Progress in Forecasting with Neural Networks? Empirical Evidence from the NN3 competition

Journal:	*International Journal of Forecasting*
Manuscript ID:	INTFOR_1010219
Manuscript Type:	Original Article
Keyword:	Time series, Forecasting competitions, Comparative studies, Neural networks, Automatic forecasting
Abstract:	This paper reports the results of the NN3 competition, a replication of the M3 competition, and an extension towards methods of neural networks (NN) and computational intelligence (CI), assessing if progress has been made in the past 10 years since M3. Two masked subsets of 111 and 11 empirical time series of monthly M3 industry data were chosen, controlling for multiple data conditions of forecasting horizons (short/medium/long), time series length (short/long) and data patterns (seasonal/non-seasonal). Relative forecasting accuracy was assessed using the metrics of M3, and latter extensions of scaled measures. The NN3 competition attracted 59 submissions from NN, CI and statistics, making it the largest competition of CI on time series data. Its main findings include: (a) only one NN outperformed Dampen Trend, but more outperformed AutomatANN of M3; (b) Ensembles of CI-approaches performed very well, (c) a novel 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, highlighting the ability of NN to handle complex data and routes for future research.

Progress in Forecasting with Neural Networks? Empirical Evidence from the NN3 competition

Abstract

This paper reports the results of the NN3 competition, a replication of the M3 competition, and an extension towards methods of neural networks (NN) and computational intelligence (CI), to assess if progress has been made in the past 10 years since M3. Two masked subsets of 111 and 11 empirical time series of the M3 monthly industry data were chosen, controlling for multiple data conditions of forecasting horizons (short / medium / long), time series length (short / long) and data patterns (seasonal / non-seasonal). Relative forecasting accuracy was assessed using the metrics of M3, and latter extensions of scaled measures. The NN3 competition attracted 59 submissions from NN, CI and statistics, making it the largest competition of CI on time series data. Its main findings include: (a) only one NN outperformed Dampen Trend on sMAPE, but more outperformed Automat ANN of M3; (b) Ensembles of CI-approaches performed very well, (c) a novel 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, highlighting the ability of NN to handle complex data beyond prior expectation, showing at the same time the way for future research.

Keywords: Time-series Forecasting; Empirical evaluation; NN3 competition; Artificial Neural Networks; Computational Intelligence;

1. Introduction

"*Neural networks: forecasting breakthrough or passing fad?"* Chatfield wondered back in (1993); and as of today the question remains largely unanswered. On the one hand, if we consider only the number of publications of artificial Neural Networks (NN) the answer would indicate a breakthrough: motivated by their proven theoretical properties of non-parametric, data driven universal approximation of any linear or nonlinear function, the last two decades have witnessed over 5000 publications in academic journals and conference proceedings on forecasting with NNs across a wide range of disciplines (Sven F. Crone & Preßmar, 2006). In a recent series of 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 contribute the single largest area of publications in Operational Research (2008) and one of the top 4 in forecasting journals (2006). Their growth in prominence appears to be easily justified: a wealth of publications indicate the competitive or even superior performance of NN, from early publications on single benchmark time series such as the popular airline passenger data (Faraway & Chatfield, 1998; Kolarik & Rudorfer, 1994; Tang & Fishwick, 1993), to subsets of established benchmarks from previous forecasting competitions(Sharda & Patil, 1992; Foster, Collopy, & Ungar, 1992; Hill, O'Connor, & Remus, 1996). Adya and Collopy (1998) found eleven studies that met the criteria for a valid and reliable empirical evaluation, and in 8 of these (73%) NNs were more accurate. However, their evaluative review of the experimental design and the implementation of NNs also raised concerns on the validity and reliability of the results in 37 of 48 studies (77%). For novel algorithms that are not evaluated following a rigorous experimental design, the results from an ex post evaluation (where the test data is 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 accuracy demonstrated by NNs, a different answer to Chatfield's question arises. In contrast to their optimistic publications, NNs have failed to provide objective evidence of their forecasting accuracy in large scale empirical evaluations in the form of forecasting competitions. The most renowned empirical investigation conducted - the M3

competition (S. Makridakis & Hibon, 2000) - indicated a comparatively poor performance from a single contestant. Consequently, the performance of NNs on batch forecasting fell far short of their presumed potential.

In contrast, forecasting competitions conducted in computer science and machine learning (e.g. the Santa Fe (Weigend, 1994) or EUNITE competition (Suykens & Vandewalle, 1998a)) attracted a large number of NN and CI algorithms. Although these demonstrated preeminent performance of NNs, 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. This ignored evidence within the forecasting field on how to design valid and reliable empirical evaluations (see, e.g., 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 the findings.

The discrepancy between NNs' preeminent theoretical capabilities, their promising accuracy in publications on known datasets and some real world application, in contrast to the lack of empirical accuracy in large scale ex ante evaluations has raised serious concerns in the forecasting domain on their adequacy for forecasting. As a consequence, Chatfield (quoted in 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 scepticism is founded only on a single contestant entering the last large scale evaluation of automatic forecasting.

In order to explore the persisting gap between the theoretical capabilities and empirical accuracy of NNs, we conducted a forecasting competition to provide valid and reliable empirical evidence on accuracy, evaluate and disseminate potential progress in modelling NNs and to determine the conditions under which different algorithms perform well. Our motivation to conduct yet another competition draws upon the same argument as the original M-competition by (S. Makridakis, et al., 1982): a full decade has passed since the start of the M3 competition, a decade that has seen the development of extended NN paradigms, theoretical advances in methodologies on specifying NNs and a range of novel computer intensive algorithms in CI becoming available for forecasting. In addition, it has seen substantial progress

in information technology (IT) that may facilitate the application of earlier algorithms and novel additions in large scale forecasting competitions that were infeasible before due to limited computational resources. As new alternatives exist, choices on selecting and using appropriate forecasting methods need to be revisited.

To evaluate progress in NNs, and to allow a comparison to the original M3-contestants over time, we utilised a subset of 111 monthly industry time series taken from the previous M3-data for which all 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 an algorithm performs well. The competition was open to all methods of NNs and CI. To limit biases we also allowed novel statistical methodologies and newer software releases to participate as benchmarks. NN3 attracted 59 submissions, making it the largest competitions in CI and forecasting to date. Results were evaluated using multiple error metrics, including the original symmetric mean absolute percent error (sMAPE), mean absolute scaled error (MASE) as proposed by Hyndman and Koehler (2006) and two non-parametric tests employed by Koning et al. (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 take into consideration all recommendations on how to conduct a valid and reliable empirical evaluation while balancing effort and resources to maximise the number of submissions in order to receive 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 implications for future research.

The rest of the paper is structured as follows: section two discusses prior forecasting competitions in forecasting and CI, with a discussion on the relevance of forecasting competitions to derive empirical evidence, guidelines for their setup and contrasts the findings of major competitions in forecasting and CI in order to justify the rationale for another one; As competitions in CI have not followed similar designs, the best practices derived in the experimental design of forecasting competitions are explored in more detail in order to disseminate them to a interdisciplinary readership. Sections three and four describe the

setup and the results of the empirical evaluation, taking these best practices into consideration. Section five 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

