# Structure Combination of Forecasting Models
## with Applications in the Energy Sector

By

Juan F. Rendon-Sanchez

Supervisors
Professor Lilian M. de Menezes
Professor ManMohan Sodhi

A dissertation submitted to City, University of London, in accordance with the requirements of the degree of

Doctor of Philosophy in Management.

City
University of London
Faculty of Management
Cass Business School

October 2016

# Contents

# List of Tables

# List of Figures

# Acknowledgements

I am deeply indebted to my supervisor, Professor Lilian De Menezes, who constantly and generously provided me with technical and intellectual guidance and support.

I owe sincere thankfulness to Professor James Taylor and Dr. Siddharth Arora for generously providing their Matlab code for the double-seasonal Holt-Winters-Taylor model and data from some of their studies. I am indebted to Dr. Eduardo Alonso, Dr. Sven Crone and Dr. Nikolay Nikolaev who helped me with their constructive comments and kindly provided valuable information and advise.

I sincerely thank City University and Cass Business School for their generosity. Specially I thank Malla Pratt for her constant support.

I profoundly thank my mother, Olga Sanchez, for her constant encouragement and support. I am also grateful to all members of my extended family who helped me at different points of this research, specially my aunt, Fanny Sanchez, and my cousins Carolina Perez and Cesar Perez.

I thank Priscila Dahiana Gato for her love, company and encouragement.

My friends, Melanie Said, Andres Villegas, Maryam Lofti, Kaizad Doctor, Marianna Russo, Christopher Benoit, Danilo Mandic, Elena Casprini, Madona Zoraida, Juan Carlos Aristizabal and Laura Mena, were a source of strength and help. I am fortunate of having found them along my way.

I am also indebted to the Free Software Foundation, the Debian, Linux Mint, Ubuntu, PostgreSQL, GIMP, GNU Emacs and LaTeX projects for their amazing tools without which I would not have been able to materialise this research.

# Declaration Of Authorship

I, Juan F. Rendon-Sanchez, declare that this thesis titled, "Structure Combination of Forecasting Models with Applications in the Energy Sector" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

# Abstract

This dissertation proposes and implements the inclusion of model structure in combining forecasts. Empirical investigations are conducted with an emphasis on neural networks and seasonal exponential smoothing models using synthetic data and real time series, from the electricity sector. It starts with a literature review on combining forecasts and ensembles of neural networks, and highlights their use in forecasting within the energy sector. Research gaps are identified and the questions to be addressed in this research are set, thus leading to three empirical studies.

The first study provides a detailed sensitivity analysis of the goodness-of-fit and forecasting performance of feed-forward neural networks on time series with different characteristics. It expands existing literature by increasing the number and variety of time series and by using graphical and statistical diagnostics to objectively judge the influence of model specification on forecasting performance. Having identified conditions for achieving stable model performance, this study facilitated the identification of suitable models for different time series characteristics, which are then useful in developing combinations (ensembles) of feed-forward neural networks.

The second study proposes structural combination methods based on clustering (CB) and genetic algorithms (GA) for forecasting time series. Clustering of neural networks using their parameter space is performed to identify a pool of forecasts to be combined. Three synthetic time series and two real time series (electricity demand and wind power production) were used to assess the performance of the two proposals against several benchmarks in univariate and multivariate forecasting problems. Structural combinations with GA were more competitive than those with CB for non-seasonal time series and the multivariate wind power forecasting application, whereas for the seasonal series, the CB tended to be more competitive.

The third study focused on forecasting univariate time series with seasonality, by structurally combining, in separate applications, multiplicative Holt-Winters and multiplicative Holt-Winters-Taylor models. Noise addition and block swapping were applied to the original time series in order to generate structurally diverse individual models. Applications were conducted using a seasonal daily peak electricity demand time series, an hourly double-seasonal electricity demand series and a half-hourly double-seasonal electricity demand series. Structural combinations worked better for the peak electricity demand and half-hourly demand time series when model variation was induced via noise addition. For the double-seasonal hourly electricity demand, block swapping, as a means for diversity in models, resulted in better forecasts.

Finally, in the last chapter of this dissertation, conclusions are drawn from this research. The contribution to the literature is assessed and a future research agenda is proposed.

# Nomenclature

%Δ wrt Avg.: Percentage difference of an error metric with respect to the forecast average from all models in the ensemble.

AFTER: Aggregated Forecast Through Exponential Re-weighting.

AHWT: Double-seasonal Holt-Winters-Taylor model in its additive form.

ARMAX: ARIMA models with exogenous variables.

Avg. Net.: forecast average from all neural network models in an ensemble.

Best Net. isMAPE: Neural network with the lowest in-sample MAPE in an ensemble.

Best Net. isMSE: Neural network with the lowest in-sample MSE in an ensemble.

DGP: Data Generating Process.

DOE: Design of Experiments.

GA: Genetic Algorithms.

IS MAE: MAE for the in-sample period.

IS MSE: MSE for the in-sample period.

IS NMAPE: NMAPE for the in-sample period.

IS RMSE: RMSE for the in-sample period.

J-T: Jonckheere-Terpstra test.

K-W: Kruskal Wallis test.

LB TEST: Ljung-Box test for serial correlation.

MAE: Mean Absolute Error.

MAP: Maximum Absolute Percentage Error.

MAPE: Mean Absolute Percentage Error.

MHW: Single-seasonal Holt-Winters model in its multiplicative form.

MHWT: Double-seasonal Holt-Winters-Taylor model in its multiplicative form.

MIMO: Multiple-Input Multiple-Output approach for multi-step ahead forecasts with NN.

MISMO: Multiple-Input Several Multiple-Outputs approach for multi-step ahead forecasts with NN.

MM5: Fifth generation Mesoscale Model.

MSE: Mean Square Error.

MSPE Mean squared prediction error.

MVE: Mean-variance estimation. Method to estimate neural network-based prediction intervals.

NMAPE: Normalised Mean Absolute Percentage Error.

NN: Neural Network.

OS MAE: MAE for the out-of-sample period.

OS MSE: MSE for the out-of-sample period.

OS NMAPE: NMAPE for the out-of-sample period.

OS RMSE: RMSE for the out-of-sample period.

RBF: Radial Basis Function.

RMSE: Root Mean Square Error.

SMAPE: Symmetric mean absolute percentage error.

SOM: Self Organising Maps.

THEIL: Theil Inequality Coefficient.

WPPT: Wind Power Prediction Tool.

# Chapter 1

# Introduction

There has been over forty years of research in forecast combinations. The interest from academics and practitioners signals the recognition of limitations of individual forecasting models and the desire to exploit the advantages of multiple models. Nowadays, people and organisations are confronted with complex problems, considerable uncertainties as well as an abundance of data and forecasting approaches. Efforts are easily discernible toward the exploitation of a wide choice of data, of forecasting models and scenarios in order to seek better forecasting performance and, if possible, to quantify the uncertainty in the forecasts (Taylor & Buizza, 2003; Stephenson et al., 2005; Taylor et al., 2009; Lemke & Gabrys, 2010). However, most of the research has focused on how to combine forecasts produced by models and experts. Little attention has been given to the combination of models based on their specification.

Variety in structure stems from differences in the functional form of models or differences in their parameter values. Consider a model $ARIMA(p, d, q)$ with parameters $\Theta$ and $\Phi$. This model would have structural differences (in terms of functional form) when compared to a model $ARIMA(p', d', q')$ if $p \neq p'$, $d \neq d'$ or $q \neq q'$. On the other hand, when several $ARIMA(p, d, r)$ models are produced with different $\Theta$ and $\Phi$ coefficients, variety in structure results from differences in parameters and not from their functional form. Analogously, a feed-forward neural network, fitted to a time series, can have different specifications, depending on the number of hidden units, the number of hidden layers, the transfer function, the training algorithm, etc. Neural networks $NN(\{w\})$ and $NN(\{w'\})$ with hidden units sets $\{w\}$ and $\{w'\}$ would differ, all other factors constant, when the number of hidden units in $\{w\}$ and $\{w'\}$ differ or the values are different, and thus they be structurally different. The main goal of this research is to use this structural information when combining forecasts.

In structural combinations, the generation of different forecasting models, previous to the calculation of the combination, arises naturally. The systematic generation and combination of forecasting models has been traditionally considered in the Neural Network literature under the term *ensembles*. This term originated in climate modelling, when it was observed that there are differences in forecasts when models are initialised with different values (Parker, 2010). The building of ensembles has evolved from the use of simple sets of models to the collective evolution of them through sophisticated computational intelligence techniques. An ensemble generally has three stages (Leith, 1974; Lorenz, 1965; Hansen & Salamon, 1990): the generation of models, their selection or pruning and their combination. It is within this framework that the present research is located and attempts to make a contribution. In doing so, this dissertation aims at bringing models that have been found to be robust in the forecasting literature to the context of ensemble building and machine learning.

The first models considered are feed-forward neural networks, because of their natural structural representation, which is related to their founding idea of imitating the brain structure (see Haykin, 1999, p.24), and their use in ensembles. Furthermore, neural networks are widely applied in forecasting problems. IEEE archives (with publications in Operations and Management starting from 1990 to 2015) reveal that they have been widely applied, with publications in transport (10), education (11), mail (5), telephone systems (14), gas (186), water (207) and specially in power systems (886)[1]. The predominance of applications in the energy sector motivates

[1] Searches performed by using `http://ieeexplore.ieee.org/search` on the 15th of December 2015.

the current research. Secondly, two statistical models are considered: the single-seasonal multiplicative Holt-Winters model and the double-seasonal multiplicative Holt-Winters-Taylor model. The former belongs to a family of models used to forecast seasonal time series due to their robustness and simplicity (Hyndman et al., 2008; Pan, 2010), and the latter is an extension that has been applied successfully to electricity load forecasting (Taylor et al., 2006; Taylor, 2010).

Given the choice of models available in the literature, the general enquiry of this dissertation is cast into the following research questions: *How can the structure of neural network forecasting models be combined? How can the structural combination be extended from neural networks to other forecasting models?*

These questions focus on the last stage of building ensembles, that is, the aggregation of forecasts after the models were generated. However, the initial stage (the specification of models) also requires attention, since one would like to combine models that are robust. The study of the behaviour of models to be included in the ensemble facilitates making decisions when building ensembles. The need for such a study is acutely felt when using neural networks, given that they are universal approximators with a structure that can be modified depending on specific needs and desired precision (Haykin, 1999). Additionally, the use of statistical tools has been essential in understanding the behaviour of neural networks in order to make a better use of them (Anders & Korn, 1999). The design of experiments is here adopted, as it allows to systematically examine the effect of different design factors in goodness-of-fit and accuracy of NNs. Although there have been studies of NNs in this direction, as for example Zhang et al. (2001) and Balestrassi et al. (2009), they were limited to one-step-ahead forecasting and have not considered the double seasonal series that are common in short-term forecasting of electricity demand. Therefore, an initial study is conducted, which focuses on sensitivity analysis, through design of experiments, as a manner to aid the selection of neural network models to build ensembles.

The exploration of structural combination can inform the literature on forecast combination, which in turn can open avenues to improve forecasting accuracy by making use of more diverse sources of information than normally considered when combining forecasts. Although most of the applications made here use data from the energy sector, it would be expected that other complex forecasting problems could benefit from structural combinations of models, specially in the context of big data and business analytics.

This dissertation investigates combinations or ensembles of forecasts, neural network sensitivity and structural model combination, in the following way:

Chapter 2 reviews relevant literature in the area of neural network ensembles and forecast combination, with an emphasis on applications in the energy sector and concludes with setting the research questions to be addressed by this research. In Chapter 3, sensitivity analyses of neural networks are conducted, using synthetic time series, and guidelines are suggested to aid model selection. In Chapter 4, a structural combination approach for neural network is proposed and applications are conducted with synthetic and real world time series (electricity demand data from Rio de Janeiro and wind power production data from the global Energy Forecasting Competition, 2012). The investigation concentrates on ensembles with structural parameter variation, while keeping the same functional form. The selection of models to include in the structural combination is conducted along the lines suggested by the analysis developed in Chapter 3. Chapter 5 explores the behaviour of the proposed structural combination focusing on the single-seasonal multiplicative Holt-Winters and the double-seasonal multiplicative Holt-Winters-Taylor models. Applications are conducted with a daily peak electricity demand time series, and two electricity demand time series (hourly observations from Rio de Janeiro and half-hourly observations from England and Wales). Finally, Chapter 6 summarises and concludes this dissertation.

# Chapter 2

# Literature Review

This chapter reviews the literature on forecast combination with an emphasis on the energy sector and neural networks (NN) ensembles. The choice of these models is motivated by their suitability for structural model combination, which is the main aim of this dissertation, and also by the extensive use of such models in forecasting electricity demand as highlighted in previous reviews of this literature (Hippert et al., 2001; Crone et al., 2011). At the end of the chapter gaps in the literature are highlighted and research questions are formulated.

## 2.1 Combination of Forecasts

### 2.1.1 The Motivation for Combining Forecasts

There are several reasons for combining forecasts (Clemen, 1989; Timmermann, 2006). If it is assumed that the information set underlying the individual forecasts is often unobserved to the forecast user, then it is not possible to gather all information and construct a "super" model. In this case, the combination of forecasts will be an attempt to optimise the use of information that is available to the forecaster.

Some models may adapt quicker than others to changes in the data generating process. As it is difficult to detect structural breaks in "real time" it is plausible that, on average, combinations of forecasts from models with different degrees of adaptability will outperform forecasts from individual models. Furthermore, individual forecast models may be subject to misspecification bias of unknown form. Combining forecasts from different models can be viewed as a way to make forecasts more robust.

In general, there is a growing consensus about the advantages of combining forecasts. Clemen (1989), while reviewing empirical evidence in forecast combination, concluded that combining multiple forecasts leads to increased forecast accuracy. Stock & Watson (2004) observed, after an empirical analysis that included data adaptivity weighting mechanisms, that forecast combination performs well with respect to autoregressive models. They also concluded that the best performing combination schemes were simple ones, such as the average and models with the most simple data adaptivity in their weighting schemes. Timmermann (2006) argued, from a theoretical perspective, that unless one can find ex ante a particular forecasting model producing smaller forecast errors than its competitors, forecast combination offers diversification gains that make it attractive to combine individual models rather than relying on a forecast from a single model.

### 2.1.2 The Most Common Methods for Combining Forecasts

Statistical approaches (such as linear combinations and clustering) and computational intelligence models (such as fuzzy logic and NN) have been used to combine forecasts.

#### 2.1.2.1 Statistical Based Approaches

Linear combination is one of the simplest forecast combination methods. The simple average is difficult to defeat (Armstrong, 2001). Che (2015) suggest improvements to the selection of models for linear combinations. The concept of entropy[1] and co-variance between forecasts are used to define the amount of common linear information between a set of forecasts and the actual value. The relevance of a random independent variable (a forecast) with respect to the dependent variable (the actual

[1] Entropy is a measure of the uncertainty of a random variable (Cover & Thomas, 2006, p. 13). For a discrete random variable $X$, it is defined as $H(X) = -\sum_{x \in \mathcal{X}} p(x) log(p(x))$

value) is defined in a similar way. The redundancy (when the common linear information is maximum) and the relevance are then used to select forecasting models for a combination; the approach minimises linear redundancy and maximises linear relevance. This procedure allows for an algorithm to find the optimal subset of all the individual models to combine without having to try all possible combinations of the individual models.

Outperformance (Bunn, 1975) is a form of combination that has the form $f_c = p'f$, where $f$ is a vector of forecasts and $f_c$ is the resulting forecast. It uses $p$, a simplex of probabilities which can be assessed and revised in a Bayesian manner. In practice, $p$ is the historical proportion that the forecaster or model has outperformed its competitors. Each individual weight is interpreted as the probability that its respective forecast will be the best (in the smallest absolute error sense) on the next occasion.

In the *optimal* approach (Bates & Granger, 1969), linear weights are calculated to minimise the error variance of the combination (assuming unbiasedness for each individual forecast). The vector of combining weights, $w$, is determined according to the formula $w = \frac{S^{-1}e}{e'S^{-1}e}$, where $e$ is the $(n \times 1)$ unit vector and $S$ is the $(n \times n)$ covariance matrix of forecast errors. The authors also proposed variations to this approach, namely, the *optimal (adaptive) with independence assumption* in which the estimate of $S$ is restricted to be diagonal, comprising just the individual forecast error variances; *optimal (adaptive) with restricted weights* with the additional restriction so that no individual weight can be outside the interval $[0, 1]$.

Participant forecasts can be used as regressors in an ordinary least squares (OLS) regression with the inclusion of a constant (Granger & Ramanathan, 1984). Regression with restricted weighs is a variant where the weighs are constrained to sum one (Granger & Ramanathan, 1984; Timmermann, 2006). Time-varying regressions could be applied when very large data sets are available. Diebold & Pauly (1987) and LeSage & Magura (1992) found such approach advantageous in dealing with structural change. Terui & van Dijk (2002) also found competitive results, when comparing with the constant weights approach, although findings are inconclusive for some series.

Trimming is a combination approach focused on selection. Instead of combining a set of forecasts, it can be advantageous to discard the models with the worst performance. Suppose that a fraction $\alpha$ of the forecasting models contain valuable information about the target variable while a fraction $1 - \alpha$ is pure noise. Then, once combination weights have to be estimated, forecasts that only add marginal information should be dropped from the combination, since the cost of their inclusion (increased parameter estimation error) is not expected to be matched by similar benefits.

Clustering of forecasts (Timmermann, 2006) is inspired by the assumption of commonalities underlying the forecasting models. It has been proposed, for example, an approach to sort forecasting models into clusters using a K-means clustering algorithm based on their past Mean Square Error (MSE) performance (see Timmermann, 2006). Alternatively, according to the author, clustering can be based on correlation patterns among the forecast errors.

Switching between different forecasts at different periods (Granger, 1993; Deutsch et al., 1994; Taylor & Majithia, 2000) is a selection strategy that makes use of information from different forecasts. It is dynamic and is based on the idea that available forecasts might vary in relevance depending on the period to forecast.

One salient feature of the approaches just outlined is of special interest for the present research: forecast combinations are based on individual point forecasts. However, other model features, besides their outputs, could be taken into account in the combination. The use of other sources of information for combination (such as subjective judgements and context information) is present in expert forecast combination (e.g  Maines, 1996; Webby & O'Connor, 1996), but when looking at model combination, the use of information beyond individual model forecasts is less explored.

#### 2.1.2.2 Computational Intelligence Approaches

Fuzzy inference systems have been tried as means to combine forecasts (Fiordaliso, 1998; Palit & Popovic, 2000; Xiong et al., 2001). Their ability to find non-linear mappings between an input and an output space is attractive as combining mechanisms.

Genetic algorithms (GA) can also be used as a combination mechanism (as in Alvarez-Diaz & Alvarez, 2005, who forecasted weekely exchange rates of Japanese Yen and Pound Sterling against the American Dollar). However, it is more common to find them in hybrid approaches as part of the optimisation process or the model specification mechanism (see for example Zhou et al., 2002; Pai & Hong, 2005).

Non-linear combinations of forecasts have tended to use NNs. As in the case of fuzzy inference systems, evidence has been found of the advantage of these models over linear combination schemes (Donaldson & Kamstra, 1996). A NN can be regarded, in isolation, as a combination device. For example, with regard to multilayer perceptrons with a single output, Crone & Kourentzes (2010) observe that each hidden node computes a non-linear autoregressive model of order $p$, $NAR(p)$, on input nodes, which are combined to $\hat{y}$ by a weighted sum of a single output node. The importance of combination when forecasting with computational intelligence models (NNs included) has been highlighted by Crone et al. (2011) in the context of the NN3 competition. Several of the highly competitive models in their review, contain a form of combination. In general, different degrees of complexity can be found, from early research done by Donaldson & Kamstra (1996) to a more recent study by Matijaš et al. (2013). In the first, the authors report the superiority of NNs to combine (two) forecasts, using daily data from stock market volatility. In the second, the authors built a framework to combine forecasting models using meta-learning algorithms and different types of NNs. Applications were conducted by using hourly electricity demand from Europe.

However, the role of NNs is not limited to the combination of forecasts produced by different models. They are frequently used in ensembles, where several NNs are systematically generated and either pruned or selected so that their forecasts are then combined using several mechanisms, which are not necessarily NNs. The test of structural combinations in the present research rests on the production of several models to be combined, thus leading to the creation of ensembles. The following section introduces the concept of ensembles, which is adopted in this research, and reviews key research in the topic. Specific applications in the energy sector are reviewed in the subsequent section.

#### 2.1.2.3 Ensembles of NN

The term ensemble originated in climate modelling. In that field scientists distinguish different types of uncertainty, as Parker (2010) described. *Structural uncertainty* refers to uncertainty about the form that the modelling equations should take; *parametric uncertainty* is uncertainty about the values that should be assigned to parameters within a set of modelling equations and *initial condition uncertainty*, which refers to the difficulty in measuring all the required variables needed in models.

> “Uncertainty regarding the choice of initial conditions became a source of concern in the context of weather forecasting several decades ago, when Ed Lorenz famously discovered that even small differences in the conditions used to initialise weather models can lead to quite large differences in the forecasts produced [...] Indeed, it was the recognition of this sen-

> sitive dependence on initial conditions that first prompted atmospheric scientists to consider ensemble approaches” (Parker, 2010, p. 264).

Ensembles were adopted in NNs by Hansen & Salamon (1990), and came to mean the use of several models, constructed with differences in one or more of their design parameters. They used ensembles of NNs for classification. In this seminal research only synthetic data are used, and superiority of the ensemble is reported in comparison to individual models. It is important to note that there is a sensitivity analysis where the performance (probability of error in classification) is measured against the number of hidden nodes. A neural network is selected when the curve of performance against the number of hidden nodes begins to flatten. ‘Better performance yet can be achieved through careful planning for an ensemble classification by using the best available parameters and training different copies on different subsets of the available database.’ (Hansen & Salamon, 1990, p. 1000)

Jacobs & Jordan (1991) built an ensemble of neural networks for pattern matching. The purpose was the determination of different tasks to be learnt and the assignment of different networks to them. The networks were then learning from different training patterns. The ensemble consisted of member networks and a gating network. The architecture of the gating networks was fixed and the architecture of the member NNs was not clearly explained. A key element of this research is the combining procedure, which is determined by a break down of the problem into tasks of different complexity, an approach that will be later used in forecasting (Bakker & Heskes, 2003).

Liu & Yao (1999) developed a procedure in which NNs were trained and combined in the same learning process (dynamic and cooperative). Networks are trained simultaneously, allowing for interactions between them and to specialise. The procedure could create negatively correlated neural networks using a correlation penalty term in the error function. Additionally, they analysed bias-variance-covariance trade-off, an extension of the bias-variance dilemma (Bishop, 1995). The Mackey-Glass time series and the Australian credit card data were used. In the ensemble, the architecture and the size of the ensemble were fixed and the input specification was not taken into account. The authors claimed that the approach can produce neural network ensembles with good generalisation ability. In this case, diversity was achieved through a dynamic selection of models such that their outputs are negatively correlated.

Liu et al. (2000) also proposed the design of an ensemble with negative correlation learning, but used an evolutionary algorithm and clustering (k-means). The logic of the approach was the following: start an ensemble, train with negative correlation, evolve individuals for a number of generations and then use k-means to find "species" and then combine them. Results were superior when compared to other algorithms. However, the architecture was fixed, preliminary analysis of the models and data are not reported and the research was limited to the classification area. The innovation in this research was the creation of diversity at different levels of complexity. An approach that could be extended to forecasting.The clustering of models is left to the end of the process. However, it could be incorporated in the optimisation stage.

Zhou et al. (2002) argued that combining some networks in an ensemble is better than combining all networks. An evolutionary algorithm was used to assign and evolve the weights of the participating networks. Finally, with the obtained weights, a subset of networks was selected. The genetic algorithm that was used worked as a form of pruning. Comparisons with bagging (Breiman, 1996) and boosting (Freund et al., 1996) showed that the approach could generate smaller ensembles with better generalisation capabilities. However, the questions of how many models to produce and how many to select for the types of problems being tackled (regression and classification) were not addressed.

Islam et al. (2003) used a constructive algorithm (without pruning) to build an ensemble. Their approach tried to minimise the ensemble error by training, adding a hidden unit to an existing NN, and, finally, by adding a new NN. The algorithm also uses negative correlation learning. Encouraging results were obtained with benchmark problems in classification and forecasting, including the Australian credit card assessment, breast cancer, diabetes, glass, heart disease, letter recognition, soybean, and Mackey-Glass time series (the frequency of the time series is not specified).

Bakker & Heskes (2003) generated many different NNs (on the order of 50) and then summarised them via a clustering procedure. Diversity was introduced by bootstrapping the training data. The authors claimed that it is not necessary to use all the models in the ensemble: those found through this summarising technique can perform equally well or better than the whole set. The authors considered both clustering of output forecasts and clustering in the parameter space of models, but implemented only the first one because they judged it had a more immediate meaning for clustering and was less computationally demanding. Their suggestion of including the models parameters in the clustering, however, is important: it hints at a possible transition from the consideration of point forecasts as inputs for combining procedures to the consideration of more complex sets of information, and is a source of inspiration for the present research.

Chen & Yao (2007) incorporated an evolutionary algorithm and negative correlation learning to automatically design and train neural network ensembles: resampling of input space, randomisation of the number of neurons in the architecture (although the number of layers is fixed) and random selection of features together with negative correlation and an evolutionary component (Gauss mutation). Excellent results were reported in comparison to random forests (Breiman, 2001), bagging (Breiman, 1996) and adaboosting (Freund et al., 1996) in different benchmark problems. Their focus was on classification and did not inform time series forecasting.

Krasnopolsky (2007) used ensembles built with fixed architecture but different initial conditions. The approach was used to produce a mapping with smooth derivatives with respect to the inputs, which might be useful in avoiding extreme outliers in forecasting models.

Adeodato et al. (2011) proposed an approach where the number of hidden units and the number of inputs in NNs are explored in specified ranges and applied it to a set of 111 monthly data sets. Two different training algorithms were used and only one architecture-training algorithm combination was selected. The resulting parameters were used to produce 15 replicas of the network. The training scheme included two stages where the validation set in stage 1 is used as part of the training set in stage 2. The median of the 15 replicas was used as the final forecast. They claimed an increased forecasting accuracy in comparison to a single MLP for multiple step-ahead forecasting (with recurrent approach). In their explorations of ranges of models an attempt was made to make a better informed decision about the models to include in the combination scheme.

In summary, NN ensembles have evolved from the use of simple sets of models to the collective evolution of them through sophisticated computational intelligence techniques. It is noticeable the relevance of the use of synthetic time series in the studies. The next section focuses on the energy sector.

## 2.2 Examples of Applications of Combinations in the Energy Sector

We now review studies devoted to forecast combination in the energy sector, with emphasis in ensembles and applications in load and wind power forecasting, where the later are significantly more volatile time series when compared to the former. NNs will be considered, because they are universal approximators (Kasabov, 1996, p. 13) and are suitable for the structural combinations, as proposed in this research.

### 2.2.1 Applications in Load Forecasting

Khotanzad et al. (1998) forecasted electricity load in the next 24 hours by using hourly data: a feed-forward NN was intended to forecast load for each hour (one output of the model mapped to one step ahead) and another NN was used to forecast the change in load, for each hour as well. Forecasts were combined through a regression with recursive least squares.

Drezga (1999) considered hourly and peak load for the next 24 and 120 hours with a small ensemble of NNs. K-nearest neighbour was used to select training sets (instead of using correlograms). The ensemble was tested using two years of hourly data from two US utilities with different weather and consumption patterns. The same set of inputs was identified for both applications. Networks were trained in parallel with an iterative approach, feeding back averaged forecasts as inputs for subsequent forecast horizons. Results in terms of MAPE were competitive, when compared with data from similar utilities and publications. The models were found to be robust when faced with sudden changes in temperature, but were limited to one-step-ahead forecasting.

By contrast, Taylor & Buizza (2002) focused on NNs and daily data to produce load forecasts from 1 to 10 days ahead based on ensembles of weather forecasts. Comparisons were made with univariate benchmarks and point forecasts produced without the ensembles. For ten lead-times, the mean of the load scenarios built with weather variables was a more accurate load forecast than that produced by the non-ensemble based procedure. This research combines the use of ensemble weather forecasts with an ensemble of rather low complexity NNs and suggests benefits from multivariate models.

Abdel-Aal (2005) used hourly load and temperature data to forecast the next day peak load for a utility in USA via ensembles of feed-forward NNs and abductive NNs. The latter have the advantage of including automatic selection of significant model inputs (Montgomery & Drake, 1991). Each network in the ensemble specialises in historical data from certain year. Comparisons were made against individual models for each year and an individual model trained with all historic data. The ensembles improved over the benchmarks, but the advantage was clearer with abductive networks. The correlation among input data for different years was acknowledged as a source of homogeneity in models that was tackled by using different structure in the networks. This study illustrates the importance of promoting model diversity when building an ensemble.

Daneshi & Daneshi (2008), by using data obtained at 5-minutes intervals, forecasted electricity load for the next 30 minutes (with steps of 5 minutes) using a scheme where data was divided into several categories depending on the time of the day, from morning to night. For each category a set of 3 NNs was trained and the output was combined using recursive least squares. The peculiarity of this research is that data variation (intended to generate model variation) was well coupled with a categorisation of data (early morning, mid-morning, noon to early night and late night), which is useful in the case of seasonal data.

Fan et al. (2009) forecasted hourly electricity load using different weather forecasts (hourly data) combined with a method called Aggregated Forecast Through Exponential Re-weighting (AFTER). Then, an ensemble with bagging[2] of NN was used to forecast the load. Comparisons were made against individual models with different input data (different combinations of weather data from different meteorological services) and the approach appeared to consistently improve accuracy.

Siwek et al. (2009) forecasted load for the next 24 hours with hourly data from the Polish power system, using different ensembles of feed-forward, self organising maps (SOM), and fuzzy SOM. The forecasts combination was made separately with simple and weighted average, Blind Source Separation (BSS) and Principal Com-

[2] A method based on subsampling that uses randomly generated training sets to obtain an ensemble of predictors (see Breiman, 1996).

ponent Analysis (PCA) decomposition. It is assumed that each participating NN generates forecasts for the following 24 hours, although it is not explicit which specific forecasting scheme was used by them. The BSS system decomposed the original stream of signals derived from NNs into components. These components were reconstructed to produce the final forecast, by analysing all possible combination of components. This was the approach that delivered better results (when comparing with the best individual predictors). Instead of seeking variety in models or data, signal decomposition and re-composition were conducted and produced promising results.

Fay & Ringwood (2010) forecasted load for the next 24 hours based partly in weather variables (with hourly frequency). Feed-forward NNs were trained taking into account the error in weather forecasts, so that both error and weather variables were included in the inputs. Every participating NN produced forecasts for different hours, but the specific multi-step-ahead forecasting approach was not mentioned in the article.Parameter estimation was split into two phases: independent and dependent of weather forecast error. The models were successful in adjusting the weighting of the sub-models to reflect the deterioration of forecasting accuracy when using weather data.

Alamaniotis et al. (2012) used 5-minutes data to forecasted load for the next 30 minutes at intervals of 5. Kernel-based Gaussian processes are ensembled. The forecasts were arranged into a linear (multi-objective) problem for which a solution was sought with GA. Performance was favourably compared against individual participating models and an ARMA model. The major innovation consisted in the use of multi objective optimisation to combine the models. Different metrics comprised the vector of objectives: MSE, Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Maximum Absolute Percentage Error (MAP) and Theil Inequality Coefficient (THEIL).

Matijaš et al. (2013) forecasted hourly load with a meta-learning framework and using hourly load data sets. They exploited the interpretation of the learning process as a link between a problem space and a solution space (Kasabov, 1996, p. 332). For different problems (data) there were different mappings (forecasting algorithms) that lead to a solution. The following models were included: Random Walk, Autoregressive Moving Average (ARMA), Similar Days, Layer Recurrent Neural Network (LRNN), Multilayer Perceptron (MLP), m-Support Vector Regression (m-SVR) and Robust LS-SVM (RobLSSVM). They were ranked with a meta-learning algorithm with rich features, based on data and the best model was used to issue forecasts. In contrast to most studies, the authors tried to expand the problems and solutions (models) representations in order to explore combination of forecasts in a wider sense.

