting of trend + series of sinusoidal error terms
B12	Contender: Statistics	Hazarika	Decomposition to random sequence basis functions and a temperature-dependent SOFTMAX combiner
B13	Contender: Statistics	Njimi, Mélard	Automatic ARIMA modeling, using TSE-AX
B03	Benchmark: Statistics	ForecastPro	ForecastPro expert selection method, Version XE 5.0.2.6. (by Stellwagen)
B04	Benchmark: Statistics	Naïve	The naïve method without any seasonality adjustment
B05	Benchmark: Statistics	Autobox	Autobox expert system forecast, version 6.0 (June 2007) (by Reily)
B06	Benchmark: Statistics	Census—X12 ARIMA	Official census method (by McElroy)
B07	Benchmark: Statistics	Theta	Exponential smoothing with decomposition, version TIFIS CM3 1.0 (by Nikolopoulos)
B14	Benchmark: Statistics	Single ES	Original M3 benchmark for the M3 competition as programmed (by Hibon)

(*continued on next page*)

Table 3 (*continued*)

Code	Classification	Name	Description
B15	Benchmark: Statistics	Holt ES	Original M3 benchmark for the M3 competition as programmed (by Hibon)
B16	Benchmark: Statistics	Dampen ES	Original M3 benchmark for the M3 competition as programmed (by Hibon)
B17	Benchmark: Statistics	Comb S-H-D ES	Original M3 benchmark—equally weighted combination of single, Holt and damped trend exponential smoothing
B00	Benchmark: NN/CI	Automat NN	Original M3 submission for the M3 competition (by Balkin & Ord)
B01	Benchmark: NN/CI	Naïve SVR	A naïve support vector regression forecasting approach (by Crone & Pietsch)
B02	Benchmark: NN/CI	Naïve MLP	A naïve multiple linear perceptron (by Crone)
C103	Benchmark: NN/CI	Ensemble of Best 3 NN/CI	Equally weighted combination of C27, C03, C46 prepared post-competition (by Hibon)
C105	Benchmark: NN/CI	Ensemble of Best 5 NN/CI	Equally weighted combination of C27, C03, C46, C50, C13 prepared post-competition (by Hibon)

series components using X-12-ARIMA. Then, 500 ESNs with reservoir sizes of between 45 and 110 hidden neurons were trained on pooled clusters of time series for each time series component. Their predictions for each time series were first recombined across components, then combined in an ensemble of all 500 ESNs using the mean of the predictions. The approach successfully outperformed all of the statistical benchmarks except for Theta, the top-performer of the M3 monthly data, which constitutes a substantial achievement and considerable progress in CI model building.

Three other CI contenders also outperformed AutomatANN and climbed into the top 10: Flores et al. (C03), who ranked 2nd for CI and 8th overall, employ a self-adaptive genetic algorithm (using conventional crossover and mutation on a fixed population of 100 individuals evolved over 500 generations) to specify the order of the autoregressive (p, P) and moving average (q, Q) terms for a seasonal ARIMA (p, d, q) $(P, D, Q)_s$ model, together with their parameter bounds and actual parameters for each time series. Chen and Yao (C46) employ an ensemble of 500 CART regression trees built on bootstrap sampling of the data and random subspace sampling of features. D'yakonov (C13) uses a simple k-nearest-neighbour (k-NN) method with a flexible window size conditional on the time series length.

The original CI benchmark, Balkin & Ord's Automat NN (B00), is ranked 5th within all submitted CI contenders, outperforming 16 (72%) of the 22 new submissions. Considering that AutomatANN was automated to run over 3003 series of different frequencies, not just 111 monthly series, and that it was developed a decade ago, it has proved its representative performance of NNs on monthly data. However, the fact that four (18%) of the submitted CI approaches outperform AutomatANN demonstrates that some progress in research has been made by Illies et al. (C27), Flores et al. (C03), Chen et al. (C46) and D'yakunov (C13). In addition, many of the CI contenders achieve accuracies which are only marginally lower than that of AutomatANN. This indicates that many algorithms and experts today are capable of predicting multiple time series at a level of accuracy similar to that of AutomatANN, an unsurprising improvement on the capabilities at the time of the M3.

Despite the enhanced performances of a few CI methods, the field of submissions in NN/CI remains wide, and many fail to outperform even basic CI benchmarks of naïve MLPs (B02) or naïve SVR (B01). Some methods even fail to outperform the naïve statistical benchmark (B04), which indicates the need for an enhanced understanding of in-sample vs. out-of-sample performances in empirical evaluations and of internal benchmarking (ideally prior to a potentially embarrassing competition performance).

It should be noted, though, that statistical approaches — whether simple or complex — are not a

Table 4
NN3 errors and ranks of errors on the complete dataset.