In the absence of a universal (theoretical or empirical) dominance of a single 'best method', competitions are an established means to provide objective evidence on the empirical ex ante accuracy of forecasting methods, and to guide rational choices between algorithms and methodologies for a given set of data conditions. Forecasting competitions have received substantial attention and initiated stimulating discussions within the academic forecasting community, opening up new areas of academic research (e.g. model selection and evaluation) and leading to best practices on valid and reliable competitions and experimental designs (Ord, Hibon, & Makridakis, 2000) An overview and discussion of the impact of empirical evaluations may be found in Fildes et al. (Fildes & Ord, 2002; Fildes & Makridakis, 1995). In contrast, competitions on time series prediction conducted in other domains, including computer science, machine learning, engineering and CI, have largely pursued different experimental designs that have ignored best practices on conducting competitions, limiting their validity and reliability. In order assess the empirical evidence provided in each field to date, and contrast the lack of dissemination of algorithms, applications and best-practices across both domains, we briefly summarize 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 substantial interest. Drawing upon criticism on earlier competitions on time series data (Reid, 1969, 1972; Newbold & Granger, 1974; Groff, 1973; S. Makridakis & Hibon, 1979) Makridakis et al. conducted a series of enlarged forecasting competitions where experts could submit predictions of their preferred algorithms, starting with the M-Competition on two datasets of 1001 or 111 time series, which took into account suggestions made at a meeting at the Royal Statistical Society (S. Makridakis, et al., 1982). A reduced subset of data was offered to allow participation of algorithms which required time and cost

intensive manual tuning through experts (e.g. ARIMA models required more than one hour per time series). The subsequent M2-competition (Spyros Makridakis, et al., 1993) focussed on non-automatic, real time judgmental forecasts of 23 time series, and hence holds less relevance for our quantitative competition design. None of the earlier competitions attracted any submissions of NNs or CI methods, as these did not emerge until the late 1980s, e.g. in the case of NNs through the (re-)discovery of the Backpropagartion algorithm (Rumelhart, Hinton, & Williams, 1994) and others such as CART (Breiman, 1984) although algorithms of Fuzzy Logic (Zadeh, 1965) and Evolutionary Computation (Fogel & Fogel, 1994) had already been developed. In 1998 the popular M3-Competition evaluated the accuracy of 24 algorithms on 3003 univariate empirical time series of historical data (S. Makridakis & Hibon, 2000), the largest dataset ever 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 variety of time series structures and different data conditions. All 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 on average outperformed all other methods: the software expert system Forecast Pro using automatic model selection and parameterisation of Exponential Smoothing (ES) and ARIMA models (Goodrich, 2000), and Theta, a decomposition approach combining exponential smoothing and regressing around a dampened trend line (Assimakopoulos & Nikolopoulos, 2000). Further statistical analysis by Koning *et al.* (Koning, et al., 2005) has revealed statistically significantly results for a group of four methods, also including Rule Base Forecasting (Adya, Armstrong, Collopy, & Kennedy, 2000) and Comb S-H-D an equal weighted combination of single, Holt's linear trend and Gardner's dampened trend ES methods (computed by Hibon) in the top performers.

Despite initial interest by various CI researchers, only one competitor finally submitted results to the competition using a NN methodology (Balkin & Ord, 2000). However, their fully automated methodology AutomatANN showed only moderate performance in comparison to the majority of the twenty statistical approaches, and was not ranked among the top performers (Table 15, S. Makridakis &

Hibon, 2000). The low participation of CI-approaches had been attributed to the high computational costs in building and parameterising these methods for each time series, but also the lack of methodologies that would allow automation beyond manual tuning by a human expert. The poor performance however was neither expected, nor explained sufficiently.

The conclusions drawn from prior M-competitions(S. Makridakis, et al., 1982; Spyros Makridakis, et al., 1993) have been confirmed in the M3-competition (see S. 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 relative performance between methods*, (H2) *Accuracy of a method depends upon the length of the forecasting horizon,* (H3) *Relative performance ranking of methods varies with the accuracy measure;* (H4) *Sampling variability of performance measures renders comparisons based on single time series unreliable; comparisons based on multiple time origins are recommended* (H5) *Accuracy of a combination of predictions performs well and often outperforms the individual methods;* (H6) *Sophisticated methods do not necessarily provide more accurate forecasts than simpler ones;* and. Consequently, valid competitions have developed a rigorous design, including the use of a representative number of time series, rolling origin design, the use of multiple robust error metrics, the comparison against 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 of particular relevance, as NNs and other computer intensive methods - just as sophisticated statistical algorithms like ARIMA before them - do not guarantee enhanced forecasting performance merely by their proven capabilities or theoretical features, and requires evaluation against simpler benchmarks. No competitions of similar scale have been conducted since (with the exception of the MH-competition on transportation data of varying time frequency, conducted in 2007 by Hibon, Young and Scaglione, and Athanasopoulos et al. (2009) on tourism forecasting (which is rather a conventional empirical study, as no call for participation was issued), of which results have not been published yet. This leaves the M3 as the last large scale evaluation in the forecasting domain to date, and explains the

impact and prominance of the disappointing yet unchallenged results of NN in empirical forecasting, based upon the single entry of the only CI-contestant AutomatANN (Balkin & Ord, 2000).

On the other hand, the findings of the M3 were not representative of NN. Despite a myriad of published NN methodologies, only a single methodology was evaluated, limiting the representativeness of the results for the class of NN (which encompasses a variety of feedforward and recurrent architectures) and for CI as a whole. Also, the M3 attracted no interest from the computer sciences, engineering or machine learning community were CI and other approaches of artificial intelligence had been advanced for years, introducing a sample selection bias of algorithms and dissemination of results (an omission caused by disseminating the CfP only trough media of the International Institute of Forecasters (IIF), i.e. IJF and the ISF conference). Consequently, the poor performance of a single NN-approach in M3 cannot be considered representative of the whole class of algorithms.

Furthermore, almost a decade has passed since M3, so that results may no longer reflect today's capabilities of NN. Evidence for substantial theoretical progress in NN exists, both in forecasting single time series (Preminger & Franck, 2007; de Menezes & Nikolaev, 2006; Terasvirta, van Dijk, & Medeiros, 2005) and on representative sets of empirical time series (see, e.g., Liao & Fildes, 2005; Zhang & Qi, 2005) applying methodologies for fully automated applications. These have not yet been evaluated in an objective empirical competition. Lastly, today's computational power is far superior to that available in 1997, when automated NNs were run for the M3 competition, also apparent in the expanding community regularly applying computationally intensive methods, which may enable a much wider participation. Thus, the results of the M3 may no longer be deemed representative. However, in the absence of more recent forecasting competitions its critical findings to NN stand unchallenged.

Outside the forecasting domain, competitions have been equally popular to determine the predictive accuracy of algorithms, and many more recent than the M3. Regular data mining competitions have been conducted, albeit focussed on classification tasks, including the annual competitions at the KDD conference, attracting over 1,000 contestants in 2008, or the recently closed Netflix competition

(www.netflixprice.com) in predicting movie choices, attracting 44,014 submissions (by awarding a price-money of 1 Million US$). As in forecasting, competitions on classification with CI generally follow a rigorous experimental design, adhere to established best practices for a valid and reliable experimental designs, and often address sophisticated modelling questions such as if domain knowledge allows better predictive accuracy than agnostic prediction (Guyon, Saffari, Dror, & Cawley, 2008) or to what extent in sample performance can be generalised for out of sample accuracy (Cawley, Janacek, Haylock, & Dorling, 2007).

In contrast, only a small number of competitions in the CI-domain were dedicated to time series data, often of small scale. A discussion of all competitions and their contributions is beyond this paper, but we will outline those most influential to identify fundamental 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 competition in CI to evaluate the ability of NNs in forecasting using a variety of exclusively nonlinear datasets recorded along time. Datasets were highly heterogeneous and included time series of a physics experiment recording oscillations and structural breaks of a NH5-Laser, and multivariate time series of tick-by-tick currency exchange rates, astrophysical data of light fluctuations from a white star, physiological data from a patient with sleep apnea, and music from Bach's last (unfinished) Fuge. given the heterogeneity of data conditions, datasets consisting only of a single time series, that most participants predicted only one of the datasets from a single origin (instead of - at least - all series), and that no statistical benchmarks were evaluated, the comparative work undertaken remains rudimentary and does not provide sufficient evidence to draw conclusions on the nonlinear algorithms' accuracy (S. Makridakis, 1994). The lack of validity seems particular disappointing considering that the authors were aware of the design and findings of the M-competitions, and that the late Clive Granger served on the competition's advisory board (Jadiz, 1995).