Kaur et al. (2014) used hour-ahead market and day-ahead market load for California Independent System Operator (CAISO) and Electric Reliability Council of Texas (ERCOT). Starting with a base forecast the residuals were modelled with generalised linear and ARIMA models with exogenous variables (ARMAX). Forecasts were combined through least squares optimisation. Ensembles were configured depending on the day of the week or the hour of the day. The approach significantly improved the forecast in super off-peak and off-peak times. The general outline of the procedure is innovative and well thought, as it uses regularities in data for the configurations of the ensembles.

Qiu et al. (2014) proposed an ensemble with deep belief networks and support vector regressions. Deep belief networks are probabilistic generative models that are composed of multiple layers of hidden units (Hinton, 2011). The main idea of support vector regression is based on the computation of a linear regression function in a high dimensional feature space where the input data are mapped via a nonlinear function (Basak et al., 2007). The NN were trained with different number of epochs and combined with support vector regression in order to produce 1-step-ahead forecasts. There was a clear division between a stage where knowledge is extracted and another where the remaining dynamics of the problem is used to issue a final forecast. The ensemble was applied to load (with half-hourly data) and synthetic series forecasting. It outperformed feed-forward NNs, support vector regression, deep belief networks and an ensemble of feed-forward NNs. Despite the complexity of the models, the research was limited to one-step-ahead point forecasts.

Burger & Moura (2015) forecasted electricity demand using building level consumption hourly series. Six hours ahead forecasts were predicted, with a combination of different models (OLS with regularisation, support vector regression with radial basis function (RBF), decision tree regression and K-nearest neighbours). After the models were trained, the validation period was used to select the model to perform the final forecast (only one model from the set). Performance was reported to be better than with the use of individual models. The way this model was selected is innovative: a mechanism was developed to predict the performance of each forecasting model. Their strategy illustrates how the process of model selection can be enriched. However, it is more likely to work with heterogeneous models. With pools of models of the same type, their probabilities of having similar performance is higher and selection is likely to be less clear. Their other selection mechanism were based on cross-validation, taking into account the RMSE during validation period, and could be applied more easily to homogeneous pools of forecasting models.

Hassan et al. (2015) used a simple ensemble of NNs with 5 different architectures. Monthly demand data from Australian Energy Market Operator (AEMO) and the New York Independent System Operator's website (NYISO) were used to derive half-hourly forecasts and test the approach. Three methods were adopted to combine the forecasts in ensembles: average, trimming and Bayesian averaging. The latter was reported to be the best performing scheme, and used the posterior probability for each model as the respective forecast combination weight.

### 2.2.2 Applications in Wind Power Forecasting

Giebel et al. (2003) reported how WPPT, a wind power forecasting system mainly developed with mathematical and statistical models, performed combinations of forecasts:

> "For both model branches the power prediction for the total region is calculated as a sum of the predictions for the sub-areas. The final prediction of the wind power production for the total region is then calculated as a weighted average of the predictions from the two model branches. A central part of this system is statistical models".

Also in describing ARMINES, another system that uses statistical tools:

> "The wind forecasting system of ARMINES integrates [...] combined forecasts: such forecasts are produced from intelligent weighting of short-term and long term forecasts for an optimal performance over the whole forecast horizon."

In general, forecast combinations are a common practice in the wind power industry since they have become a building block in forecasting systems (Martí et al., 2006).

Sanchez (2006) combined different autoregresssion models to forecast wind power, based on hourly data, for the next 48 hours. Each model had different components (choosing between wind speed, wind direction and wind power). A subset of models was chosen to perform the final combination, in the form of a linear expression, with the weight assigned to them changing through time and the number of models combined varying as well. Different sets of parameters were used for different steps ahead. This work emphasised the statistical treatment of forecast combinations, and can be seen as a source of a general framework to analyse computing intelligence models.

Sanchez (2008) expanded on the previous research and used two ways of combining forecasts: *combination for improvement*, in which the objective is to find the best (constrained) linear combination of a set of forecasts, and *combination for adaptation* which aims to perform as well as the best individual procedure, trying to track the best available predictor. The author proposed the following:

1. Do *combination for improvement* with various methods. Do *combination for adaptation* with various methods.

2. Do *combination for adaptation* to further combine the combinations produced in the previous step. The recombination seems to give good results according to some authors (e.g. Gunter & Aksu, 1989; Yang, 2004).

The procedure above described was tested with mean hourly power generated in two wind farms. For each case, a set of four forecasts were available (for various hours ahead), which were provided by an independent professional forecaster. The investigation aimed at forecasting up to 18 hours ahead. The authors reported promising results, which were based on Mean Squared Prediction Error (MSPE).

Salcedo-Sanz et al. (2009) forecasted wind speed, which was used in forecasting wind power. Using coarse information from weather data, a down-scaling was performed to obtain forecasts at an hourly resolution to find the speed of wind at specific locations. They used information from global prediction systems which give predictions of weather variables at certain altitudes and spacial resolutions. A combination of global models with different parameterizations gave a pool of data sources to feed the different structures of neural networks. The global model-parameterizations were taken into account when setting up the NN models. Their approaches, including combinations, were the following:

1. Set up a NN for each global model-parameterization combination. Select the best from the set.

2. Set up a NN for each global model-parameterization combination, and aggregate the outputs from all combinations in a single output unit.

3. Set up a NN for each global model-parameterization combination, and then aggregate all outputs using a hidden followed by a single output unit.

Two hourly series of wind speed at two points in a wind park were used. Other predictors used included wind direction and a measure of temperature in one of the points. These values were based on results a fifth generation Mesoscale Model (MM5) at a given height, equal for all the wind turbines[3]. Other input variables were two time series measuring the solar cycle. The authors claimed that the bank of neural networks obtained better results than the best of the models with a single neural network, for the specific site in question (located in Spain). Statistical diagnoses were not reported, though the authors highlighted how combinations are useful to cope with the uncertainty in weather data.

Li et al. (2011) used a hybrid approach to combine 1-step-ahead forecasts with NNs in a first stage and a Bayesian combination in a second stage. The models were applied to forecast hourly wind speed. Their results suggested that the inconsistencies in performance between different kinds of NNs can be overcome with the use of Bayesian averaging.

Wang & Hu (2015) forecasted wind speed from two wind farms in China by using 15-minutes and 30-minutes data. Signal processing was used at the beginning and several forecasting models of different nature were used afterwards (ARIMA, support

[3] According to Salcedo-Sanz et al. (2009), the MM5 "is a limited area model, which solves the Navier-Stokes equations which modeled the behavior of the atmosphere (similar to the global models), but without including ocean-land interactions and other important variables of the global forecasting models."

vector, least square support vector machine and extreme learning machine[4]). The final combination was made with a Gaussian process regression model. Comparisons, were made between the combining mechanism proposed and individual models. The contribution was in the manner of arranging models in a process.

Ren et al. (2015) reviewed recent literature on ensembles for wind power and solar power forecasting. Their classification of methods distinguished two types of approaches. Cooperative ensemble forecast divides prediction into several sub-tasks and selects appropriate predictors for each sub-task based on their characteristics. The final forecast is a sum of all the outputs of the base predictors. Competitive ensembles train different predictors individually, with different data sets or different parameters, and the prediction is obtained by summarising forecasts of all base predictors. The authors evaluated three ensemble forecasting methods, with real wind speed and solar irradiance data sets, and concluded that the competitive ensemble forecasting method (Bagging-Back-Propagation) had better performance on longer forecasting horizons and the cooperative ensemble forecasting method (Wavelet Transform-Back-Propagation) had better performance on shorter forecast horizons. Their findings are difficult to generalise due to the limited number of approaches evaluated. However, they might have implications for ensemble approaches, since switching between types of ensembles depending on the forecast horizon could be a sound strategy. In general, their review focuses on point forecasts combination.

### 2.2.3 Other Applications

Khotanzad et al. (2000) forecasted daily gas consumption. Two NNs (a simple

[4] Extreme learning machine (ELM) is a new formulation for training single hidden layer feed-forward neural networks. ELM is formulated as a linear-in-the-parameter model which transform the training into solving a linear system. Compared to traditional feed-forward learning methods, ELM is very efficient and has desirable properties in convergence (Huang et al., 2015).

feed-forward and a functional link network[5]) were used in different combinations to forecast consumption in the next day: average, recursive least squares, fuzzy logic, feed-forward NN, functional link NN, a partition of the temperature space (an external variable)[6], a linear programming algorithm and a mixture of local experts[7]. The best performance was obtained with a NN as a combination mechanism. The forecasting task was limited to the total daily gas consumption one step ahead by using daily data. Perhaps the conditions for this experiment were comfortable for NN models with the smoothing of the data. Investigations of multiple steps ahead would have rendered the study more complete.

Yu et al. (2008) explored time series decomposition in conjunction with a NN combination (the networks having 3 layers with an unspecified number of units) in forecasting crude oil spot price. Empirical mode decomposition (EMD) was adopted, where, by using daily data, the components are modelled with independent networks (one for each sub-series). The individual forecasts were aggregated through a linear NN. The number of NNs in the ensemble varied with the number of sub-series derived from the original data. The model showed good results when compared to ARIMA models and other variants of their proposal.

Alessandrini et al. (2015) forecasted hourly power data from three solar farms located in Italy by using ensembles in conjunction with hourly data. They incorporated techniques that have been used in weather and wind power forecasting. Their main idea was to compare the predictions for a given time horizon with the corresponding past observations and establish a measure of similarity (through distance). By using a ranking procedure a set of past observations were selected (based on calculated distances) and constituted an ensemble for the given horizon. The pro-

[5] See Pao & Takefuji (1992).

[6] A scheme where a space of the estimated and actual temperature was divided and a regression was done with the models falling into each area.

[7] The approach used a third module, called gating network, to learn to assign different parts of the input space to different local experts.

posed approach compared well against other ensemble methods and point forecasts (including NN[8]). The research is relevant because it incorporates the concept of similarity (based on distance calculations), which can be applied to forecasting models in general, and not just point forecasts or ensembles of point forecasts. In this way, it highlights potential ways to enrich forecast combinations.

From the variety of approaches described above, some similarities and differences can be identified in the way ensembles have been devised. The following section addresses key aspects in ensembles and highlights limitations that lead to the focus of this dissertation.

## 2.3 The Current State of the Literature on Ensembles for Time Series Forecasting

It is noticeable that ensembles involve three main tasks, as depicted in Figure 2.1: generation, pruning and combination. Generation involves the creation of different instances of models. Pruning comprises a selection of them, which is optional (Mendes-Moreira et al., 2012). The final stage performs the combination of forecasts. Other factors to be considered, besides these three, are: the type of model to ensemble, the way the individual models are produced, the level of automation of the process and the way combinations are made. Therefore the construction of ensembles is far more complex than the construction of a single model.

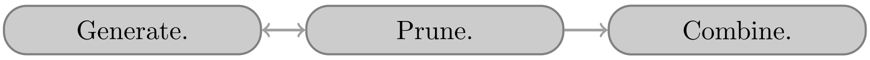


Figure 2.1: General steps in ensemble generation.

There are sequential approaches where the generation of models is followed by a pruning stage and finished with a combination stage (see for example Zhou et al.,

[8] In their approach, the authors included the forecasting horizon into the input patterns and therefore every network produced forecasts for each horizon.

2002). But in other approaches this sequence is not entirely observed: the evolution (or generation) of the individual models can be done in parallel, so that the information of the training stage can be shared and used to modify the collective estimation of parameters (see Liu et al., 2000; Chen & Yao, 2007).

The type of model used and the generation process are interrelated. The most common types of models in the literature are feed-forward NNs (for example Hansen & Salamon, 1990; Fan et al., 2009), although RBF (Yu et al., 2008; Maqsood et al., 2004), Elman recurrent network and Hopfield (Maqsood et al., 2004), Deep belief network (Qiu et al., 2014) and Abductive networks (Abdel-Aal, 2005) can also be found.

Once the type of model is chosen, a question arises about how to specify the individual models. In the case of feed-forward NN, which is the most used in the literature, the creation of a single model for forecasting implies the specification of several parameters. They can be determined by trial and error, heuristic rules or by systematic approaches. For example, Anders & Korn (1999) conduct model selection with strategies based on sequential statistical tests, information criteria and cross validation. Zhang et al. (2001) and Balestrassi et al. (2009) use design of experiments to identify an appropriate network configuration. Crone & Kourentzes (2010) emphasise input selection as part of model specification and propose a methodology for seasonal components. Nevertheless, no universal guidelines exist on how to select the most appropriate model (Kourentzes et al., 2014).

For ensembles of NNs, the selection of participant models tends to be simple. Sometimes the structure of the models is determined automatically (for example through randomisation of structural parameters, as in Chen & Yao, 2007) without an initial analysis of the potential choices of models. In some cases, there is a hint of a preliminary sensitivity analysis (e.g. Hansen & Salamon, 1990; Adeodato et al., 2011). Different networks are evaluated, with architectures varying in a given range and in combination with two different training algorithms. The best are selected from that step to then generate different networks to later combine. However, preliminary analysis can be carried out in order to support the ensemble generation in a more direct way. Kourentzes et al. (2014) use a systematic approach with step-wise regression for inputs and minimisation of the validation MSE for the number of hidden units. Their focus is in the selection of optimal design parameters (inputs and hidden units), but the variation in performance is not studied in detail for different specifications, which could reveal instabilities or error patterns that could inform model selection.

A sensitivity analysis of the relationship between specified parameters and model forecasting performance as conducted by Zhang et al. (2001) and Balestrassi et al. (2009), using design of experiments, and as suggested by Hansen & Salamon (1990), would be useful in the construction of ensembles. Even if the selection of models is done automatically in the ensemble algorithms, such a preliminary analysis could help to screen regions of the input space from which models have particular difficulties in learning or regions of the parameter space that show special instability.

In a subject area that relies on the availability of diverse models it is important to understand variation in models, in order to make informed decisions about the base components to use in the model generation stage. Other methodologies, including those which help in the selection of inputs to the forecasting system, could be explored in conjunction with a sensitivity analysis. Additionally, model specification and the building of model ensembles interact: the decisions about the structure of individual models affect the ensemble performance. Hence, there is need for detailed study of both topics and their interactions.

In ensemble construction, once the characteristics of the models are established, variety is usually introduced by modifying initial random weights (Adeodato et al., 2011), or by randomising training samples (Zhou et al., 2002; Bakker & Heskes, 2003). The randomisation of the feature space (Chen & Yao, 2007) could count both as a strategy for model variation and as a strategy for model specification.

In the final block in Figure 2.1, models are combined to produce the output of the ensemble. The methods that have been proposed in the literature (focusing on forecasting rather than on classification) include: a gating network (Jacobs & Jordan, 1991), a simple or weighted average (Krasnopolsky, 2007; Liu et al., 2000; Islam et al., 2003; Maqsood et al., 2004), a nonlinear average through another NN (Krasnopolsky, 2007), a feed-forward NN (Yu et al., 2005, 2008), a RBF (Yu et al., 2008), the median of forecasts (Adeodato et al., 2011; Fan et al., 2009) and the mode (Kourentzes et al., 2014). It can be seen that the complexity and effectiveness of the ensemble approaches are only partially related to the combining procedure at the end of the process, because there are other steps involved.

In order to summarise key ensemble characteristics from this body of literature, Table 2.1 shows the main aspects in NN ensemble creation: the type of ensemble approach (column *Ensemble approach*), the type of model (column *Network type*) and the way the individual models are constructed (column *Architecture selection*). Additionally, the area of application is given. In the case of ensembles characterised by fixed NN architectures, sequential approaches and the existence of clusters of models, there are key elements to take into account: the number of NN generated, the number of clusters used, the number of networks selected (per cluster or in general if no clustering is done) and the final model combination. These aspects are summarised in the last columns of the same table.

Table 2.1: Ensemble generation schemes.

| Reference | Ensemble | Architecture | Area of | Network | Num. NN | Num. | Num. NN sel. | Num. NN sel. | Final |
|---|---|---|---|---|---|---|---|---|---|
| | approach | selection | application | type | generated | clusters | per cluster | if no clustering | combination |
| Hansen & Salamon (1990) | Sequential. | Architecture selected after a small sensitivity analysis. | Classification. | Feedforward. | Up to 15 | NA | NA | All. | Majority consensus, plurality consensus. |
| Jacobs & Jordan (1991) | Dynamic, competitive. Networks compete to learn from different parts of the input space. | Fixed architecture. | Pattern matching (vision and control). | Unspecified. | Unspecified. | NA | NA | Unspecified | Through a gating network. |
| Liu & Yao (1999) | Dynamic, cooperative. Networks are trained simultaneously, allowing for interactions between them and to specialise. | Fixed architecture (1 hidden layer). | Forecasting and classification. | Multilayer perceptron. | Unspecified. | NA | NA | Unspecified. | Average, winner take all for classification. |
| Lu & Ito (1999) | Sequential. | Different numbers of hidden nodes and hidden layers are tried, but not very systematically. | Classification. | Multilayer perceptron or multilayer quadratic perceptron. | Unspecified. | NA | NA | All. | Based on min and max operators. |
| Drezga (1999) | Sequential. | Trial and error. Unespecified range of parameters. | Load forecasting. | Feedforward NN. | 2 | NA | NA | All. | Average. |
| Siwek et al. (2009) | Sequential. | Trial and error. | Load forecasting. | Multilayer perceptron layer (MLP), self organizing map (SOM), and fuzzy SOM. | 3 | NA | NA | All. | Simple and weighted average, Blind Source Separation (BSS), Principal Component Analysis (PCA) decomposition. |

Ensemble generation schemes (continued).

| Reference | Ensemble approach | Architecture selection | Area of application | Network type | Num. NN generated | Num. clusters | Num. NN sel. per cluster | Num. NN sel. if no clustering | Final combination |
|---|---|---|---|---|---|---|---|---|---|
| Liu et al. (2000) | Dynamic. Generation and pruning are simultaneous. Negative correlation is used. | Fixed architecture. Random initial weights restricted to a small range. | Classification. | Feedforward. | 25 (initial population) | Min. 3 and max. 25 | 1 | NA | Average, majority voting and winner-takes-all. |
| Zhou et al. (2002) | Sequential. A genetic algorithm is used for pruning. | Fixed architecture (1 hidden layer). | Regression (predefined functions) and classification (benchmark problems). | Feedforward. | Unspecified. | NA | NA | All resulting from the evolutionary algorithm. | Average for forecasting, maximisation procedure for classification. |
| Taylor & Buizza (2002) | Sequential. | Fixed architecture. | Load forecasting. | Feedforward NN. | 51 | NA | NA | All. | Average. |
| Bakker & Heskes (2003) | Sequential. | Fixed architecture. | Forecasting. | Feedforward. | 50 models per ensemble (10 ensembles), | Determined by an algorithm based on Annealing. | 1 | NA | Weighted average. |
| Islam et al. (2003) | Constructive: sequential, with negative correlation, no pruning. | Number of nodes in NNs in the ensemble is increased gradually as part of the optimisation. | Forecasting and classification with benchmark problems. | Multilayer peceptron. | Unspecified. | NA | NA | All. | Average. |
| Maqsood et al. (2004) | Sequential. | Architecture is determined by trial and error. | Classification and forecasting of weather variables. | Feedfoward. | Unspecified. | NA | NA | All. | Average, weighted average, winner takes all. |
| Yu et al. (2005) | Sequential. | Architecture of NN is determined through trial and error. | Forecasting foreign exchange rates. | Feedforward. | Dependent on the decomposition of the series | NA | NA | Depending on the result of PCA | Another NN. |
| Abdel-Aal (2005) | Sequential. | Not specified. | Load forecasting. | Feedforward NN, abductive network. | 3 | NA | NA | All. | Weighted average based on variance. |

Ensemble generation schemes (continued).

| Reference | Ensemble approach | Architecture selection | Area of application | Network type | Num. NN generated | Num. clusters | Num. NN sel. per cluster | Num. NN sel. if no clustering | Final combination |
|---|---|---|---|---|---|---|---|---|---|
| Krasnopolsky (2007) | Sequential. | Fixed architecture. | Data assimilation, environmental models. | Feedforward. | From 8 to 12 | NA | NA | All. | Linear averaging and nonlinear averaging (with another NN). |
| Chen & Yao (2007) | Dynamic (evolutionary). | Randomisation of the number of neurons and the number of features. The number of layers is fixed (3). | Classification (benchmark problems). | Feedforward. | 200 (initial population) | NA | NA | Resulting population of the evolutionary algorithm. | Major voting. |
| Yu et al. (2008) | Sequential. | The number of layers is fixed (3), with unspecified number of units. | Forecasting crude oil price. | Feedfworard. | Dependent on the number of sub-series derived from the original series. | NA | NA | All. | NN of linear type. |
| Yu et al. (2008) | Sequential. | Fixed (unspecified) architecture | Foreign exchange rates forecasting. | RBF. | Unspecified (high) | NA | NA | Dependent on minimisation of conditional generalised variance (CGV) | NN of RBF type. |
| Fan et al. (2009) | Sequential. | Fixed architecture. The number of hidden nodes is determined adding neurons and stopping when the error is minimum. | Load forecasting. | Feedforward with 3 layers. | Unspecified | NA | NA | All. | Median. |
| Salcedo-Sanz et al. (2009) | Sequential. | Fixed architecture. | Wind-power forecasting. | Feedforward NN. | 9 | NA | NA | 1 (if lowest error used); all if aggregation used. | Selection of best NN, another NN or a hidden layer. |

Ensemble generation schemes (continued).

| Reference | Ensemble approach | Architecture selection | Area of application | Network type | Num. NN generated | Num. clusters | Num. NN sel. per cluster | Num. NN sel. if no clustering | Final combination |
|---|---|---|---|---|---|---|---|---|---|
| Adeodato et al. (2011) | Sequential. | Number of hidden units in the range 1-30 and number of inputs in the range 12 (one year) to M where M is determined using autocorrelation or Fourier analysis. | Forecasting with series from NN3 competition. | Multilayer perceptron. | 15 | NA | NA | All. | Median. |
| Matijaš et al. (2013) | Sequential. | NA | Load forecasting. | Random walk (RW) algorithm, autoregressive moving average (ARMA), similar days algorithm, layer recurrent neural network (LRNN), MLP, v-Support vector regression (v-SVR), and robust LS-SVM (RobLSSVM). | 7 | NA | NA | All. | Ranking through meta-learning. |

Ensemble generation schemes (continued).

| Reference | Ensemble approach | Architecture selection | Area of application | Network type | Num. NN generated | Num. clusters | Num. NN sel. per cluster | Num. NN sel. if no clustering | Final combination |
|---|---|---|---|---|---|---|---|---|---|
| Kourentzes et al. (2014) | Sequential. | Step-wise regression (for inputs) and minimisation of the validation MSE (for num. neurons) | Forecasting economic and retail series. | Feedforward. | 10 to 100 | NA | NA | Varying (dependent on median or mode) | Average, median, mode. |
| Qiu et al. (2014) | Sequential. | Fixed architecture | Load forecasting | Deep belief network | 20 | NA | NA | All. | Support vector regression. |
| Burger & Moura (2015) | Sequential. | NA | Load forecasting | Ordinary Least Squares with $l_2$ Regularisation (Ridge),<br>K nearest neighbour. | 16 | NA | NA | 1 (if lowest error used); all if aggregation used. | Gating (selection of best model) |
| Hassan et al. (2015) | Sequential. | Fixed set of architectures | Load forecasting. | Feedforward | 100 | NA | NA | NN with best validation MAPE; optimal trimming; all. | Average, trimmed, Bayesian model average. |

In the combining stage, the literature tends to focus on alternative ways to aggregate point forecasts. However, one could argue that combinations, or consensus, in a broader sense, is a fundamental feature of reality and science, which suggests potential avenues for research. Perhaps, by considering forecasting models as objects, with different components and outcomes, a forecast combination can be extended from the standard aggregation of point forecasts to take into account differences in model specifications, and, in particular cases of neural networks, their internal components, i.e, their structure.

The exploration of such extensions in forecast combination can have an impact in related concepts, as for example, forecast encompassing. This concept helps in assessing whether one forecast, or set of forecasts, includes all information present in another forecast or set. Encompassing tests are relevant in situations where no dominant model can be identified, and, therefore, combination is preferred over a single forecast (Timmermann, 2006; Makridakis & Winkler, 1983). The tests can be done between pairs of forecasts or between sets of more than two (Harvey & Newbold, 2000). Some encompassing tests are based on regression analysis (Fisher & Wallis, 1990; Fair & Shiller, 1990; Cooper & Nelson, 1975). Other studies focused on the conditions under which encompassing tests can be applied, the implications of having multi-step ahead forecasts and the presence of non-normality or heterokedasticity in forecast errors (see Newbold & Harvey, 2007).

In summary, two gaps are identified. The first is related to the study conducted prior to the selection of models for an ensemble. The second is related to the way in which the combination of NN models in ensembles is done. The following section formulates the related research questions.

## 2.4 Research Questions

As discussed above, a preliminary analysis of how sensitive are forecasts to design parameters would be useful in the construction of ensembles. Research conducted with design of experiments is limited and can be expanded to aid the selection of models for an ensemble. The first question to be addressed in this dissertation is thus,

– How can a sensitivity analysis, based on design of experiments, be used to aid the selection of NN models for a forecasting ensemble?

From the literature review on forecast combination and NN ensembles, it is also clear that very sophisticated ways of performing combination of forecasts have emerged. However, there is limited research on combination approaches which consider the internal characteristics of the models. Bakker & Heskes (2003) considered a research direction that would consist in using clustering of structural parameters and summarising models based on such clusters. Matijaš et al. (2013) exploited the interpretation of a learning process as a link between a problem and a solution space and tried to use such representations to explore combination of forecasts in a wider sense. The general idea of expanding the model information used in combining forecasts can be further explored. The specific case of using clustering and structural parameters can be explored with different clustering techniques and time series. As NNs have a clear structural representation, which is intended to store some *knowledge* about the problem at hand (pattern matching), it is important to investigate the extent to which the inclusion of the structure of the models into the forecast combinations improves the accuracy of predictions. Additionally, we lack knowledge concerning the use of model structure when combining statistical models as well. Therefore the following questions emerge:

– How can the structure of neural networks be combined?

- How do the proposed models perform in forecasting?

- How can the structural combination approach be extended from NN to other forecasting models?

### 2.4.1 Research Objectives

Given the research questions outlined above, the objectives of this dissertation are the following:

- To gain further insights into the choice of models, when developing ensembles, in terms of the types of time series and forecast horizons, over existing research on the performance of NNs in order to better understand their behaviour given different data generating processes.

- To use the knowledge gained through the sensitivity analysis of NN performance in the construction of ensembles.

- To implement a NN forecasting model combination that incorporates model structural information.

- To assess the forecasting performance of such combination scheme.

- To propose a form of model structural combination for statistical forecasting models and assess their performance.

In order to fulfil these objectives, first a thorough exploration of NN structures with synthetic time series that cover a wide range of processes is undertaken in Chapter 3. Subsequently, two chapters are devoted to the development of a structural combination approach, first with NNs (Chapter 4) and then with a statistical model (Chapter 5) that has been found to outperform a range of models when predicting electricity demand (Taylor et al., 2006). Chapter 6 will summarise, assess the implications of the research and conclude this dissertation. Overall, the main contributions of this dissertation are concerned with the first and last stages depicted in Figure 2.1, which are the generation and combination of forecasting models.

# Chapter 3

# A Sensitivity Analysis of the Performance of Feed-Forward Neural Networks

## 3.1 Abstract

As highlighted in the previous chapter, forecasting time series with ensembles of models is a promising research avenue (Crone et al., 2011). Neural networks (NN) ensembles comprise 3 stages: the generation of the models, the pruning or selection of models and the integration, as described in section section 2.1.2.3. During the first stage, it is important that the modeller understands which NN models are worth combining. Given that literature suggests that NN can be very volatile (e.g. Geman et al., 1992; Breimanet al. , 1996; Dietterich, 1997; Teräsvirta et al., 2005; Medeiros et al., 2006), the present study examines individual NN model behaviour in order to aid the selection of models for ensembles. Simple feed-forward neural networks are explored, for which different configurations of key parameters (sample size, number of inputs and number of neurons) are used to model simulated time series data of different complexity and assess the sensitivity of forecasting performance to the chosen parameters.

The design of experiments (Montgomery, 2008) is used to evaluate the influence of different factors on the performance of NNs. Results show that there are significant effects of different factors on forecasting performance. Graphic and statistical databases that were created in the study have also facilitated a more objective assessment of models to be combined.

## 3.2 Introduction

In forecasting with NN ensembles, sometimes the structure of the models is determined automatically, for example through re-sampling (see Chen & Yao, 2007), cooperative training (Islam et al., 2003), negative correlation training (Liu & Yao, 1999; Chen & Yao, 2009) and evolutionary algorithms (Chandra & Yao, 2006; Yao & Islam, 2008). When using automated or partly automated schemes, the absence of an initial analysis seems to be justifiable. However, in most studies, no preliminary analysis is reported. As shown in Table 2.1, several design factors in ensembles tend to be set constant, whereas others vary without clear justifications concerning the modeller's choice. Understanding how the choice of parameters may impact performance is helpful in order to assess the limitations of individual models and different architectures. Even if models were to be automatically selected, it would be helpful to identify regions in the input space where models have difficulties in learning, or regions of the parameter space that show greater volatility. Consequently, an assessment of how model fit and performance are sensitive to the specification of NN is critical for the development of ensembles.

This chapter develops a sensitivity analysis, based on design of experiments (DOE), which is a standard methodology in industrial assessments of new products. DOE provides guidelines for planning and conducting experiments and analysing the results so that objective conclusions are obtained (Montgomery, 2008, p. 1). DOE has been less used in simulation studies, despite the potential benefits it can offer (Balestrassi et al., 2009).

The analysis presented here will aid the selection of NN models for a forecasting ensemble. It uses graphical summaries and statistical tests, which allow the modeller to identify better behaved models. The modeller might want to limit the volatility of NN in the ensemble for different tuning tasks. She might consider ensembles with different network architectures. A detailed study, as presented here, can be a basis upon which to make decisions concerning the specification of ensembles.

Sequential approach for model generation is adopted, where each NN is trained independently. Figure 3.1 shows a general process with DOE. The goal is to use knowledge gained through DOE to choose individual models for an ensemble.

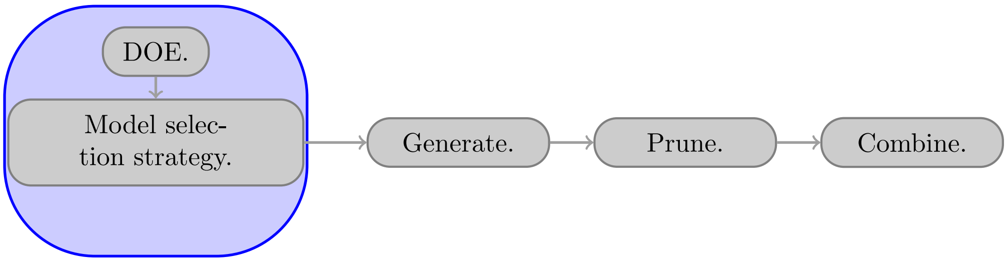


Figure 3.1: General steps in ensemble generation with DOE.

Previous research on NN specification and DOE have focused on the most appropriate models for a given forecasting problem. Zhang et al. (2001) and Balestrassi et al. (2009) conducted experimental studies to evaluate the effect of several configuration parameters of feed-forward networks on performance metrics. The first study considered non-seasonal series and examined three factors: number of inputs, number of neurons and sample size. For each combination of factor levels, 30 different experiments were conducted. Each experiment corresponded to a time series replication (obtained through its generating process), which was fitted with a network specified according to the given combination of factor levels. The second study included a seasonal component in some series and expanded the factors considered. However, given the quick growth in the number of experiments needed (and the corresponding increase in computing time), a strategy of screening, the Taguchi approach (Taguchi & Yokoyama, 1993), was adopted in order to have fewer experiments per factor combination. Khadem & Dillon (2012) also resorted to an abbreviated procedure: instead of using a full factorial design, they used an orthogonal design. The full factorial design requires $L^f$ experiments[1], with $f$ being the number of factors and $L$ the number of levels per factor, while the orthogonal design requires only $L * f$ experiments. Such approach was used to define neural network (NN) parameters for a single forecasting problem (traffic flow). The main outcome was the visualisation of main effects of different factors over a single error metric[2]. Crone & Dhawan (2007) also used design of experiments to analyse different lengths of seasonality. They concluded that the results of their sensitivity analysis may serve only as a guidance for future modelling of seasonal time series. In order to establish NNs as an alternative to statistical methods, they considered as important the extension of the sensitivity analysis to a full factorial design, including the use of statistical tests such as ANOVA and multiple performance metrics.