		Average errors				Rank across all methods				Rank across NN/CI contender				Class[a]
		sMAPE	MdRAE	MASE	AR	sMAPE	MdRAE	MASE	AR	sMAPE	MdRAE	MASE	AR	
B09	Wildi	14.84	0.82	1.13	17.3	1	1	1	1	–	–	–	–	Stat C
B07	Theta	14.89	0.88	1.13	17.8	2	3	1	2	–	–	–	–	Stat B
C27	Illies	15.18	0.84	1.25	18.4	3	2	11	4	1	1	4	1	NN C
B03	ForecastPro	15.44	0.89	1.17	18.2	4	4	3	3	–	–	–	–	Stat B
B16	DES	15.90	0.94	1.17	18.9	5	14	3	6	–	–	–	–	Stat B
B17	Comb S-H-D	15.93	0.09	1.21	18.8	6	5	7	5	–	–	–	–	Stat B
B05	Autobox	15.95	0.93	1.18	19.2	7	11	5	7	–	–	–	–	Stat B
C03	Flores	16.31	0.93	1.20	19.3	8	11	6	8	2	5	1	2	NN C
B14	SES	16.42	0.96	1.21	19.6	9	16	7	12	–	–	–	–	Stat B
B15	HES	16.49	0.92	1.31	19.5	10	9	16	9	–	–	–	–	Stat B
C46	Chen	16.55	0.94	1.34	19.5	11	14	18	9	3	7	9	3	NN C
C13	D'yakonov	16.57	0.91	1.26	20.0	12	7	12	15	4	3	5	6	NN C
B00	AutomatANN	16.81	0.91	1.21	19.5	13	7	7	9	5	3	2	3	NN B
C50	Kamel	16.92	0.90	1.28	19.6	14	5	13	12	6	2	6	5	NN C
B13	Njimi	17.05	0.96	1.34	20.2	15	16	18	18	–	–	–	–	Stat C
C24	Abou-Nasr	17.54	1.02	1.43	21.6	16	26	27	25	7	14	16	14	NN C
C31	Theodosiou	17.62	0.96	1.24	20.0	17	16	10	15	8	8	3	6	NN C
B06	Census X12	17.78	0.92	1.29	19.6	18	9	14	12	–	–	–	–	Stat B
B02	nMLP	17.84	0.97	2.03	20.9	19	19	37	19	–	–	–	–	NN B
C38	Adeodato	17.87	1.00	1.35	21.2	20	22	20	20	9	11	10	9	NN C
C26	de Vos	18.24	1.00	1.35	21.7	21	22	20	27	10	11	10	15	NN C
B01	nSVR	18.32	1.06	2.30	21.6	22	29	38	25	–	–	–	–	NN B
C44	Yan	18.58	1.06	1.37	21.2	23	29	23	20	11	15	13	9	NN C
C11	Perfilieva	18.62	0.93	1.57	20.1	24	11	32	17	12	5	19	8	NN C
C37	Duclos	18.68	0.99	1.30	21.5	25	20	15	24	13	9	7	13	NN C
C49	Schliebs	18.72	1.06	1.37	21.9	26	29	23	28	14	15	13	16	NN C
C59	Beliakov	18.73	1.00	1.36	21.4	27	22	22	23	15	11	12	12	NN C
C20	Kurogi	18.97	0.99	1.31	21.3	28	20	16	22	16	9	8	11	NN C
B10	Beadle	19.14	1.04	1.41	22.1	29	28	25	30	–	–	–	–	Stat C
B11	Lewicke	19.17	1.03	1.43	21.9	30	27	27	28	–	–	–	–	Stat C
C36	Sorjamaa	19.51	1.13	1.42	22.5	31	33	26	31	17	18	15	17	NN C
C15	Isa	20.00	1.12	1.53	23.3	32	32	31	33	18	17	18	19	NN C
C28	Eruhimov	20.19	1.13	1.50	23.2	33	33	30	32	19	18	17	18	NN C
C51	Papadaki	22.60	1.27	1.77	25.0	34	35	34	35	20	20	21	20	NN C
B04	Naïve	22.69	1.00	1.48	24.2	35	22	29	34	–	–	–	–	Stat B
B12	Hazarika	23.72	1.34	1.80	25.6	36	36	35	37	–	–	–	–	Stat C
C17	Chang	24.09	1.35	1.81	26.3	37	37	36	38	21	21	22	22	NN C
C30	Pucheta	25.13	1.37	1.73	25.3	38	38	33	36	22	22	20	21	NN C
C57	Corzo	32.66	1.51	3.61	26.9	39	39	39	39	23	23	23	23	NN C

[a] Stat C = statistical contender; Stat B = statistical benchmark; NNC = NN/CI contender; NNB = NN/CI benchmark.

panacea either: the performances of other novel statistical contenders such as X-12 (B06), composite forecasts (B10) and the Paracaster software (B11) are average at best, with random sequence basis functions (B12) even failing to outperform the naïve statistical benchmark (B04). Also, the weaker contestants in the M3 were not included as benchmarks, biasing the perception of the relative rankings of the CI contenders and the benchmarks, to the disadvantage of NNs; in fact, many of the contestants outperformed established methods from the M3, but we were most interested in the progress at the top of the field relative to AutomatANN.

As with the M3-competition, where Hibon computed Comb S-H-D as a novel contender, we sought to assess the accuracy of combining heterogeneous CI-algorithms. From the submissions, two ensembles were created, combining the forecasts of the top three (C27, C03, C46) and the top five (C27, C03, C46, C13, C50) CI methodologies, respectively, using the arithmetic mean. Both of the CI benchmarks performed outstandingly well: with an sMAPE of 14.89, the ensemble of the top three CI-algorithms would have ranked third overall—tied with Theta (B07) and better than echo state neural networks (C27). Even more convincing, with a sMAPE of 14.87, the ensemble of the top five (C105) would have ranked 2nd only to Wildi (B09), outperforming Theta and all of the other statistical and BI methods (the methods are both listed in Table 5). Although this ex-post combination of the best methods does not represent a valid “ex ante” accuracy (it may be overcome by a quasi-ex ante model selection), it once again underlines the potential of combining heterogeneous predictions. While Illies et al.’s (C27) performance obviously contributed significantly to the performances of the two CI-ensembles, the combination increases the accuracy beyond that of each individual contender, an effect which is well documented (in addition to the second benefit of a decreased error variance). More importantly, by including the top five instead of the top three CI algorithms, essentially introducing more inferior forecasts into an ensemble, the overall accuracy was increased even further. Therefore, it seems that further increases in accuracy are feasible for CI by combining diverse base-algorithms into heterogeneous ensembles, a finding which is well documented for statistical algorithms in prior forecasting competitions and which promises further potential in improving forecasting accuracy due to the vast and heterogeneous model classes available in CI which were not evaluated here.

4.2. Significance of the findings

Regardless of the recent and vivid discussion about statistical significance within the forecasting community (Armstrong, 2007a,b; Goodwin, 2007), we computed two non-parametric tests, replicating the analysis of the M3 by Koning et al. (2005): ANOM and MCB, both of which are based upon the average ranks of 41 methods (including both CI ensembles) over 111 series and 18 horizons (see Figs. 1 and 2).

For ANOM, only the ensemble of the top 5 (C105) and the methodology by Wildi (B09) prove to be statistically significantly better than average. On the other side, four CI approaches (those by Chang (C17), Pucheta (C30), Papadaki (C51) and Corzo (C57)) and one statistical contender, that by Hazarika (B12), perform significantly worse than the average.

The findings of MCB are similar to those of ANOM: the ensemble of the top five (C105) and Wildi (B09) are identified as the two best approaches, while the same four CI (C17, C30, C51, C57) and one statistical contender (B12), plus the naïve (B04), are significantly worse than the best. Despite the limited differences in statistical significance, it is worth mentioning that even a small gain in accuracy, e.g. 1%, is often amplified in operational benefits, and could result in manifold savings in safety stocks. Thus, accuracy results in term of average metrics should never be ignored, as they are often operationally significant (Syntetos, Nikolopoulos, & Boylan, 2010). It should be noted that there are more adequate tests available today for assessing significant differences between the relative performances of algorithms, see, e.g., Demsar (2006); however, they were omitted here to allow for coherence with the previous M3 analysis. As an indication of the limitations of these tests, the Theta method — which was previously better than other algorithms in the competition — is no longer significantly better than other algorithms, indicating the sensitivity of the test to the sample size and structure (as for all tests), adding further to the discussion of tests.