The largest CI-competition on time series to date was organised by Suykens and Vanderwalle (Suykens & Vandewalle, 1998a) in 2000 for the European Network on Intelligent Technologies for Smart

Adaptive Systems (EUNITE, www.eunite.org no longer online) attracted 24 submissions from 16 contestants (only 29% of those 56 registered to compete). It evaluated the accuracy in predicting one time series of maximum electrical load 31 days into the future from a single origin, using 2 years of half-hourly electrical load data. Data was provided by the Eastern Slovakian Electricity Corporation including explanatory variables of past temperature and holidays. The best contestant used Support Vector Regression (Chen, Chang, & Lin, 2004) to outperform contestants of CI and one 'statistical' contender using regression on decomposed time series componentts. Although all algorithms were published in a monograph (Sincák, Strackeljan, Kolcun, Novotný, & Szathmáry, 2002) it has received limited attention outside the electrical load literature.

A range of smaller competitions has been run at conferences on computational intelligence, often without publications, including the Competition on Artificial Time Series (CATS) using multiple samples of synthetic data (Lendasse, Oja, Simula, & Verleysen, 2007), held at the 2004 IEEE International Joint Conference on Neural Networks (IJCNN), the Predictive Uncertainty Competition at the 2006 IJCNN (Cawley, et al., 2007) on environmental data, the 2003 and 2006 Business Intelligence Cup on predicting time series of sugar and retail sales, organised by Richard Weber held 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 by Suykens and Vanderwalle (Suykens & Vandewalle, 1998b) held at the International Workshop on Advanced Black-Box Techniques for Nonlinear Modeling in 1998 on synthetic data (see also McNames, Suykens, & Vandewalle, 1999). Table 1 provides a structured summary of prior time series competitions, both in forecasting and CI, and contrasts differences in the experimental design in both domains to assess their contributions.

Table 1: Competitions design in Forecasting and Computational Intelligence

Competition Name	Data Properties			# of algorithms		Conditions evaluated♣			
	# of series	# of observ.	Data type	Statistics	NN & CI	Multiple Metrics*	Multiple horizons*	Muliple types*	Time Frequency♦
Empirical Accuracy[1]	111	?	univariate	22	0	X	X	X	Y, Q, M
M1[2]	1001	15 to 150	univariate	24	0	X	X	X	Y, Q, M
M3[3]	3003	20 to 144	univariate	24	1	X	X	X	Y, Q, M
MH/Transport[4]	278	19 to 1502	univariate	3-10	1	X	X	X	Y, Q, M W, D, H
Santa Fe	6	1,000 to 300,000	2 univariate 4 multivariate	0	14	-	-	-	Synthetic
KULeuven [6]	1	2000	Univariate	0	17	-	-	-	Synthetic
2001 EUNITE [7]	1	35,040	multivariate	1	24	-	-	-	30m
ANNEXG[8]	1	1,460	Multivariate	0	12	-	-	-	360m
BI Cup 2003 [9]	1	365	Multivariate	0	10	-	-	-	D
CATS 2005 [10]	1	4,905	univariate	0	25	-	-	-	Synthetic
Predictive Uncertainty, [11]	4	380 to 21,000	1 univariate 3 multivariate	0	20	-	-	-	Synthetic D, 3D
BI Cup 2006 [12]	1	1325	1 univariate	0	?	-	-	-	15m

[1] (S. Makridakis & Hibon, 1979); [2] (S. Makridakis, et al., 1982); [3] (S. Makridakis & Hibon, 2000); [4] unpublished; [5] (Weigend, 1994); [6] (Suykens & Vandewalle, 1998a); [7] (Suykens & Vandewalle, 1998b); [8] (Dawson, et al., 2005); [9] unpublished; [10] (Lendasse, et al., 2007); [11] (Cawley, et al., 2007); [12] unpublished

* X indicates the use of multiple error metrics, multiple forecasting horizons and data types, - indicates the absence

♦ Y=yearly data; Q = quarterly data; M=monthly data; W=weekly data; D=daily data; H=Hourly data; m=minutes; ? indicates non disclosed information

Few similarities emerge, bare one: both domains favour and evaluate almost exclusively its preferred family of algorithms: forecasting competitions evaluate only statistical methods (and expert systems configuring these), with the exception of the sole NN contender at the M3 and the unpublished MH-competitions, while CI-competitions have not evaluated statistical algorithms at all.

Another commonality shared across domains remains the evaluation across a single time origin (due to the time consuming nature of stepwise release of data), with the exception of CATS, which provided 5 origins of a continuous time series simultaneously (and the M2 not mentioned here). However, as the test data was withheld as 5 gaps of one time series, contestants used heuristics to connect end and beginning of the series in addition to fore- and backcasting in the sense of imputing consecutive missing

values rather than forecasting. Furthermore, as the data was synthetic and not empirically motivated, the competition holds limited relevance to assess accuracy in time series prediction.

More noticeably, differences and discrepancies in the design of all CI-competitions become evident, which seriously impairs their contribution. While all competitions of the forecasting domain used representative sample sizes of hundreds or even thousands of time series, CI-competitions evaluated accuracy only on a single time series. Those competitions that evaluate multiple time series, such as the Santa Fe and the Predictive Uncertainty competition, did so in distinct groups 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 permitted somewhat replicable results. In contrast, the same authors applied different methodologies for each dataset, even within a competition, leading to distinctly different models and preventing any comparisons.

Also, none of the CI-competitions compare results against established benchmark methods, be they Naive methods, simple statistical benchmarks or non-statistical methods in the same family of algorithms, e.g. a simple single layer MLP 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, in particular the use of multiple similar time series, have been ignored by the CI community. Makridakis' (2000) original criticism holds: just as theoretical statisticians before them, researchers in NN have concentrated efforts on building more sophisticated models without regard to assessing their accuracy and objective empirical verifications, successfully ignoring the strong empirical evidence of the M-competition and the ground rules it has laid out on how to asses them. This substantially limits the validity and reliability of the evidence from CI-competitions, which can not challenge the authority of the earlier M3-competition which failed to show benefits of NNs in terms of accuracy.

With competitions in both domains limited in their coverage of algorithms, results of the M3 competition not, or at least no longer representative of CI, and more recent competitions in CI unreliable, the gap between the theoretical capabilities and empirical accuracy of NNs remains unexplored. In order

to evaluate potential progress in the development of NN and CI-approaches, a new competition seemed most suitable to provide valid and reliable empirical evidence on their accuracy, the conditions under which different algorithms perform well, and to disseminate potential progress in modelling NNs. For the sake of consistency, it seemed a natural choice to use the original setup and a homogeneous subset of the M3 data in the form of a replication, which will be discussed in detail in the next session.

In reviewing table 1, we note a further discrepancy in data conditions explored. Forecasting competitions have been focussed exclusively on low time series frequencies of yearly, quarterly, or at most monthly data in a univariate context. Although this adequately reflects the persisting theme of operational forecasting set out by Makridakis' series of M-competitions (S. Makridakis, et al., 1982), it does not allow us to generalise these findings to unseen data conditions, in particular towards high-frequency data of weekly, daily, hourly or shorted time intervals (and the resulting longer time series) on which NN have been evaluated (represented in the dissimilar data conditions of CI competitions). It seems Armstrong's (2006) criticism of NN is not only limited in evidence due to a single contestant of M3, but more importantly limited due to a substantial omission of empirical data conditions, for which - following his arguments - no evidence exists. As the omitted conditions represent those on which NN are regularly - and successfully - employed in practice, it might also yield an explanation for the simultaneous scepticism and euphoria re NN between disciplines, and provide the motivation to close the gap in more representative competitions on novel data conditions.

3. Design and Organisation of the *NN3 competition*

3.1 Objectives

Following the rationale provided, we sought to explore the current forecasting performance of NN and CI methods. The M-competitions explicitly focussed on a particular set of data conditions, which Makridakis coined in the context of forecasting for operations. To assess progress in relation to M3, we will keep this tradition and constrain our competition to the operations context of monthly industry data, although other conditions for forecasting with NNs exist that may promise different results.