The studies cited above show that the number of combinations of parameters in the specification of networks is considerably high and at some point the computing time needed goes beyond practical means. Consequently, there is a trade-off between scope and detail. That is, either the study considers a reduced number of factors (parameters or design decisions in networks) with a modest number of levels and keeps a reasonable number of replications per combination, or the number of factors is increased at the expense of reducing the number of replications. Overall, previous studies show that such experiments allow for the simultaneous visualisation of different factors that affect the forecast error.

Simple seasonality was considered by Balestrassi et al. (2009) in their synthetic series and was generated through the addition of a time lag of 24 to the generating processes. Their idea was to use data with characteristics present in electricity load, daily electricity prices or water consumption time series. In the present study, seasonality is analysed, and is also extended to double seasonal series, which are

[1] Assuming the same number of levels for all factors.

[2] A main effect is the effect of one factor on the outcome variable, normally displayed in a graph (Montgomery, 2008, p. 5)

characteristic of short-term electricity demand (daily: peak and low demand; weekly, as consumption during weekdays differs from weekends). Its inclusion expands the studies by Zhang et al. (2001) and Balestrassi et al. (2009) and Crone & Dhawan (2007).

In the present study, a trade-off was made between the number of factors and the number of experiments per factor combination. The factors considered by Zhang et al. (2001) were kept, and summary graphs displaying the effect of a factor on a forecast error metric were adopted, that also included 95% confidence bands. Additionally, a version of summary graphs was constructed using cross-validation (Bishop, 1995). With such procedure, the training is conducted with a time series divided in $n$ segments. For a given experiment, the network is trained $n$ times, each time omitting one segment, which is used for testing. A summary statistic of performance metric obtained from the $n$-fold cross-validation is used to construct the graphs, instead of the single value obtained without cross-validation. The graphs obtained without cross-validation highlight the variance in forecasting performance of the actual networks, whereas the graphs based on cross-validation highlight the more general tendency. In contrast to previous studies, which focused on one-step-ahead forecasts, multiple step-ahead forecasts are here considered. Additionally, the number of experiments per factor combination was increased to 100 (30 were used by Zhang et al., 2001). The aim is to observe the sensitivity of forecast accuracy with respect to some structural parameters. Feed-forward neural networks were selected due to the simplicity of their architectures, popularity, and the greater likelihood of obtaining manageable training times for the ensembles.

In Section 3.3, the time series used in the present study are described. Section 3.4 describes the design used for the sensitivity analysis. Section 3.5 presents numerical and graphical results. Section 3.6 discusses the results and, finally, Section 3.7 states the main conclusions.

## 3.3 The Synthetic Time Series

The models used to generate synthetic time series are listed below. For each group (non-seasonal and seasonal), the level of non-linear complexity is in ascending order.

### Non-seasonal time series

The generating processes for these series are the same used by Zhang et al. (2001).

- Sign autoregressive (SAR) model:

$$\begin{aligned} y_t &= sign(y_{t-1}) + \varepsilon_t, \\ sign(x) &= 1 \text{ if } x > 0, \\ &= 0 \text{ if } x = 0 \text{ and} \\ &= -1 \text{ if } x < 0 \end{aligned} \tag{3.1}$$

- Bilinear model 1 (BL1) :

$$y_t = 0.7y_{t-1}\varepsilon_{t-2} + \varepsilon_t \tag{3.2}$$

- Bilinear model 2 (BL2) :

$$y_t = 0.4y_{t-1} - 0.3y_{t-2} + 0.5y_{t-1}\varepsilon_{t-1} + \varepsilon_t \tag{3.3}$$

- Threshold autoregressive (TAR) model :

$$\begin{aligned} y_t &= 0.9y_{t-1} + \varepsilon_t \text{ for } |y_{t-1}| \leq 1 \text{ and} \\ &= -0.3y_{t-1} - \varepsilon_t \text{ for } |y_{t-1}| > 1 \end{aligned} \tag{3.4}$$

- Nonlinear autoregressive (NAR1) model :

$$y_t = \frac{0.7|y_{t-1}|}{(|y_{t-1}| + 2)} + \varepsilon_t \tag{3.5}$$

- Nonlinear moving average (NMA) model :

$$y_t = \varepsilon_t - 0.3\varepsilon_{t-1} + 0.2\varepsilon_{t-2} + 0.4\varepsilon_{t-1}\varepsilon_{t-2} - 0.25\varepsilon_{t-2}^2 \tag{3.6}$$

- Smooth transition autoregressive (STAR1) model :

$$y_t = 0.8y_{t-1} - 0.8y_{t-1}(1 + e^{-10y_{t-1}})^{-1} + \varepsilon_t \tag{3.7}$$

- Smooth transition autoregressive (STAR2) model :

$$y_t = 0.3y_{t-1} + 0.6y_{t-2} + (0.1 - 0.9y_{t-1} + 0.8y_{t-2})(1 + e^{-10y_{t-1}})^{-1} + \varepsilon_t \tag{3.8}$$

For all these processes $\varepsilon_t \sim NID(0, 1)$.

## Single seasonal and double seasonal time series

- Single seasonal synthetic (Synthetic-1S) :

$$\begin{aligned} y_t(k) &= l_t + w_{t-s_2+k} + \phi^k(y_t - (l_{t-1} + w_{t-s_2})) + \varepsilon_t \qquad (3.9) \\ l_t &= \lambda(y_t - w_{t-s_2}) + (1-\lambda)l_{t-1} \\ w_t &= \omega(y_t - l_{t-1}) + (1-\omega)w_{t-s_2} \end{aligned}$$

$y_t(k)$ is the simulated series value at time $t + k$, $l_t$ denotes the smoothed level and $w_t$ denotes the seasonal index. $\varepsilon_t \sim NID(0, \sigma^2)$, with $\sigma^2$ being a constant variance. Parameters are $\lambda = 0.2$; $\omega = 0.01$; $\phi = 0.943$; $s_2 = 12$.

This single-seasonal series was simulated through the following steps:

1. Generate an initial pattern for the seasonal cycle, thus covering values for $t = 1, \ldots, s_2$ of equation

$$y_t(k) = l_t + w_{t-s_2+k} + \phi^k(y_t - (l_{t-1} + w_{t-s_2})) + \varepsilon_t \tag{3.10}$$

$$\begin{aligned} l_t &= \lambda(y_t - w_{t-s_2}) + (1-\lambda)l_{t-1} \\ w_t &= \omega(y_t - l_{t-1}) + (1-\omega)w_{t-s_2} \end{aligned}$$

2. Use the expression to generate the values starting from the last point generated previously. This implies that information in one cycle is used to generate the next, as illustrated in Figure 3.2.

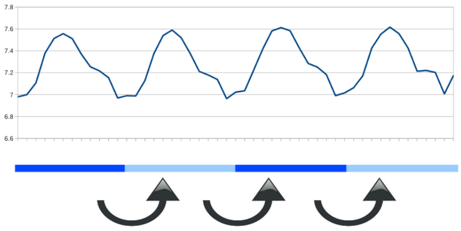


Figure 3.2: Illustration of single-seasonal synthetic time series generation.

3. Repeat step 2 until a length of 10000 is reached.
4. Take a subset of the series according to the length needed to fit and test the forecasting models.

- Double seasonal synthetic (Synthetic-2S) exponential smoothing:

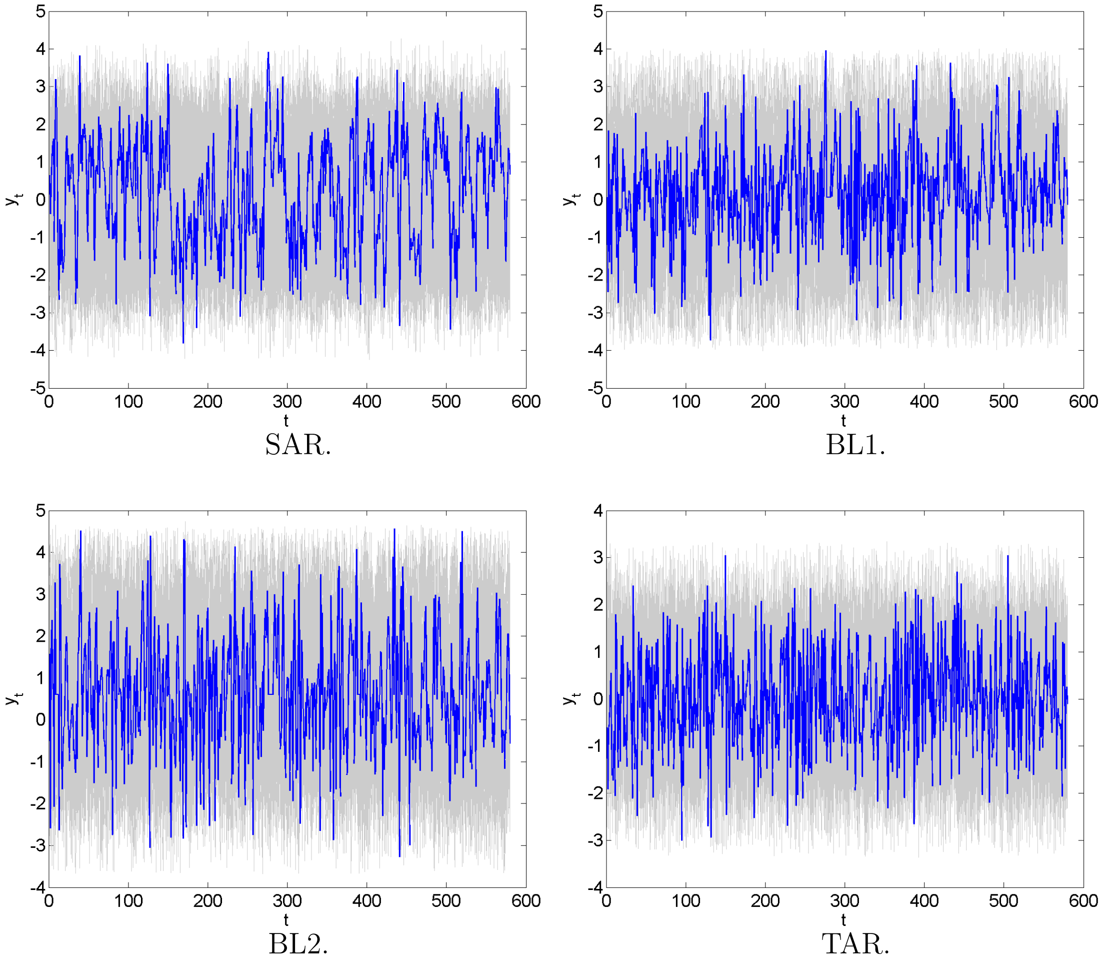
$$\begin{aligned} y_t &= l_{t-1} + d_{t-s_1} + w_{t-s_2} + \phi(y_{t-1} - (l_{t-2} + d_{t-s_1-1} + w_{t-s_2-1})) + \varepsilon_t \quad (3.11) \\ l_t &= \lambda(y_t - d_{t-s_1} - w_{t-s_2}) + (1-\lambda)l_{t-1} \\ d_t &= \delta(y_t - l_{t-1} - w_{t-s_2}) + (1-\delta)d_{t-s_1} \\ w_t &= \omega(y_t - l_{t-1} - d_{t-s_1}) + (1-\omega)w_{t-s_2} \end{aligned}$$

Here $l_t$ denotes the smoothed level, $w_t$ denotes the long cycle seasonal index and $d_t$ denotes the short cycle seasonal index. Parameters are $\lambda = 0.2$; $\delta =$

0.13; $\omega = 0.3$; $\phi = 0.5$; $s_1 = 3$; $s_2 = 12$. After generating a series of length 10000, a subset is taken, according to the length needed to fit and test the forecasting models.

Initial values for the time series were not based on an actual load curve. This selection had the purpose of testing the forecasting model combinations in different conditions. This choice of having quarterly cycle within an yearly cycle allowed us to preserve the factor combination in the design of experiments (and the corresponding number of simulations) within feasible computing cost boundaries at the time of this research.

Figures 3.3 and 3.4 show different replications of the series and illustrate the range (with the set of series in grey) and the shape of each time series. STAR2 series is particular because it is more predictable than the other series (clearer patterns are visible in the time series when the noise is smaller). Although this regularity is not appreciated in the series generated with $\sigma^2 = 1$ (adopted here following Zhang et al., 2001), it is, nevertheless, more predictable than all other non-seasonal series.



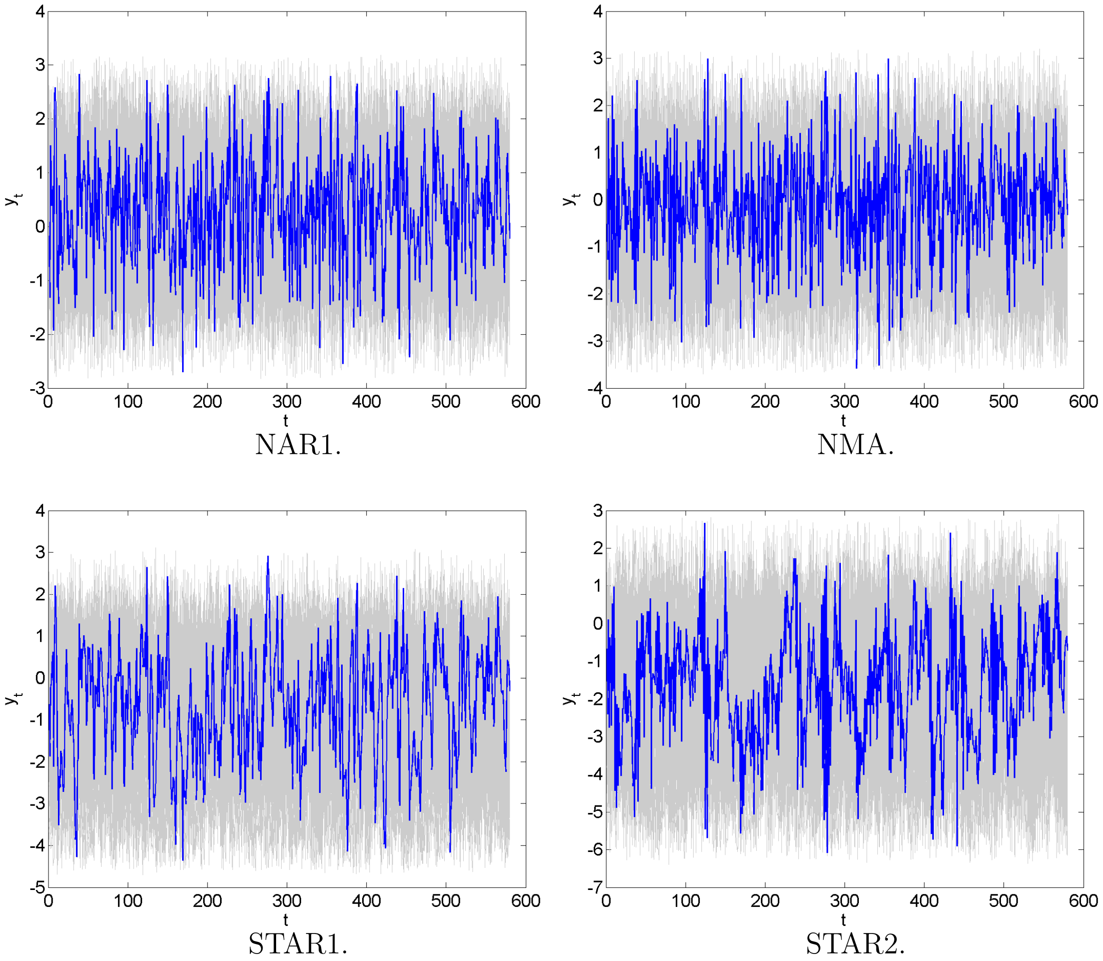
SAR.

BL1.

BL2.

TAR.

Figure 3.3: Simulated series (non-seasonal). From 100 replications, in grey colour, one is plotted in blue.



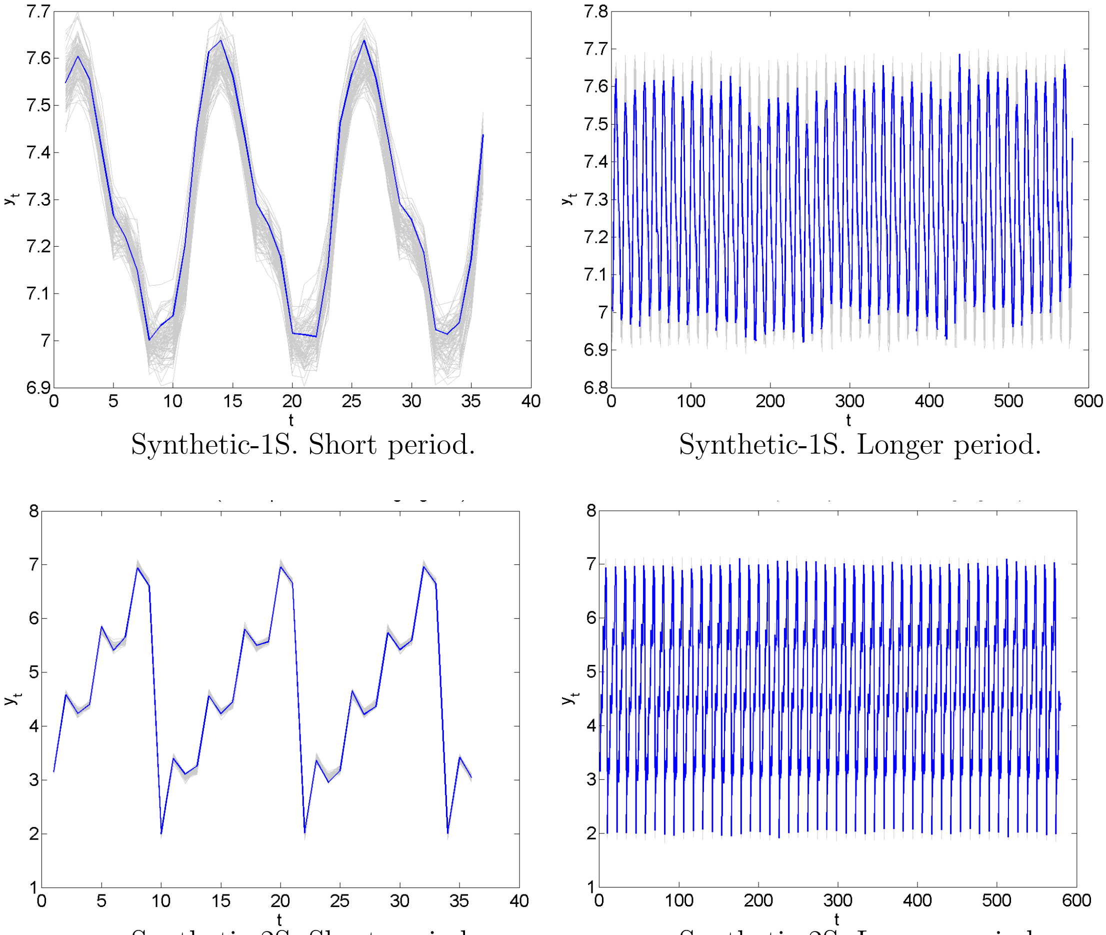
NAR1.

NMA.

STAR1.

STAR2.

(Continued) Simulated series (non-seasonal). From 100 replications, in grey colour, one is plotted in blue.



Synthetic-1S. Short period.

Synthetic-1S. Longer period.

Synthetic-2S. Short period.

Synthetic-2S. Longer period.

Figure 3.4: Simulated series (seasonal). From 100 replications, in grey colour, one is plotted in blue.

## 3.4 Design

The study is focused on assessing the influence of key design parameters in the forecasting accuracy of neural networks. The type of network and the multi-step-ahead forecasts approach influence the way the problem is treated. The networks are restricted to the feed-forward type and the multi-step ahead forecast approach is direct (a review of approaches is provided in Appendix A). That is, a separate NN is used for each forecast horizon. This approach allows for networks to specialise in a specific forecast horizon (Gouriveau & Zerhouni, 2012) and, consequently, for a division of the training task between multiple machines. In sum, the aim of the study focuses on forecasting synthetic time series with feed-forward NNs and the results are expected to show the influence of several design parameters (number of lagged inputs, number of hidden units and sample size) on the forecasting performance of such models for every time series considered.

The use of design of experiments has an impact in the use of some forms of ensemble generation. Negative correlation, an approach mentioned in the introduction, creates pools of NNs with negatively correlated errors during the training period. It has two forms: simultaneous training and separate training. Simultaneous training would imply a relation between NNs during training because the algorithm is global and takes into account the correlation of models errors during the optimisations (see for example Liu & Yao, 1999). However, the sensitivity analysis presupposes independence (independent experiments, that is, independent training). The approach followed here was to generate the NNs independently to facilitate the statistical analysis.

### 3.4.1 Choice of Factors, Levels, and Ranges

The factors selected here are the number of inputs for the network (NI), which correspond to lagged values of the series; the number of hidden units, or neurons

(NU); the sample size (SS) and the presence or absence of pruning (PU) in the weights of hidden units. The levels were chosen to achieve a good granularity. Table 3.1 provides a list of the factors considered and specifies which varied and their respective ranges. Further details are summarised as follows:

- Number of inputs: this factor has been found to impact the structure of the models (Zhang et al., 2001; Crone & Kourentzes, 2010). Given that the underlying processes for the non-seasonal series do not use information beyond the previous two lags, a maximum of 6 lags was chosen for these types of series.
- Number of hidden layers: following previous literature (Zhang et al., 2001; Crone & Dhawan, 2007) single layered networks were used.
- Number of hidden units: following Zhang et al. (2001),it was set as twice the maximum number of inputs.
- Activation functions: following Zhang et al. (2001) it was kept constant.
- Initial values for the weights: are in the range $[-2, 2]$, according to the Nguyen-Widrow algorithm (Nguyen & Widrow, 1990), which generates them randomly within several constraints in order to speed up the training process.
- Combination coefficient ($\mu$): The Levenberg-Marquardt algorithm is a combination of the steepest descent algorithm and the Gauss-Newton algorithm, switching between the two during the training process (Yu & Wilamowski, 2010). The default initial value of the combination coefficient in the Matlab® toolbox is set to 0.001, which starts the training closer to the Gauss-Newton algorithm (Haykin, 1999, p. 148).
- Training algorithm: The Levenberg-Marquardt algorithm was chosen due to speed. One of the main features is its combination coefficient, $\mu$, which can be

interpreted (if it is very big) as the learning coefficient in the steepest descent method $\alpha = 1/\mu$ (Yu & Wilamowski, 2010).

- Stopping criteria: The training algorithm stopping criteria (a standard routine in Matlab® NN implementation) takes into account several conditions: the maximum number of epochs, the maximum amount of time for training, the performance goal, the performance gradient, the upper limit for $\mu$ and the number of times the performance in the validation period has increased since the last time it decreased.

- Input scaling: linear scaling of inputs in the interval $[-1, 1]$ is performed (additional common configurations can be found in Zhang et al., 1998).

- Sample size: sample sizes of 340, 580, 1060 observations were used.

- Data configuration for training, testing and validation: 100 observations were used for out-of-sample testing leaving 240, 480 and 960 for in-sample (training + validation) in each sample size. The number of observations for validation was 10% of the training size, that is 24, 48 and 96 observations. Data division is sequential, following Adeodato et al. (2011) and Adya & Collopy (1998).

- Data usage: a rolling window was used to train the NNs for different forecast horizons. Figure 3.5 exemplifies the partition with a small data set.

- Treatment of extreme values: In the non-seasonal series the outliers were replaced by the average of the series. The identification of outliers was done using an algorithm proposed by Janczura et al. (2013). For the single-seasonal and double-seasonal series no treatment of extreme values was applied as the realisations of the processes were very regular. This was achieved by selecting a level of noise in the generating process that preserved the characteristics of the time series.

Table 3.1: Factors.

| **Factor** | **Symbol** | **Levels** |
|---|---|---|
| Number of inputs | NI | $\mathbf{1}, \ldots, \mathbf{6}$ (non-seasonal) |
| Number of hidden layers | NL | 1 |
| Number of hidden units | NU | $\mathbf{1}, \ldots, \mathbf{2N}$, where $N$ is the number of inputs |
| Activation function for hidden nodes | AF1 | Tangent Sigmoid |
| Activation function for the output node | AF2 | Linear |
| Initial values for the weights | W0 | Values in the range [-2 2] established by the Nguyen-Widrow algorithm (there is a degree of randomness) |
| Training algorithm | TA | Back-propagation with Levenberg-Marquardt optimisation. |
| Stopping criteria | SC | * The maximum number of epochs (repetitions) is reached: 4000.<br>* The maximum amount of time is exceeded: $\infty$<br>* Performance is minimised to the goal: 0<br>* The performance gradient falls below $min_{grad} : 10^{-10}$<br>* $\mu$ exceeds $\mu_{max} = 10^3$<br>* Validation performance has increased more than $max_{fail}$ times since the last time it decreased (when using validation): 6 |
| Data normalisation | DN | Yes |
| Combination coefficient ($\mu$) | MU | 0.001 |
| Prune units | PU | **Yes, No** |
| Prune input variables | PI | No |
| Sample size | SS | **240, 480, 960.** $SS = 240$ corresponds to setting with training and validation sizes totalling 240 plus a testing size of 100 for a total of 340. The same applies for the other two configurations. |
| Data configuration for training, validation and testing (training+validation=in-sample period; testing=out-of-sample period) | DC | Conf. 1: $Ntr = 216(63.53\%)$, $Nva = 24(7.06\%)$ $Nte = 100(29.41\%)$; conf. 2: $Ntr = 432(74.48\%)$, $Nva = 48(8.28\%)$, $Nte = 100(17.24\%)$; conf. 3: $Ntr = 864(81.51\%)$, $Nva = 96(9.06\%)$, $Nte = 100(9.43\%)$ |
| Extreme values treated | EV | In some of the series. |
| Sampling method | SM | block, cross-validated |
| Forecast approach | FA | Direct: a separate model for each step ahead |

In bold are the factors which vary in the study.

- Sampling method: block and cross-validated sampling were used. The first corresponds to sequential data division as mentioned above and the second is conducted with a 10-fold cross-validation procedure (Bishop, 1995).

- Prune units: Pruning was conducted in the hidden layer weights of the networks. A threshold of tr=0.05 was used (following Balestrassi et al., 2009).

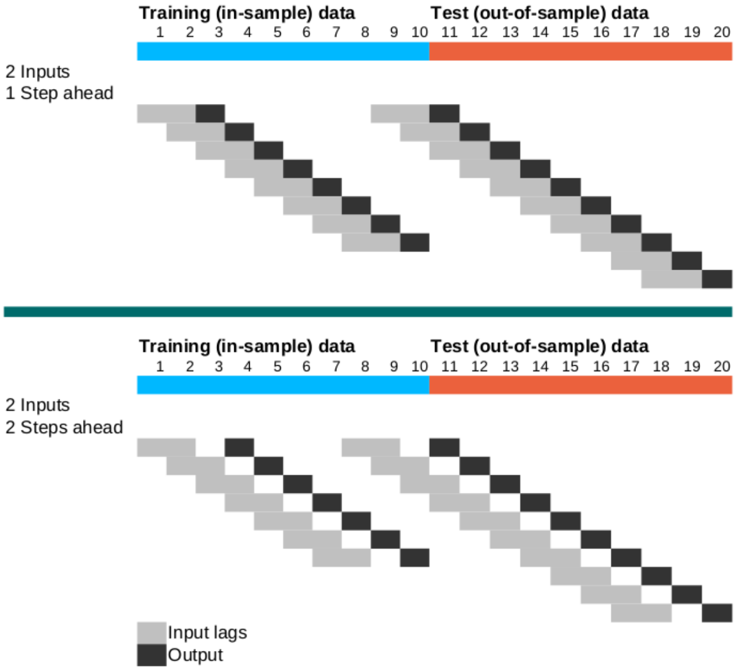


Figure 3.5: Rolling window for training (fitting with in-sample data) and testing with out-of-sample data.

### 3.4.2 Selection of the Response Variable

Given the purpose of the study, the in-sample and out-of-sample mean squared error (MSE) and mean absolute error (MAE) were chosen as the response variables. The first metric was used in the sensitivity analysis by Zhang et al. (2001) and the latter is included as an alternative metric. Percentage-based metrics were omitted as they were noticed to be unstable for the time series whose values are close to zero.

### 3.4.3 Choice of Experimental Design

A full factorial design was chosen for the number of inputs, number of neurons, and sample size.

### 3.4.4 Methodology to Perform the Experiments

For a given time series, 100 experiments are conducted per factor combination. For each experiment (corresponding to a realisation of the generating process), there are several results: the in-sample MSE and MAE and the out-of-sample MSE and MAE. These measures are used to conduct statistical tests (ANOVA, Kruskal Wallis, and Jonckheere-Terpstra, explained in the following section) and generate graphs of factor effects.

For a given combination of factors, the set of time series replications ($n = 100$) is used for training (fitting) 100 NNs. The in-sample MSE is examined and the training is repeated for synthetic series that lead to extreme MSE values, that is when $MSE > q_2 + 3 * Iqr$, where $q_2$ is the upper quartile and $Iqr$ is the inter-quartile range for the set of 100 MSEs during training. The procedure is conducted while extreme MSEs persist, or until 10 iterations are reached. The size of the problem for this routine is determined by $N_h$ (the number of forecast horizons), $N_{NI}$ (the number of values used for factor $NI$), $N_{NU}$ (the number of values used for factor $NU$) and $N_{SS}$ (the number of values used for factor $SS$). Factor $PU$ is treated in a reprocessing algorithm excluding training iterations. For this study, $N_{NU} = 2N_{NI}$ and $N_{NI} \approx N_h$. Therefore the order of the routine is $O(N_{SS} \cdot N_h^3)$. The computational cost is therefore very high, but the algorithm can be run in parallel in different machines.

### 3.4.5 Statistical Analysis of Design Factors

A factorial design can produce many results, depending on the number of factors and levels. In our case, there are 4 varying factors (NI, UN, SS, PU) comprising 1728 combinations of levels. For each case there are 12 models, one for each step ahead, that are iterated 100 times, totalling 2073600 trials. A succinct visualisation of the multidimensional data produced is unfeasible. Therefore, the analysis, although based on the full set of results, will focus on sub-samples.

The first step of the analysis uses graphic information to study sensitivity and variability of the metrics to the factors NI and NU under conditions determined by factors PU, SS, and SM. That is, the behaviour of the metrics is assessed for NI and NU, separately, for cases $(PU = 0, PU = 1) \times (SS = 240, SS = 480, SS = 960) \times (SM = block, SM = cross - validated)$. This space comprises the combination of pruned and non-pruned models, different sample sizes and non-cross validated vs. cross-validated conditions. A graph depicts the behaviour of a performance metric against one of the factors. Confidence bands are added in order to examine variability. Another graph assesses serial correlation in forecast errors. It is comprised of cells that display the number of times (out of 100) the Ljung-Box test finds evidence of serial correlation, for a given combination $NI * NU$. Colour arrangement in the cells helps in identifying regions of the $NI * NU$ combinations with specific patterns regarding a test. The second step is a non-parametric analysis of variance (ANOVA) with main effects graphs to study the influence of factors NI, NU, PU and SS on the error metrics.

ANOVA requires homogeneous variance in the groups (defined by factors). In our case, a comparison between two groups can be made if, for example, the mean in-sample MSE (IS MSE) is compared between a group of NNs trained with 2 inputs and a group trained with 3 inputs. It was found that the variance was different in many cases. Consequently, two non-parametric tests were considered: Kruskal

Wallis (K-W) and Jonckheere-Terpstra (J-T). The first helps to assess the influence of the factors on the metrics and the second gives additional information about the direction of the influence: absolute values of the standard J-T statistic greater than 1.65 indicate a significant difference in the medians of the groups (formed with the factor under study). If the value is positive, then it indicates a trend of ascending medians; if it is negative, it indicates a trend of descending medians (see Field, 2009). Having a test based on medians is convenient in the case of data distributions with extreme values (significant kurtosis), as obtained in the present investigation.