4.3. Analysis of data conditions

Next, we analyse the data conditions under which the different algorithms perform well. As it is not feasible to present all 24 tables of rankings for each error measure and data subset, Table 5 summarizes the results of the top five performers for both the complete and reduced datasets (111 and 11 series), and for the conditions of long and short time series lengths (50 series each), seasonal and non-seasonal time series patterns (50 series each), and the combination of both conditions (25 series each). Table 6 shows the top five performers by sMAPE across the different forecasting horizons. In order to facilitate the replication and external analysis of the results, all of the tables for

Table 5
NN3-competition results across data conditions on sMAPE, MdRAE, and MASE.

Error metrics	Complete dataset (incl. reduced)	Reduced dataset	Data conditions				Combined data conditions			
			Short	Long	Seasonal	Non-seas.	Short		Long	
							Non-seas.	Seasonal	Non-seas.	Seasonal
# of series	111	11	50	50	50	50	25	25	25	25
sMAPE	*B09*	B05	**C27**	B03	*B09*	**C105**	**C27**	*B09*	B03	B03
	C105	B03	**C105**	*B09*&B16	**C105**	B07	**C105**	**C27**	B16	*B09*
	B07&**C103**	**C44**	**C103**	–	**C103**	**C103**	**C103**	**C105**	B07&*B09*	B17
	–	B07	*B09*	B14	B07	B03	B17	**C103**	–	B14
	C27	**C59**	B07	B07	**C27**	B14	B13	**C50**	B00	B16
MdRAE	**C105**	C38	**C27**	B03	*B09*	**C105**	**C27**	**C27**	B00	B16
	B09&**C103**	**C105**	*B09*	*B09*&B15	**C27**	B00	**C105**	*B09*	B03&*B09*	B14&B17
	–	**C11**	**C105**	B16&B17	**C103**&**C105**	**C27**	*B09*&B14	**C103**&**C105**	–	–
	C27	**C103**	**C50**&**C103**	–	–	*B09*&**C50**	B17	–	B16&**C105**	B03
	B07	B03	–	–	B07	–	–	B05	–	B07
MASE	**C105**	B05&**C59**	**C105**	B14	B09	B14	*B09*	**C27**	B14	B14–B17
	B07&*B09*	–	**C27**&**C103**	B16	B07	**C105**	**C103**&**C105**	**C103**&**C105**	B00	–
	–	B03	–	B07	**C105**	B00	–	–	B16	*B09*&B16
	B03&B16	**C44**	*B09*	B17&**C105**	**C103**	B07&B16	**C27**	B07–B17	B04	–
	–	**C18**	**C50**	–	**C27**	–	**C50**	–	**C105**	B03

Bold: CI contenders; *Italics*: Statistical contenders; Normal: Benchmarks; Underlined: AutomatANN M3 benchmark.

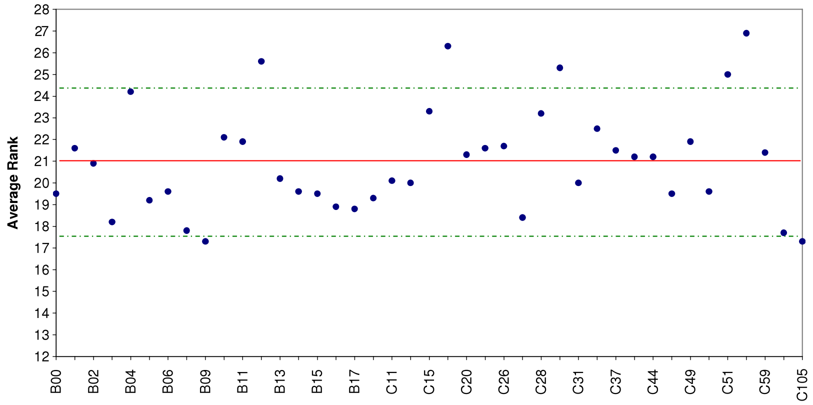

Fig. 1. Analysis of means on the complete dataset.

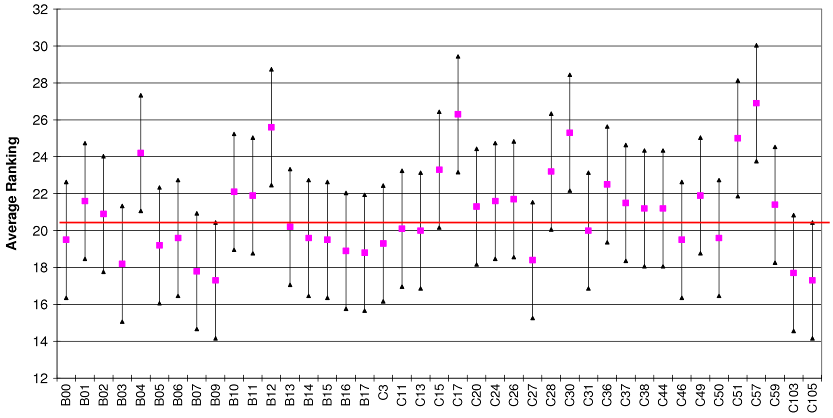

Fig. 2. Multiple comparisons with the best on the complete dataset.

sMAPEs (Tables 6–15), MdRAEs (Tables 16–20), MASEs (Tables 21–25), ARs for all methods and for CI contenders separately (Tables 26–27), ANOM (Table 28), and MCB based upon AR (Table 29) will be provided online on the journal website (www.forecasters.org/ijf/).

On the complete dataset (first column of Table 5), the ranking of all algorithms is identical to the results provided in Table 4, identifying the top performers in the NN3 according to sMAPE, namely Wildi (B09), ensemble of the top five CI (C105), Theta (B07) in a draw with the ensemble of the top three CI (C103) and Illies et al. (C27). In comparison, different algorithms performed well on the reduced dataset of 11 time series which were deemed to be hard to forecast: the statistical expert system Autobox (B05) was ranked 1st by sMAPE, playing out its strengths in modeling pulse interventions, level shifts, local time trends and seasonal pulses. ForecastPro (B03) ranked 2nd and Theta (B07) ranked 4th. Two new CI contestants enter the top five of the reduced dataset: Yan (C44), ranked 3rd on sMAPE across all methods and 1st for CI methods, employs three sets of 18 generalized regression NNs per time series, each of which is trained separately to predict for a forecasting horizon $h = (1, 2, \ldots, 18)$ with three distinct parameter settings, recombining the predictions to give one trace forecast, then combining the predictions of the three architectures in an ensemble, hence the name 'multiple model fusion'.