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

3.1 Datasets, Working Hypothesis and Data Conditions

The M3 dataset yielded substantial insights, but proved challenging for CI-methods: the sample of 3,003 time series was large given the computational resources available in the 1990s, and the heterogeneity of time series frequencies and data domains required multiple candidate methodologies (and human intervention at many stages), which limited automation and may have prevented many computationally intensive methods to participate. In order to attract a representative number of contestants and algorithms we sought to limit the number of time series used and heterogeneity of data conditions (and resulting insights), yet enough to derive reliable results. A set of 111 time series was selected randomly from the M3 subset of monthly industry data, representative of the M-competition's original focus of forecasting for operations. Data from a single time frequency was chosen in order to limit the competition's complexity to a single methodology for monthly data We also hoped that a sample would further mask the origin of the NN3 competition data and prevent biases in results through prior knowledge.

Four working hypothesis 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 NN, we allowed participation on two (disguised) datasets

of different size. Contestants were asked to predict either a reduced dataset of 11 or a complete set of 111 time series (which included the reduced set) as accurately as possible. As a fully automated methodology, as required for operational forecasting, could be applied to datasets of larges size, more submissions on the reduced dataset would indicate the need for manual tuning, limitations of the automation or extremely computational intensive approaches, indicating the need for further research in methodologies.

A second working hypothesis seeks to assess the relative accuracy of NN and statistical approaches on longer forecasting horizons, where statistical algorithms have outperformed NN in past studies (Hill, et al., 1996). Each contestant is required to predict multiple forecast y_{t+h} of $h = (1,...,18)$ steps into the future, which are later analysed for short (1-3 months), medium (3-12 months) and long (13-18 months) forecasting horizons to assess differences in the results.

Two further working hypothesis address the data conditions under which different methods perform well. First, following the widespread belief that NNs are data hungry and require long time series, balanced stratified samples were taken by time series length n, resulting in 50 long ($n > 100$) vs. 50 short time series ($n < 50$). Second, to evaluate recent publications which conclude that NNs cannot forecast seasonal time series (Nelson, Hill, Remus, & O'Connor, 1999; Zhang & Qi, 2005; Curry, 2007) stratified samples were taken to reflect time series patterns of 50 seasonal vs. 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 to derive (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 to the complete set. The reduced dataset contained 11 time series which we classified as difficult to forecast, 4 of which were seasonal and the remaining 7 non-seasonal (including outliers, structural breaks), to identify if non-automated methodologies are capable of predicting across different data conditions. Table 1 summarises the time series conditions of both datasets.

Table 1. NN3 Datasets with data conditions of time series length and seasonality.

	Complete Dataset				
			Reduced Dataset		
	Short	Long	Normal	Difficult	SUM
Non-Seasonal	25 (SN)	25 (LN)	4 (NN)	3 (DN)	57
Seasonal	25 (SS)	25 (LS)	4 (NS)	-	54
SUM	50	50	8	3	111

The conditions within the reduced dataset were not meant to be statistically explored due to the limited amount of series (3-3-4) that would result in insignificant results. Nonetheless, findings on the reduced dataset would still yield results of increased validity and reliability in comparison to previous CI competitions using only a single time series.

3.2 Evaluation and Error Metrics

In order to evaluate the performance of the NN3 submissions, and to ensure consistency with the results of the M3-competition, we employed three metrics used in the M3 competition (all averaged across all series): Symmetric Mean Absolute Percent Error (sMAPE), Median Relative Absolute Error (MdRAE) and Average Ranking (AR) (S. Makridakis & Hibon, 2000). We also estimated two non-parametric tests proposed by Koning *et. al.* (2005) in a follow up analysis: Analysis of the Mean (ANOM) and Multiple Comparisons to the Best (MCB), both using the average ranking as criterion. Finally, in order to align with the current literature, we have calculated the Mean Absolute Scaled Error (MASE) as proposed by Hyndman and Koehler (2006). In order to ensure a consistent computation of errors, we collaborated with one of the original investigators of the M3 competition who computed all metrics as in the original competition.

Average sMAPE was announced beforehand as the metric used to determine the "winner", in order to allow those CI-methods capable of using non squared error cost functions to align their approaches with the final criterion (see, e.g., the discussion with Zellner (1986) following the M3). Despite the shortcomings of sMAPE (Goodwin & Lawton, 1999), it was chosen as it served as the

primary criterion in the M3 competition and to make the NN3 results accessible to practitioners, where the predominant error metric is MAPE. As the NN3 time series contained no zero, negative or small values, we anticipate only limited potential for bias.

3.2 Methods and Benchmarks

The competition invited contributions from all areas of 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 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 NN3.

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

In order to assess 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, the collaboration with one of the original investigators 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 (B14), Holt's linear trend (B15) and Gardner's dampened trend ES (B16), and their combination to Comb S-H-D (B17). Predictions for Theta (B7) were recomputed by the organisers, using an identical setup as in the M3 competition.

In addition, we computed various CI-benchmarks to provide additional levels of comparison of the entries, including a Naïve Support Vector Regression approach (S. F. Crone & Pietsch, 2007, B01)

and a Naïve Mulitlayer Perceptron (B02), which replicate novice model building mistakes as a lower bound. A novel NN-extension of the successful Theta-Method named Theta-AI (B08) by Nikolopoulos and Bougioukos that determined optimal nonlinear weights for the Theta-lines was withdrawn in order not to bias results, as it was based on Theta that was known a-priori to perform well on the NN3 data.

3.4 Process of organising the Competition

The competition design and feasibility was pre-tested in a small scale trial competition using two time series (held at the 2005 ISF, San Antonio, USA), which facilitated feedback by 9 contestants and external experts, including a panel of IIF judges for a grant. The NN3 competition was first announced at the ISF 2006 in Santander and was open for eight months from October 2007 to May 2008. Contestants were required to submit predictions and a full description of their methodology, which are both published on the competition website in order to facilitate replication. Contestants could withhold their identity prior to disclosing the final results, in order to limit negative publicity for software vendors and participants. Following submission, each methodology was classified to distinguish CI-contenders eligible to "win" the competition (identified by consecutive IDs C01-C59) from other 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). Some contributors requested to withhold their identity, but their results are included in the tables with their original submission IDs to ensure consistency with previously disclosed results.

In order to limit sample selection biases in the participation through timing, location and audience of the conference where the competition was run, multiple special sessions were advertised and conducted at conferences throughout 2007, and across the domains of forecasting, computational intelligence and electrical engineering, data mining and machine learning. These included the 2007 International Symposium on Forecasting (ISF'07), in New York, USA, the 2007 IEEE International Joint Conference on Neural Networks (IJCNN'07) in 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 NN*3 competition*

4.1 Results on the complete dataset

The competition attracted 46 contestants using NN and CI-methods and 17 benchmark methods, making it the largest empirical evaluation in NN, CI and forecasting to date.

All 46 contenders submitted predictions for the reduced set of the 11 time series, but only 22 contenders predicted the 111 time series of the complete set. With less than half of the contestants (47%) able to predict more than 11 series, it provides evidence that the need for manual tuning and human intervention still dominates most methodologies. This reflects our experience, both in academia and practice, and is supported by the lack of commercial CI-software for automatic time series forecasting.

Table 2 presents the names of the NN3 contestants, a summary of the algorithm and a consecutive ID (assigned during the competition) that provided forecasts for the 111 series of the complete dataset. A discussion of all submissions is not feasible within the scope of this paper, so we will limit our discussion to those methods that have stood out on some or all of the data conditions we analysed. Detailed descriptions of all methodologies, including the 24 contenders that provided forecasts only for the 11 series of the reduced dataset, are available on the NN3 competition website www.neural-forecasting-competition.com for further reading.