The first step in the analysis helps to identify the conditions under which changes in the error metric happen, whereas the second step helps to numerically assess the observed patterns found in the first one: the influence of a factor over an error metric might not be graphically discernible in some occasions, but the statistical tests help in better judging such cases. This is done by looking into the differences in mean and median values for the error metrics, under different conditions (levels of factors).

In the second step, the number of inputs (NI), the number of hidden units (NU) and the sample size (SS) were considered with the full set of replications, whereas the pruning (PU) factor was examined only on those models where pruning was conducted: as mentioned above, a threshold was applied to the hidden layer of NNs (a model's weight was pruned if $WeightValue < 0.05$). That is, only the models which resulted in their weights being pruned were considered for the statistical tests. Comparisons were made between the performance of such models before and after pruning.

An additional step can be applied to identify models with promising forecasting capacity, which can then be combined. It is a simple heuristic, which is the following.

*Base procedure to aid model selection*

1. Screen for serially correlated residuals and forecast errors: look for regions in the NI*NU space where models have better behaved residuals and errors according to Ljung-Box test.

2. Screen for performance: consider the plots of OS MSE (MSE for out-of-sample period) vs. NI, OSMSE vs. NU, main effects graphs and tests (ANOVA, K-W and J-T). Look for instabilities and inflection points (where the metric begins to deteriorate). Together with findings from 1, select the most parsimonious models with better behaved residuals and forecast errors and lower error metric.

3. Compare results based on different sample sizes in order to determine which size provides stability.

In this context, for the selection of models, the out-of-sample period would function as the cross-validation part of the in-sample period. If pruning reduces the error metric in the selected models, it can be taken into account and the models with pruned weights are preferred.

The use of the experimental design described is subject to the ensemble design. If small networks are needed, the study would indicate which are the preferred models with such constraints. If a specific number of inputs is needed, this restriction can be used to examine the graphs and statistical results in order to make the selection.

## 3.5 Results

The analysis of a process is reported in detail, and illustrates the way in which results were analysed. Findings concerning the remaining processes are summarised in Table 3.2. Additional information can be requested from the author.

The generating process for SAR series, in Equation 3.1, has a simple non-linearity with a change in level (structural shift) between -1 and 1. The process works in such a way that there are sequences of either positive or negative numbers (autoregression in sign). A set of 100 replications of the process revealed that the length of the sequence of positive or negative values is on average 6 to 7. The nature of this dependency is reflected in the plot of median MSE vs. forecast horizon produced with data from the fitted NNs (Figure 3.6), where it can be seen how performance erodes up until $h = 6$. For longer forecast horizons, the graph is flat.

Figure 3.6 also includes the plot of average MSE for the best performing NN architectures in each forecast horizon (with 95% confidence bands). The models with lowest average MSE have up to 3 neurons. The number of inputs (lags required) is usually one or two for the first horizon and higher for some of the subsequent horizons.

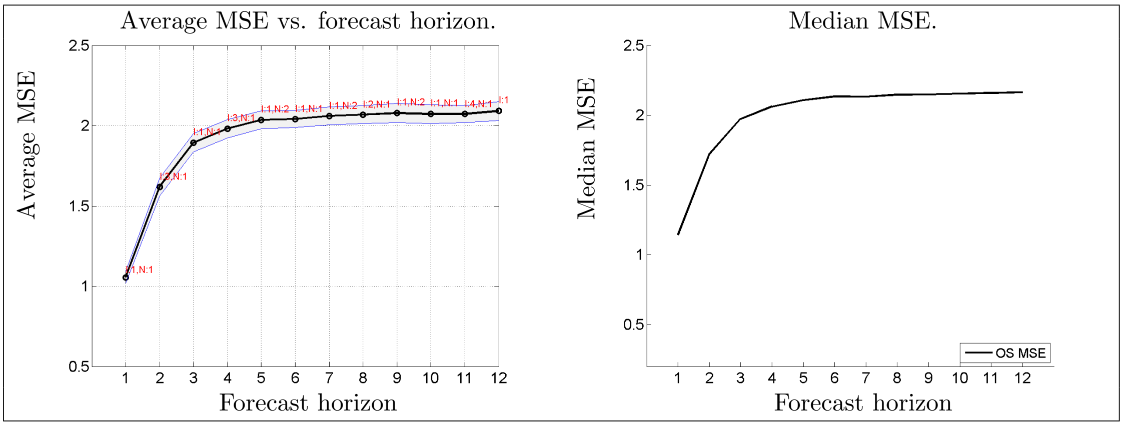
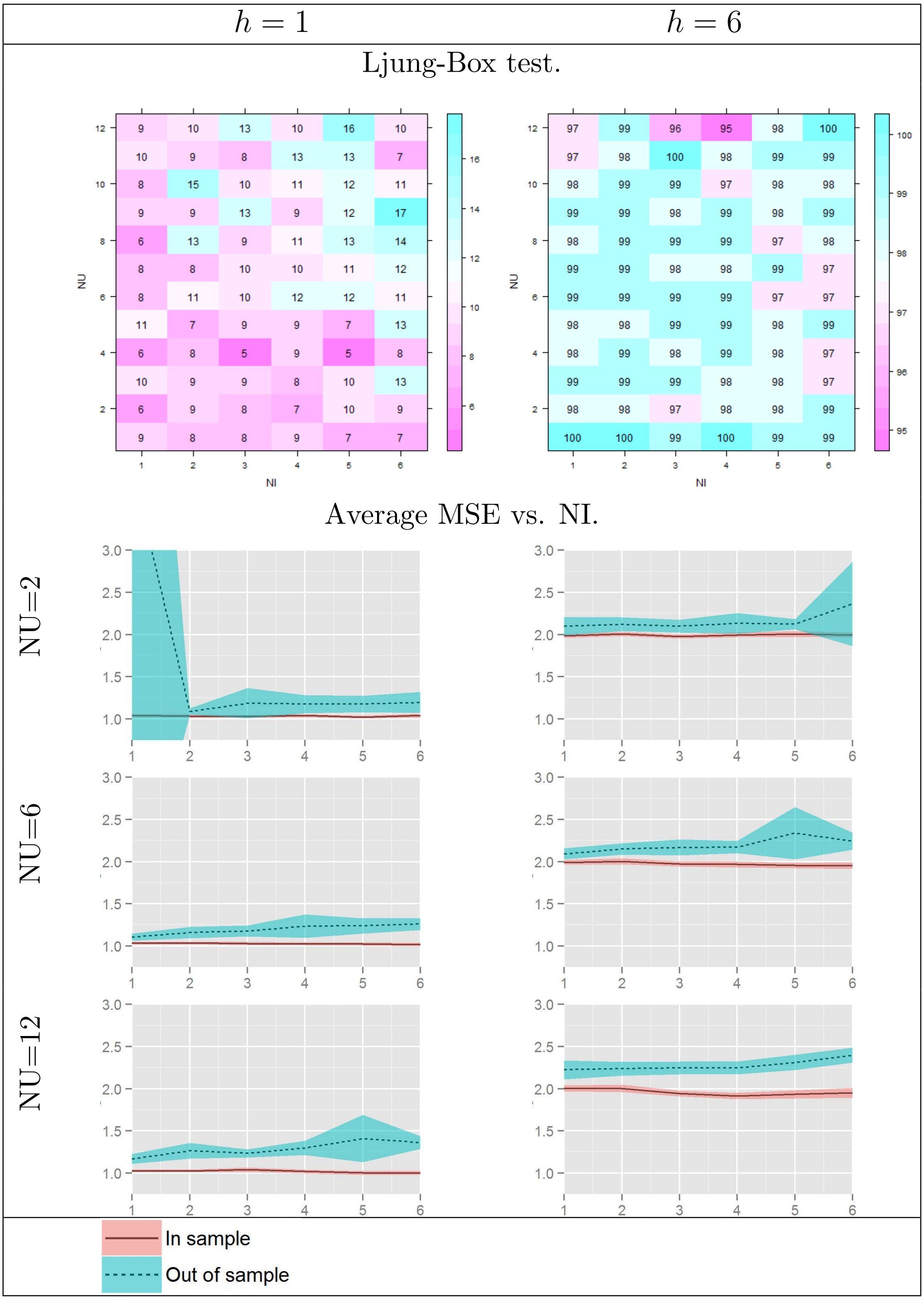


Figure 3.6: SAR series summary.

A sample of graphs depicting the average MSE for the in-sample and out-of-sample periods (with 95% confidence bands) is included in Figure 3.7[3]. It was observed that the in-sample fit, in terms of MSE, is less sensitive to the addition of past lags of the time series and neurons, when compared to the out-of-sample period. In-sample error tends to decrease whereas the out-of-sample worsens with

[3] The complete data-set includes other forecast horizons and the cross-validated version of such graphs.



| NU | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 12 | 9 | 10 | 13 | 10 | 16 | 10 |
|  | 10 | 9 | 8 | 13 | 13 | 7 |
| 10 | 8 | 15 | 10 | 11 | 12 | 11 |
|  | 9 | 9 | 13 | 9 | 12 | 17 |
| 8 | 6 | 13 | 9 | 11 | 13 | 14 |
|  | 8 | 8 | 10 | 10 | 11 | 12 |
| 6 | 8 | 11 | 10 | 12 | 12 | 11 |
|  | 11 | 7 | 9 | 9 | 7 | 13 |
| 4 | 6 | 8 | 5 | 9 | 5 | 8 |
|  | 10 | 9 | 9 | 8 | 10 | 13 |
| 2 | 6 | 9 | 8 | 7 | 10 | 9 |
|  | 9 | 8 | 8 | 9 | 7 | 7 |

| NU | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 12 | 97 | 99 | 96 | 95 | 98 | 100 |
|  | 97 | 98 | 100 | 98 | 99 | 99 |
| 10 | 98 | 99 | 99 | 97 | 98 | 98 |
|  | 99 | 99 | 98 | 99 | 98 | 99 |
| 8 | 98 | 99 | 99 | 99 | 97 | 98 |
|  | 99 | 99 | 98 | 98 | 99 | 97 |
| 6 | 99 | 99 | 99 | 99 | 97 | 97 |
|  | 98 | 98 | 99 | 99 | 98 | 99 |
| 4 | 98 | 99 | 98 | 99 | 98 | 97 |
|  | 99 | 99 | 99 | 98 | 98 | 97 |
| 2 | 98 | 98 | 97 | 98 | 98 | 99 |
|  | 100 | 100 | 99 | 100 | 99 | 99 |

Figure 3.7: SAR series summary.

increased model complexity, thus implying over-fitting. The behaviour is similar for all sample sizes (240, 480 and 960), and extreme values in the error metric are more frequent in the smaller samples.

The forecast errors were assessed for serial correlation through the Ljung-Box test. A sample graph is shown in the summary below (Figure 3.7, top). The horizontal axis represents the number of inputs (NI) while the vertical axis represents the number of neurons (NU). Each cell contains the number of times (out of 100) the tests found evidence of serially correlated errors (with a significance level of 0.05) for the corresponding NI*NU combination.

These tests show that the NNs are incapable of capturing the full dynamics of the time series, as there is consistent evidence of correlation in forecast errors for horizons $h \geq 2$. A further investigation of the models was conducted, by also considering previous literature.

In their study with synthetic time series for ensembles, Barrow et al. (2010) found that the size of the random error added to the time series affected the performance of the NNs. The authors used a generating process for seasonal data which produced series with noise at three levels: low, medium and high. In the low noise level a standard deviation of 1 was used, (corresponding to approximately 4% of the interquartile range of the series); in the medium level the standard deviation was 5 (corresponding to 21% of the interquartile range) and for the high level of noise a standard deviation of 10 was used (corresponding to 39% of the interquartile range). The SAR series used by Zhang et al. (2001) is produced by a generating process that adds a noise with a standard deviation of 1, but in this case, it corresponds to approximately 47% of the interquartile range. This level of noise probably affects the capacity of NNs to capture the dynamics of the series, in line with observations made by Barrow et al. (2010) of a negative influence of the noise level on the performance of NNs.

Additional experiments for this and other time series were performed in order to assess the effect of the level of noise added to the process. Although the effect of the level of noise is not the same as in more regular series (like those used in Barrow et al., 2010), there is an improvement of the forecast accuracy for lower levels of noise, within a range that preserves the general structure of the series (when the standard deviation of the noise added to the time series falls to 0.3 the sign AR process structure is no longer clear).

The main effect summary graphs (a sub-sample of which is provided in Figure 3.8, for MSE ander factors NI and NU) indicates insensitivity of the MSE metric to the number of inputs (NI), number of hidden units (NU), sample size (SS), and pruning (PU) in the in-sample period, whereas in the out-of-sample period there is a tendency for the metric to deteriorate as model complexity increases (higher NI and NU), and a tendency of the metric to improve with higher SS. Apparent insensitivity to PU persists in this period.

In general, ANOVA, Kruskal Wallis (K-W) and Jonckheere-Terpstra (J-T) tests show that there is sensitivity of the fit and forecasting accuracy to the factors, with the exception of pruning (PU) in some cases. A deterioration of the forecasting accuracy (out-of-sample metrics) for complex models is present in the first horizons, although for subsequent horizons there appear to be improvements in accuracy when models are more complex.

For SAR and the rest of the series, results are summarised in Table 3.9. It contains the characteristics of the generating processes and the main findings, obtained through the graphs of average MSE vs. NI and NU (a sub-sample of graphs of the average MSE vs. NI is provided in Figure 3.9), the correlation maps, the main effects graphs and the statistical tests.

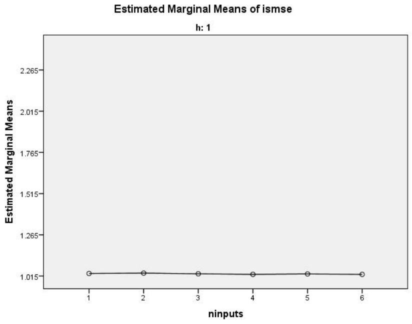
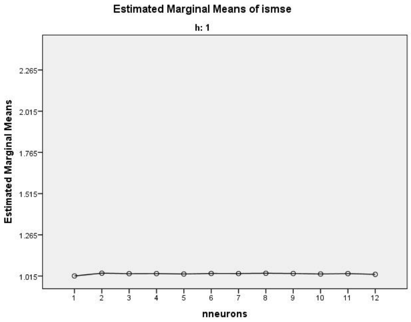
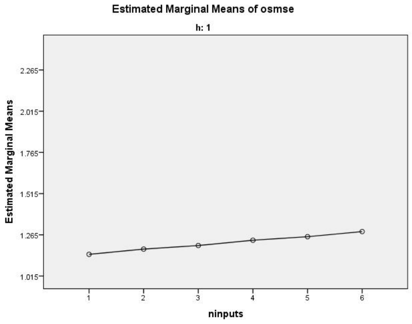
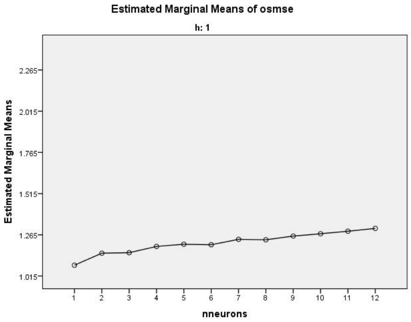
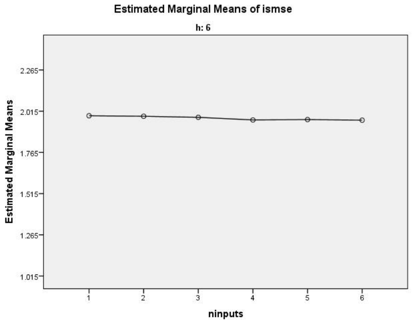
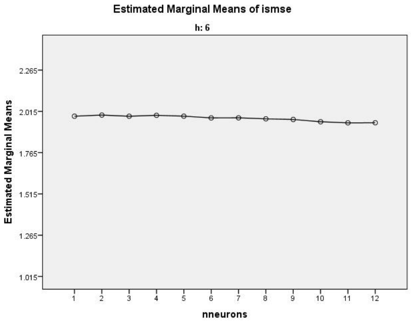
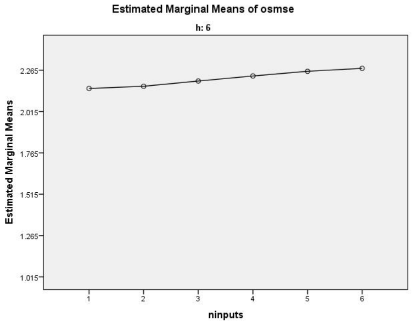
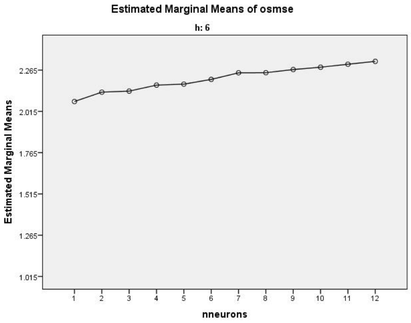
<table>
<tr><th></th><th></th><th>MSE<br>NI</th><th>NU</th></tr>
<tr><td>$h = 1$</td><td>IS</td><td>
</td><td>
</td></tr>
<tr><td></td><td>OS</td><td>
</td><td>
</td></tr>
<tr><td>$h = 6$</td><td>IS</td><td>
</td><td>
</td></tr>
<tr><td></td><td>OS</td><td>
</td><td>
</td></tr>
</table>

Figure 3.8: Main effects. Series: SAR. Estimated marginal means vs. NI (first column of graphs) and NU (second column). IS stands for in-sample; OS stands for out-of-sample. $h$ refers to the forecast horizon.

Table 3.2: Summary of findings for the sensitivity analysis.

| Series | Characteristics | Results |
|---|---|---|
| SAR | The generating process for this series, in Equation 3.1, has a simple non-linearity with a change in level (structural shift) between -1 and 1. There are sequences of positive or negative numbers (autoregressiveness in sign). With a set of 100 replications of the process it can be seen that the length of the sequence of positive or negative values is on average 6 to 7. | There is sensitivity of the fit and forecasting accuracy to the factors, with the exception of pruning (PU) in some cases. A deterioration of the forecasting accuracy for complex models is present in the first horizons, although for subsequent horizons there appear to be improvements in accuracy when models are more complex. |
| BL1 | The process (equation 3.2) has a multiplicative term that differentiates it from the SAR series. The direct relationship between $y_t$ and $y_{t+h}$ is $y_{t+h} = (0.7^h)y_t e_{t+h-2} e_{t+h-3} e_{t+h-4}$ and can make it considerably more volatile and, consequently, more difficult to forecast. Simulating this dependency for $h = 1, ..., 12$ using 100 replications for each horizon, $h$, it is observed that it significantly drops in magnitude when $h = 5$ and stays on values of the same order of magnitude or lower afterwards. | In general, the behaviour of the error metrics for the first horizons (up to 4) is different to the behaviour for subsequent steps. This roughly coincides with the dynamics observed on the process and also the serial correlation in the residuals and forecast errors (they are better behaved for $h = 1$ to $h = 3$). The models are more adequate to forecast this series than the previous (SAR), but the dynamics of the series are still not entirely captured. |
| BL2 | The BL2 series is slightly more complex than BL1 in terms of lags and non-linear behaviour. The direct dependency of $y_{t+h}$ on $y_t$ has the form $y_{t+h} = (0.4^h)y_t + P$, where $P$ comprises interaction terms, random noise and other lags $(y_{t-1}, y_{t-2} \ldots)$. The dependency is considerably weaker for $h >= 5$ than for previous horizons, which tallies with the change in the average error pattern. | The behaviour of MSE error metric for BL2 series is similar to the behaviour of the same error metric for BL1. In terms of the autocorrelation maps, the in-sample and out-of-sample Ljung-box tests for BL2 series highlight the need of more than one input in order to have well behaved residuals and forecast errors for horizon 1. For other forecast horizons (in the in-sample period) it is noticed how the behaviour of residuals changes drastically depending on the sample size. In the out-of-sample period it is noticed how for most forecast horizons and sample sizes, the area where the Ljung-Box test shows better behaved errors is roughly defined by $NU > NI$. |

(Continued) Summary of findings for the sensitivity analysis.

| Series | Characteristics | Results |
|---|---|---|
| TAR | This series is characterised by the existence of a threshold: $y_t$ has different behaviour for $\|y_{t-1}\| \leq 1$ and $\|y_{t-1}\| > 1$. Performing 100 simulations, $y_t$ stays within $\|y_t\| \leq 1$ for 2.68 consecutive steps on average. And stays in $\|y_t\| > 1$ for 1.6 consecutive steps on average. The strong dependency on the previous value of the series holds for 1 step ahead only (at most 2 for the case $\|y_t\| \leq 1$). Consequently, for an autoregressive model, capturing the time series dynamics beyond step 2 becomes more difficult. | The MSE is sensitive to the addition of neurons for the first forecast horizon only. The strong dependency between $y_t$ and $y_{t-1}$ allows the NNs to take advantage of model complexity to better capture the dynamics of the time series at this horizon, whereas for other horizons the limitations of the NNs in capturing the temporal dependency appears to be manifested in a deterioration of the performance.<br>The analysis of main effects implies a significant influence of the different factors on the error metrics. Forecast accuracy does not improve with the addition of inputs for the first forecast horizon, but improves with the addition of neurons, corroborating findings based on the average MSE. Adding more lags of the time series to models for $h > 1$ was beneficial in the serial correlation tests. |
| NAR1 | Considering the dependency pattern of the process for several steps, it can be seen that<br>$y_{t+h} \approx \frac{0.7^h\|y_t\| + f(\|y_t\|, e_{t+1}, ..., e_{t+h-1})}{0.7^{h-1}\|y_t\| + g(\|y_t\|, e_{t+1}, ..., e_{t+h-1})}$<br>The influence of $y_t$ on $y_{t+h}$ tends to be similar for different horizons, as $f()$ and $g()$ have the same order of complexity. This is confirmed by the tendency of the performance to be very homogeneous in all horizons. | Different factors have significant influence on the error metrics with a very similar behaviour across forecast horizons. |
| NMA | The generating process has complex relations between $y_t$ and $y_{t+h}$ (interaction terms are observed when developing the temporal dependency of the series and solving for $y_t$). For $h = 6$ the coefficient of $y_t$ would be $0.2{*}0.3{*}0.2{*}0.3{*}0.2 = 0.00072$. At this horizon the direct dependency between $y_t$ and $y_{t+h}$ becomes very weak. By analysing the behaviour of the MSE error metric it is noticed that after $h = 6$ the NNs are unable to improve performance markedly. | The sensitivity of the error metrics for $h = 1$ is different when compared to other horizons. There is a marked improvement of forecast accuracy from NI=1 to NI=2 (i.e, with the inclusion of the second lag). This is expected as the generating process is based on information from the previous two lags. The sensitivity of error metrics for the rest of horizons has a pattern of marginal deterioration in forecast accuracy measured by MSE and MAE. Correlation maps also suggest the inadequacy of NNs to forecast this series, specially for $h > 1$. |

(Continued) Summary of findings for the sensitivity analysis.

| Series | Characteristics | Results |
|---|---|---|
| STAR1 | The series involves a logistic function that produces oscillations in the dependency between $y_t$ and $y_{t+h}$, thus adding complexity to the forecasting problem. | The series is insensitive to structural parameters and forecast accuracy deteriorates beyond the first horizon. The statistical tests and main effects graphs confirm this observation for both MSE and MAE. |
| STAR2 | This series also involves a logistic function that produces oscillations in the dependency between $y_t$ and $y_{t+h}$. | This series presents a clear sensitivity in forecast accuracy with respect to structural parameters (NI and NU) for $h = 1$ and $h = 3$. A preferred number of lagged inputs can be identified in the main effects graphs for MSE and MAE, when the average error metric reaches a minimum. For other horizons, the structural parameters, along with SS, are influential, but no remarkable patterns nor improvement are found regarding forecast accuracy. The complex NNs tend to perform poorly for longer horizons. |
| Synthetic-1S | A single-seasonal time series. | Forecast accuracy improves with model complexity. When the number of inputs is low, there is volatility in the error metric due to misspecification. For most forecast horizons, models with better behaved forecast errors are structurally complex. In some cases, specially for $h \geq 8$, there is stronger influence of the number of inputs in producing better behaved forecast errors. Other factors also influence the error metrics: pruning deteriorates the accuracy of the forecasts, while the greater the sample size the better are the models. |
| Synthetic-2S | A double-seasonal time series. | Abrupt changes in the error metrics were found, with greater volatility when the number of inputs or neurons is small. The number of neurons, when increased, tends to improve the accuracy, as measured by both MSE and MAE. The limit where the number of inputs (NI) stops improving accuracy is clearly visible in the graphs ($NI = 3$), which tallies with the length of the shorter cycle in the simulated series. For this series, the benefit in added complexity is noticed in improved forecast error behaviour (lower serial correlation), when compared with simpler models. |

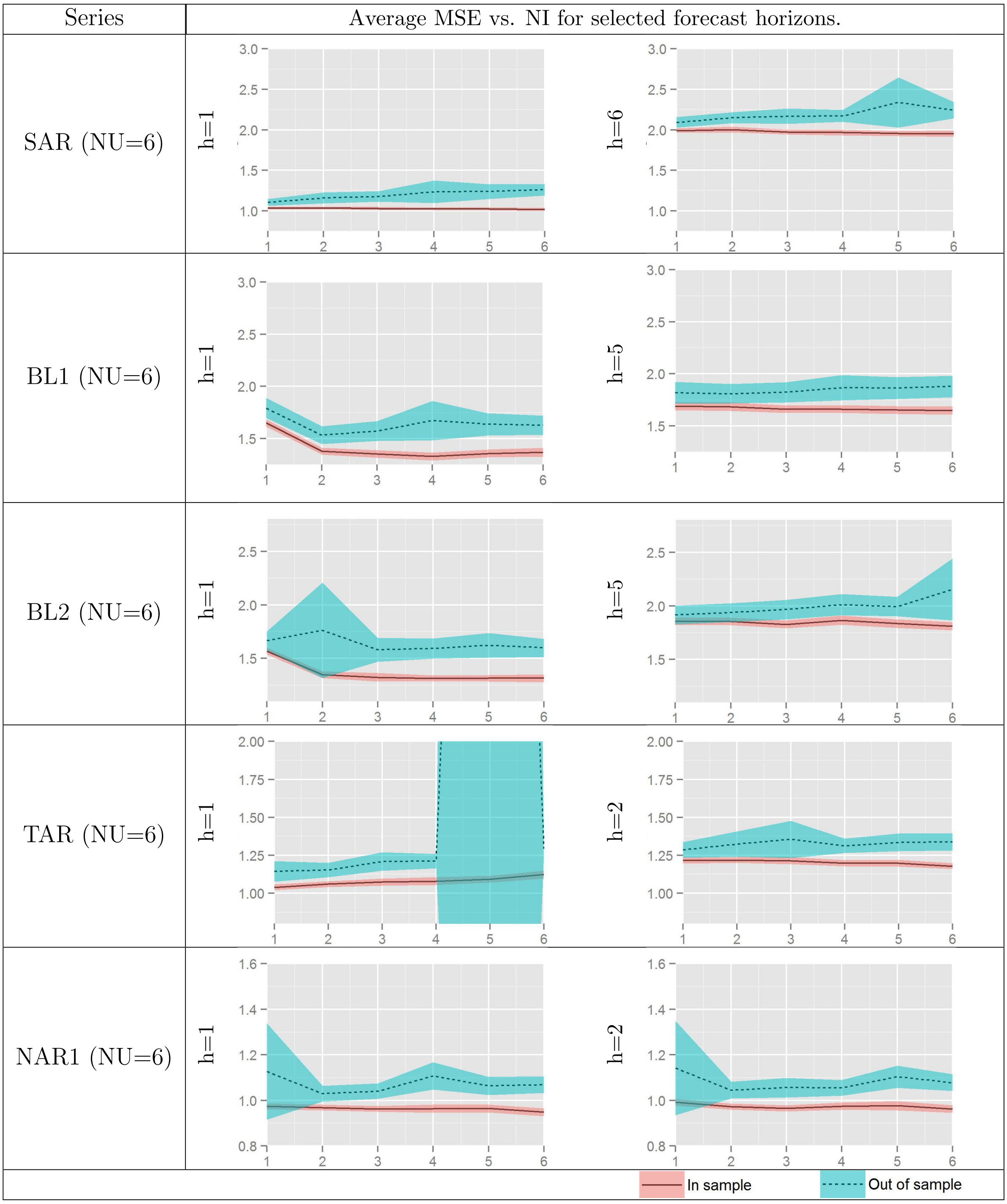
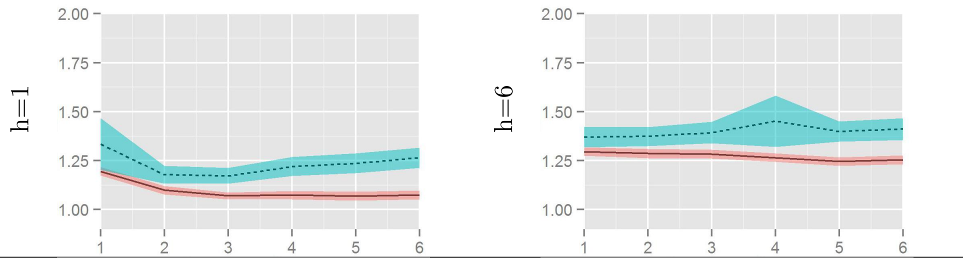
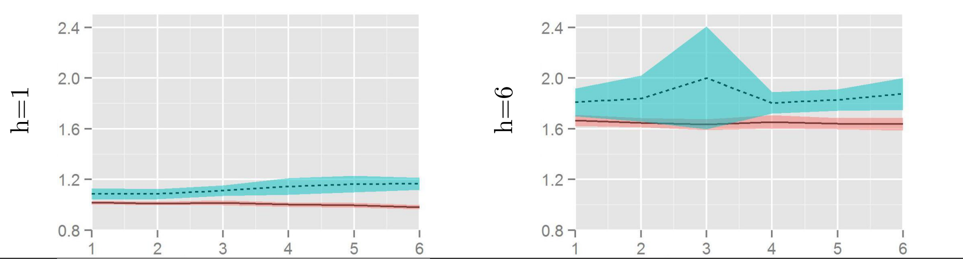
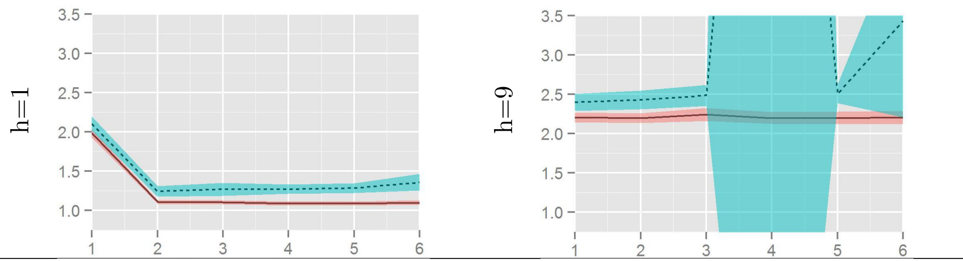
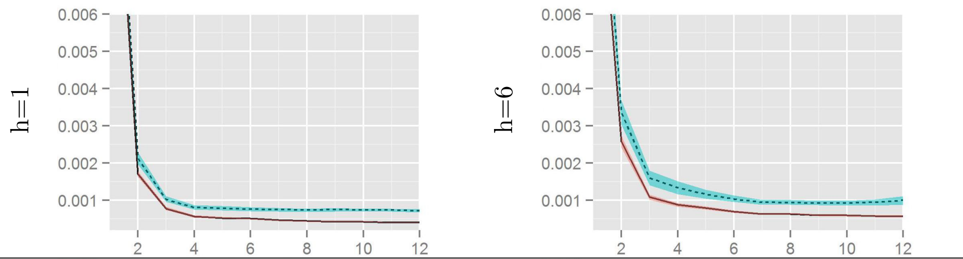
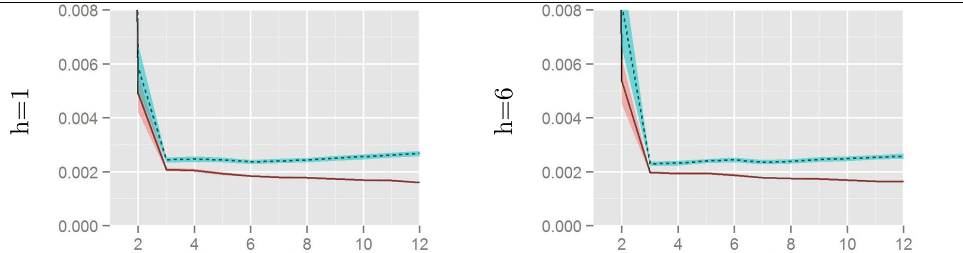


Figure 3.9: Summary of Average MSE vs. NI (number of lagged inputs).