Using the MdRAE, other CI contenders enter the top five: Adeonato et al. (C38), using ensembles of 15 MLPs, and Perfilieva (C11), forecasting using fuzzy transformations, indicating that the results on only a

Table 6
NN3 results of sMAPE across short, medium, long and all forecasting horizons.

Error metrics	Complete dataset (incl. reduced)	Reduced dataset	Combined data conditions			
			Short		Long	
			Non-seas.	Seasonal	Non-seas.	Seasonal
# of series	111	11	25	25	25	25
Short (h = 1–3)	B07	**C20**	**C105**	**C27**	B07	B16–B17
	B09	*B10*	**C27**	*B09*	B03	–
	B03–**C105**	**C08**	**C50**	B00	B16	B03
	–	B03	*B09*	B05	B06	B14
	C103	**C59**	**C59**	**C50**	B17	B15
Medium (h = 4–12)	**C105**	**C44**	**C27**	*B09*	B03	*B09*
	B09	**C50**	B17	**C50**	**C3**	B06
	C103	**C46**	**C105**	**C105**	B07	B03
	B07	B07	**C103**	**C27**	B16	B16
	C27	B05	B14	**C103**	B14	B17
Long (h = 13–18)	**C103**	B05	**C27**	*B09*	**C105**	B17
	B07	**C38**	**C46**	**C27**	**C103**	B14
	C105	B03	**C103**	**C103**	*B09*	B03
	B09	**C18**	B13	B07	**C13**	B07
	C27	**C59**	B14	**C105**	B00	B16
All (h = 1–18)	*B09*	B05	**C27**	*B09*	B03	B03
	C105	B03	**C105**	**C27**	B16	*B09*
	B07–**C103**	**C44**	**C103**	**C105**	B07–*B09*	B17
	–	B07	B17	**C103**	–	B14
	C27	**C59**	B13	**C50**	B00	B16

Bold: CI contenders; *Italics*: Statistical contenders; Normal: Benchmarks; *Underlined*: AutomatANN M3 benchmark.

few series are not as reliable across error measures as for the complete set. This does, however, show that there is the potential for specialised statistical and CI algorithms which are tuned (or robust) to particular time series properties to outperform other approaches, though at the same time it questions the ability of these CI methodologies to generalise to larger datasets than the ones they were originally tailored to.

Next, we analyse the results across the data conditions of time series length and seasonality. Wildi's (B09) new statistical approach ranks well under all data conditions and metrics, with the exception of short & non-seasonal series on sMAPE, indicating that some of its success is derived from capturing seasonality well (1st for all metrics). Variants of ES (B14, B15, B16 and their combination B17) make frequent appearances on long & seasonal time series, indicating that the decomposition approach used for M3 — *Deseasonalise + Extrapolate + Reseasonalise* — works competitively. Similarly, the expert system ForecastPro (B03), which selects amongst these methods, outperforms them on long series of both seasonal and non-seasonal data, confirming that industry still does well to rely on this family of methods for these typical data conditions. The related Theta (B07) appears in the top performers on all aggregate conditions, but its combinations do not, verifying its robustness across many data conditions by a consistent level of accuracy, but not winning in any particular category.

For CI, multiple CI contenders enter the top five under different conditions, while the M3 benchmark AutomatANN (B00) is absent across all categories and metrics (with the exception of *Long + Non-seasonal* data using sMAPE). In the light of earlier research, the most striking result of NN3 comes in the *Short + Non-seasonal* subset, which, judging by recent publications, is one of the most difficult conditions for CI methods. Echo state networks by Illies et al. (C27) achieved the *colpo grosso* and won this category, as well as that of the broader 50 short series, which we speculate is an effect of training on pooled clusters of time series. CI ensembles of three (C103) and five (C105) CI algorithms performed equally well across data conditions of short + seasonal and short + non-seasonal series, ranking 2nd/3rd and 3rd/4th respectively, but less so across long series with and without seasonality (unsurprisingly, as C27 was contained in them). Of the remaining CI competitors, only Kamel (C50) made an appearance in the *Short + Seasonal* category, combining MLPs with Gaussian process regression.

These results on short time series challenge prior beliefs in NN modeling (in accordance with working hypothesis WH3) that a significant number of historic observations are a prerequisite for the sufficient initialization, training, validation, evaluation and generalisation of CI approaches (see for example Haykin, 1999). Furthermore, across time series patterns, more CI are ranked highly on seasonal data than on non-seasonal data, a second fundamental contradiction to prior research in the form of working hypothesis WH4, which had identified problems in predicting seasonal time series with NNs and proposed prior deseasonalisation (e.g., Zhang & Qi, 2005). While these results provide no insights into the reasons for this improved performance, they do demonstrate that novel CI-paradigms can yield competitive performances beyond their traditional application domain, and that systematic replications of earlier studies should be conducted in order to challenge prior findings. However, the majority of CI approaches are absent across datasets and conditions, on the one hand demonstrating consistent results, but on the other indicating that only a few algorithms have the capacity to perform well.

The results across forecasting horizons seem to confirm earlier findings by Hill et al. (1996): ES methods (B07 and B09) appear to perform best for short term forecasting, but with an increasing forecasting horizon the CI approaches take the lead, although it remains unclear whether this contribution stems from the forecast combinations in ensembles, or the underlying methods' performances improving with the horizon (see also working hypothesis WH2).

However, for CI, the accuracy levels achieved across horizons show a surprising degree of consistency. On the complete dataset, the contenders which are ranked highly overall are also consistently ranked amongst the top five across all horizons of short, medium and long term forecasts, with only minor changes in rankings. This is also confirmed across data conditions, where the relative performances remain consistent across different horizons: CI methods perform well for short time series with and without seasonality across all forecasting horizons, and in particular Illies' (C27) and the ensembles C105 and C103. Similarly, for long time series, ES methods perform consistently well across all horizons, again without significant changes in rankings. The only noticeable change appears for long + non-seasonal data, where ES dominates for short horizons, and CI for long. Results across horizons for a particular data subset remain more stable than expected, given prior findings. For example, Wildi's (B09) approach, which is optimised specifically for multiple horizons of a trace forecast, performs consistently well across all horizons for short + seasonal time series, as was intended by the algorithm.