Table 2. NN3 participant IDs, names and method descriptions for the complete dataset of 111 series

Code	Classification	Name	Description
C03	NN/CI Contender	Flores, Anaya, Ramirez, Morales	Automated Linear Modeling of Time Series with Self Adaptive Genetic Algorithms
C11	NN/CI Contender	Perfilieva, Novak, Pavliska, Dvorak, Stepnicka	Combination of two techniques: fuzzy transform and perception-based logical deduction
C13	NN/CI Contender	D'yakonov	Simple kNN-Method for Times Series Prediction
C15	NN/CI Contender	Isa	Growing fuzzy inference neural network
C17	NN/CI Contender	Chang	K-Nearest-Neighbor and Support-Vector Regression
C20	NN/CI Contender	Kurogi, Koyama, Tanaka, Sanuki	Using First-Order Difference of Time Series and Bagging of Competitive Associative Nets
C24	NN/CI Contender	Abou-Nasr	Recurrent Neural Networks
C26	NN/CI Contender	de Vos	Multi-Resolution Time Series Forecasting Using Wavelet Decomposition
C27	NN/CI Contender	Illies, Jäger, Kosuchinas, Rincon, Sakenas, Vaskevcius	Stepping forward through echoes of the past: forecasting with Echo State Networks
C28	NN/CI Contender	Eruhimov, Martyanov, Tuv	Windowed wavelet decomposition and Gradient Boosted Trees
C30	NN/CI Contender	Pucheta, Patino, Kuchen	Neural Networks-Based Prediction Using Long and Short Term Dependence in the Learning Process
C31	NN/CI Contender	Theodosiou, Swamy	A hybrid approach: Structural Decomposition, Generalised Regression Neural Networks and Theta model
C36	NN/CI Contender	Sorjamaa, Lendasse	A non-linear approach (Self-Organized Maps) combined with a linear one (Empirical Orthogonal Functions)
C37	NN/CI Contender	Duclos-Gosselin	Fully-recurrent neural network learned with M.A.P (Bayesian), leventberg and genetic algorithm.
C38	NN/CI Contender	Adeodato, Vasconcelos, Arnaud, Chunha, Monteiro	Multilayer Perceptron Networks
C44	NN/CI Contender	Yan	Multiple-Model Fusion for Robust Time-Series Forecasting
C46	NN/CI Contender	Chen, Yao	Ensemble Regression Trees
C49	NN/CI Contender	Schliebs, Platel, Kasabov	Quantum Inspired Feature Selection and Neural Network Models
C50	NN/CI Contender	Kamel, Atiya, Gayar, El-Shishiny	A Combined Neural Network/Gaussian Process Regression Time Series Forecasting System
C51	NN/CI Contender	Papadaki, Amaxopolous	Dynamic Architecture for Artificial Neural Networks
C57	NN/CI Contender	Corzo, Hong	Global neural networks ensembles with M5 prime model trees
C59	NN/CI Contender	Beliakov & Troiano	Time series forecasting using Lipschitz optimal interpolation
B00	NN/CI Benchmark	Automat NN	Original M3 Submission for the M3 competiiton by Balkin & Ord
B01	NN/CI Benchmark	Naive SVR	A naïve Support Vector Regression forecasting approach by Crone & Pietsch
B02	NN/CI Benchmark	Naive MLP	A naïve Multiple Linear Perceptron by Crone
B03	Stat. Benchmark	ForecastPro	The Expert method of the renown software. Version XE 5.0.2.6.
B04	Stat. Benchmark	Naive	The Naïve method without any seasonality adjustment
B05	Stat. Benchmark	Autobox	Forecast provided directly from David Reily with Version 6.0 of the software (June 2007)
B06	Stat. Benchmark	Census - X12 ARIMA	Official Cencus method, prepared by McElroy (www.census.gov/ts/x12a/v03/x12adocV03.pdf)
B07	Stat. Benchmark	Theta	Exponential Smoothing with Decomposition, prepared by Nikolopoulos, Version TIFIS CM3 1.0
B14	Stat. Benchmark	Single ES	Original M3 Benchmark for the M3 competiton as programmed by Dr M Hibon
B15	Stat. Benchmark	Holt ES	Original M3 Benchmark for the M3 competiton as programmed by Dr M Hibon
B16	Stat. Benchmark	Dampen ES	Original M3 Benchmark for the M3 competiton as programmed by Dr M Hibon
B17	Stat. Benchmark	Comb S-H-D ES	Original M3 Benchmark - Equal Weighted combination of Single, Holt and Dampen Expl Smoothing
B09	Stat C	Wildi	An Adaptive Robustified Multi-Step-Ahead Out-Of-Sample Forecasting Combination Approach
B10	Stat C	Beadle	Composite Forecasting Strategy Using Seasonal Schemata
B11	Stat. Contender	Lewicke (Parabolic Systems)	Paracaster Software -fitting equations consisting of a trend plus a series of sinusoidal error terms
B12	Stat. Contender	Hazarika	Decomposition to Random Sequence Basis Functions and a Temperature-Dependent SOFTMAX Combiner
B13	Stat. Contender	Njimi, Mélard	Automatic ARIMA modelling, using TSE-AX
C103	NN/CI Benchmark	Ensemble of Best 3 NN/CI	Equal Weighted combination of C27, C03, C46 prepared post competition by Hibon
C105	NN/CI Benchmark	Ensemble of Best 5 NN/CI	Equal Weighted combination of C27, C03, C46 ,C50, C13 prepared post competition by Hibon

Table 3 shows the results on the complete dataset as average sMAPE, MdRAE, MASE and AR across 111 time series and 18 forecasting horizons. Relative ranks by error measure are given across all methods and for CI contestants alone (NN C).

Table 3. NN3 errors and ranks of errors on the complete dataset

		Average Errors				Rank across all methods				Rank across NN C				
		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	Naive	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

Stat.C = statistical contender; NN C = NN/CI contender; Stat.C = statistical contender; NN C = NN/CI contender

Has progress been made, within CI and in comparison to statistical methods? Regardless of accuracy, the ability of 22 contestants to predict a large number of time series with CI indicates

unsurprising progress, both on the development of methodologies that facilitate automation and / or increased computational resources.

On accuracy, the top 10 algorithms indicate some progress in accuracy, but not enough to proclaim Chatfield's breakthrough for NN. Unsurprisingly, the top contenders of the M3 monthly data are also ranked high on 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 joined the best performers. These algorithms will be briefly introduced as they have not been published elsewhere.

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

More in line with the competition theme, the team 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 neural networks with random connections in a reservoir of hidden neurons. Each time series was categorised into 6 clusters by time series length (ignoring different data properties in each cluster and the unrelated nature of most series, that was unknown to the contestants) and decomposed into its time series components using X-12-ARIMA. 500 ESNs with reservoir sizes of 45 to 110 neurons were trained on pooled clusters of time series for each component, and their predictions per time series first recombined across components, and then averaged across all 500 ESNs using the mean. The approach successfully outperformed all statistical benchmarks with the exception of Theta, the top-performer of the M3 monthly data, which constitutes a substantial achievement and progress in CI-model building.

Three more CI-contenders outperformed AutomatANN and breached 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 an SARIMA model form, parameter bounds and parameters for each time series. Chen and Yao (C46) employ an ensemble of 500 CART-regression trees built upon bootstrap sampling of the data and random subspace sampling of features. D'yakonov (C13) used a simple k-nearest neighbour (k-NN) method with 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 frequency, not just 111 monthly series, and that it was developed a decade ago, it has legitimated its representative performance of NN on monthly data. However, that 4 (18%) of the submitted CI-approaches outperforming AutomatANN, shows progress in research through Illies et al. (C27), Flores et al. (C03), Chen et al. (C46) and D'yakunov (C13). In addition, many CI-contenders achieve accuracy only marginally lower than AutomatANN, indicating that in contrast to the time of M3 now many researchers are capable of predicting large scale a level of accuracy similar to AutomatANN.

Despite enhanced performance of a few, the field of submissions in CI remains wide, and many fail to outperform basic CI-benchmarks of a naive multilayer perceptron (B02) or SVR (B01). Some methods even fail to outperform the Naive benchmark (B04), indicating the need of enhanced understanding of empirical evaluations for internal benchmarking (ideally prior to a potentially embarrassing competition performance).