A medium number of neurons (6 for non-seasonal and 12 for seasonal time series) and two forecast horizons (h) are reported (the first and a subsequent one, where the dynamics of the series notably changes). For each series a selected number of hidden units (NU) is displayed in parenthesis.

| Series | Average MSE vs. NI for selected forecast horizons. |
|---|---|
| NMA (NU=6) |  |
| STAR1 (NU=6) |  |
| STAR2 (NU=6) |  |
| Synthetic-1S (NU=6) |  |
| Synthetic-2S (NU=12) |  |
|  | 🟥 In sample 🟩 Out of sample |

(Continued) Summary of Average MSE vs. NI (number of lagged inputs).

A medium number of neurons (6 for non-seasonal and 12 for seasonal time series) and two forecast horizons (h) are reported (the first and a subsequent one, where the dynamics of the series notably changes). For each series a selected number of hidden units (NU) is displayed in parenthesis.

## 3.6 Discussion

The present study is conducted in line with previous literature (Zhang et al., 2001; Balestrassi et al., 2009; Crone & Dhawan, 2007). The time series processes used, however, are based on the first study.

Zhang et al. (2001) used linear regressions to asses the influence of factors in NN performance. An attempt to conduct a similar study was done here, but coefficients were extremely small and the model assumptions were not well respected. Therefore ANOVA and non-parametric tests were preferred, as for example, the Jonckheere-Terpstra test. It assesses the existence of a trend of ascending or descending medians with respect to a factor, and therefore is useful in identifying the general direction of the contribution of a factor.

A summary of the direction in which factors affect the error metrics for the first forecast horizon is given in Table 3.3. Considering model fit and forecast accuracy, the effect on MSE and MAE is very similar for all factors. The single and double seasonal synthetic time series were better captured by complex NNs, which is shown in increased forecast accuracy.

Considering all forecast horizons, there is homogeneity in most factors with respect to fit (in-sample) and mixed results in terms of forecast accuracy (out-of-sample). Table 3.4 provides the number of times a factor is significant out of the 12 forecast horizons, along with an indication of the predominant direction of the trend in medians (+ indicates that for 50% or more horizons the factor appears to produce an ascending trend of medians).

Mixed results are observed in the case of SAR, BL1, TAR, NAR1, and NMA. Results are more homogeneous for the series that resulted in better behaved NNs. For example, STAR2 series has dynamics that are better captured by the NNs, even when the series are very volatile (for lower levels of volatility, the forecast accuracy

Table 3.3: Direction of contribution of a factor.

| MSE | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | NI | | NU | | SS | | PU | |
| Series | IS | OS | IS | OS | IS | OS | IS | OS |
| SAR | - | + | Non-sig | + | - | - | + | + |
| BL1 | - | - | - | - | + | - | + | + |
| BL2 | - | Non-sig | - | + | + | - | + | + |
| TAR | + | + | - | Non-sig | - | - | + | + |
| NAR1 | - | + | - | + | + | - | + | + |
| NMA | - | + | - | + | + | - | + | + |
| STAR1 | - | + | - | + | + | - | + | + |
| STAR2 | - | - | - | + | + | - | + | + |
| Synthetic-1S | - | - | - | - | + | - | + | + |
| Synthetic-2S | - | - | - | - | + | - | + | + |
| MAE | | | | | | | | |
| | NI | | NU | | SS | | PU | |
| Series | IS | OS | IS | OS | IS | OS | IS | OS |
| SAR | - | + | - | + | - | - | + | + |
| BL1 | - | Non-sig | - | - | + | - | + | + |
| BL2 | - | Non-sig | - | + | + | - | + | + |
| TAR | + | + | - | - | - | - | + | + |
| NAR1 | - | + | - | + | + | - | + | + |
| NMA | - | + | - | + | + | - | + | + |
| STAR1 | - | + | - | + | + | - | - | - |
| STAR2 | - | - | - | + | + | - | + | + |
| Synthetic-1S | - | - | - | - | + | - | + | + |
| Synthetic-2S | - | - | - | - | + | - | + | + |

Direction of factor influence according to J-T test for forecast horizon 1. IS: in-sample; OS: out-of-sample.

Table 3.4: Summary of factor influence.

| MSE | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | NI | | NU | | SS | | PU | |
| Series | IS | OS | IS | OS | IS | OS | IS | OS |
| SAR | 12 (-) | 11 (+) | 12 (-) | 11 (+) | 12 (+) | 12 (-) | 12 (+) | 9 (+) |
| BL1 | 12 (-) | 11 (+) | 12 (-) | 11 (+) | 12 (+) | 12 (-) | 12 (+) | 9 (+) |
| BL2 | 12 (-) | 11 (+) | 12 (-) | 12 (+) | 12 (+) | 12 (-) | 12 (+) | 12 (+) |
| TAR | 11 (-) | 12 (+) | 12 (-) | 11 (+) | 11 (+) | 12 (-) | 12 (+) | 12 (-) |
| NAR1 | 12 (-) | 12 (+) | 12 (-) | 12 (+) | 12 (+) | 12 (-) | 12 (+) | 12 (+) |
| NMA | 12 (-) | 12 (+) | 12 (-) | 12 (+) | 12 (+) | 12 (-) | 12 (+) | 12 (+) |
| STAR1 | 12 (-) | 12 (+) | 12 (-) | 12 (+) | 12 (+) | 12 (-) | 12 (+) | 7 (+) |
| STAR2 | 12 (-) | 10 (+) | 12 (-) | 12 (+) | 11 (+) | 12 (-) | 12 (+) | 12 (+) |
| Synthetic-1S | 12 (-) | 12 (-) | 12 (-) | 8 (-) | 12 (+) | 12 (-) | 12 (+) | 12 (+) |
| Synthetic-2S | 12 (-) | 11 (-) | 12 (-) | 12 (-) | 12 (+) | 12 (-) | 12 (+) | 12 (+) |
| MAE | | | | | | | | |
| | NI | | NU | | SS | | PU | |
| Series | IS | OS | IS | OS | IS | OS | IS | OS |
| SAR | 10(-) | 12(+) | 12(-) | 12(+) | 10(+) | 12(-) | 12(+) | 9(+) |
| BL1 | 12(-) | 11(+) | 12(-) | 11(+) | 12(+) | 12(-) | 12(+) | 12(+) |
| BL2 | 12(-) | 11(+) | 12(-) | 12(+) | 12(+) | 12(-) | 12(+) | 12(+) |
| TAR | 11(-) | 12(+) | 12(-) | 11(+) | 11(+) | 12(-) | 12(+) | 12(+) |
| NAR1 | 12(-) | 12(+) | 12(-) | 12(+) | 12(+) | 12(-) | 12(+) | 12(+) |
| NMA | 12(-) | 12(+) | 12(-) | 12(+) | 12(+) | 12(-) | 12(+) | 12(+) |
| STAR1 | 12(-) | 12(+) | 12(-) | 12(+) | 12(+) | 12(-) | 9(+) | 10(+) |
| STAR2 | 12(-) | 10(+) | 12(-) | 12(+) | 12(+) | 12(-) | 12(+) | 12(+) |
| Synthetic-1S | 12(-) | 11(-) | 12(-) | 9(-) | 12(+) | 12(-) | 12(+) | 12(+) |
| Synthetic-2S | 12(-) | 6(-) | 12(-) | 12(-) | 12(+) | 12(-) | 12(+) | 12(+) |

Number of horizons for which a factor is significant and predominant influence, according to J-T test. IS: in-sample; OS: out-of-sample.

improves). As the single-seasonal and the double-seasonal series are more regular, NNs appear to be capable of producing better forecasts.

In general, it was observed that the time series that were better captured by the feed-forward NNs used in this study are characterised by a regular pattern, a generating process with a strong dependence of $y_t$ on the previous values and a mild effect of the level of noise (the regular pattern is discernible in the series even with the addition of noise).

This was manifested in two clear patterns in the effect of the number of inputs and the number of neurons over the fit and forecasting error. It was found that single-seasonal and double-seasonal time series produced decreasing patterns for the fit error and decreasing or approximately U-shaped patterns for the forecasting error. On the other hand, non-seasonal time series showed decreasing fit error but a rapidly growing forecasting error. Therefore, for seasonal time series the over-fitting is less apparent, complexity in terms of lagged-inputs and neurons is beneficial and the limit of this benefit can be observed. Apart from these, no other patterns are clear enough to be generalisable.

Zhang et al. (2001) reported a significant impact of input nodes (NI in this study) on MSE and MdAPE (median absolute percentage error) for both training and test sets across different sample sizes for one-step-ahead forecasts. The experiments conducted in the present study confirm these findings for the in-sample period and almost all series. Additionally, the trend in contribution that was observed is almost always negative for the first horizon (except SAR and TAR series which have positive influence of NI, as seen in Table 3.3, meaning that the number of inputs appears to increase the error metric). For other forecast horizons the findings are mixed, but generally an increased number of lagged inputs (NI) improves the model fit (see Table 3.4). For the out-of-sample period it was found a significant effect of NI in most cases, but the direction of the influence is mixed.

Zhang et al. (2001) also reported that while the number of hidden nodes is significant on training (in-sample) MSE, it is not significant judging from training and test MdAPE. In the present study, the factor NU was found significant in all the forecast horizons and series, for both MSE and MAE during the training (in-sample) period, having a negative effect (improving fit as the metric is lowered). In the out-of-sample period the influence of factors on the error metrics is significant for most of forecast horizons and series. However, the direction of the effect for non-seasonal time (increased error) series might signalls problems of over-fitting while for the single-seasonal and double-seasonal there is evidence of improvement when the number of neurons is increased.

Balestrassi et al. (2009) reported NU as significant for all the series considered in their study (SAR, BL1, BL2, TAR, NAR1, NMA, STAR1 and STAR2). Factor NI was fixed in their univariate time series approach (an additional approach with dummy variables is not comparable with the present study). SS was also found significant, which coincides with findings obtained in the present study (Table 3.4). On the other hand, the simple pruning of NNs weights conducted here led to mixed results, with a common tendency to worsen the forecast accuracy.

It is evident how the best models in this study tend to be rather simple for non-seasonal series. Results obtained by Zhang et al. (2001) suggest this is expected, as the generating processes rely on a few lagged values and NNs are generally capable of identifying a number of inputs related to such lags. For the single-seasonal and double-seasonal series, which are generated by using several lagged values, the added complexity (in terms of inputs and neurons) clearly has an impact on model fit and forecast accuracy.

Model selection for single NNs has been studied by Anders & Korn (1999), Balkin & Ord (2000), Crone & Kourentzes (2010) and model selection for ensembles has been studied by, for example, Chen & Yao (2007), through re-sampling; Islam et al. (2003), through cooperative training; Liu & Yao (1999) and Chen & Yao (2009), via negative correlation training and Chandra & Yao (2006) and Yao & Islam (2008), through evolutionary algorithms. Differing from these authors, the approach followed here is an application of DOE to the selection of models for ensembles. Its immediate benefit is in visualising the behaviour of error metrics and their assessment through statistical tests. Given that the ensembles are generally summarised through averages, and DOE as applied here looks into the behaviour of average metrics, this approach gives an approximate view of the potential behaviour of ensembled models. Here an attempt has been made to explore the benefit in using DOE when building ensembles of NNs rather than comparing it with established model selection mechanisms. A study performing a comparison of strategies for NN model selection, including DOE-based approaches, would be desirable and is included in the research agenda.

The combination of models with modest accuracy and negatively correlated errors can be more productive than the combination of models with high accuracy. The study of this issue is rare in the context of sensitivity analysis of NN performance, but it has been studied in the context of ensembles of NN (see Liu & Yao, 1999; Liu et al., 2000).

Regarding a sensitivity analysis, as performed here, the models for every factor combination (including number of lagged inputs and hidden units) share a similarity in specification that can lead to positively correlated forecasting errors. Negative correlation could be studied by performing a filtering of the models in the simulations depending on a correlation threshold. That is, for a combination of factors $NI$, $NU$, $SS$ and $PU$, negative correlation training would be repeated until the correlation between errors (for the set of 100 models) reaches a given level (which could be another factor under study). This poses challenges in the statistical analysis of data, given that the generation of one NN with a given configuration is related to the generation of the previous ones, as the correlation was taken into account. This would imply that the independence between experiments does not hold. However, with a proper analysis, results would be helpful in indicating the effect of the level of correlation allowed in the forecasting performance. This would be a way to systematically study the idea of having negatively correlated errors in order to improve accuracy. This could also inform later stages that depend on the sensitivity analysis, such as the production and combination of NNs.

## 3.7 Conclusions and Further Research

Sensitivity analysis of some sort was suggested as a useful practice since the inception of NN ensembles (Hansen & Salamon, 1990). The practicality of conducting such an analysis depends on the objective of the study, the design of the experiments (related to the objective), the theoretical computing time needed to explore the combinations of factors and the availability of time and computing power.

Here the focus has been on design of experiments. The aim was to study the influence of key parameters in forecasting accuracy of neural networks and to use such information to aid the selection of models to include in an ensemble. Different views of the metrics are given, based on extensive simulations. The main contributions of this study to the literature are the inclusion of more trials per factor combination than other studies, the use of a wide range of plots and tests, the extension of types of synthetic series to include double-seasonal time series and, most importantly, the assessment of multi-step-ahead forecasts.

Results show a significant sensitivity of performance metrics to the number of inputs (past lags), number of neurons and sample size. Pruning is less significant. The sensitivity patterns observed by Zhang et al. (2001) for the series reported by them (STAR2) coincides with findings in the present study. For other non-seasonal series, results are mixed. Such differences might be due to interacting factors. One is the presence of a high level of noise in the generating processes (Zhang et al., 2001, added a noise equivalent to a high proportion of the interquartile range). Performing experiments with configurations used by Barrow et al. (2010) suggests that high volatility impacts the forecast accuracy of NNs. Other factors affecting the performance of NNs are the non-linearity of the processes and the inherent limitations of feed-forward NNs.

In general, two clear patterns were observed in the effect of the number of inputs and the number of neurons over the fit and forecasting error. It was found that long memory processes produced decreasing patterns for the fit error and decreasing or approximately U-shaped patterns for the forecasting error. On the other hand, short memory processes showed decreasing fit error but a rapidly growing forecasting error. Therefore, when the generating process has long memory, complexity in terms of lagged-inputs and neurons is beneficial and the limit of this benefit can be observed. Apart from these, no other patterns are clear enough to be generalisable.

Design of experiments can be used to aid model selection in ensembles by identifying the effect of design factors on forecast performance. If the ensemble is to be built with models that have the same structure, this means that the sensitivity analysis allows the making of informed decisions about the *base* model. Subsequent stages of the development of ensembles involve the generation and, optionally, pruning of the models thus generated to finally combine forecasts. If the ensemble includes models with different structures, the analysis can also help in assessing these structures, by allowing the modeller to see their particular behaviour. In summary, the selection of models in ensemble development should include the following steps: establish the characteristics of the ensemble (big or small models, same or different structure, etc.), design the experiments, create a model performance data-base and use this data-base to select base models for the ensembles.

Additionally, a sensitivity analysis with a full factorial design should be conducted in an incremental way, starting with a modest number of factors and levels and gradually augmenting them, depending on findings and needs. In this way, the cost in time of the experiments can be controlled.

More complex networks, such as recursive, and more sophisticated approaches for NN specification can be examined in a similar procedure. This appears to be a natural extension of the research, but the quick growth of factor combination has to be taken into account, probably through more sophisticated designs, so that the analysis is computationally feasible.

The comparison of DOE, for ensemble design, with established model selection strategies (intended to be used in ensembles or not) can be explored. Such study would shed light into interactions and combinations of DOE and model selection strategies.

# Chapter 4

# Structural Combination of Neural Network Forecasting Models [1]

## 4.1 Abstract

Forecasts combinations normally use point forecasts that were obtained from different models or sources (Newbold & Granger, 1974; Clemen, 1989; Timmermann, 2006). This chapter explores the incorporation of internal structure parameters of NN models as an approach to combine their forecasts via ensembles. This is done, first, by developing a clustering-based approach, such that the generated NN models that could be part of the ensembles are subject to a clustering algorithm that uses the structure parameters and, from each of the clusters obtained, a small set of models is selected and combined. Secondly, in an alternative and simpler implementation, a subset of the generated NN models is selected by using several reference points in the model structure parameter space. The choice of the reference points is optimised through a genetic algorithm and the models selected are averaged. Synthetic time series, hourly multivariate time series data from wind power production and electricity demand time series are used to assess multi-step ahead forecasting performance for up to 12 hours ahead. Results are compared against several statistical benchmarks, the average of the individual forecasts and the best models in the ensembles. Results show that structural combination with genetic algorithms (GA) outperforms the average more easily than cluster-based (CB) combination for non-seasonal time series, whereas for the double-seasonal series the CB do better in

[1] Initial proposal for this research was presented at the ISF (2012), 24-27 June 2012, Boston, USA. First results were presented at WIPFOR (2013), 5-7 June 2013, Paris, France.

outperforming such benchmark.

## 4.2 Introduction

In Chapter 2, it was highlighted how there is limited research on forecast combination approaches which consider the internal characteristics of the models involved. Therefore this chapter addresses the question: how can the structure of neural networks be combined?

The most common types of NN models in the literature of ensemble development are feed-forward NNs (for example Hansen & Salamon, 1990; Fan et al., 2009). Such NNs are common in forecasting (Crone & Dhawan, 2007; Crone & Kourentzes, 2010) and, specifically, in the electricity sector (i.e. Khotanzad et al., 1998; Drezga, 1999; Abdel-Aal, 2005). Multi-layer perceptrons are the most frequently applied (Crone & Dhawan, 2007) and, therefore, constitute the starting point in answering the question formulated above. Figure 4.1 depicts a multi-layer perceptron.

Its algebraic representation is

$$\hat{y}_{t+1} = \beta_0 + \sum_{k=1}^{H} \beta_k \varphi \left( \gamma_{k0} + \sum_{i=1}^{I} \gamma_{ki} x_i \right) \tag{4.1}$$

where $\hat{y}_{t+1}$ is the one-step-ahead forecast produced from input variables $x_i$ (lagged series values, for example); $\beta = [\beta_1, \ldots, \beta_H]$ and $\gamma = [\gamma_{11}, \ldots, \gamma_{HI}]$ denote the network weights for the output and hidden layers, respectively. The $\beta_0$ and $\gamma_{k0}$ are the biases of each neuron and $\varphi(\cdot)$ is the activation function.

In the context of the above models, there are two forms in which the structure can be used in a forecast combination. One possibility is the exploration of relationships between components of the model and the forecast produced by it. The identified relationships can then be used to combine the outputs of several models. For example, Garson (1991) and Goh (1995) propose an approach for the identification of the importance of an input variable by studying the weighted connections

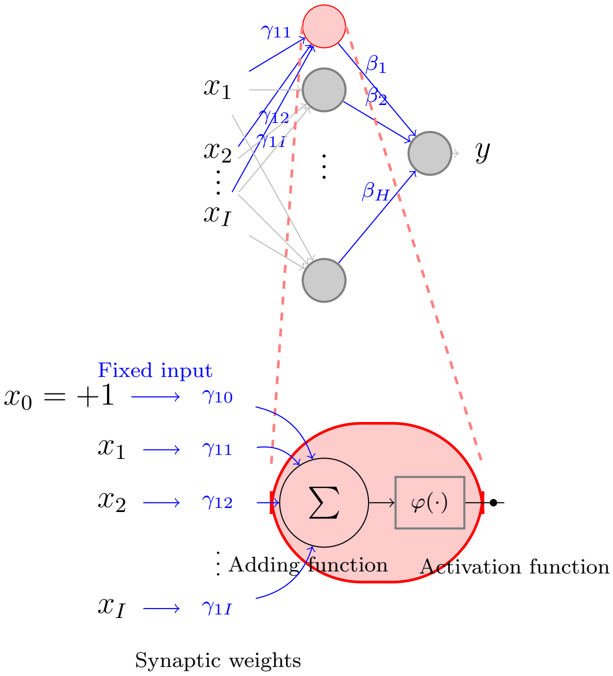


Figure 4.1: Multi-layer perceptron with $I$ inputs and $H$ hidden nodes. The bias node is displayed only for the detail of the neuron.

between nodes of interest from the input to the output. In this way mappings of importance can be constructed, which can be extended and used to produce forecast combinations.

Another possibility is to identify patterns in the internal components of a set of models and use these identified patterns to develop forecast combinations. A neural network can be represented as a vector containing different model parameters, such as the weights of the synaptic connections and training parameters, as well as descriptions of the activation functions, measurements of error and other features that are needed in specifying a NN. With this vector and maintaining the architecture of the NNs fixed, different models can be represented as points in a $N$-dimensional space, where $N$ is the total number of parameters. Different clustering algorithms can be adopted, such as K-means, nearest-neighbour, fuzzy C-means or Gustafson Kessel (see Jang et al., 1997). For example, in K-means, a set of vectors

$\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n}$, each containing parameters of a model, is organised in $c$ ($c \leq n$) clusters, $G_i$ $i = 1, \ldots, c$, such that a measure of dissimilarity between a vector $\mathbf{x_k}$ in group $j$ and the corresponding cluster centre $\mathbf{c_i}$ is minimised:

$$J = \sum_{i=1}^{c} \left( \sum_{k, x_k \in G_i} \|\mathbf{x_k} - \mathbf{c_i}\|^2 \right) \tag{4.2}$$

Clustering would provide a set of points that minimise the distance between objects within groups, thus yielding, in a way, simplified versions of the models in every group or cluster. Bakker & Heskes (2003) follows an approach, based in K-means clustering and deterministic annealing [2], that is a source of inspiration for this study, whose methodology is described in the next section.

The forecasting approach comprises different stages: data pre-processing, model generation, model combination, forecasting, and assessment of uncertainty in forecasts. In subsequent sections, studies with synthetic and real world time series are summarised, followed by a discussion. Finally, conclusions and implications for future studies are drawn in the last section.

## 4.3 Methodology

Figure 4.2 describes the modelling process. A base NN model, which has been selected through a preliminary process (such as a sensitivity analysis), is used to fit the data using different models with the same structure. Model parameter diversity is introduced through the randomisation of input-output patterns for the neural networks. That is, if $\mathbf{x_i}$ is a set of input lagged variables of the series and $y_i$ is the next $h$-th corresponding observation in the series (with $h = 1, 2, \ldots, H$, being the forecast horizon), the patterns $P_i = (\mathbf{x_i}, y_i)$, $i = 1, \ldots, n$ comprise the training (in-sample) set, which can be shuffled so that sets of $P_i$ are presented in a different order

[2] Concepts of physics and fuzzy logic are incorporated into in a clustering technique aimed to avoid local minima.

to each NN, e.g. $\{P_5, P_3, P_1, \ldots, P_n, \ldots, P_k\}$ instead of $\{P_1, P_2, P_3, \ldots, P_k, \ldots, P_n\}$. Once the ensemble is generated, the models are combined taking their structure into consideration and, finally, forecasts are produced and their uncertainty assessed.

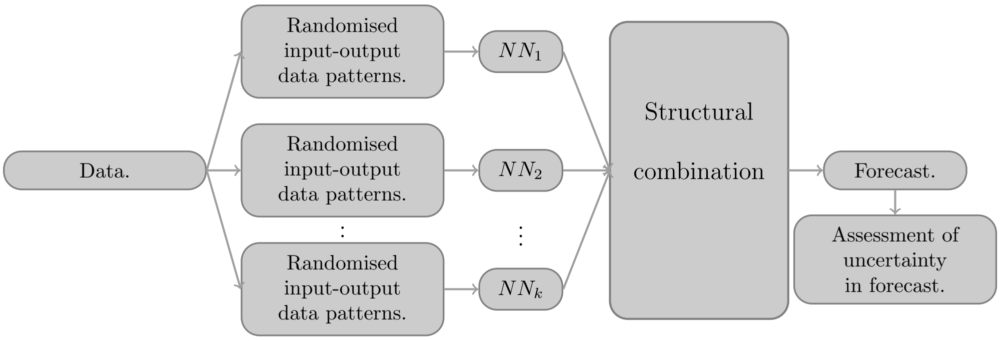

Data.
Randomised input-output data patterns.
Randomised input-output data patterns.
Randomised input-output data patterns.
$NN_1$
$NN_2$
$NN_k$
Structural combination
Forecast.
Assessment of uncertainty in forecast.

Figure 4.2: Modelling process.

In general in this study several time series are forecasted with feed-forward NNs and forecast combination are calculated through the structural proposed approach. Results are classified according to the maximum number of clusters allowed in the process of combining the models (which will be explained in sub-section 4.3.2). Results are compared against the naïve benchmark, the simple average of forecasts produced by NN and standard benchmarks from other studies (see sub-section 4.4.1). It is expected to have different results given the different nature of the time series and to have a picture of the benefit of structurally combining forecasting models for these specific time series.

This section focuses on the implementation of the structural combination based on clustering. A simplified genetic algorithm implementation will be described in a subsequent subsection.

### 4.3.1 Randomisation of Input-Output Patterns

As described above randomisation of input-output patterns in the training period enables the creation of diverse NN models (with different parameter sets), thus leading to different clusters in the parameter space. Hence, it is an initial stage in the combination procedure.

### 4.3.2 Structural Combination Based on Clustering (CB)

In the structural combination stage in Figure 4.2 is implemented through clustering, the combination is a mechanism that considers the structure of models and finds groups in the space defined by such structure. The idea is to widen the sources of model diversity, by using model structural representation in the combining process. Clustering algorithms facilitate this form of combination, but in general any method capable of representing and aggregating objects through their features has the potential to be useful.

Differing from Bakker & Heskes (2003), a fuzzy C-means algorithm was chosen, because, as models forecast the same process, they should be similar. Fuzzy C-means is thus attractive, because it allows for models to have different likelihoods of belonging to distinct clusters. The basic building block is the fuzzy set, which follows the definition of Jang et al. (1997, p.14):

*If X is a collection of objects denoted by x, then a fuzzy set A in X is defined as a set of ordered pairs*

$$A = \{(x, \mu_A(x)) \mid x \in X\} \tag{4.3}$$

where $\mu_A(x)$ is called membership function for fuzzy set $A$. This function makes a mapping between each element in $X$ and a degree of membership in $[0, 1]$.

The fuzzy set is used to express vagueness as, for example in the context of economics, when asserting that *demand is high.* There can be different values for a variable *demand* that can vaguely be classified as *high.* Here, the concept of *high* would be called a linguistic label and would be described by a fuzzy set. Such fuzzy sets are used to form inference systems that consist of rules of the type *If Demand is high then y=f(x)*, where $y$ can be a value used to take decisions, depending on the level of demand[3]. The nonlinear mappings between an input and an output space that can be achieved using such inference systems have been applied in different areas, including engineering, control and forecasting (see for example, Kasabov, 1996; Jang et al., 1997).

Fuzzy C-Means is an algorithm that partitions a collection of vectors into $c$ fuzzy groups and finds a cluster centre in each group so a cost function of dissimilarity is minimised (see Jang et al., 1997). The algorithm has interesting features. One is its ability to produce centres that do not necessarily correspond to data points in the set. Centres are $m$-dimensional vectors that are not necessarily close to the $m$-dimensional data points used in the algorithm. A second feature, originated from fuzzy systems, is its use of a degree of membership to clusters (between 0 and 1), instead of a binary membership (0 or 1, equivalent to *non-member* or *member*). Therefore, a given data point may belong to several groups with different degrees of belongingness defined by grades between 0 and 1. Normalisation can be imposed, such that the summation of degrees of belongingness of a data set always equals unity (as described by Jang et al., 1997, p. 426):

$$\sum_{i=1}^{c} u_{ij} = 1, \forall j = 1, \ldots, n \tag{4.4}$$

Where $u_{ij}$ is the degree of belongingness of $j$th data point to the $i$th cluster. The cost function is:

$$J(U, \mathbf{c_1}, \ldots, \mathbf{c_c}) = \sum_{i=1}^{c} \sum_{j}^{n} u_{ij}^{m} d_{ij}^{2} \tag{4.5}$$

Where $U$ is the matrix of all $u_{ij}$, $\mathbf{c_i}$ is the cluster centre of fuzzy group $i$, $d_{ij} = \|\mathbf{c_i} - \mathbf{x_j}\|$ is the Euclidean distance between $i$th cluster centre and $j$th data point. The parameter $m \in [1, \infty)$ is a weighting exponent.

[3] The example refers to a specific kind of inference system: the Takagi-Sugeno.

The necessary conditions for Equation 4.5 to reach a minimum are:

$$\mathbf{c_i} = \frac{\sum_{j=1}^{n} u_{ij}^{m} \mathbf{x_j}}{\sum_{j=1}^{n} u_{ij}^{m}} \tag{4.6}$$

and

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{d_{ij}}{d_{kj}}\right)^{2/(m-1)}} \tag{4.7}$$

The fuzzy C-means algorithm is an iterative procedure satisfying the necessary conditions described above. In a batch mode, the steps are the following:

Step 1: Initialise the membership matrix $\mathbf{U}$ with random values between 0 and 1 such that the constraints in Equation 4.4 are satisfied.

Step 2: Calculate $\mathbf{c}$ fuzzy cluster centres $\mathbf{c_i}$, $i = 1, \ldots, c$, using Equation 4.6.

Step 3: Compute the cost function according to Equation 4.5. Stop if either it is below a certain tolerance value or its improvement over previous iteration is below a certain threshold.

Step 4: Compute a new $\mathbf{U}$ using Equation 4.7. Go to step 2.

As highlighted above, the attraction of a fuzzy C-means approach to the present study is its use of a degree of membership of elements to clusters (between 0 and 1), instead of a binary membership (0 or 1, equivalent to *non-member* and *member* in K-means). Therefore, a given element (a NN model in this case) can belong to several groups with different degrees of belongingness in the interval $[0, 1]$ (Jang et al., 1997, p. 426).

However, C-means produces non-deterministic partitions or clusters. A variant of the algorithm is used here, based on Friedman (1991), which uses a recursive partitioning of elements space that helps in producing a deterministic partition. The next subsection describes the model in detail.

#### 4.3.2.1 The Forecasting Model

The clustering-based algorithm uses structure information of models, that is, their collection of synaptic weights ($\beta$ and $\gamma$ in Figure 4.1) and produces forecasts based on the data and this information. The one-step-ahead forecast output for time $t+1$ obtained from a set of inputs (e.g lagged values of the series to be forecasted), $x_t = \{y_t, y_{t-1}, \ldots, y_{t-p}\}$, is:

$$\hat{y}_{t+1} = \sum_{i=1}^{n} \phi_i \hat{y}_{C_i}(x_t) \tag{4.8}$$

Where $\hat{y}_{C_i}$ is the output from cluster $i$:

$$\hat{y}_{C_i}(x_t) = \alpha_{i,0} + \alpha_{i,1}\hat{y}_{i,1}(x_t) + \alpha_{i,2}\hat{y}_{i,2}(x_t) + \ldots + \alpha_{i,L}\hat{y}_{i,L}(x_t) \tag{4.9}$$

$\hat{y}_{i,1}, \hat{y}_{i,2}, \ldots$ are the forecasts produced by models selected within cluster $i$, and $\alpha_{i,1}, \alpha_{i,2}, \ldots$ are the coefficients obtained via OLS using $\hat{y}_{i,1}, \hat{y}_{i,2}, \ldots$ as regressors and the $y_i$ as the independent variable. $L$ models are selected, with $L$ varying between 1 and 5. For each model, its structural representation is a vector comprising $\beta$ and $\gamma$ (highlighted in blue in Figure 4.1).