5. Discussion

The NN3 competition has contributed empirical evidence in the tradition of the M-competitions, with a particular emphasis on extending the findings of the M3 competition towards a current and complete range of CI methods. The NN3 seems to have succeeded in this, having attracted contestants from all major paradigms, including feed-forward and recurrent NNs, fuzzy logic, genetic algorithms and evolutionary computation, and hybrid systems. In addition, the results of this replication and extension of the M3 allow us to evaluate the six hypotheses of the original M-competition (see Section 2), and to determine whether the findings conform to the established wisdom or add novel insights to the body of knowledge. First, we will review hypotheses H1, H2 and H3, as they allow us to assess the similarity of the M3 and its replication, and allow a verification of the NN3 competition design. (H4 cannot be assessed, as the NN3 — like various other forecasting and CI competitions — chose to employ only a single hold-out evaluation set rather than multiple test sets of rolling time origins for a time dependent k-fold cross validation, which would require a prohibitive amount of resources both to conduct and to take part in the competition. However, the implications of H4 were considered in setting the competition design to 111 time series.)

(H1) '*Data characteristics determine relative performances*?' The results of the NN3 across data conditions (Table 5) confirm those of the earlier M3: the data characteristics have a substantial influence on the relative performances of algorithms in statistics and CI alike. Different algorithms perform well on seasonal vs. non-seasonal and short vs. long time series. Here, NN3 contributes further to the discussion by providing objective evidence that NNs are capable of predicting seasonal time series (in contrast to Zhang & Qi, 2005, for example), and of predicting short time series (in contrast to Hill et al., 1996, for example) accurately, contrary to the findings of previous studies, thus indicating the need for further research. However, as was demonstrated by Illies et al. (C27), the pooling of data across different conditions may yield robust algorithms which are capable of accurate forecasting across data characteristics.

(H2) '*Accuracy depends upon the forecasting horizon*?' The relative performance varies across forecasting horizons (Table 6), and different methods perform best for different horizons, which confirms the findings of M3. Also, the efficacy of CI methods relative to statistical methods increases for longer forecasting horizons, as was identified in previous studies (Hill et al., 1996). However, for the best CI algorithms, the accuracy remained almost constant for increasing forecasting horizons, with good performances for short horizons as well. Further research is needed to determine whether methods incorporating trace errors in their modelling (e.g. Wildi (B09) or Yan (C44)) can overcome this limitation, as first indications seem to suggest.

(H3) '*Performance ranking varies by metric*?' The rankings of the NN3 contestants based upon the sMAPE, MdRAE, MASE and AR each result in different relative performances of the algorithms, across all datasets and data conditions (see Table 5). However, many methods in the upper deciles of the field perform consistently well on multiple metrics, and vice versa, increasing the confidence in their relative performances and predictive capabilities.

Next, we will review H5 and H6, which consider the relative accuracies of the algorithms, which is the main topic of this extension of the M3 competition.

(H5) '*Combinations outperform individual methods*?' Reviewing the common properties of the top performers (Table 5), the success of combinations stands out. With the exception of the five original submissions to the M3 (ForecastPro, Autobox, SES, DES, and HES), all three of the leading statistical methods in the top 10 use forecast combinations (most notably Wildi (B09) across all conditions, Comb S-H-D (B17) for long series, and Theta (B07), which essentially employs a weighted forecast combination of a linear trend and ES). Also, with the exception of Flores (C03), all CI methodologies in the top 10 employ forecast combinations (Illies (C27), Chen (C46), ensemble of the top five (C105), and ensemble of the top three CI/NN (C103)). The ensembles (C105, C103) dominate our results, but also indicate the positive effect of increasing the coverage and diversity in an ensemble (i.e., the heterogeneity of the base learner), which thus warrants more research effort across disciplines. As sophisticated 'ensembles' in the form of boosting, bagging, arcing, etc., are more widespread in CI classifications than in statistical modelling, and time series prediction in particular, we see some potential for cross-disciplinary research here.

(H6) '*Sophisticated methods are not better than simpler methods*?' Seeing that the majority of CI approaches have failed to outperform simple ES (B14), and four performed worse than naïve (B04) (see Tables 6–15 online), we could not disagree. However, NN3 has introduced a novel univariate method, and provided evidence of its ability to outperform established statistical benchmarks, including the respective winners on the monthly M3 data (dampen ES, Theta and ForecastPro), and all CI contenders to date. Although the algorithm by Wildi (B07) is statistical in nature and not based upon CI, the method cannot be classified as anything other than complex, as it combines various innovations in estimation and model selection to automatically tune it to the data. This conflicts with H6, and with the common belief that complex methods cannot significantly outperform simple ones. Similarly, NN3 provides evidence that some complex methods are capable of outperforming all statistical methods from the M3, showing a substantial improvement in accuracy. To provide further evidence, with the submissions of Wildi, Theta, ForecastPro and Autobox for statistics, and with Illies and Flores representing CI, four of the top five (80%) and six of the top 10 methods (60%) must reasonably be classified as complex methods. As such, we have provided objective evidence that does not support H6. Rather than refuting H6 on the basis of a few algorithms, we seek to reverse the hypothesis to challenge the established wisdom:

(H6.b) *Simple methods are not better than sophisticated methods*.
Despite the fact that the content is identical, H6 all too easily suggested that no benefits arise from sophistication, and allowed the misinterpretation that '*Simpler is better*'. We conclude that the complex methods of CI/NN and statistics have caught up, and, overall, simple statistical methods can no longer claim to outperform CI methods without a proper empirical evaluation.

As with every empirical study, the findings only hold for the properties of the empirical dataset provided, and as such, the NN3 competition does not aim to be representative of all data properties in operational forecasting. However, our competition is still prone to certain limitations and biases that must be reviewed critically. These include the obvious shortcomings that are endogenous to most competitions: no rolling origin design (due to the challenge of organising such a setup; see H4), the limited representativeness of the datasets in size, structure and heterogeneity, and the exclusion of certain performance metrics that assess the final impact on decision making, e.g., the inventory costs arising from operational forecasting (Timmermann & Granger, 2004). As with prior M-competitions, our assessment considered only the empirical accuracy of the algorithms, and neglected robustness, interpretability, and efficiency through the computational resources required, all important aspects in forecasting for operations. Because expert software systems such as Autobox and ForecastPro contain much faster forecasting engines than CI (i.e., we received the submission of Autobox almost instantaneously following the release of the data), algorithms and systems employing efficient statistical methods may still remain the first choice in operations.