It should be noted though, that statistical approaches - simple or complex - are also not a panacea: other novel statistical contenders such as X-12 (B06) B10 and B11 perform at best average, with B12 even failing to outperform the Naive (B04). Also, weaker contestants of the M3 were not included as benchmarks, biasing the perception of relative ranking of CI contenders and benchmarks to the

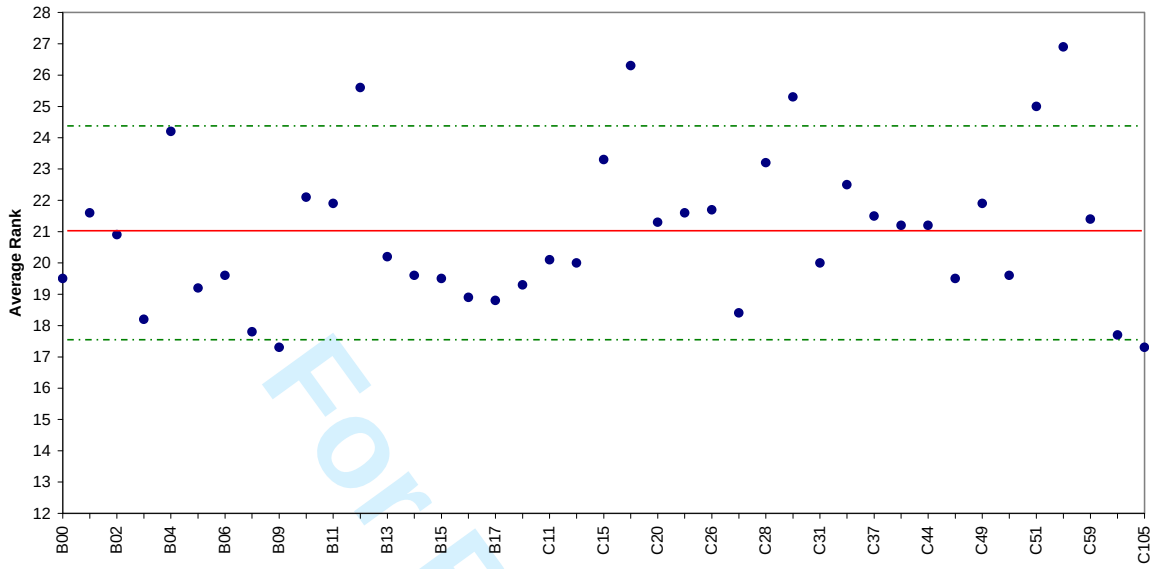
disadvantage of NN; in fact, many contestants had outperformed established methods of the M3, but we were most interested in progress at the top of the field in comparison to AutomatANN.

As for 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. Following 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 using the arithmetic mean. Both CI benchmarks performed outstandingly well: with an sMAPE of 14.89 the ensemble of the top 3 CI-algorithms would have ranked overall third - 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 5 (C105) would have ranked 2nd only to Wildi (B09), outperforming Theta and all other methods. Although this ex-post combination of best methods does not represent a valid "ex ante" accuracy (it may be overcome with a quasi-ex ante model selection), it underlines once more the potential of combining heterogeneous predictions. While Illies' et. al (C27) performance obviously contributed significantly to the performance of the two CI-ensembles, the combination increases accuracy beyond that of each individual contender, an effect well documented (in addition to a second benefit of decreased error variance). More importantly, by including 5 instead of the top 3 CI-algorithms, essentially introducing more inferior forecasts into an ensemble, the overall accuracy was increased even further. Therefore, it seems, further increases in accuracy are feasible for CI using simple means, well documented in the forecasting domain.

4.2 Significance of the findings

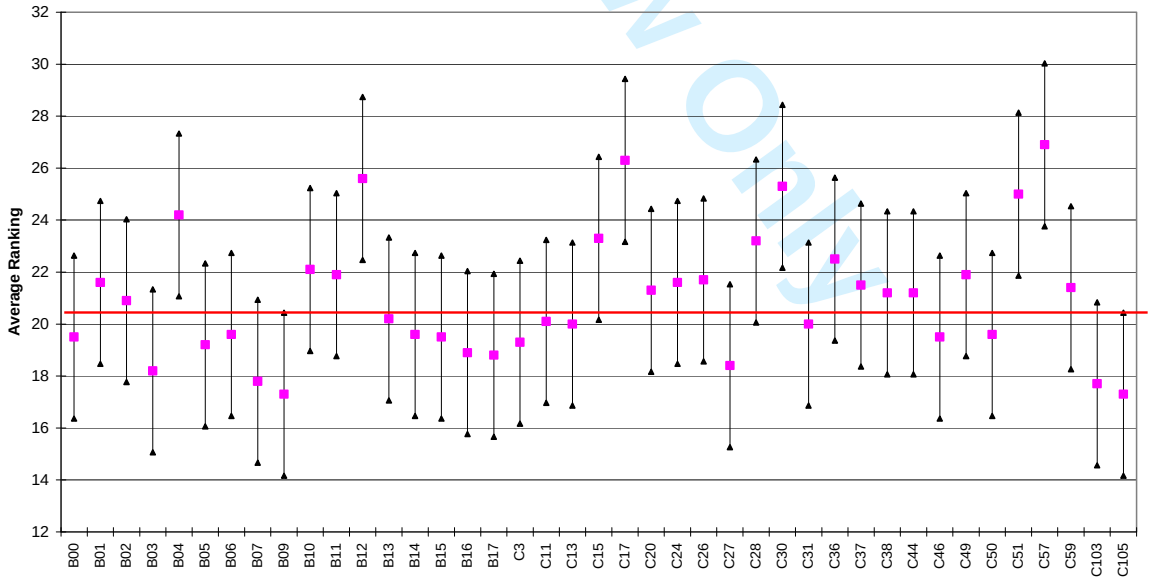
Regardless of the recent and vivid discussion about the statistical significance of published accuracy results within the forecasting community (Armstrong, 2007a, 2007b; Goodwin, 2007) we computed two non-parametric tests, replicating the analysis by Koning *et. al.* (2005)on the M3: Analysis of the Mean (ANOM) and Multiple Comparisons to the Best method (MCB), both based upon average ranks of 41 methods (including both CI-ensembles) over 111 series and 18 horizons (see fig. 1 and 2).

Figure 1. Analysis of Means on the complete dataset



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

Figure 2. Multiple Comparisons with the Best on the complete dataset



MCB confirms similar findings as ANOM: the ensemble of the top 5 (C105) and Wildi (B09) are identified as the two best approaches; in comparison to them, the same four CI approaches (C17, C30, C51, C57) and a statistical contender (B12) plus the Naïve (B04) are significantly worse than the best.

Despite limited differences in statistical significance, it is worth mentioning that even a small accuracy gain, e.g. of 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). As an indication of the tests' limitations, the Theta method - previously better than other algorithms in the competition - is no longer significantly better, indicating the sensitivity of the test to sample size and structure (as of all tests), adding further to the discussion of tests.

4.3 Analysis of Data Conditions

Next, we analyse the data conditions under which different algorithms perform well. As it is infeasible to present all 24 tables of ranking per error measure and data subset, Table 4 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 5 shows the top 5 performers by sMAPE across different forecasting horizons. In order to facilitate replication and external analysis of the results, all tables of sMAPE (Tables 6-15), MdRAE (Tables 16-20), MASE (Tables 21-25), AR for all methods and for CI contenders separately (Tables 26-27), and ANOM (Table 28) and MCB based upon AR (Table 29) will be provided online on the competition and journal homepage.

Table 4. NN3-competition results across data conditions on sMAPE, MdRAE, and MASE

Error Metrics	Complete dataset (incl.reduced)	Data Conditions					Combined Data Conditions			
		Reduced dataset	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

On the complete dataset (first column of table 4), the ranking of all algorithms in is identical to results provided in table 3, identifying the top performers of NN3 by sMAPE, id est: Wildi (B09), ensemble of top 5 CI (C105), Theta (B07) in a draw with the ensemble of top 3 CI (C103) and Illies et al. (C27). In comparison, different algorithms performed well on the reduced dataset of 11 harder to forecast time series: the statistical expert system Autobox (B05) ranks 1st by sMAPE, playing out its strengths in modeling pulse interventions, level shifts, local time trends and seasonal pulses. ForecastPro (B03) ranks 2nd and Theta (B07) ranked 4th. Two new CI-contestants enter the top 5 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 trained separately to predict for a forecasting horizon h = 1, 2, ..., 18 with three distinct parameters settings, recombining the predictions to one trace forecast, and combining the predictions of the three architectures, hence called 'multiple model fusion'.