In several types of fuzzy systems, fuzzy sets are used to form inference systems that consist of rules of the type *If Demand is high then* $y = f(x)$, where *Demand* is a variable, *high* is a fuzzy set and $y$ can be a value used to take decisions, depending on the level of *Demand*. This rule has a firing strength, depending on how high *Demand* is. In the case implemented here, the fuzzy set included in the rule is comprised by models in a cluster and their Euclidean distances (in the structural space) to the cluster centroid are used to calculate the equivalent of the firing strength of the rule in the example above. Such rule has the form *If* $A_i(x)$ *then* $y = f(\hat{y}_{C_i}; \phi_i)$, where $A_i$ is a fuzzy set formed with models from cluster $i$. Calculations are made by adapting the concept from Jang et al. (1997), p. 85, to take into account elements around the centroid rather than the centroid itself:

$$u_i(v) = e^{-\frac{D_i^2(v)}{\sum_{j=1}^{n} D_j^2(v)}} \tag{4.10}$$

$$w_i(v) = \frac{u_i(v)}{\sum_{j=1}^{n} u_j(v)} \tag{4.11}$$

$$\phi_k = \frac{\sum_{m \in C_k} w_m(v_{C_k})}{N_k} \tag{4.12}$$

$C_k$ denotes cluster $k$, $v_{C_k}$ is the centre of such cluster and $N_k$ is the number of models in it. $u_i(v)$ is the membership of $v$ to cluster $i$ (being $v$ again a model represented in the form of a vector with its synaptic weights). The squared distance between $v$ and the $i$-th centre is divided by the sum of squared distances from $v$ to all centres. Subsequently, an exponential transformation is taken in order to allow for the membership of a vector (or model) to a cluster to decrease, as long as the distance from the centre increases. $w_i(v)$ is the normalised membership degree of $v$ to cluster $i$. $\phi_k$ are calculated as an average of the normalised membership degree of models selected within cluster $k$.

The clustering algorithm partitions the parameter space of models and in each iteration of this partitioning, it performs the following tasks:

1. Calculates an OLS regression of the forecasts produced by models selected in each cluster, according to Equation 4.9, in order to obtain $\alpha$ coefficients.

2. Calculates coefficients $\phi$ as an average of the degree of belongingness of models in each cluster, according to Equations 4.12, 4.11 and 4.10.

Therefore, the models in each cluster are summarised in a single forecast and then are further combined. Figure 4.3 illustrates the case when the number of models per cluster is 5 and the models are in a two-dimensional space.

The partitioning, through a forward step, grows partitions in the form of a tree and prunes, in a backwards step, those regions that do not improve fit function (MSE). When the growing and pruning steps have stopped, an optimisation is made to carry out steps 1 and 2 above, using a non-linear optimisation routine (*fmincon* in Matlab).

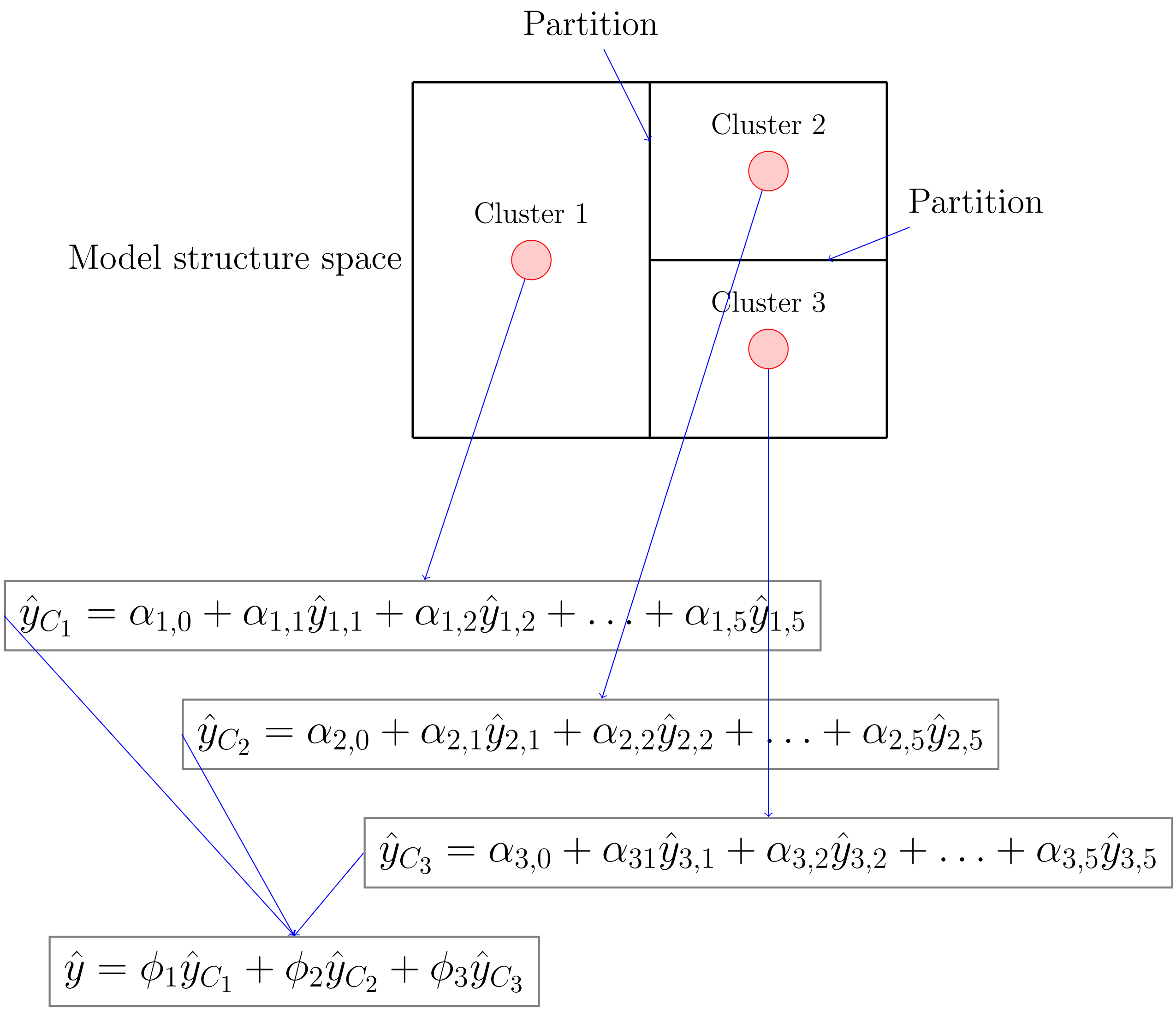

Partition
Cluster 2
Partition
Cluster 1
Model structure space
Cluster 3
$\hat{y}_{C_1} = \alpha_{1,0} + \alpha_{1,1}\hat{y}_{1,1} + \alpha_{1,2}\hat{y}_{1,2} + \ldots + \alpha_{1,5}\hat{y}_{1,5}$
$\hat{y}_{C_2} = \alpha_{2,0} + \alpha_{2,1}\hat{y}_{2,1} + \alpha_{2,2}\hat{y}_{2,2} + \ldots + \alpha_{2,5}\hat{y}_{2,5}$
$\hat{y}_{C_3} = \alpha_{3,0} + \alpha_{31}\hat{y}_{3,1} + \alpha_{3,2}\hat{y}_{3,2} + \ldots + \alpha_{3,5}\hat{y}_{3,5}$
$\hat{y} = \phi_1\hat{y}_{C_1} + \phi_2\hat{y}_{C_2} + \phi_3\hat{y}_{C_3}$

Figure 4.3: Structural combination based on clustering.

The regions used by the partitioning routines are defined as base functions:

$$B_m\left(\mathbf{Z}\right) = \prod_{k=1}^{K_m} H\left(s_{km} \cdot \left(z_{v(k,m)} - t_{km}\right)\right) \tag{4.13}$$

$$H(\eta) = \begin{cases} 1 \text{ if } \eta \geq 0 \\ 0 \text{ otherwise} \end{cases} \tag{4.14}$$

For a vector $z$, the function $B_m(z)$ establishes if $z$ belongs to the $m$-th region. If so, the function takes the value 1. If $z$ does not belong to the region, the function would have the value 0. $K_m$ is the number of partitions in the space that define the region. $s_{km}$ is a constant that takes the values 1 or -1, signalling if the partition is to the right or to the left of the value $t_{km}$. The variable $z_v$ is the dimension, in the space of parameters, in which a partition is made.

#### 4.3.2.2 Assessment of Uncertainty in the Forecast

Forecast intervals can be calculated via different techniques (for a comprehensive review see Khosravi et al., 2011). Some have been mainly applied to individual models (for example the Delta and MVE[4] methods) and some to ensembles (the bootstrap method).

The Delta method calculates forecast intervals for a single model by interpreting a NN as a non-linear regression model. It relies on the difference between the estimated weights of the network $\hat{w}$ and the optimal ones $w*$ (Khosravi et al., 2011). The MVE method estimates the mean and variance of the dependent variable by using separate NNs. The bootstrap method relies on generating forecasts by different models obtained through the re-sampling of training (in-sample) data. The empirical distribution of forecasts thus produced are used to estimate the mean and variance, needed to estimate (normal) confidence intervals, its main drawback is that it is more complex than other methods to implement (Khosravi et al., 2011).

[4] Mean-variance estimation.

The empirical method proposed by Lee & Scholtes (2014) is simpler and has lower computing demands, and is applied to the whole system (ensemble) as a unit. It has a parametric and a non-parametric version. The latter makes use of quantiles and therefore does not assume normally distributed forecast errors. This version is used here, in order to allow for the non-normality in forecasting errors that has been observed in Chapter 3.

The series of out-of-sample forecast errors $\hat{e}_{t,\tau} = y_{t+\tau} - \hat{y}_{t,\tau}$ for a horizon $\tau$, are used as a proxy for the true post-sample forecast errors and are generated through a rolling window over the out-of-sample period. If $\hat{o}(r)_{k,\tau}$ denotes the $r$th order statistic of the $k$ empirical forecast errors for a given lead time $\tau$, the non-parametric empirical forecast error quantile is then $\hat{Q}_\tau(p) = \hat{o}(r)_{k,\tau}$, where $r = \lfloor kp \rfloor + 1$, and $\lfloor s \rfloor$ is the largest integer $m$ such that $m \leq s$. The empirical prediction interval is given by

$$\left[\hat{L}_{n,\tau}, \hat{U}_{n,\tau}\right] = [\hat{y}_{n,\tau} + \hat{o}(r_L)_{k,\tau}, \hat{y}_{n,\tau} + \hat{o}(r_U)_{k,\tau}] \tag{4.15}$$

where $r_L = \lfloor k(1-\alpha)/2 \rfloor + 1$ and $r_U = \lfloor k(1+\alpha)/2 \rfloor + 1$

### 4.3.3 Structural Combination Based on Genetic Algorithms (GA)

A genetic algorithm based structural combination (GA) is proposed and is illustrated in Figure 4.4. A series of reference points in the NN parameter space is generated, which work in a similar way to cluster centres. From each point, $P_i$, five NN models are selected, as those having the smallest euclidean distance to it. The forecasts from these models are averaged, thus producing forecasts for each reference point. The final forecast combination ($\hat{y}_{Avg}$) is the average of these reference points forecasts. Genetic algorithms routines are used to select the reference points such that the MSE of the $\hat{y}_{Avg}$ in-sample one-step-ahead forecasts is minimised.

A GA combination can be viewed as a structurally informed average: it selects models based on their closeness to different points in the parameter space and then performs an average. The algorithm is run over the same NN pool that is used to perform the cluster-based structural combination. It was implemented in Matlab® 2010 using *ga* routine, with a maximum number of generations equal to 3000.

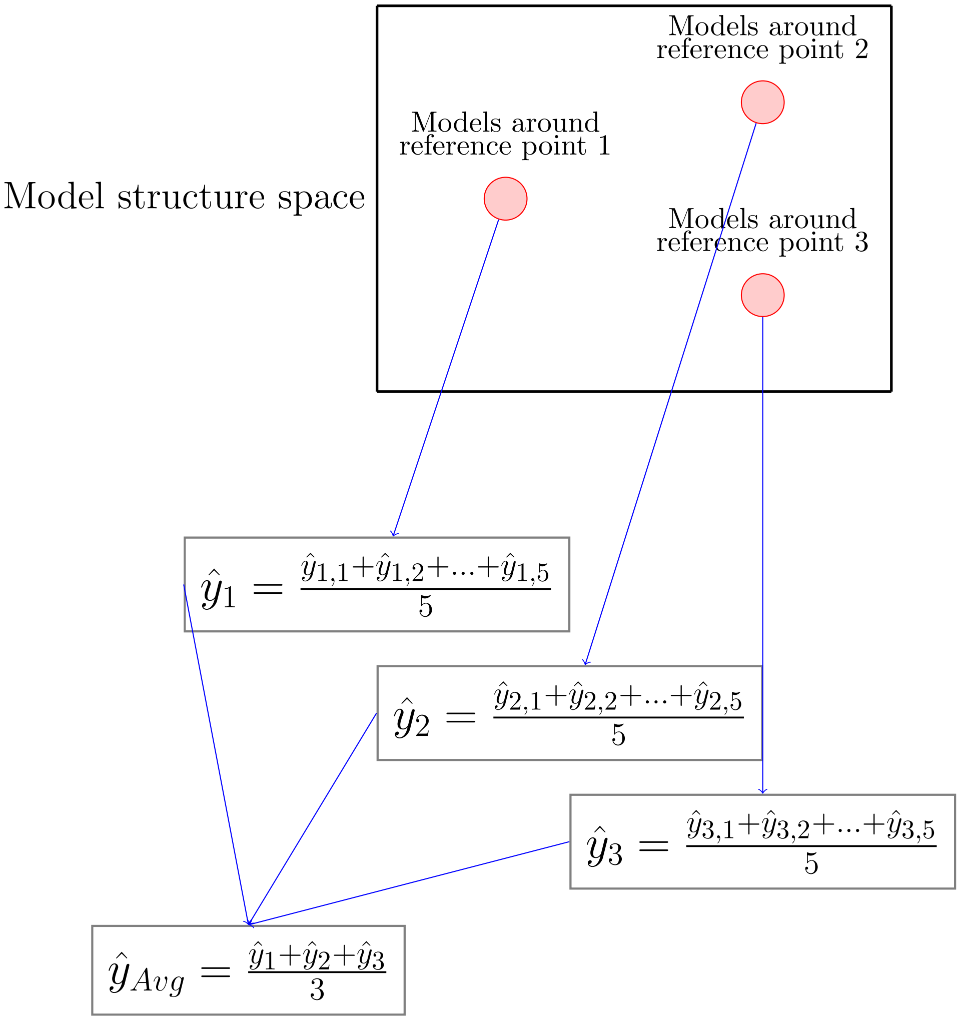

Model structure space
Models around reference point 1
Models around reference point 2
Models around reference point 3
$\hat{y}_1 = \frac{\hat{y}_{1,1}+\hat{y}_{1,2}+\ldots+\hat{y}_{1,5}}{5}$
$\hat{y}_2 = \frac{\hat{y}_{2,1}+\hat{y}_{2,2}+\ldots+\hat{y}_{2,5}}{5}$
$\hat{y}_3 = \frac{\hat{y}_{3,1}+\hat{y}_{3,2}+\ldots+\hat{y}_{3,5}}{5}$
$\hat{y}_{Avg} = \frac{\hat{y}_1+\hat{y}_2+\hat{y}_3}{3}$

Figure 4.4: Structural combination based on genetic algorithms.

## 4.4 The Empirical Studies

Three series from Chapter 3 were selected, and present different levels of complexity. STAR2 (Equation 3.8) is a relatively complex non-seasonal time series. The single-seasonal (Synthetic-1S) and double-seasonal (Synthetic-2S) series allow for the assessment of the proposed models with seasonal data and different degrees of complexity (see Equations 3.9 and 3.11). Secondly, two real time series are used: hourly observations for electricity demand in Rio de Janeiro and wind generation multivariate data from one of the wind farms included in the global Energy Forecasting Competition 2012 - Wind Forecasting (Kaggle, 2012). In all cases, synthetic and real data, models were subject to a preliminary sensitivity analysis by using 100 replications of the time series. When performing the forecast combination, 50 NNs were newly fitted to the original series, with randomised input-output patterns.

Table 4.1 shows the configuration of the experiments for all three synthetic series. The number of models in the ensemble was chosen in accordance with Bakker & Heskes (2003). For each time series, an ensemble was built and used in a structural combination with three levels of $MaxC$, the maximum number of clusters allowed. In all cases, $MperC$, the number of models per cluster, was set to 5. In this way, the maximum number of models selected from the ensemble would be at least 20% of the total and at most 80%, thus following findings by Zhou et al. (2002) which suggest that it is better to ensemble many available NNs but not all. The feed-forward NN models included in the ensembles have the same architecture, which was determined based on the sensitivity analysis that was described in Chapter 3 (for more information on this selection, see Appendix B).

The decomposition of error into bias and variance components (Geman et al., 1992; Bishop, 1995) gives insight into how to decrease the generalisation error in NNs. Bias appears when the model is far too simple to represent the underlying generating function. Variance appears in a model when it fits very well the data (due to high model complexity, for example), but still misses the true model. As pointed out by Bishop (1995), there is a trade-off between the two components. A model which closely fits the data will tend to have high variance, which can be lowered by reducing complexity to allow for a smoother approximation to the underlying function, but if taken too far, it can generate large bias and errors.

Due to averaging (see Equation 4.8) and to training using randomised data sets, which by promoting parameter diversity decreases deviations from the true model, the resulting forecast is expected to have less variation than that from a single model based on the fit of a static data set. The reduction of bias, on the other hand, depends on model complexity at both the individual level (single NN) and the global level (structural combination method).

Individual configuration parameters of networks are listed in Table 4.2. The ensembles are all implemented in Matlab® 2010 and ran using two PCs, each with two 2.2 GHz cores and 2 GB of RAM.

The proposed clustering combination approach relies on an algorithm that exhaustively explores all dimensions that describe the objects to be clustered. In the case of NNs, the number of dimensions grows with the size of the individual models involved and, consequently, the computing time increases at a rapid rate[5]. In light of this, the use of small models was favoured in the selection scheme, as suggested by the findings in Chapter 3.

The clustering algorithm works with a maximum number of clusters as its starting point and performs a pruning in a later stage. This means that, for example, if it starts with 4 as a maximum number of clusters, it could finish with 2. When the number of final clusters is increased, the degree of belongingness of a model to them tends to be similar: it is more likely that a model is close to several clusters

[5] The partitioning routine used to build the clusters has an order of $F \times O(MaxC \times V)$, where F is a factor depending on the density of models in the parameter space.

Table 4.1: Configuration of the clustering combination algorithm.

| **Factor** | **Symbol** | **Levels** |
|---|---|---|
| Number of models | NM | 50 |
| Num. Max. clusters | MaxC | 2, 4, 8 |
| Models per cluster | MperC | 5 |
| Final combination | FC | Linear Combination |
| Randomised input-output patterns | | Yes |
| Structural content to represent individual models | | All synaptic weights in each NN. |

Table 4.2: Configuration of individual networks.

| **Parameter** | **Value** |
|---|---|
| Number of hidden layers | 1 |
| Number of Inputs | Determined by sensitivity analysis |
| Number of hidden units | Determined by sensitivity analysis |
| Activation function for hidden nodes | Tangent Sigmoid |
| Activation function for the output node | Linear |
| Initial values for the weights | Values in the range [-2 2] established by the Nguyen-Widrow algorithm (there is a degree of randomness) |
| Training algorithm | Back-propagation with Levenberg-Marquardt optimisation. |
| Stopping criteria | * The maximum number of epochs (repetitions) is reached: 4000.<br>* The maximum amount of time is exceeded: $\infty$<br>* Performance is minimised to the goal: 0<br>* The performance gradient falls below $min_{grad} : 10^{-10}$<br>* $\mu$ exceeds $\mu_{max} = 10^3$<br>* Validation performance has increased more than $max_{fail}$ times since the last time it decreased (when using validation): 6 |
| Data normalisation | Yes |
| Initial combination coefficient ($\mu$) | 0.001 |

at the same time. Therefore the experiments conducted here limited the maximum number of clusters to 2, 4 and 8. After the CB models are fitted, a selection can be made based on comparative performance and the selection can be further supported by cluster validity measures. These measures are intended to evaluate the quality of the clusters found by an algorithm. Several have been developed for fuzzy-clustering (two reviews can be found in Wu & Yang, 2005; Zhang et al., 2014). The modified partition coefficient (MPC) is a simple measure that varies in the interval $[0, 1]$ and implies a well-performing partition when it approaches 1. Additionally, Abdallatif et al. (2016) proposed a measure focused on separation of clusters, called MDO (membership degree optimum). A proportion of elements with a degree of belongingness to some cluster superior to a threshold (55%) measures how well-partitioned are the elements into the clusters. These measures are here provided to supplement the analysis of results.

The MPC index is calculated as follows:

$$\begin{aligned} PC &= \frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{C} w_{i,j}^2 \\ MPC &= 1 - \frac{C}{1-C}(1-PC) \end{aligned} \tag{4.16}$$

The MDO measure is defined as follows:

$$\begin{aligned} &MDO = \frac{\text{Numer of elements for which } max(w_{i,j}) > 0.55}{N} \\ &1 \leq i \leq C \end{aligned} \tag{4.17}$$

, where $w_{i,j}$ is the degree of belongingness of element $j$ to cluster $i$, $C$ is the number of clusters and $N$ the number of elements clustered. The quantity $max(w_{i,j})$, for $1 \leq i \leq C$, refers to the maximum membership degree for an element $j$ across all clusters.

### 4.4.1 Analysis Procedure

The main error metrics used to evaluate the forecast performance of the proposed combination model are Mean Squared Error (MSE) and Mean Absolute Percentage Error (MAPE) metrics. For the time horizon $h$ they are calculated as follows:

$$MSE_h = \frac{1}{N-h-IS+1} \sum_{Fo=IS}^{N-h} (x_{Fo+h} - \hat{x}_{Fo+h})^2 \tag{4.18}$$

$$MAPE_h = \frac{1}{N-h-IS+1} \sum_{Fo=IS}^{N-h} \left| \frac{x_{Fo+h} - \hat{x}_{Fo+h}}{x_{Fo+h}} \right| \tag{4.19}$$

Where $N$ is the length of the time series, $IS$ is the index of the last in-sample observation, $Fo$ is the forecast origin, $x_i$ is the observed value and $\hat{x_i}$ is the forecasted value. Other error metrics (MdAPE, SMAPE, MAE, NMAPE or RMSE) are added to the analysis, when significant differences are observed between MAPE and MSE, when MAPE becomes unestable due to a time series having values close to zero or when a different metric facilitates comparisons with existing literature[6].

During the performance evaluation, comparisons are made against the following models:

- For non-seasonal series:
  - Naïve benchmark: the current observation is used to issue a forecast for $h$ steps ahead.
  - NN with the lowest in-sample MSE: a NN model is selected from the ensemble constructed for forecast horizon $h$, having the lowest in-sample MSE from all the models.

[6] $MdAPE = median(\{|(X_i - F_i)/X_i|\}_{i=1}^{n})$, $SMAPE = 100(1/n)\sum_{t=1}^{n} \frac{|X_t - F_t|}{(|X_t|+|F_t|)/2}$, $NMAPE = \frac{1}{n}\frac{\sum_{i=1}^{n}|X_i - F_i|}{(1/n)\sum_{i=1}^{n} X_i}$, where $X_t$ and $F_t$ are the actual and forecasted values, respectively.

  - NN with the lowest in-sample MAE or MAPE: a NN model is selected from the same ensemble, having the lowest in-sample MAE (when the time series had values close to zero) or MAPE (in all other cases) from all the models.
  - Average of point forecasts of all NN in the ensemble.
  - ARIMA: benchmarks from this family of models were obtained though the automatic identification routine provided in *forecast* R package (see Hyndman & Khandakar, 2008; Hyndman, 2015).

- For seasonal series:

  All benchmarks described above are used, with the Naïve, seasonal and double-seasonal statistical models being the following:

  - Naïve benchmark: the forecast for time period $t$ and lead time $k$ is $\hat{y}_t(k) = y_{t+k-S}$, where $S$ is the longest seasonal cycle.
  - Seasonal ARIMA and seasonal Holt-Winters model.
  - Double-Seasonal Holt-Winters-Taylor model based on Taylor (2010).

The best fit (minimum MSE) models and the average are based on the NN pool from which the structural combination is performed. The use of a single model as benchmark is well established (Yu et al., 2008; Fan et al., 2009), but its selection criterion is generally subjective. In our case, the MSE is adopted, as this is the most common fit measure that is found in the literature. The use of the simple average of forecasts as a benchmark is common and justified by its robustness (De Menezes et al., 2000).

Serial correlation in forecast errors is assessed via the Ljung-Box test and normality of the forecast error distribution by using Lilliefors and Jarque-Bera tests, which are available in different software packages and are commonly found in text books[7]. In the following sections results and analysis are presented.

## 4.5 Studies with Synthetic Series

### 4.5.1 STAR2 Series

The generating process of this series, is the following:

$$y_t = 0.3y_{t-1} + 0.6y_{t-2} + (0.1 - 0.9y_{t-1} + 0.8y_{t-2})(1 + e^{-10y_{t-1}})^{-1} + \varepsilon_t \tag{4.20}$$

The simulated series is depicted in Figure 4.5. By using the screening procedure proposed in section 3.4.5 a series of suitable models was obtained and the selected structures are listed in Table 4.3.

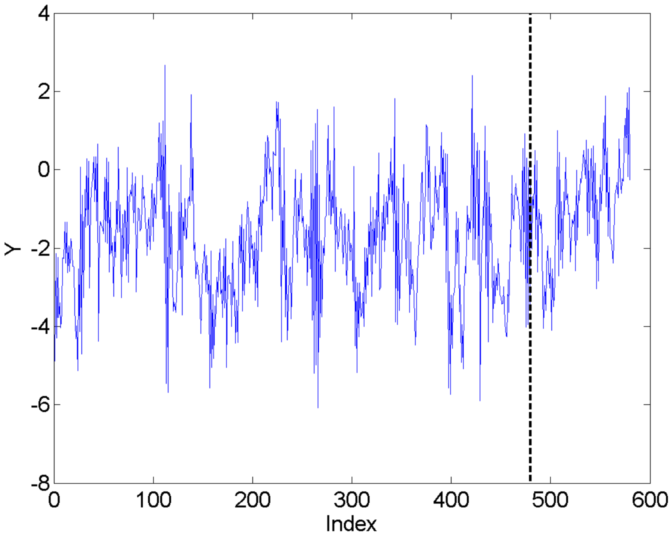


Figure 4.5: STAR2 series. The dashed line separates the in-sample from the out-of-sample period.

[7] Models (and ensembles) are used to forecast separately for each horizon. Forecasts are produced in a rolling window fashion: $(y_{t-k}, \ldots y_t)$ are used to obtain $\hat{y}_{t+h}$, then $(y_{t-k+1}, \ldots y_{t+1})$ are used to obtain $\hat{y}_{t+h+1}$. Therefore the Ljung-box test was used to assess the serial correlation of errors $(y_{t+h} - \hat{y}_{t+h})$, $(y_{t+h+1} - \hat{y}_{t+h+1}), \ldots$

Table 4.3: Selected NN models for STAR2 series.

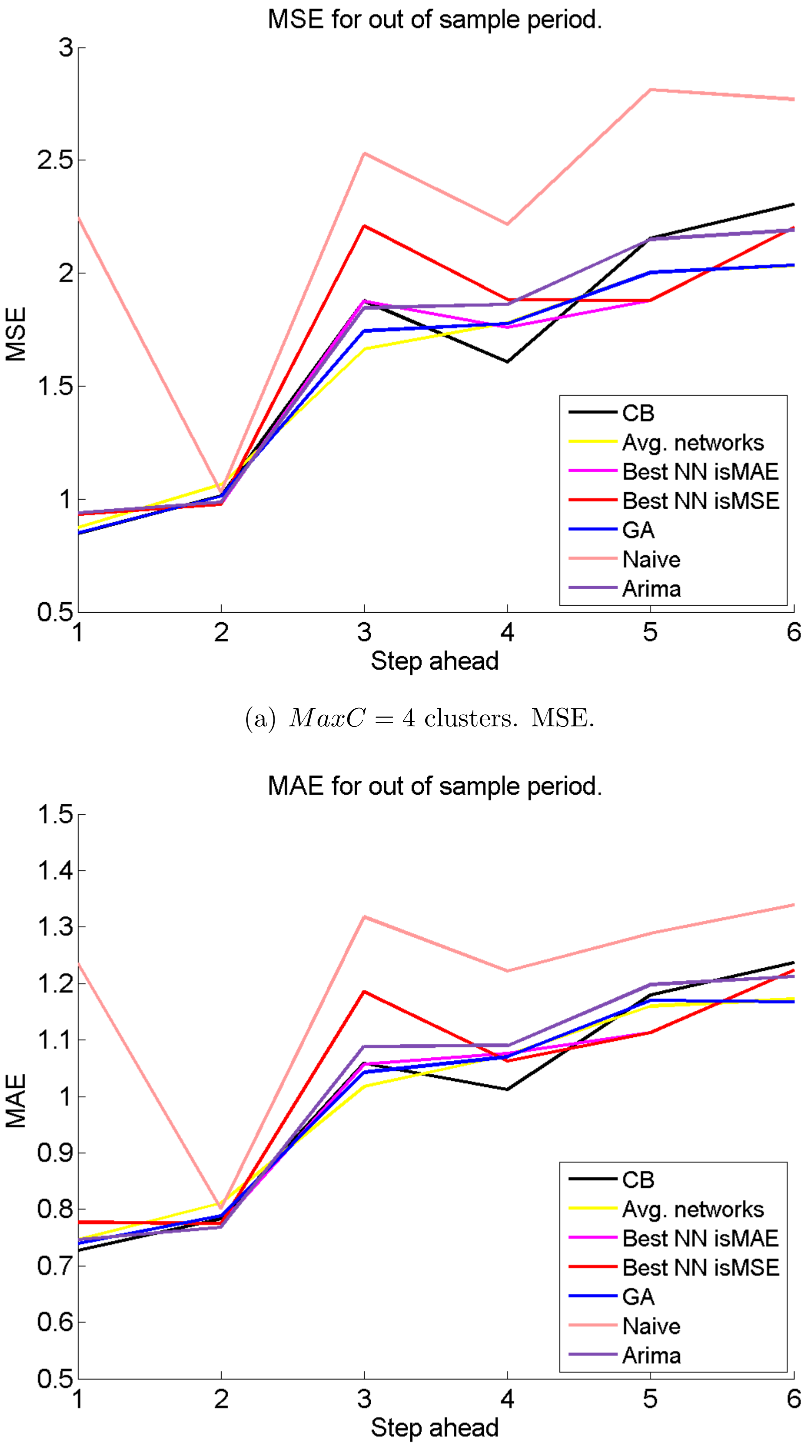
| h | NI | NU |
|---|---|---|
| 1 | 2 | 3 |
| 2 | 2 | 1 |
| 3 | 3 | 5 |
| 4 | 2 | 5 |
| 5 | 2 | 1 |
| 6 | 2 | 3 |

From the three sample sizes used during the sensitivity analysis conducted in Chapter 3, 480 was chosen for the assessment of the ensemble approach, since middle and upper sizes (960) were found to be appropriate. From this sample, 10% of observations were used for validation, corresponding to 48 observations, leaving 432 observations for training. Out-of-sample evaluation was conducted using 100 observations.

Figure 4.6 shows the out-of-sample MSE and MAE for models estimated using the proposed clustering algorithm and the selected benchmarks, when the maximum number of cluster was 4 (the full set of graphs, comprising all levels of clusters can be found in Section C.1 from Appendix C). Additionally, Figure 4.7 provides detail for a subset of models. Table 4.4 provides a ranking of models for every forecast horizon, with the first position corresponding to the lowest error metric, and a percentage of error difference with respect to the forecast average of all the NNs in the ensemble. Considering benchmarks for this series, the following ARIMA model was obtained though the automatic specification routine *auto.arima* available in R *forecast* package (Hyndman & Khandakar, 2008; Hyndman, 2015):

$$y_t = -1.8277 - 0.6631y_{t-1} + 0.5983y_{t-2} + 0.4449y_{t-3} + e_t + 0.7827e_{t-1} + 0.2217e_{t-2} \tag{4.21}$$

The error metrics for this series (MSE and MAE) behave similarly. GA structural combinations (simpler than CB) perform similarly to CB, and their forecast accuracy



(a) $MaxC = 4$ clusters. MSE.