Despite our efforts, biases in the representativeness of the algorithms may exist. In tailoring the NN3 to NN and CI algorithms, we may have biased the sample of contestants by attracting more CI contestants than statistics contestants. Furthermore, the majority of the submissions came from researchers in CI, while professionals and (possibly advanced) software companies in NN, CI and AI (e.g., Siemens, Alyuda, Neuro Dimensions, and SAS) chose not to participate, despite personal invitations. Also, more participation from econometrics and forecasting software vendors which are active in forecasting for operations (e.g. SAP, Oracle, John Galt, Smart, etc.) would have increased the validity of results; however, they likewise did not accept personal invitations to participate. Nevertheless, we tried to be as objective and inclusive as possible, taking into consideration the design suggestions of prior competitions and reaching out to the communities which had previously been omitted. Therefore, we are confident that NN3 provides a more comprehensive and up-to-date assessment of the performances of CI methods in predicting monthly time series than M3, as well as providing more valid and reliable evidence than previous CI competitions.

One fundamental flaw — grounded in the nature of a replication — lies in the prior availability of the data, although its origin was undisclosed and masked in a sample. Although we are convinced of the integrity of all contestants, this is a reminder of the importance of true ex-ante evaluations on unknown data for future competitions, to avoid any data snooping.

6. Conclusions

Replicating and extending the prominent M3 competition, NN3 aspired to challenge prior evidence on the inferior forecasting accuracy of NN approaches in operational forecasting. The final results assess the accuracies of over 60 forecasting algorithms, the largest assessment of different methods on time series data to date. Ex ante accuracies were evaluated on either 111 or 11 empirical time series using multiple established error metrics and following a rigorous competition design, while the conditions examined include the presence of seasonality, the length of the series, and the forecasting horizon.

The objective of the NN3, namely to extend the M3 competition to NN and CI algorithms, was successfully achieved by attracting 46 CI contestants and novel statistical benchmarks, making it the largest empirical evaluation on time series data in the areas of NN, CI and forecasting to date. The main findings confirm prior hypotheses, but also initiate new research discussions. New algorithms are feasible, in CI, NN and statistics alike. The competition assessed a novel — and complex — statistical method by Wildi (B9), which performed exceptionally well for both datasets. Illies et al. (C27) introduced a NN methodology which outperformed damped trend ES, but still did not perform as well as the Theta method across all series. This algorithm also outperformed all other algorithms on 25 short and seasonal time series, the most difficult subset of the competition, while Yan (C44) outperformed all others on a subset of 11 complex/difficult series. These achievements are surprising, considering prior beliefs on the data properties required when using NN methods on empirical data, and demand further attention. Overall, we hope that the success of complex algorithms on such a well-established dataset will at least rekindle the discussion of innovative, sophisticated algorithms for time series extrapolation in forecasting, econometrics and statistics.

The results of the NN3 suggest that NN and CI methods can perform competitively relative to established statistical methods in time series prediction, but still cannot outperform them. However, in the absence of any (statistically significant) differences between algorithms, we can no longer assume that they are inferior either. Considering the results of the M3, we have consciously included the top-performers of ForecastPro, Theta, and Comb S-H-D as hard benchmarks for NN to compete against. As such, we expected that the ES methods, the workhorses of operational forecasting in practice for over 30 years, would be serious contenders that would prove challenging to outperform—after all, they did outperform most others methods in the original M3. It should, however, be noted that the other 20 statistical methods in M3 performed less admirably, and would not be expected to do better than many CI contestants. We feel that CI has closed in on the established benchmarks, showing a range of different algorithms which are capable of predicting both datasets as accurately as AutomatANN, the only CI contestant in the M3 some 10 years ago, thus indicating that there have been improvements in the feasibility and empirical accuracy of forecasting with NNs, and hence motivating further research.

Disappointingly, it does not seem possible to provide any more focussed guidance as to promising routes for future CI research, as no common ‘best practises’ can be identified for the top NN or CI contenders. Each submission was unique, both conceptually and methodologically, combining freely (and often seemingly arbitrarily) from the repository of algorithms and techniques which are available to machine learning today, and without any evaluation of the contribution each fragment of the methodology made to increasing the accuracy. For example, for Illies et al. it is still not clear whether the accuracy stems from pooling time series for training, combining predictions in ensembles, or the echo state neural networks algorithm itself. In an attempt to generalise, only the paradigm of forecast combinations seemed to drive the accuracy, an observation which has been well established before. Ensembles of CI and statistical algorithms performed very well, but again no consensus on the meta-parameters of ensemble size or combination metric could be determined, although the heterogeneity of its base learners seemed to have a positive effect on the accuracy. As no two algorithms are alike, it then becomes impossible to attribute a positive performance to a particular modelling choice, thus allowing an evaluation of composite yet distinct algorithms, but not providing any guidance as to promising areas for future research. Without such insights, progress in CI may be slow and undirected. If this heterogeneity cannot be overcome, only a meta-learning analysis could yield insights to partial contributions, linking the properties of algorithms and data conditions in order to guide future research effort.

The NN3 competition has proven a stimulating exercise that has attracted, engaged and unified researchers from the areas of forecasting, informatics, machine learning, data mining and engineering. We therefore hope that the NN3 will provide a means to disseminate best practices not only on CI-methods, but also, more importantly, on competition design beyond the forecasting community. We conclude that the findings of the NN3 competition provide encouraging evidence of the capabilities of NN and CI methods in time series prediction, even for a well established domain such as monthly time series prediction. The promising results of NN3 thus motivate us to run future competitions in order to add to the knowledge on modelling neural networks for time series prediction. Already, it has sparked a resurgence of interest in CI competitions, with regular competition tracks having been held at the ESTSP, IJCNN, DMIN and WCCI conferences since. For future competitions, we see the need to evaluate novel application domains that are empirically important but have previously been omitted, and in particular those of high frequency data, where NNs are regularly employed in practice. Still, no method will be a true panacea. However, only by extending competition designs to novel data conditions, beyond those of the M-style competitions, will we be able to determine the data for which the application of neural networks is indeed either a breakthrough or a passing fad.

Acknowledgments

We are grateful to SAS and the International Institute of Forecasters, who have provided us with the funding to conduct the two-year NN3 project under an IIF/SAS grant funded project entitled "*Automatic Modelling and Forecasting with Artificial Neural Networks—A forecasting competition evaluation*" in 2006. The authors would like to thank Prof. Robert Fildes for his continuous advice in the setup and running of this competition. In particular, we would like to thank all the contestants of NN3 for their participation. Dr. Hibon joined the project at the later stages to ensure the exact replication of the benchmarks, error metrics and analyses used in the M3 competition.

References

Adya, M., Armstrong, J. S., Collopy, F., & Kennedy, M. (2000). An application of rule-based forecasting to a situation lacking domain knowledge. *International Journal of Forecasting*, *16*, 477–484.