On MdRAE other CI-contenders enter the top 5, Adeonato et al. (C38) using ensembles of 15 MLPs and Perfilieva (C11), forecasting using fuzzy transforms, indicating that the results on only a few

series do not yield the same level of reliability across error measures as on the complete set. It does however show the potential that specialised statistical and CI-algorithms tuned (or robust) to particular time series properties can outperform other approaches, but at the same time questions the ability of these CI methodologies to generalise on 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 derives 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 this family of methods for these typical data conditions. The related Theta (B07) appears on all aggregate conditions but not its combinations, verifying its robustness across many data conditions by a consistent level of accuracy, but not winning any particular category.

For CI, multiple CI-contenders enter the top 5 on different conditions, while the M3 benchmark AutomatANN (B00) is absent across all categories and metrics (with the exception of sMAPE on *Long+Non-seasonal* data). In the light of earlier research, the most striking result of NN3 comes in the *Short+Non-seasonal* subset, judging by recent publications 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, we speculate as an effect of training on pooled clusters of time series. CI ensembles C103 and C105 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 (unsurprising as C27 was contained in them). From the remaining CI competitors only Kamel (C50) made an appearance in the *Short+Seasonal* category, combining MLPs with Gaussian Process Regression.

These results challenge prior beliefs in NN-modeling that a significant amount of historic observations are a prerequisite for sufficient initialization, training, validation, evaluation and generalisation of CI approaches (see, e.g., 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 which had identified problems in predicting seasonal time series with NNs and proposed prior deseasonalisation (e.g., Zhang & Qi, 2005). While these results reveal no insight into the reasons for the increased performance, it does demonstrate that novel CI-paradigms can yield competitive performance beyond their traditional application domain and that systematic replications of earlier studies should be conducted to challenge prior findings. However, the majority of CI approaches is absent across datasets and conditions, one the one hand demonstrating consistent results, but on the other indicating that only few algorithms have the capacity to perform well.

Table 5. NN3 results of sMAPE across short, medium, long and all forecasting horizons

	Complete dataset (incl.reduced)	Combined Data Conditions				
		Reduced dataset	Short		Long	
Error Metrics			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

Results across forecasting horizons seem to confirm earlier findings by Hill et al(Hill, et al., 1996): for short term forecasting, methods of ES (B07 and B09) appear to perform best, but with an increasing forecasting horizon the CI approaches take the lead, although it remains unclear whether the contribution stems from the forecast combinations in ensembles, or the underlying method's performance increasing with the horizon.

However, for CI the accuracy achieved across horizons show surprising consistency. On the complete dataset those contenders ranked highly overall are also ranked consistently amongst the top 5 across all horizons of short, medium and long forecasts, with only minor changes in ranks. This is confirmed across data conditions, where the relative performance remains consistent across different horizons: CI-methods perform well for short time series with and without seasonality across all forecasting horizons, in particular Illies' (C27) and the ensembles C105 and C103. Similarly for long time series, methods of ES perform consistently well across all horizons, again without significant changes in ranks. The only noticeable change appears for long & non-seasonal data, where ES dominates on short, and CI on long horizons. It stands to argue, that results across horizons for a particular data subset, remain more stable than expected given prior findings. As an example, Wildi's (B09) approach, which is optimised specifically for multiple horizons of a trace forecast, performs consistently across all horizons for short & seasonal time series, as was intended by the algorithm.

5. Discussion

The NN3 competition contributed empirical evidence in the tradition of the M-competitions, with a particular emphasis on extending the findings of M3 towards a current and complete range of CI-methods. In that the NN3 has succeeded, attracting contestants from all mayor paradigms, including feed-forward and recurrent NN, Fuzzy Logic, Genetic Algorithms and Evolutionary Computation and hybrid systems. In addition, the results of this replication and extension of M3 allow us to evaluate the six hypothesis of the original M-competition (see section 2), and to determine if the findings conform to established wisdom, or add novel insights to the body of knowledge. First, we will review hypothesis H2,

H1 and H3 as they allow us to assess the similarity of M3 and its replication, and allow a verification of the NN3 competition design.

(H1) '*Data characteristics determine relative performance?'* The results of N3 across data conditions (table 4) confirm that of earlier M3. Data characteristics substantially influence the relative performance of algorithms in statistics and CI alike. Here, NN3 contributes to the discussion by providing objective evidence that NNs are capable to predict seasonal time series (in contrast to, e.g., Zhang & Qi, 2005), and to predict short time series (in contrast to, e.g., Hill, et al., 1996) accurately, contrary to prior publications and indicating the need for further research.

(H2) *'Accuracy depends upon forecasting horizon?'* Across forecasting horizons (table 5), relative performance varies, different methods perform best in different horizons, confirming the finidngs of M3. Also, the efficacy of CI-methods in comparison to statistical methods increases for longer forecasting horizons, as identified in prior studies (Hill, et al., 1996) However, for the best algorithms in CI the accuracy remained almost consistent for increasing forecasting horizons. Further research is needed to determine if those methods incorporating trace errors in their modelling (e.g. B09, C44) can overcome this limitation, as first indications seem to suggest.

(H3) *'Performance ranking varies with metric?'*. NN3 rankings based upon sMAPE, MdRAE, MASE and AR result in a different relative performance of algorithms, across all datasets and data conditions (see table 4). However, many methods in the upper deciles of the field perform consistently well on multiple metrics, and vice versa, building increasing confidence on their relative performance and predictive capabilities.

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

(H5) *Combinations outperform individual methods?* Reviewing common properties of the top performers (table 4), the success of combinations stands out. With the exception of the five original submissions to the M3 (ForecastPro, Autobox. SES, DES, and HES) the three leading statistical methods in the top 10 use forecast combinations (most notably Wildi (B09) across all conditions, Comb S-H-D (B017 for long series, and also Theta (B07) which essentially employs a weighted forecast combination of 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 5 (C105), and Ensemble of the top 3 NN (C103)). C105 and C103 dominate our results, but also signify the effect of increasing coverage in the ensemble (i.e. heterogeneity of the base learner), which warrants more research effort across disciplines. As sophisticated 'ensembles' in the form of boosting, bagging, arcing etc. are more widespread in CI classification than in statistical modelling and time series prediction in particular, we see potential for cross-disciplinary research.

(H6) *Sophisticated methods are not better than simpler methods?* Seeing the majority of CI approaches have failed to outperform Simple ES (B14), and four performed worse than Naïve (B04) (see Tables 6-15) we could not disagree. However, NN3 has introduced a novel univariate method, and provided the evidence of its capability to outperform established statistical benchmarks, including the respective winners of 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, combining various innovations in estimation and model selection to automatically tune it to the data. This stands in conflict to H6 and to common belief that complex methods do not significantly outperform simple ones. Similarly, NN3 provide evidence that some complex methods are capable to outperform all statistical methods of the M3, showing a substantial improvement in accuracy. To provide further evidence, with the submissions of Wildi, Theta, ForecastPro, Autobox for statistics, and Illies and Flores representing CI, 4 of the top 5 (80%) and 6 of the top 10 methods (60%) could be classified

as complex methods. As such, we have provided objective evidence that does not support H6. Short of refuting H6 on the basis of a few algorithms, we seek to reverse it to challenge established wisdom: (H6.b) *Simple methods are not better than sophisticated methods.* Despite identical content, the prior connotation of H6 all too easily suggested that no benefit arose from sophistication, and allowed misinterpretation that *'Simpler is better'*. We conclude that complex methods of CI/NN and statistics have caught up, and overall simple statistical methods can no longer claim to outperform CI methods without proper empirical evaluation.