(b) $MaxC = 4$ clusters. MAE.

Figure 4.6: Out-of-sample MSE and MAE for STAR2 series.

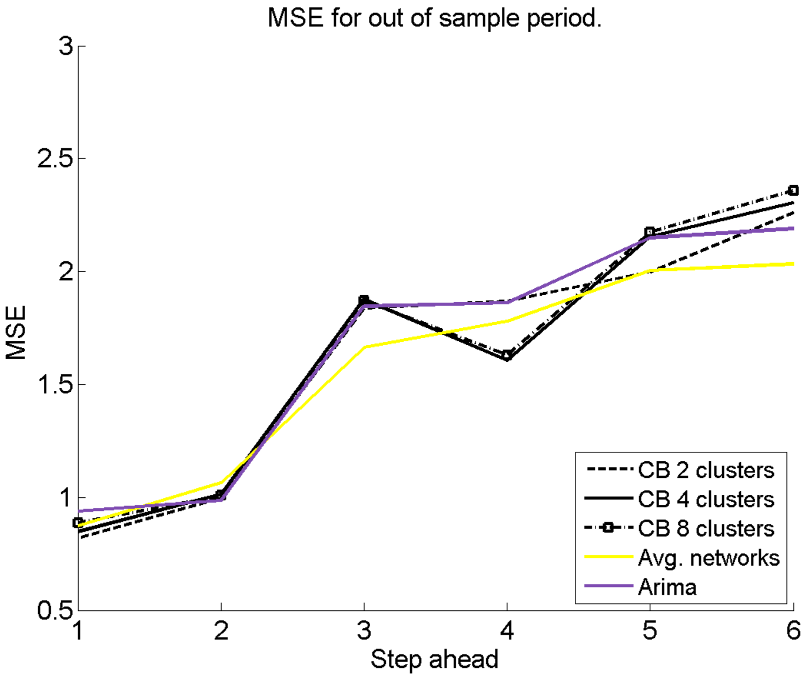


Figure 4.7: Out-of-sample MSE for STAR2 series.

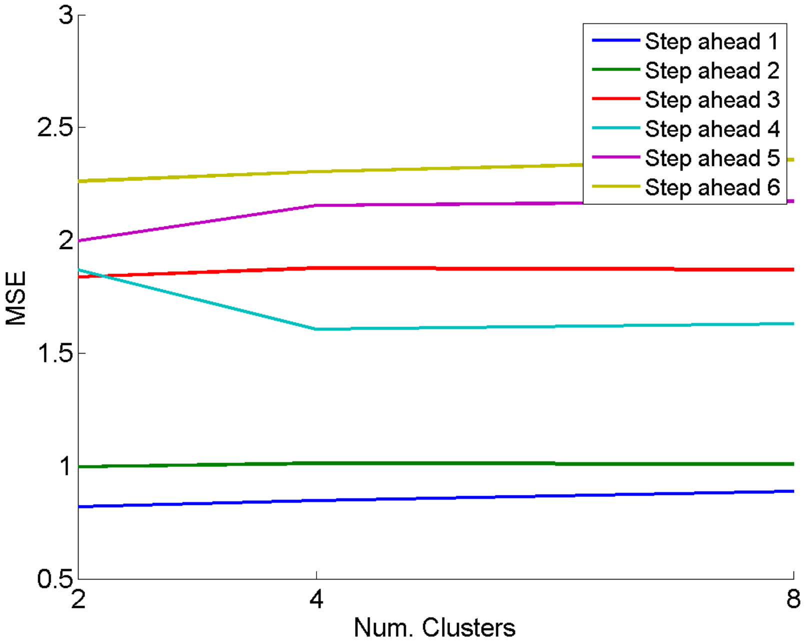


Figure 4.8: Out-of-sample MSE vs. number of clusters. STAR2 series.

tend to be more stable throughout forecast horizons.

There is relative insensitivity of the MSE to the number of clusters, as seen in Figure 4.8. Further analysis indicated a similar behaviour for MAE. Among CB combinations, CB4 (with a maximum of 4 clusters) has a more stable MSE and MAE throughout forecast horizons. Parameters for this model are detailed in Table 4.5. Diversity in the contribution of different clusters to the forecasts is observed in the different weighting of cluster outputs ($\Phi$ parameter). Variability is noticed in the $\alpha_j$ parameters, probably due to differences in performance among individual models in specific clusters, which in turn might come from the volatility in the time series. Similar characteristics are observed in CB2 and CB8 models (details available from the author upon request).

Table 4.6 indicates serially independent forecast errors for horizons 1 and 2, while subsequently, there is evidence of serial correlation. This is in line with findings from Chapter 3, where a better behaviour of individual NN models was observed for short horizons[8]. Additionally, the generating process for this series is based on two previous values, which are inputs in the NNs for all ensembles. A better ability to capture the dynamics of the series is then expected for these horizons. Lilliefors and Jarque-Bera normality tests confirm normally distributed forecast errors. Uncertainty in forecasts for CB combinations with different number of clusters is comparable, as seen in Figure 4.9, and is consistent with the high volatility of the time series.

[8] Correlation maps like those provided in summary Figure 3.7 were used in this assessment.

Table 4.4: Forecasting performance. STAR2 series.

| Forecast horizon | Model | MSE | Rank | %Δ wrt Avg. | MAE | | %Δ wrt Avg. |
|---|---|---|---|---|---|---|---|
| h=1 | CB2 | 0.81853 | 1 | -6.32% | 0.71167 | 1 | -4.66% |
| | CB4 | 0.84685 | 4 | -3.08% | 0.72675 | 2 | -2.64% |
| | CB8 | 0.88707 | 7 | 1.52% | 0.74157 | 6 | -0.65% |
| | Avg. Net. | 0.87378 | 6 | 0.00% | 0.74643 | 8 | 0.00% |
| | Best Net. IsMAE | 0.93347 | 8 | 6.83% | 0.77682 | 10 | 4.07% |
| | Best Net. isMSE | 0.93347 | 9 | 6.83% | 0.77682 | 9 | 4.07% |
| | GA2 | 0.84198 | 3 | -3.64% | 0.73425 | 4 | -1.63% |
| | GA4 | 0.84947 | 5 | -2.78% | 0.73923 | 5 | -0.96% |
| | GA8 | 0.83923 | 2 | -3.95% | 0.73356 | 3 | -1.72% |
| | Naive | 2.24582 | 11 | 157.02% | 1.23528 | 11 | 65.49% |
| | ARIMA(3,0,2) | 0.93897 | 10 | 7.46% | 0.74633 | 7 | -0.01% |
| h=2 | CB2 | 0.99557 | 4 | -6.30% | 0.77460 | 2 | -4.45% |
| | CB4 | 1.01240 | 8 | -4.72% | 0.78334 | 6 | -3.37% |
| | CB8 | 1.00860 | 5 | -5.07% | 0.78191 | 5 | -3.55% |
| | Avg. Net. | 1.06250 | 11 | 0.00% | 0.81067 | 11 | 0.00% |
| | Best Net. IsMAE | 0.97734 | 1 | -8.02% | 0.77483 | 4 | -4.42% |
| | Best Net. isMSE | 0.97734 | 2 | -8.02% | 0.77483 | 3 | -4.42% |
| | GA2 | 1.01845 | 9 | -4.15% | 0.79083 | 9 | -2.45% |
| | GA4 | 1.01200 | 7 | -4.75% | 0.78882 | 8 | -2.70% |
| | GA8 | 1.01021 | 6 | -4.92% | 0.78860 | 7 | -2.72% |
| | Naive | 1.03055 | 10 | -3.01% | 0.80059 | 10 | -1.24% |
| | ARIMA(3,0,2) | 0.98545 | 3 | -7.25% | 0.76837 | 1 | -5.22% |
| h=3 | CB2 | 1.83800 | 5 | 10.50% | 1.05640 | 6 | 3.80% |
| | CB4 | 1.87720 | 9 | 12.86% | 1.05900 | 8 | 4.06% |
| | CB8 | 1.87060 | 7 | 12.46% | 1.04780 | 5 | 2.96% |
| | Avg. Net. | 1.66330 | 1 | 0.00% | 1.01770 | 1 | 0.00% |
| | Best Net. IsMAE | 1.87623 | 8 | 12.80% | 1.05666 | 7 | 3.83% |
| | Best Net. isMSE | 2.20830 | 10 | 32.77% | 1.18570 | 10 | 16.51% |
| | GA2 | 1.74767 | 4 | 5.07% | 1.04716 | 4 | 2.89% |
| | GA4 | 1.74426 | 2 | 4.87% | 1.04260 | 2 | 2.45% |
| | GA8 | 1.74454 | 3 | 4.88% | 1.04266 | 3 | 2.45% |
| | Naive | 2.52912 | 11 | 52.05% | 1.31771 | 11 | 29.48% |
| | ARIMA(3,0,2) | 1.84573 | 6 | 10.97% | 1.08851 | 9 | 6.96% |

$\%\Delta = 100(M_{model} - M_{Avg})/M_{Avg}$ with $M_i$ being the metric for model $i$. Negative % values indicate improvement over the average. CB and GA refer to clustering based and genetic algorithm based structural combinations with the corresponding number of clusters. Bst. isMAE and Bst. isMSE denote the NNs with the lowest in-sample MAE and MSE respectively.

Forecasting performance (continued). STAR2 series.

| Forecast horizon | Model | MSE | Rank | %Δ wrt Avg. | MAE | | %Δ wrt Avg. |
|---|---|---|---|---|---|---|---|
| h=4 | CB2 | 1.86860 | 9 | 4.94% | 1.08820 | 9 | 1.44% |
| | CB4 | 1.60620 | 1 | -9.79% | 1.01170 | 2 | -5.70% |
| | CB8 | 1.63030 | 2 | -8.44% | 1.00400 | 1 | -6.41% |
| | Avg. Net. | 1.78060 | 7 | 0.00% | 1.07280 | 7 | 0.00% |
| | Best Net. IsMAE | 1.76128 | 4 | -1.09% | 1.07604 | 8 | 0.30% |
| | Best Net. isMSE | 1.88170 | 10 | 5.68% | 1.06270 | 3 | -0.94% |
| | GA2 | 1.76397 | 5 | -0.93% | 1.06658 | 5 | -0.58% |
| | GA4 | 1.77568 | 6 | -0.28% | 1.07005 | 6 | -0.26% |
| | GA8 | 1.75204 | 3 | -1.60% | 1.06466 | 4 | -0.76% |
| | Naive | 2.21601 | 11 | 24.45% | 1.22139 | 11 | 13.85% |
| | ARIMA(3,0,2) | 1.86225 | 8 | 4.59% | 1.08993 | 10 | 1.60% |
| h=5 | CB2 | 1.99900 | 4 | -0.28% | 1.16170 | 5 | 0.19% |
| | CB4 | 2.15420 | 9 | 7.46% | 1.17940 | 9 | 1.72% |
| | CB8 | 2.17360 | 10 | 8.43% | 1.17220 | 8 | 1.10% |
| | Avg. Net. | 2.00460 | 6 | 0.00% | 1.15950 | 4 | 0.00% |
| | Best Net. IsMAE | 1.87812 | 2 | -6.31% | 1.11278 | 1 | -4.03% |
| | Best Net. isMSE | 1.87810 | 1 | -6.31% | 1.11280 | 2 | -4.03% |
| | GA2 | 1.98273 | 3 | -1.09% | 1.15316 | 3 | -0.55% |
| | GA4 | 2.00338 | 5 | -0.06% | 1.16968 | 6 | 0.88% |
| | GA8 | 2.00817 | 7 | 0.18% | 1.16977 | 7 | 0.89% |
| | Naive | 2.81245 | 11 | 40.30% | 1.28869 | 11 | 11.14% |
| | ARIMA(3,0,2) | 2.14999 | 8 | 7.25% | 1.19797 | 10 | 3.32% |
| h=6 | CB2 | 2.26110 | 8 | 11.15% | 1.21930 | 6 | 4.06% |
| | CB4 | 2.30370 | 9 | 13.24% | 1.23640 | 9 | 5.52% |
| | CB8 | 2.35710 | 10 | 15.87% | 1.24270 | 10 | 6.06% |
| | Avg. Net. | 2.03430 | 3 | 0.00% | 1.17170 | 4 | 0.00% |
| | Best Net. IsMAE | 2.20135 | 6 | 8.21% | 1.22391 | 8 | 4.46% |
| | Best Net. isMSE | 2.20140 | 7 | 8.21% | 1.22390 | 7 | 4.46% |
| | GA2 | 1.99902 | 1 | -1.73% | 1.15856 | 1 | -1.12% |
| | GA4 | 2.03467 | 4 | 0.02% | 1.16724 | 3 | -0.38% |
| | GA8 | 2.01675 | 2 | -0.86% | 1.16299 | 2 | -0.74% |
| | Naive | 2.76895 | 11 | 36.11% | 1.33940 | 11 | 14.31% |
| | ARIMA(3,0,2) | 2.19070 | 5 | 7.69% | 1.21249 | 5 | 3.48% |

Table 4.5: Coefficients for structural combination of NNs. STAR2 series.

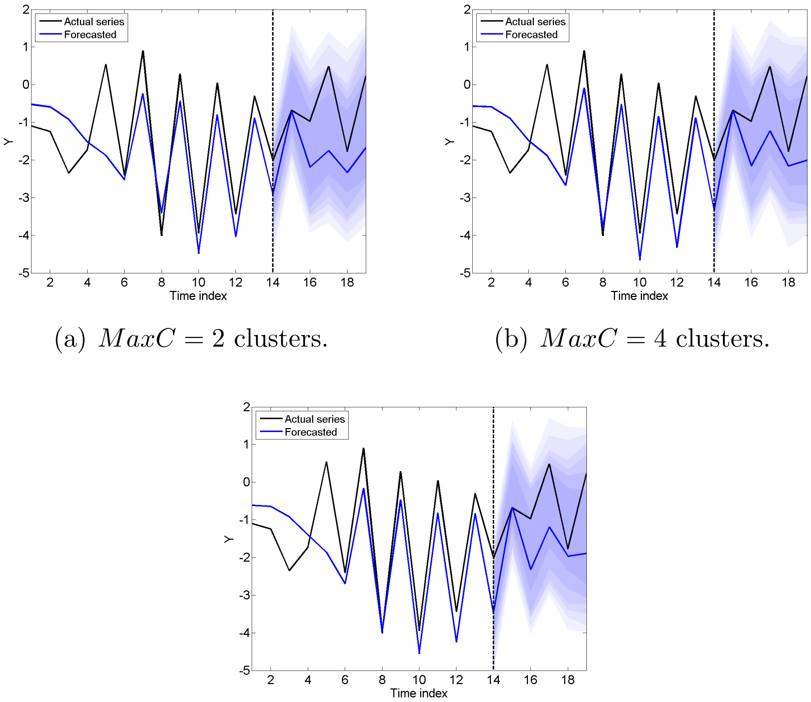
| h | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1 | $\alpha_1$ | 0.0024 | 0.5177 | -0.1897 | 0.4398 | 0.3582 | 0.6047 |
| | $\alpha_2$ | -0.0386 | -0.1514 | -0.8881 | 0.6099 | 1.2979 | -0.6475 |
| | $\alpha_3$ | -0.1054 | -0.2808 | 0.5041 | -0.8643 | 0.5226 | 0.7118 |
| | $\Phi$ | 0.4209 | 0.1694 | 0.4171 | | | |
| 2 | $\alpha_1$ | -0.4498 | 6.6383 | 9.5711 | -3.7736 | -3.5243 | -8.1883 |
| | $\alpha_2$ | 1.8857 | 3.1845 | -1.2678 | -0.5329 | 2.6914 | -2.1761 |
| | $\alpha_3$ | 1.1315 | 3.2374 | 3.7720 | -2.0936 | -4.2302 | 0.7674 |
| | $\Phi$ | 0.3700 | 0.2733 | 0.3511 | | | |
| 3 | $\alpha_1$ | -0.0905 | 0.9923 | 1.1698 | 0.2526 | 0.7650 | -0.9240 |
| | $\alpha_2$ | -0.0859 | -0.6742 | 0.7467 | 0.0395 | 0.4493 | 0.3494 |
| | $\alpha_3$ | 0.2564 | -0.7611 | 0.5774 | 0.1967 | -0.2842 | -0.0286 |
| | $\alpha_4$ | -0.0808 | -1.0623 | 0.1789 | 0.7821 | -0.1564 | 1.2046 |
| | $\Phi$ | 0.2628 | 0.2747 | 0.2543 | | | |
| 4 | $\alpha_1$ | -0.0708 | -0.1028 | 0.1286 | -0.6521 | -0.8088 | 0.2282 |
| | $\alpha_2$ | -0.1797 | -0.7864 | -1.1721 | -0.3896 | 1.6441 | 1.4353 |
| | $\alpha_3$ | -0.5508 | -0.2370 | -0.2068 | 0.7653 | 0.5905 | 2.4214 |
| | $\alpha_4$ | -0.3565 | 1.5184 | -2.6119 | 0.5516 | 0.9769 | 0.0943 |
| | $\Phi$ | 0.2877 | 0.2426 | 0.2689 | | | |
| 5 | $\alpha_1$ | -1.3582 | 1.8220 | 0.1425 | -2.5384 | -0.0434 | 3.0505 |
| | $\alpha_2$ | 3.9163 | 0.3606 | -2.2459 | 5.1728 | -4.5140 | 5.4884 |
| | $\alpha_3$ | 0.9281 | 2.7350 | -0.1212 | -7.0898 | 1.4585 | 2.1944 |
| | $\Phi$ | 0.2216 | 0.4327 | 0.3909 | | | |
| 6 | $\alpha_1$ | -0.2165 | -1.1697 | -1.5304 | 0.5813 | -0.3042 | 0.8755 |
| | $\alpha_2$ | 0.0909 | -2.5596 | 1.1675 | 0.4293 | 1.4063 | -0.4797 |
| | $\alpha_3$ | 0.0609 | 0.9739 | 0.2828 | 1.1587 | 1.1072 | 1.1011 |
| | $\Phi$ | 0.3841 | 0.3129 | 0.3797 | | | |

$MaxC = 4$ is the maximum number of clusters. $h$ denotes the forecast horizon, $\alpha_i$ are the coefficients applied to point-forecasts from models in cluster $i$ and $\Phi$ are the weights applied to the outputs from clusters.

Table 4.6: Ljung-Box test. Series: STAR2.

| | h | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| 2C | CB | | | * | * | * | * |
| | GA | | | * | * | * | * |
| 4C | CB | | | * | * | * | * |
| | GA | | | * | * | * | * |
| 8C | CB | | | * | * | * | * |
| | GA | | | * | * | * | * |
| | Avg. Net. | | | * | * | * | * |
| | Best Net. isMAE | | | * | * | * | * |
| | Best Net. isMSE | | | * | * | * | * |

Ljung-Box test for serial correlation (with 95% confidence level) for STAR2 series. The rejection of the hypothesis of independent forecast errors is indicated with *.



(a) $MaxC = 2$ clusters.

(b) $MaxC = 4$ clusters.

(c) $MaxC = 8$ clusters.

Figure 4.9: Forecast intervals for STAR2 time series.

The graphs cover the period for $t - 12 \leq t \leq t + H$ where $t$ is the last observation of the in-sample period and $H = 6$ is the number of forecast horizons. The shades, from lighter to darker, correspond to $\alpha$ levels 0.95, 0.90, 0.85, 0.80, 0.75 and 0.60.

Table 4.7: Cluster validity indexes. Series: STAR2.

| | Maximum number of clusters: 2 | | | |
|---|---|---|---|---|
| *h* | PC | MPC | MDO | Final num. clusters |
| 1 | 1.0000 | NA | 1.0000 | 1 |
| 2 | 1.0000 | NA | 1.0000 | 1 |
| 3 | 1.0000 | NA | 1.0000 | 1 |
| 4 | 0.5036 | 0.0072 | 0.3000 | 2 |
| 5 | 0.5961 | 0.1923 | 1.0000 | 2 |
| 6 | 1.0000 | NA | 1.0000 | 1 |
| | Maximum number of clusters: 4 | | | |
| 1 | 0.3757 | 0.0635 | 0.0000 | 3 |
| 2 | 0.3399 | 0.0099 | 0.0000 | 3 |
| 3 | 0.2508 | 0.0010 | 0.0000 | 4 |
| 4 | 0.2522 | 0.0029 | 0.0000 | 4 |
| 5 | 0.3591 | 0.0386 | 0.0000 | 3 |
| 6 | 0.3379 | 0.0068 | 0.0000 | 3 |
| | Maximum number of clusters: 8 | | | |
| 1 | 0.2136 | 0.0170 | 0.0000 | 5 |
| 2 | 0.2009 | 0.0011 | 0.0000 | 5 |
| 3 | 0.1671 | 0.0005 | 0.0000 | 6 |
| 4 | 0.1687 | 0.0025 | 0.0000 | 6 |
| 5 | 0.2135 | 0.0169 | 0.0000 | 5 |
| 6 | 0.1679 | 0.0015 | 0.0000 | 6 |

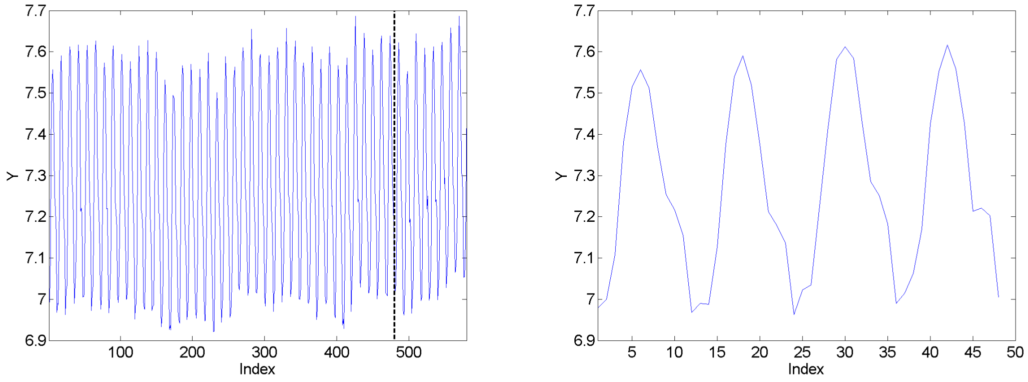
*h* denotes the forecast horizon, PC denotes the Partition Coefficient, MPC denotes the Modified Partition Coefficient and MDO denotes the Membership Degree Optimum. Values closer to 1 are preferable.

### 4.5.2 Synthetic-1S Series

The generating process of this series, as in Chapter 3, is

$$\begin{aligned} y_t(k) &= l_t + w_{t-s_2+k} + \phi^k(y_t - (l_{t-1} + w_{t-s_2})) + \varepsilon_t \\ l_t &= \lambda(y_t - w_{t-s_2}) + (1-\lambda)l_{t-1} \\ w_t &= \omega(y_t - l_{t-1}) + (1-\omega)w_{t-s_2} \end{aligned} \tag{4.22}$$

$y_t(k)$ is the simulated series value at time $t+k$, $l_t$ denotes the smoothed level and $w_t$ denotes the seasonal index. $\varepsilon_t \sim NID(0, \sigma^2)$, with $\sigma^2$ constant. The simulated series was generated with parameters $\lambda = 0.2$; $\omega = 0.01$; $\phi = 0.943$ and $s_2 = 12$ and is depicted in Figure 4.10. Table 4.8 lists the NN model specifications for the ensemble, selected though the sensitivity analysis and guidelines from Chapter 3.



(a) Series.

(b) Subset.

Figure 4.10: Synthetic-1S series.

Data partitioning is the same as above: 432 observations for training, 48 for validation and 100 for testing. Training and validation data comprise the in-sample set and testing data comprise the out-of-sample set.

For this series, the following SARIMA benchmark was obtained with the *auto.arima*

Table 4.8: Selected NN models for Synthetic-1S series.

| h | NI | NU | h | NI | NU |
|---|---|---|---|---|---|
| 1 | 7 | 9 | 7 | 6 | 9 |
| 2 | 7 | 9 | 8 | 4 | 7 |
| 3 | 6 | 10 | 9 | 3 | 8 |
| 4 | 7 | 9 | 10 | 3 | 8 |
| 5 | 6 | 9 | 11 | 2 | 7 |
| 6 | 7 | 8 | 12 | 1 | 8 |

routines available in the *forecast* R package:

$$(1 - \phi_1 B - \phi_2 B^2 - \phi_3 B^3 - \phi_4 B^4)(1 - \Phi_1 B^{12})(y_t - \mu) = (1 + \theta_1 B + \theta_2 B^2 + \theta_3 B^3)e_t$$

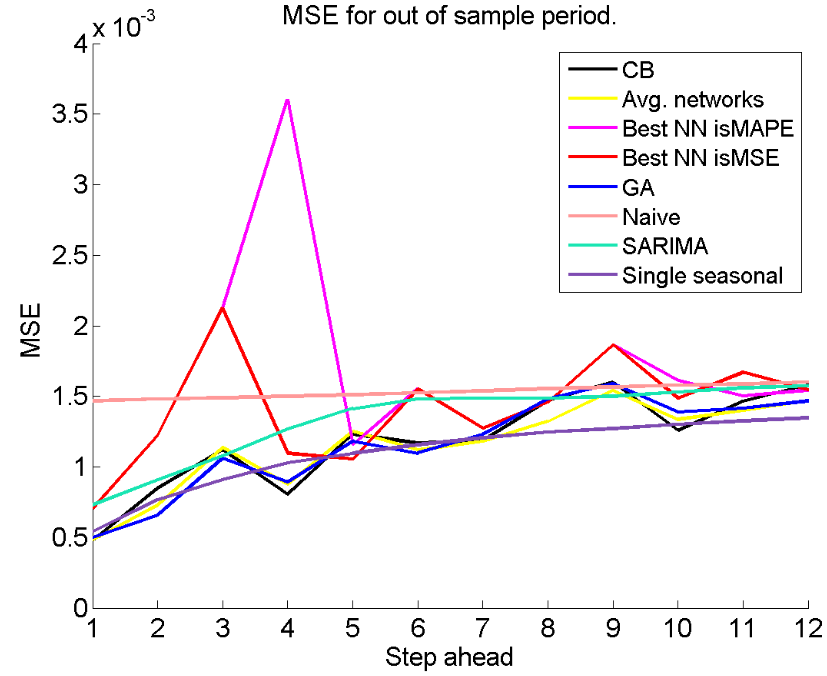
where $\phi_1 = 0.8057, \phi_2 = -0.4618, \phi_3 = -0.3330, \phi_4 = 0.3485, \Phi_1 = 0.9872,$

$\theta_1 = -0.4685, \theta_2 = 0.5563, \theta_3 = 0.4718$ and $\mu = 7.2677$

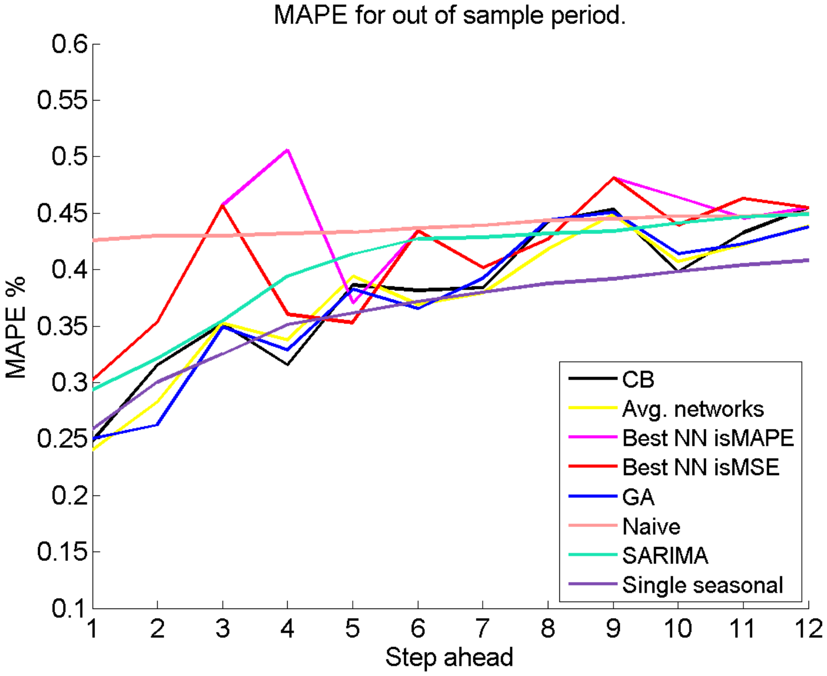
A single seasonal benchmark model was also built, following Equation 4.22, which resulted in parameters $\lambda = 0.0418$, $\omega = 0.5221$ and $\phi = 0.5218$. Although the parameters of the fitted model and the generating process are different, the forecasts are equivalent as judged by plots and RMSE.

Figure 4.11 depicts the out-of-sample MSE and MAPE of benchmarks and CB models with a maximum of 4 clusters (additional graphs can be found in Figure C.3 from Appendix C). Further detail is available in Figure 4.12, which compares the cluster-based combinations with the average and statistical benchmarks. Additionally, Table 4.9 summarises performance according to the error metrics, the ranking of models and the percentage of improvement with respect to the average forecast.

In the ranking of models (Table 4.9) it is noticed how in every forecast horizon, at least one of the CB combinations ranks among the first three, and the set of CB models is generally followed by GA combinations. Despite the slightly better forecast performance for CB combinations compared to GA combinations, the later seem to be more stable throughout forecast horizons (Figures 4.11 and 4.12). The ARIMA benchmark is outperformed comfortably in most horizons by both structural



(a) $MaxC = 4$ clusters. MSE.



(b) $MaxC = 4$ clusters. MAPE.

Figure 4.11: Out-of-sample MSE and MAPE for Synthetic-1S series.

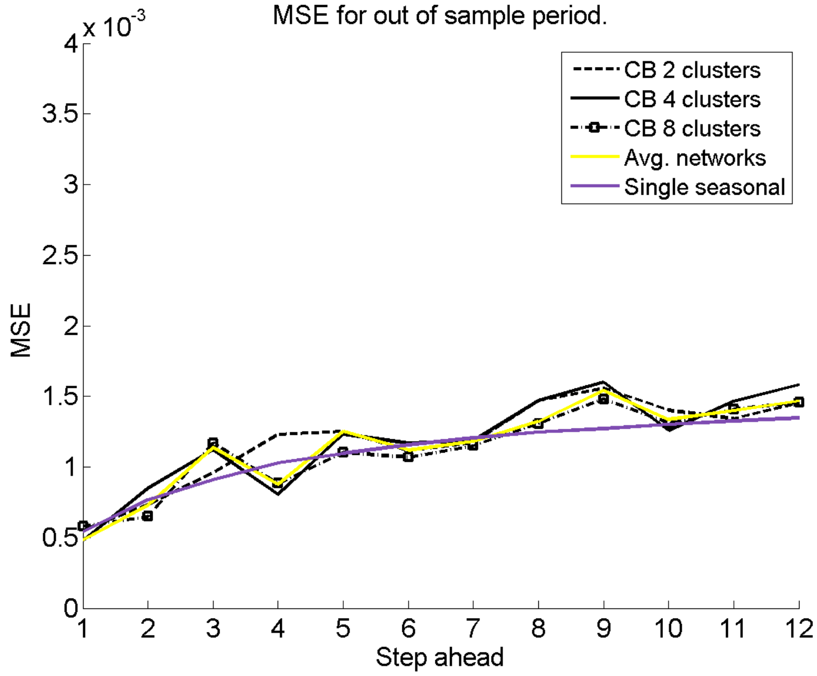
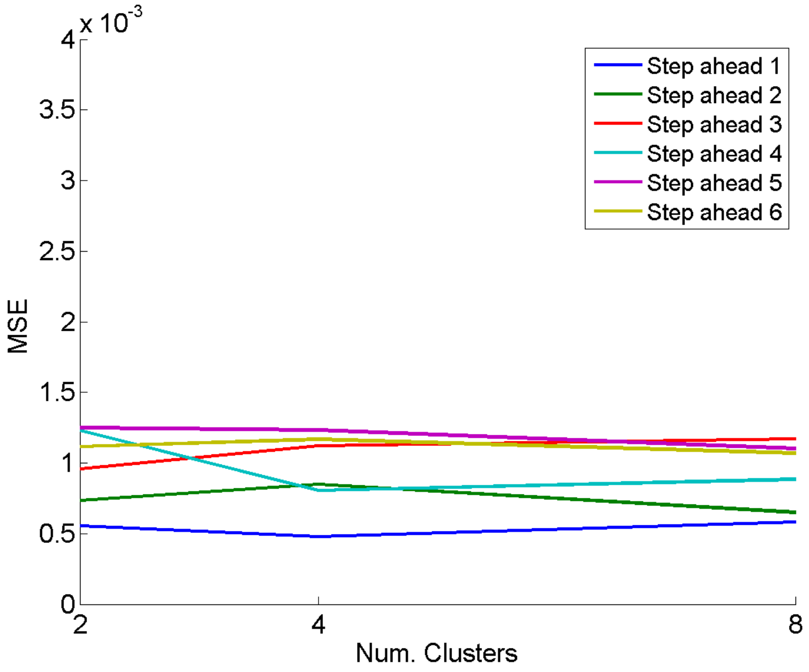


Figure 4.12: Out-of-sample MSE for Synthetic-1S series.