Adya, M., & Collopy, F. (1998). How effective are neural networks at forecasting and prediction? A review and evaluation. *Journal of Forecasting*, *17*, 481–495.

Armstrong, J. S. (1985). *Long-range forecasting: from crystal ball to computer* (2nd ed.) New York: Wiley.

Armstrong, J. S. (2006). Findings from evidence-based forecasting: methods for reducing forecast error. *International Journal of Forecasting*, *22*, 583–598.

Armstrong, J. S. (2007a). Significance tests harm progress in forecasting. *International Journal of Forecasting*, *23*, 321–327.

Armstrong, J. S. (2007b). Statistical significance tests are unnecessary even when properly done and properly interpreted: reply to commentaries. *International Journal of Forecasting*, *23*, 335–336.

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

Athanasopoulos, G., Hyndman, R. J., Song, H., & Wu, D. C. (2011). The tourism forecasting competition. *International Journal of Forecasting*, *27*, 822–844.

Balkin, S. D., & Ord, J. K. (2000). Automatic neural network modeling for univariate time series. *International Journal of Forecasting*, *16*, 509–515.

Billah, B., King, M. L., Snyder, R. D., & Koehler, A. B. (2006). Exponential smoothing model selection for forecasting. *International Journal of Forecasting*, *22*, 239–247.

Breiman, L. (1984). *Classification and regression trees*. Belmont, Calif: Wadsworth International Group.

Cawley, G. C., Janacek, G. J., Haylock, M. R., & Dorling, S. R. (2007). Predictive uncertainty in environmental modelling. *Neural Networks*, *20*, 537–549.

Chatfield, C. (1993). Neural networks: forecasting breakthrough or passing fad? *International Journal of Forecasting*, *9*, 1–3.

Chen, B. J., Chang, M. W., & Lin, C. J. (2004). Load forecasting using support vector machines: a study on EUNITE competition 2001. *IEEE Transactions on Power Systems*, *19*, 1821–1830.

Crone, S. F., & Kourentzes, N. (2010). Feature selection for time series prediction—a combined filter and wrapper approach for neural networks. *Neurocomputing*, *73*, 1923–1936.

Crone, S. F., & Pietsch, S. (2007). A naïve support vector regression benchmark for the NN3 forecasting competition. In *2007 IEEE international joint conference on neural networks. Vols. 1–6* (pp. 2453–2458).

Crone, S. F., & Preßmar, D. B. (2006). An extended evaluation framework for neural network publications in sales forecasting. In *AIA'06 proceedings of the 24th IASTED international conference on artificial intelligence and applications*.

Curry, B. (2007). Neural networks and seasonality: some technical considerations. *European Journal of Operational Research*, *179*, 267–274.

Dawson, C. W., See, L. M., Abrahart, R. J., Wilby, R. L., Shamseldin, A. Y., Anctil, F., et al. (2005). A comparative study of artificial neural network techniques for river stage forecasting. In *Proceedings of the international joint conference on neural networks*: *Vols. 1–5* (pp. 2666–2670).

de Menezes, L. M., & Nikolaev, N. Y. (2006). Forecasting with genetically programmed polynomial neural networks. *International Journal of Forecasting*, *22*, 249–265.

Demsar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, *7*, 1–30.

Faraway, J., & Chatfield, C. (1998). Time series forecasting with neural networks: a comparative study using the airline data. *Applied Statistics*, *47*, 231–250.

Fildes, R. (1992). The evaluation of extrapolative forecasting methods. *International Journal of Forecasting*, *8*, 81–98.

Fildes, R. (2006). The forecasting journals and their contribution to forecasting research: citation analysis and expert opinion. *International Journal of Forecasting*, *22*, 415–432.

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

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., Nikolopoulos, K., Crone, S. F., & Syntetos, A. A. (2008). Forecasting and operational research: a review. *Journal of the Operational Research Society*, *59*, 1150–1172.

Fildes, R., & Ord, K. (2002). Forecasting competitions: their role in improving forecasting practice and research. In M. P. Clements, & D. F. Hendry (Eds.), *A companion to economic forecasting* (pp. 322–353). Malden, Mass: Blackwell.

Fogel, D. B. (1994). An introduction to simulated evolutionary optimization. *IEEE Transactions on Neural Networks*, *5*, 3–14.

Foster, W. R., Collopy, F., & Ungar, L. H. (1992). Neural network forecasting of short, noisy time-series. *Computers and Chemical Engineering*, *16*, 293–297.

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, *55*, 119–139.

Goodrich, R. L. (2000). The ForecastPro methodology. *International Journal of Forecasting*, *16*, 533–535.

Goodwin, P. (2007). Should we be using significance tests in forecasting research? *International Journal of Forecasting*, *23*, 333–334.

Goodwin, P., & Lawton, R. (1999). On the asymmetry of the symmetric MAPE. *International Journal of Forecasting*, *15*, 405–408.

Groff, G. K. (1973). Empirical comparison of models for short range forecasting. *Management Science, Series A—Theory*, *20*, 22–31.

Guyon, I., Saffari, A., Dror, G., & Cawley, G. (2008). Analysis of the IJCNN 2007 agnostic learning vs. prior knowledge challenge. *Neural Networks*, *21*, 544–550.

Haykin, S. S. (1999). *Neural networks: a comprehensive foundation* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

Hill, T., O'Connor, M., & Remus, W. (1996). Neural network models for time series forecasts. *Management Science*, *42*, 1082–1092.

Hippert, H. S., Pedreira, C. E., & Souza, R. C. (2001). Neural networks for short-term load forecasting: a review and evaluation. *IEEE Transactions on Power Systems*, *16*, 44–55.

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, *22*, 679–688.

Jaeger, H., & Haas, H. (2004). Harnessing nonlinearity: predicting chaotic systems and saving energy in wireless communication. *Science*, *304*, 78–80.

Kolarik, T., & Rudorfer, G. (1994). Time series forecasting using neural networks. In *Proceedings of the international conference on APL. Antwerp, Belgium* (pp. 86–94).

Koning, A. J., Franses, P. H., Hibon, M., & Stekler, H. O. (2005). The M3 competition: statistical tests of the results. *International Journal of Forecasting*, *21*, 397–409.

Lendasse, A., Oja, E., Simula, O., & Verleysen, M. (2007). Time series prediction competition: the CATS benchmark. *Neurocomputing*, *70*, 2325–2329.

Liao, K. P., & Fildes, R. (2005). The accuracy of a procedural approach to specifying feedforward neural networks for forecasting. *Computers and Operations Research*, *32*, 2151–2169.