As every empirical study, findings hold only for the properties of the empirical dataset provided, and as such the NN3 competition did not aim to be representative of all data properties in operational forecasting. Still our competition was prone to certain limitations and biases that must be critically reviewed. These include the obvious shortcomings that are endogenous to most competitions: no rolling origin design (due to the challenge of organising such as setup), limited representativeness of datasets in size, structure and heterogeneity, and the exclusion of certain error metrics that assess the final impact on decision making, i.e. inventory costs arising from operational forecasting (Timmermann and Granger, 2004). As in prior M competitions, our assessment considered only empirical accuracy and neglected computational resources required, an important aspect in forecasting for operations. As Expert software systems such as Autobox and ForecastPro contain much faster forecasting engines than CI (i.e. we received the submission of Autobox almost instantaneous after the release if 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 algorithms may exist. In tailoring the NN3 to algorithms of NN and CI we may have biased the sample of contestants by attracting more CI contestants than those from statistics. Furthermore, the majority of submissions came from researchers in CI, while professionals and (possibly advanced) software companies in NN, CI or AI (e.g. Siemens, Alyuda, NeuroDimensions, and SAS) chose not to participate despite personal invitations. Also, more participation from econometrics and forecasting software vendors active in forecasting for operations (e.g.

SAP, Oracle, John Galt, Smart etc.) would have increased the validity of results. However, we tried to be as objective and inclusive as we could, taking into consideration the design suggestions of prior competitions and reaching out to the communities omitted before. Therefore we are confident that NN3 provides a more comprehensive and up-to-date assessment of the performance of CI-methods in predicting monthly time series than M3, as well as more valid and reliable evidence than that of prior competitions in CI. One fundamental flaw - grounded in the nature of a replication - exists 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, it reminds us of the importance of true ex-ante evaluations on unknown data to avoid any data snooping for future competitions.

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 accuracy of over 60 forecasting algorithms, the largest assessment of methods on time series data to date. Ex ante accuracies were evaluated on 111 or 11 empirical time series using multiple established error metrics and following a rigorous competition design; conditions examined include the presence of seasonality, the length of the series, and the forecasting horizon.

The NN3 objective, extending the M3 competition towards NN and CI algorithms, was successfully achieved in attracting 46 CI contestants and novel statistical benchmarks, making it the largest empirical evaluation in NN, CI and forecasting to date. The main findings confirm prior hypothesis, but also initiate new research discussions. New algorithms are feasible, in CI, NN and statistics alike. The competition assessed a novel statistical - and complex - method by Wildi (B9), which showed exceptional performance on both datasets. Illies' et al. (C27) introduced a NN-methodology that outperformed Dampen Trend ES, but still lacked the performance of the Theta across all series. This algorithm also outperfomed all other algorithms on 25 short and seasonal time series, the most difficult subset of the competition, and Yan (C44) outperformed all others on a subset of complex/difficult series. These achievements are surprising considering prior beliefs on the data properties required to use NN

methods on empirical data, and demand further attention. Overall, we hope the success of complex algorithms on such an established dataset will at least rekindle the discussion of innovative, sophisticated algorithms for time series extrapolation.

Results of NN3 suggest that methods of NN and CI can perform competitively 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 methods of ES, the workhorses of operational forecasting in practice for over 30 years, would be serious contenders that would prove challenging to outperform - after all they had outperformed most others 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 in comparison to many CI contestants. We feel that CI has closed in on established benchmarks, showing a range of different algorithms capable to predict both datasets as accurate as AutomatANN, the only contestant entering the M3 some 10 years ago, thus indicating improvements in feasibility and empirical accuracy to forecast with NN, and hence a motivation for further research.

Disappointingly, it seems impossible to provide more focussed guidance as to promising routes of future CI research, as no common 'best practises' can be identified for the top NN or CI contenders. Each submission was unique conceptually and methodologically, combining freely (and often seemingly arbitrarily) from the repository of algorithms and techniques available to machine learning today, without any evaluation of the contribution each fragment of the methodology made to increasing accuracy. As an example, for Illies' et al. it remains unclear if the accuracy stems from pooling time series for training, combining predictions in ensembles, or the algorithm of Echo State Networks itself. In an attempt to generalise, only the paradigm of forecast combinations seemed to drive accuracy, an observation made before. Ensembles of CI and statistical algorithms performed very well, but again no consensus on its meta-parameters of ensemble size or combination metric could be determined, although heterogeneity of

its base learners seemed to positively affect accuracy. As no two algorithms are alike, it becomes impossible to attribute positive performance to a particular modelling choice, allowing only an evaluation of composite yet distinct algorithms but not to derive guidance for promising areas of future research. Without insight, progress in CI may be slow and undirected. If this heterogeneity cannot be overcome, only a meta-learning analysis could yield insights in partial contributions, linking properties of algorithms and data conditions to guide future research effort.

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

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. (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. (2009). The tourism forecasting competition. In *Monash University Working Paper*: Monash University.

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

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., & Pietsch, S. (2007). A naive support vector regression benchmark for the NN3 forecasting competition. *2007 Ieee International Joint Conference on Neural Networks, Vols 1-6*, 2453-2458.

Crone, S. F., & Preßmar, D. B. (2006). An extended evaluation framework for publications on artificial neural networks in sales forecasting. In IASTED (Ed.), *AIA'06* (Vol. 1). Innsbruck: IASTED Press.

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., Belbachir, A. N., Bowden, G., Dandy, G., Lauzon, N., Maier, H., & Mason, G. (2005). A comparative study of artificial neural network techniques for river stage forecasting. *Proceedings of the International Joint Conference on Neural Networks (IJCNN), Vols 1-5*, 2666-2670.

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

Faraway, J., & Chatfield, C. (1998). Time series forecasting with neural networks: A comparative study using the airline data. *Applied statistics, 47*, 20.

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 *A companion to economic forecasting*. Malden, Mass. [u.a.]: Blackwell.

Fogel, D. B., & Fogel, L. J. (1994). EVOLUTIONARY COMPUTATION. *Ieee Transactions on Neural Networks, 5*, 1-1.

Foster, W. R., Collopy, F., & Ungar, L. H. (1992). Neural Network Forecasting of Short, Noisy Time-Series. *Computers & Chemical Engineering, 16*, 293-297.

Goodrich, R. L. (2000). The Forecast Pro 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.

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

Kolarik, T., & Rudorfer, G. (1994). Time Series Forecasting Using Neural Networks. In *Proceedings of the international conference on APL* (pp. 86-94). Antwerp, Belgium.

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 & Operations Research, 32*, 2151-2169.

Makridakis, S. (1994). Book Review: Time Series Predicition - Forecasting the Future and Understanding the Past - Weigend, A.S.,Gershenfeld, N.A. . *International Journal Of Forecasting, 10*, 463-466.

Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, E., & Winkler, R. (1982). The Accuracy Of Extrapolation (Time-Series) Methods - Results Of A Forecasting Competition. *Journal Of Forecasting, 1*, 111-153.

Makridakis, S., Chatfield, C., Hibon, M., Lawrence, M., Mills, T., Ord, K., & Simmons, L. F. (1993). The 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.
Reid, D. J. (1969). *A comparative study of time series prediction techniques on economic data.* Unpublished PhD thesis, University of Nottingham, Nottingham, UK.
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: Operations 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.
Suykens, J. A. K., & Vandewalle, J. (1998a). *The K.U.Leuven time series prediction competition.*
Suykens, J. A. K., & Vandewalle, J. (1998b). Nonlinear Modeling: advanced black-box techniques. In. Boston: Kluwer Academic Publishers.
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.
Weigend, A. S. (1994). *Time series prediction: forecasting the future and understanding the past. proceedings of the NATO Advanced Research Workshop on Comparative Time Series Analysis held in Santa Fe, New Mexico, May 14 - 17,1992* (1. printing ed.). Reading: Addison-Wesley.
Zadeh, L. A. (1965). FUZZY SETS. *Information and Control, 8*, 338-&.
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.