Makridakis, S. (1994). Book review: "Time series predicition—forecasting the future and understanding the past" by A.S. Weigend & N.A. Gershenfeld. *International Journal of Forecasting*, *10*, 463–466.

Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., et al. (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., et al. (1993). The M2-competition: a real-time judgmentally based forecasting study. *International Journal of Forecasting*, *9*, 5–22.

Makridakis, S., & Hibon, M. (1979). Accuracy of forecasting—empirical investigation. *Journal of the Royal Statistical Society, Series A—Statistics in Society*, *142*, 97–145.

Makridakis, S., & Hibon, M. (2000). The M3-competition: results, conclusions and implications. *International Journal of Forecasting*, *16*, 451–476.

McNames, J., Suykens, J. A. K., & Vandewalle, J. (1999). Winning entry of the K.U. Leuven time-series prediction competition. *International Journal of Bifurcation and Chaos*, *9*, 1485–1500.

Nelson, M., Hill, T., Remus, W., & O'Connor, M. (1999). Time series forecasting using neural networks: should the data be deseasonalized first? *Journal of Forecasting*, *18*, 359–367.

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

Ord, K., Hibon, M., & Makridakis, S. (2000). The M3-competition. *International Journal of Forecasting*, *16*, 433–436.

Preminger, A., & Franck, R. (2007). Forecasting exchange rates: a robust regression approach. *International Journal of Forecasting*, *23*, 71–84.

Qi, M., & Zhang, G. P. (2001). An investigation of model selection criteria for neural network time series forecasting. *European Journal of Operational Research*, *132*, 666–680.

Reid, D. J. (1969). *A comparative study of time series prediction techniques on economic data*. Ph.D. thesis. University of Nottingham, Nottingham, UK (unpublished).

Reid, D. J. (1972). A comparison of forecasting techniques on economic time series. In M. J. Bramson, I. G. Helps, & J. A. C. C. Watson-Grady (Eds.), *Forecasting in action*. Birmingham, UK: Operational Research Society.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1994). Learning representations by back-propagating errors (from Nature 1986). *Spie Milestone Series Ms, 96*, 138.

Sharda, R., & Patil, R. B. (1992). Connectionist approach to time-series prediction—an empirical test. *Journal of Intelligent Manufacturing*, *3*, 317–323.

Sincák, P., Strackeljan, J., Kolcun, M., Novotný, D., & Szathmáry, P. (2002). Electricity load forecast using intelligent technologies. In *EUNITE European Network of Intelligent Technologies*.

Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, *14*, 199–222.

Suykens, J. A. K., & Vandewalle, J. (1998a). The K.U. Leuven time series prediction competition. In J. A. K. Suykens, & J. Vandewalle (Eds.), *Nonlinear modeling: advanced black-box techniques* (pp. 241–253). Kluwer Academic Publishers.

Suykens, J. A. K., & Vandewalle, J. (Eds.) (1998b). *Nonlinear modeling: advanced black-box techniques*. Boston: Kluwer Academic Publishers.

Suykens, J. A. K., & Vandewalle, J. (2000). The K.U. Leuven competition data—a challenge for advanced neural network techniques. In *European symposium on artificial neural networks* (pp. 299–304).

Syntetos, A. A., Nikolopoulos, K., & Boylan, J. E. (2010). Judging the judges through accuracy-implication metrics: the case of inventory forecasting. *International Journal of Forecasting*, *26*, 134–143.

Tang, Z. Y., & Fishwick, P. A. (1993). Feed-forward neural nets as models for time series forecasting. *ORSA Journal on Computing*, *5*, 374–386.

Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: an analysis and review. *International Journal of Forecasting*, *16*, 437–450.

Terasvirta, T., van Dijk, D., & Medeiros, M. C. (2005). Linear models, smooth transition autoregressions, and neural networks for forecasting macroeconomic time series: a re-examination. *International Journal of Forecasting*, *21*, 755–774.

Timmermann, A., & Granger, C. W. J. (2004). Efficient market hypothesis and forecasting. *International Journal of Forecasting*, *20*, 15–27.

Weigend, A. S., & Gershenfeld, N. A. (1994). Time series prediction: forecasting the future and understanding the past. In *Proceedings of the NATO advanced research workshop on comparative time series analysis held in Santa Fe, New Mexico, May 14–17, 1992* (1st printing ed.). Reading: Addison-Wesley.

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, *8*, 338–353.

Zellner, A. (1986). Bayesian estimation and prediction using asymmetric loss functions. *Journal of the American Statistical Association*, *81*, 446–451.

Zhang, G. P., & Qi, M. (2005). Neural network forecasting for seasonal and trend time series. *European Journal of Operational Research*, *160*, 501–514.

Sven F. Crone is an Assistant Professor of Management Science at Lancaster University Management School, and the deputy director of the Lancaster Research Centre for Forecasting. His research focuses on forecasting, time series prediction and data mining in business applications, frequently employing methods from computational intelligence such as neural networks and support vector machines. His research has been published in the *European Journal of Operational Research, Journal of Operational Research Society* and *International Journal of Forecasting*. Sven is the competition chair of the IEEE CIS Data Mining Technical Committee and has co-organised the 2007 Neural Network Forecasting Competition (NN3), cosponsored by the IIF, NSF and SAS, as well as the 2008 NN5 and the current 2009 IEEE Grand Challenge on Time Series Prediction with Computational Intelligence.

Michèle Hibon is an Emeritus Lecturer and Senior Research Fellow at INSEAD, France. Originally a graduate in Physics from the University of Paris, she has been working in the area of forecasting methods and forecasting accuracy since the late 70s, and also, together with Spyros Makridakis, conducted the M, M2 and M3 forecasting competitions. She is joint author of several articles published in the *International Journal of Forecasting*. Her research interests lie in forecasting competitions, accuracy of forecasting methods, and combination of forecasts.

Konstantinos Nikolopoulos is a Professor at Bangor University. He has received his Ph.D. in December 2002 from the National Technical University of Greece (NTUA) in the research field of business forecasting information systems under the supervision of Prof. V. Assimakopoulos, Director of the FSU in NTUA. He has published in various referred academic journals (IMDS, IJSEKE, IJF, AEL, JCIS) and international conference proceedings (DSI, ISF, ERES). His research interests are in the fields of time series forecasting, statistics, logistics, econometrics, neural networks, geometry and software engineering. He is currently Research Officer for the EPSRC Forecasting Support Systems project, Book Reviewer for the *International Journal of Forecasting* (since 2/2002) and *Interfaces* (since 10/2004) and a member of the International Institute of Forecasters (since 6/2001).