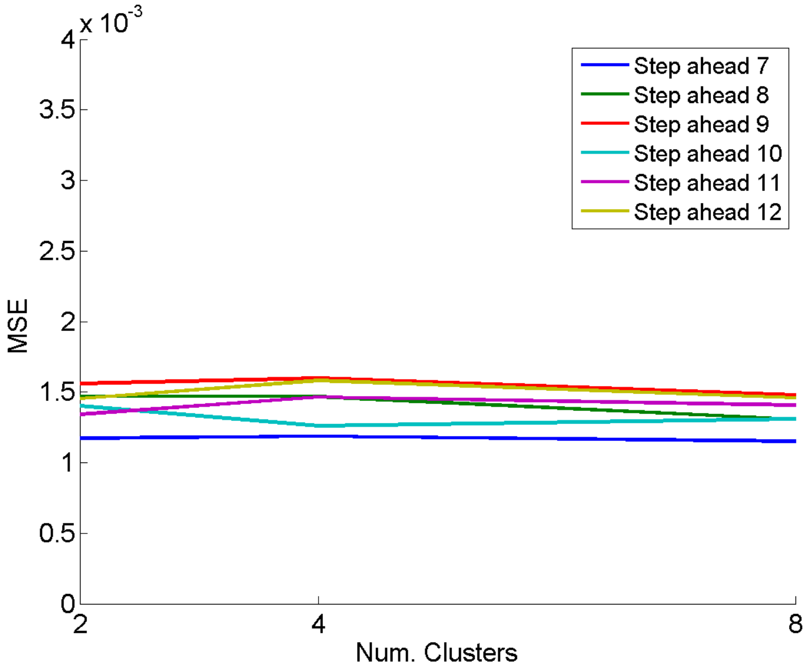
combinations, but the Single seasonal model is outperformed in only 4 horizons. The superiority of such benchmark is not surprising as it belongs to the family of models from which the synthetic series was simulated. Forecast accuracy in terms of MSE and MAPE is comparable.

No pattern is evident regarding the sensitivity of CB error metric to the number of clusters (Figure 4.13). The assessment of serial correlation revealed the presence of independent forecast errors for the first forecast horizon in all the cluster based models. Lilliefors and Jarque-Bera normality tests confirmed normality of errors for most horizons.

CB8, a cluster-based combination with a maximum of 8 clusters, outperforms more consistently the average of forecasts from the ensemble than other CB combinations. Parameters for that model are listed in Table 4.10 (details of other CB models and forecast horizons are omitted).



(a) Steps 1 to 6.



(b) Steps 7 to 12.

Figure 4.13: Out-of-sample MSE vs. number of clusters. Synthetic-1S series.

Table 4.9: Forecasting performance. Synthetic-1S series.

| Forecast horizon | Model | MSE | Rank | %Δ wrt Avg. | MAPE | | %Δ wrt Avg. |
|---|---|---|---|---|---|---|---|
| h=1 | CB2 | 0.000553 | 7 | 14.66% | 0.2609% | 7 | 8.55% |
| | CB4 | 0.000480 | 1 | -0.62% | 0.2482% | 2 | 3.30% |
| | CB8 | 0.000581 | 8 | 20.39% | 0.2684% | 8 | 11.71% |
| | Avg. Net. | 0.000483 | 2 | 0.00% | 0.2403% | 1 | 0.00% |
| | Best Net. isMAPE | 0.000707 | 9 | 46.42% | 0.3020% | 10 | 25.67% |
| | Best Net. isMSE | 0.000707 | 9 | 46.42% | 0.3020% | 10 | 25.67% |
| | GA2 | 0.000503 | 4 | 4.24% | 0.2518% | 4 | 4.79% |
| | GA4 | 0.000496 | 3 | 2.79% | 0.2503% | 3 | 4.16% |
| | GA8 | 0.000511 | 5 | 5.90% | 0.2528% | 5 | 5.20% |
| | Naive | 0.001467 | 12 | 204.04% | 0.4256% | 12 | 77.11% |
| | SARIMA | 0.000732 | 11 | 51.70% | 0.2937% | 9 | 22.22% |
| | Single seasonal | 0.000542 | 6 | 12.32% | 0.2582% | 6 | 7.47% |
| h=2 | CB2 | 0.000735 | 6 | 0.87% | 0.2909% | 6 | 2.89% |
| | CB4 | 0.000849 | 8 | 16.48% | 0.3154% | 8 | 11.57% |
| | CB8 | 0.000649 | 1 | -10.88% | 0.2639% | 2 | -6.65% |
| | Avg. Net. | 0.000729 | 5 | 0.00% | 0.2827% | 5 | 0.00% |
| | Best Net. isMAPE | 0.001225 | 10 | 68.17% | 0.3535% | 10 | 25.05% |
| | Best Net. isMSE | 0.001225 | 10 | 68.17% | 0.3535% | 10 | 25.05% |
| | GA2 | 0.000717 | 4 | -1.59% | 0.2747% | 4 | -2.83% |
| | GA4 | 0.000657 | 2 | -9.82% | 0.2623% | 1 | -7.22% |
| | GA8 | 0.000694 | 3 | -4.74% | 0.2696% | 3 | -4.64% |
| | Naive | 0.001482 | 12 | 103.40% | 0.4297% | 12 | 52.00% |
| | SARIMA | 0.000908 | 9 | 24.63% | 0.3213% | 9 | 13.65% |
| | Single seasonal | 0.000768 | 7 | 5.41% | 0.3004% | 7 | 6.26% |
| h=3 | CB2 | 0.000959 | 2 | -15.74% | 0.3243% | 1 | -8.07% |
| | CB4 | 0.001122 | 7 | -1.39% | 0.3520% | 5 | -0.19% |
| | CB8 | 0.001170 | 9 | 2.80% | 0.3588% | 9 | 1.74% |
| | Avg. Net. | 0.001138 | 8 | 0.00% | 0.3527% | 6 | 0.00% |
| | Best Net. isMAPE | 0.002127 | 11 | 86.93% | 0.4569% | 11 | 29.54% |
| | Best Net. isMSE | 0.002127 | 11 | 86.93% | 0.4569% | 11 | 29.54% |
| | GA2 | 0.000977 | 3 | -14.13% | 0.3335% | 3 | -5.45% |
| | GA4 | 0.001062 | 4 | -6.66% | 0.3489% | 4 | -1.08% |
| | GA8 | 0.001088 | 6 | -4.38% | 0.3545% | 7 | 0.51% |
| | Naive | 0.001488 | 10 | 30.79% | 0.4298% | 10 | 21.87% |
| | SARIMA | 0.001078 | 5 | -5.26% | 0.3546% | 8 | 0.54% |
| | Single seasonal | 0.000907 | 1 | -20.28% | 0.3251% | 2 | -7.84% |
| h=4 | CB2 | 0.001229 | 9 | 40.16% | 0.3922% | 9 | 16.18% |
| | CB4 | 0.000807 | 1 | -8.05% | 0.3154% | 1 | -6.58% |
| | CB8 | 0.000883 | 3 | 0.70% | 0.3382% | 5 | 0.16% |
| | Avg. Net. | 0.000877 | 2 | 0.00% | 0.3376% | 4 | 0.00% |
| | Best Net. isMAPE | 0.003604 | 12 | 310.94% | 0.5059% | 12 | 49.85% |
| | Best Net. isMSE | 0.001097 | 8 | 25.03% | 0.3603% | 8 | 6.72% |
| | GA2 | 0.000971 | 6 | 10.71% | 0.3471% | 6 | 2.81% |
| | GA4 | 0.000893 | 5 | 1.81% | 0.3287% | 2 | -2.64% |
| | GA8 | 0.000885 | 4 | 0.90% | 0.3342% | 3 | -1.01% |
| | Naive | 0.001501 | 11 | 71.14% | 0.4320% | 11 | 27.97% |
| | SARIMA | 0.001269 | 10 | 44.68% | 0.3939% | 10 | 16.68% |
| | Single seasonal | 0.001029 | 7 | 17.32% | 0.3512% | 7 | 4.03% |

$\%\Delta = 100(M_{model} - M_{Avg})/M_{Avg}$ with $M_i$ being the metric for model $i$.
Negative % values indicate improvement over the average.

(Continued) Forecasting performance. Synthetic-1S series.

| Forecast horizon | Model | MSE | Rank | %Δ wrt Avg. | MAPE | | %Δ wrt Avg. |
|---|---|---|---|---|---|---|---|
| h=5 | CB2 | 0.001251 | 9 | -0.11% | 0.3874% | 9 | -1.66% |
| | CB4 | 0.001233 | 8 | -1.52% | 0.3863% | 8 | -1.93% |
| | CB8 | 0.001103 | 3 | -11.92% | 0.3634% | 3 | -7.73% |
| | Avg. Net. | 0.001252 | 10 | 0.00% | 0.3939% | 10 | 0.00% |
| | Best Net. isMAPE | 0.001157 | 5 | -7.56% | 0.3703% | 4 | -6.00% |
| | Best Net. isMSE | 0.001057 | 1 | -15.62% | 0.3528% | 1 | -10.44% |
| | GA2 | 0.001156 | 4 | -7.68% | 0.3738% | 5 | -5.10% |
| | GA4 | 0.001181 | 6 | -5.68% | 0.3821% | 6 | -2.99% |
| | GA8 | 0.001209 | 7 | -3.44% | 0.3830% | 7 | -2.76% |
| | Naive | 0.001510 | 12 | 20.63% | 0.4331% | 12 | 9.96% |
| | SARIMA | 0.001413 | 11 | 12.85% | 0.4135% | 11 | 4.98% |
| | Single seasonal | 0.001098 | 2 | -12.31% | 0.3618% | 2 | -8.14% |
| h=6 | CB2 | 0.001115 | 4 | -0.40% | 0.3724% | 6 | 0.83% |
| | CB4 | 0.001168 | 8 | 4.36% | 0.3814% | 8 | 3.27% |
| | CB8 | 0.001071 | 1 | -4.33% | 0.3601% | 1 | -2.50% |
| | Avg. Net. | 0.001119 | 5 | 0.00% | 0.3693% | 3 | 0.00% |
| | Best Net. isMAPE | 0.001551 | 11 | 38.59% | 0.4342% | 10 | 17.58% |
| | Best Net. isMSE | 0.001551 | 11 | 38.59% | 0.4342% | 10 | 17.58% |
| | GA2 | 0.001127 | 6 | 0.71% | 0.3726% | 7 | 0.89% |
| | GA4 | 0.001097 | 2 | -1.97% | 0.3653% | 2 | -1.09% |
| | GA8 | 0.001107 | 3 | -1.07% | 0.3695% | 4 | 0.05% |
| | Naive | 0.001526 | 10 | 36.34% | 0.4365% | 12 | 18.20% |
| | SARIMA | 0.001480 | 9 | 32.26% | 0.4272% | 9 | 15.68% |
| | Single seasonal | 0.001157 | 7 | 3.40% | 0.3716% | 5 | 0.61% |
| h=7 | CB2 | 0.001173 | 2 | -0.69% | 0.3846% | 5 | 1.48% |
| | CB4 | 0.001186 | 4 | 0.41% | 0.3839% | 4 | 1.29% |
| | CB8 | 0.001150 | 1 | -2.64% | 0.3745% | 1 | -1.18% |
| | Avg. Net. | 0.001181 | 3 | 0.00% | 0.3790% | 2 | 0.00% |
| | Best Net. isMAPE | 0.001275 | 9 | 7.96% | 0.4016% | 9 | 5.96% |
| | Best Net. isMSE | 0.001275 | 9 | 7.96% | 0.4016% | 9 | 5.96% |
| | GA2 | 0.001230 | 6 | 4.11% | 0.3928% | 7 | 3.65% |
| | GA4 | 0.001231 | 7 | 4.20% | 0.3921% | 6 | 3.46% |
| | GA8 | 0.001245 | 8 | 5.38% | 0.3956% | 8 | 4.39% |
| | Naive | 0.001539 | 12 | 30.29% | 0.4390% | 12 | 15.83% |
| | SARIMA | 0.001486 | 11 | 25.78% | 0.4284% | 11 | 13.04% |
| | Single seasonal | 0.001207 | 5 | 2.17% | 0.3796% | 3 | 0.17% |
| h=8 | CB2 | 0.001469 | 7 | 11.16% | 0.4424% | 8 | 5.87% |
| | CB4 | 0.001469 | 7 | 11.16% | 0.4424% | 8 | 5.87% |
| | CB8 | 0.001307 | 2 | -1.07% | 0.4130% | 2 | -1.15% |
| | Avg. Net. | 0.001321 | 3 | 0.00% | 0.4178% | 3 | 0.00% |
| | Best Net. isMAPE | 0.001457 | 5 | 10.31% | 0.4269% | 4 | 2.16% |
| | Best Net. isMSE | 0.001457 | 5 | 10.31% | 0.4269% | 4 | 2.16% |
| | GA2 | 0.001499 | 11 | 13.46% | 0.4457% | 12 | 6.67% |
| | GA4 | 0.001480 | 9 | 12.02% | 0.4440% | 11 | 6.26% |
| | GA8 | 0.001441 | 4 | 9.07% | 0.4379% | 7 | 4.80% |
| | Naive | 0.001556 | 12 | 17.75% | 0.4436% | 10 | 6.18% |
| | SARIMA | 0.001487 | 10 | 12.55% | 0.4320% | 6 | 3.39% |
| | Single seasonal | 0.001246 | 1 | -5.69% | 0.3875% | 1 | -7.26% |

(Continued) Forecasting performance. Synthetic-1S series.

| Forecast horizon | Model | MSE | Rank | %Δ wrt Avg. | MAPE | | %Δ wrt Avg. |
|---|---|---|---|---|---|---|---|
| h=9 | CB2 | 0.001559 | 6 | 1.00% | 0.4442% | 4 | -1.00% |
| | CB4 | 0.001598 | 10 | 3.53% | 0.4534% | 10 | 1.06% |
| | CB8 | 0.001479 | 2 | -4.14% | 0.4419% | 3 | -1.52% |
| | Avg. Net. | 0.001543 | 4 | 0.00% | 0.4487% | 7 | 0.00% |
| | Best Net. isMAPE | 0.001866 | 11 | 20.94% | 0.4812% | 11 | 7.24% |
| | Best Net. isMSE | 0.001866 | 11 | 20.94% | 0.4812% | 11 | 7.24% |
| | GA2 | 0.001551 | 5 | 0.51% | 0.4468% | 6 | -0.42% |
| | GA4 | 0.001591 | 9 | 3.10% | 0.4505% | 8 | 0.41% |
| | GA8 | 0.001580 | 8 | 2.39% | 0.4511% | 9 | 0.54% |
| | Naive | 0.001566 | 7 | 1.48% | 0.4448% | 5 | -0.86% |
| | SARIMA | 0.001501 | 3 | -2.73% | 0.4339% | 2 | -3.29% |
| | Single seasonal | 0.001271 | 1 | -17.63% | 0.3918% | 1 | -12.67% |
| h=10 | CB2 | 0.001401 | 8 | 4.89% | 0.4123% | 5 | 1.32% |
| | CB4 | 0.001261 | 1 | -5.62% | 0.3978% | 1 | -2.25% |
| | CB8 | 0.001311 | 3 | -1.88% | 0.4086% | 4 | 0.42% |
| | Avg. Net. | 0.001336 | 4 | 0.00% | 0.4069% | 3 | 0.00% |
| | Best Net. isMAPE | 0.001613 | 12 | 20.75% | 0.4638% | 12 | 13.98% |
| | Best Net. isMSE | 0.001486 | 9 | 11.24% | 0.4394% | 9 | 7.99% |
| | GA2 | 0.001386 | 6 | 3.77% | 0.4146% | 8 | 1.88% |
| | GA4 | 0.001388 | 7 | 3.92% | 0.4142% | 7 | 1.79% |
| | GA8 | 0.001385 | 5 | 3.70% | 0.4136% | 6 | 1.64% |
| | Naive | 0.001580 | 11 | 18.31% | 0.4472% | 11 | 9.90% |
| | SARIMA | 0.001531 | 10 | 14.63% | 0.4412% | 10 | 8.42% |
| | Single seasonal | 0.001302 | 2 | -2.52% | 0.3984% | 2 | -2.11% |
| h=11 | CB2 | 0.001341 | 2 | -4.26% | 0.4128% | 2 | -2.12% |
| | CB4 | 0.001465 | 8 | 4.58% | 0.4326% | 8 | 2.57% |
| | CB8 | 0.001408 | 4 | 0.54% | 0.4231% | 6 | 0.32% |
| | Avg. Net. | 0.001400 | 3 | 0.00% | 0.4218% | 3 | 0.00% |
| | Best Net. isMAPE | 0.001502 | 9 | 7.23% | 0.4456% | 9 | 5.65% |
| | Best Net. isMSE | 0.001668 | 12 | 19.14% | 0.4631% | 12 | 9.79% |
| | GA2 | 0.001424 | 6 | 1.69% | 0.4227% | 5 | 0.22% |
| | GA4 | 0.001416 | 5 | 1.11% | 0.4226% | 4 | 0.20% |
| | GA8 | 0.001428 | 7 | 1.97% | 0.4262% | 7 | 1.05% |
| | Naive | 0.001587 | 11 | 13.32% | 0.4474% | 11 | 6.07% |
| | SARIMA | 0.001560 | 10 | 11.40% | 0.4467% | 10 | 5.91% |
| | Single seasonal | 0.001326 | 1 | -5.31% | 0.4039% | 1 | -4.24% |
| h=12 | CB2 | 0.001455 | 3 | -0.96% | 0.4361% | 3 | -0.56% |
| | CB4 | 0.001582 | 11 | 7.75% | 0.4549% | 12 | 3.71% |
| | CB8 | 0.001459 | 4 | -0.64% | 0.4366% | 4 | -0.46% |
| | Avg. Net. | 0.001469 | 6 | 0.00% | 0.4386% | 7 | 0.00% |
| | Best Net. isMAPE | 0.001543 | 8 | 5.06% | 0.4549% | 10 | 3.70% |
| | Best Net. isMSE | 0.001543 | 8 | 5.06% | 0.4549% | 10 | 3.70% |
| | GA2 | 0.001473 | 7 | 0.30% | 0.4382% | 6 | -0.09% |
| | GA4 | 0.001467 | 5 | -0.11% | 0.4375% | 5 | -0.25% |
| | GA8 | 0.001452 | 2 | -1.13% | 0.4355% | 2 | -0.71% |
| | Naive | 0.001602 | 12 | 9.08% | 0.4500% | 9 | 2.59% |
| | SARIMA | 0.001573 | 10 | 7.11% | 0.4490% | 8 | 2.37% |
| | Single seasonal | 0.001348 | 1 | -8.21% | 0.4080% | 1 | -6.99% |

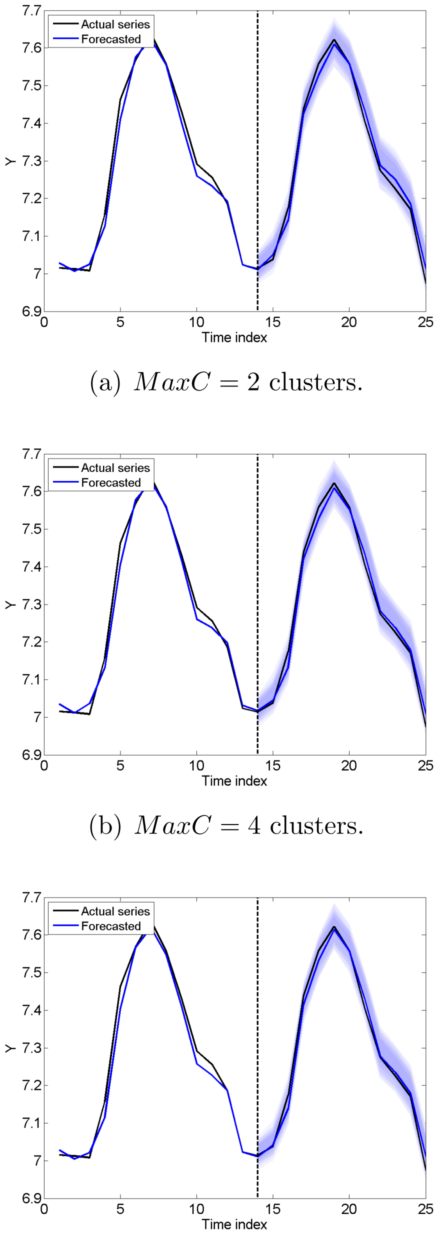
The $\Phi$ coefficients reveal a generally homogeneous contribution of clusters to the final forecast and, in general, model parameters seem to be stable. Cluster validity indexes (Table 4.11) suggest that clusters are not very different and, consequently, diversity cannot be exploited.

Finally, the assessment of uncertainty in forecasts (Figure 4.14) reveals an homogeneous behaviour between the three CB configurations, which reflects the stability of the models and their similar performance (although GA combinations are even more stable). These results could be explained by the use of separate models for every horizon and the regularity in the time series. Such conditions facilitate the specialisation of an ensemble in a specific horizon and potentially produces lower uncertainty than in the case of a single model used for all horizons.

Table 4.10: Coefficients for structural combination of NN for Synthetic-1S series.

| h | $\Phi$ | 0.2514 | 0.2362 | 0.2621 | 0.2625 | | |
|---|---|---|---|---|---|---|---|
| 1 | $\alpha_1$ | -0.0132 | 0.3187 | -0.0081 | 0.2678 | 0.0086 | 0.4021 |
| | $\alpha_2$ | 0.0070 | -0.0492 | 0.2900 | 0.3174 | 0.2973 | 0.1317 |
| | $\alpha_3$ | -0.0341 | 0.0488 | 0.1399 | 0.2044 | 0.3155 | 0.2828 |
| | $\alpha_4$ | 0.0013 | 0.0683 | 0.2181 | 0.1440 | 0.3027 | 0.2535 |
| | $\Phi$ | 0.2514 | 0.2362 | 0.2621 | 0.2625 | | |
| 2 | $\alpha_1$ | -0.0235 | 0.1917 | 0.1881 | 0.3554 | 0.1976 | 0.0643 |
| | $\alpha_2$ | 0.0108 | 0.3497 | 0.0244 | 0.1033 | 0.2751 | 0.2405 |
| | $\alpha_3$ | -0.0575 | 0.0396 | 0.4646 | 0.0748 | 0.3498 | 0.0737 |
| | $\alpha_4$ | -0.0295 | 0.2402 | 0.0319 | 0.3258 | 0.4015 | |
| | $\Phi$ | 0.2734 | 0.2414 | 0.2273 | 0.2633 | | |
| 6 | $\alpha_1$ | -0.0010 | 0.0201 | 0.1802 | 0.1719 | 0.2595 | 0.3531 |
| | $\alpha_2$ | 0.1080 | -0.1279 | 0.1691 | 0.2198 | 0.4353 | 0.2731 |
| | $\alpha_3$ | -0.0737 | 0.1359 | 0.3192 | 0.2606 | 0.3130 | -0.0337 |
| | $\alpha_4$ | -0.0037 | 0.0081 | 0.0804 | 0.2320 | 0.4401 | 0.2239 |
| | $\alpha_5$ | -0.0015 | -0.0664 | 0.0929 | 0.2691 | 0.2574 | 0.4318 |
| | $\alpha_6$ | -0.0850 | 0.0817 | 0.1470 | 0.3135 | 0.1832 | 0.2703 |
| | $\Phi$ | 0.1635 | 0.1732 | 0.1661 | 0.1729 | 0.1683 | 0.1720 |
| 12 | $\alpha_1$ | 0.0713 | 0.4002 | 0.0513 | 0.3758 | 0.2505 | -0.1017 |
| | $\alpha_2$ | 0.0629 | -0.0002 | 0.1829 | 0.6662 | -0.0441 | 0.1728 |
| | $\alpha_3$ | -0.0459 | 0.0152 | 0.8243 | 0.1587 | | |
| | $\alpha_4$ | 0.0354 | 0.1537 | -0.1368 | 0.0287 | 0.5972 | 0.3384 |
| | $\alpha_5$ | -0.1302 | 0.7279 | -0.2851 | 0.0700 | 0.4688 | 0.0224 |
| | $\alpha_6$ | 0.0140 | 0.2526 | 0.3597 | 0.3771 | | |
| | $\Phi$ | 0.1693 | 0.1691 | 0.1623 | 0.1693 | 0.1697 | 0.1726 |

$MaxC = 8$ is the maximum number of clusters. $h$ denotes the forecast horizon, $\alpha_i$ are the coefficients applied to point-forecasts from models in cluster $i$ and $\Phi$ are the weights applied to the outputs from clusters.



(a) $MaxC = 2$ clusters.

(b) $MaxC = 4$ clusters.

(c) $MaxC = 8$ clusters.

Figure 4.14: Forecast intervals for Synthetic-1S time series.

The graphs cover the period for $t - 12 \leq t \leq t + H$ where $t$ is the last observation of the in-sample period and $H = 12$ is the number of forecast horizons. The shades, from lighter to darker, correspond to $\alpha$ levels 0.95, 0.90, 0.85, 0.80, 0.75 and 0.60.

Table 4.11: Cluster validity indexes. Series: Synthetic-1S.

| | Maximum number of clusters: 2 | | | |
|---|---|---|---|---|
| $h$ | PC | MPC | MDO | Final num. clusters |
| 1 | 1.0000 | NA | 1.0000 | 1 |
| 2 | 1.0000 | NA | 1.0000 | 1 |
| 3 | 1.0000 | NA | 1.0000 | 1 |
| 4 | 1.0000 | NA | 1.0000 | 1 |
| 5 | 0.5026 | 0.0052 | 0.1000 | 2 |
| 6 | 1.0000 | NA | 1.0000 | 1 |
| 7 | 1.0000 | NA | 1.0000 | 1 |
| 8 | 1.0000 | NA | 1.0000 | 1 |
| 9 | 0.5013 | 0.0027 | 0.1000 | 2 |
| 10 | 1.0000 | NA | 1.0000 | 1 |
| 11 | 0.5016 | 0.0032 | 0.1000 | 2 |
| 12 | 1.0000 | NA | 1.0000 | 1 |

| | Maximum number of clusters: 4 | | | |
|---|---|---|---|---|
| $h$ | PC | MPC | MDO | Final num. clusters |
| 1 | 0.3360 | 0.0040 | 0.0000 | 3 |
| 2 | 0.2511 | 0.0015 | 0.0000 | 4 |
| 3 | 0.3345 | 0.0018 | 0.0000 | 3 |
| 4 | 0.3593 | 0.0390 | 0.0000 | 3 |
| 5 | 1.0000 | NA | 1.0000 | 1 |
| 6 | 0.5012 | 0.0024 | 0.1000 | 2 |
| 7 | 0.3342 | 0.0012 | 0.0000 | 3 |
| 8 | 1.0000 | NA | 1.0000 | 1 |
| 9 | 0.3349 | 0.0023 | 0.0000 | 3 |
| 10 | 0.2657 | 0.0209 | 0.0000 | 4 |
| 11 | 0.2502 | 0.0003 | 0.0000 | 4 |
| 12 | 0.5021 | 0.0042 | 0.1000 | 2 |

| | Maximum number of clusters: 8 | | | |
|---|---|---|---|---|
| $h$ | PC | MPC | MDO | Final num. clusters |
| 1 | 0.2505 | 0.0007 | 0.0000 | 4 |
| 2 | 0.2515 | 0.0020 | 0.0000 | 4 |
| 3 | 0.1254 | 0.0004 | 0.0000 | 8 |
| 4 | 0.1736 | 0.0083 | 0.0000 | 6 |
| 5 | 0.2546 | 0.0061 | 0.0000 | 4 |
| 6 | 0.1668 | 0.0002 | 0.0000 | 6 |
| 7 | 0.1668 | 0.0002 | 0.0000 | 6 |
| 8 | 0.2013 | 0.0017 | 0.0000 | 5 |
| 9 | 0.2010 | 0.0013 | 0.0000 | 5 |
| 10 | 0.1465 | 0.0043 | 0.0000 | 7 |
| 11 | 0.1668 | 0.0001 | 0.0000 | 6 |
| 12 | 0.1669 | 0.0003 | 0.0000 | 6 |

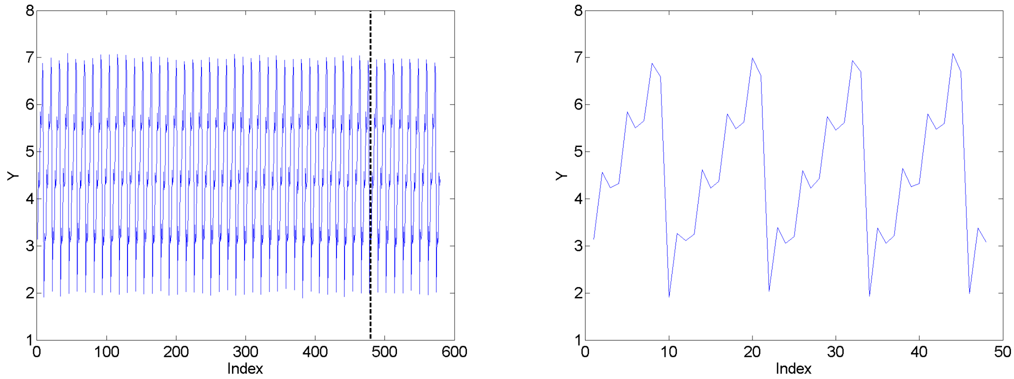
$h$ denotes the forecast horizon, PC denotes the Partition Coefficient, MPC denotes the Modified Partition Coefficient, and MDO denotes the Membership Degree Optimum. Values closer to 1 are preferable.

### 4.5.3 Synthetic-2S Series

The generating process for this series is:

$$\begin{aligned}
y_t &= l_{t-1} + d_{t-s_1} + w_{t-s_2} + \phi(y_{t-1} - (l_{t-2} + d_{t-s_1-1} + w_{t-s_2-1})) + \varepsilon_t \qquad (4.23)\\
l_t &= \lambda(y_t - d_{t-s_1} - w_{t-s_2}) + (1-\lambda)l_{t-1}\\
d_t &= \delta(y_t - l_{t-1} - w_{t-s_2}) + (1-\delta)d_{t-s_1}\\
w_t &= \omega(y_t - l_{t-1} - d_{t-s_1}) + (1-\omega)w_{t-s_2}
\end{aligned}$$

where $y_t$ is the simulated series, $l_t$ denotes the smoothed level, $w_t$ denotes the long cycle seasonal index and $d_t$ denotes the short cycle seasonal index. Parameters are $\lambda = 0.2$; $\delta = 0.13$; $\omega = 0.3$; $\phi = 0.5$; $s_1 = 3$; $s_2 = 12$. It is displayed in Figure 4.15.



(a) Series.

(b) Subset.

Figure 4.15: Synthetic-2S series.

The selected models for the NN ensemble, based on the corresponding sensitivity analysis made in Chapter 3 are listed in Table 4.12. Data partitioning is the same as for the previous series.

The double seasonal benchmark model (denoted as AddDblSeasonal) follows the model in Equation 4.23 with parameters $\lambda = 0.0084$; $\delta = 0.0037$; $\omega = 0.0706$; $\phi = 0.0771$;

Table 4.12: Selected NN models for Synthetic-2S series.

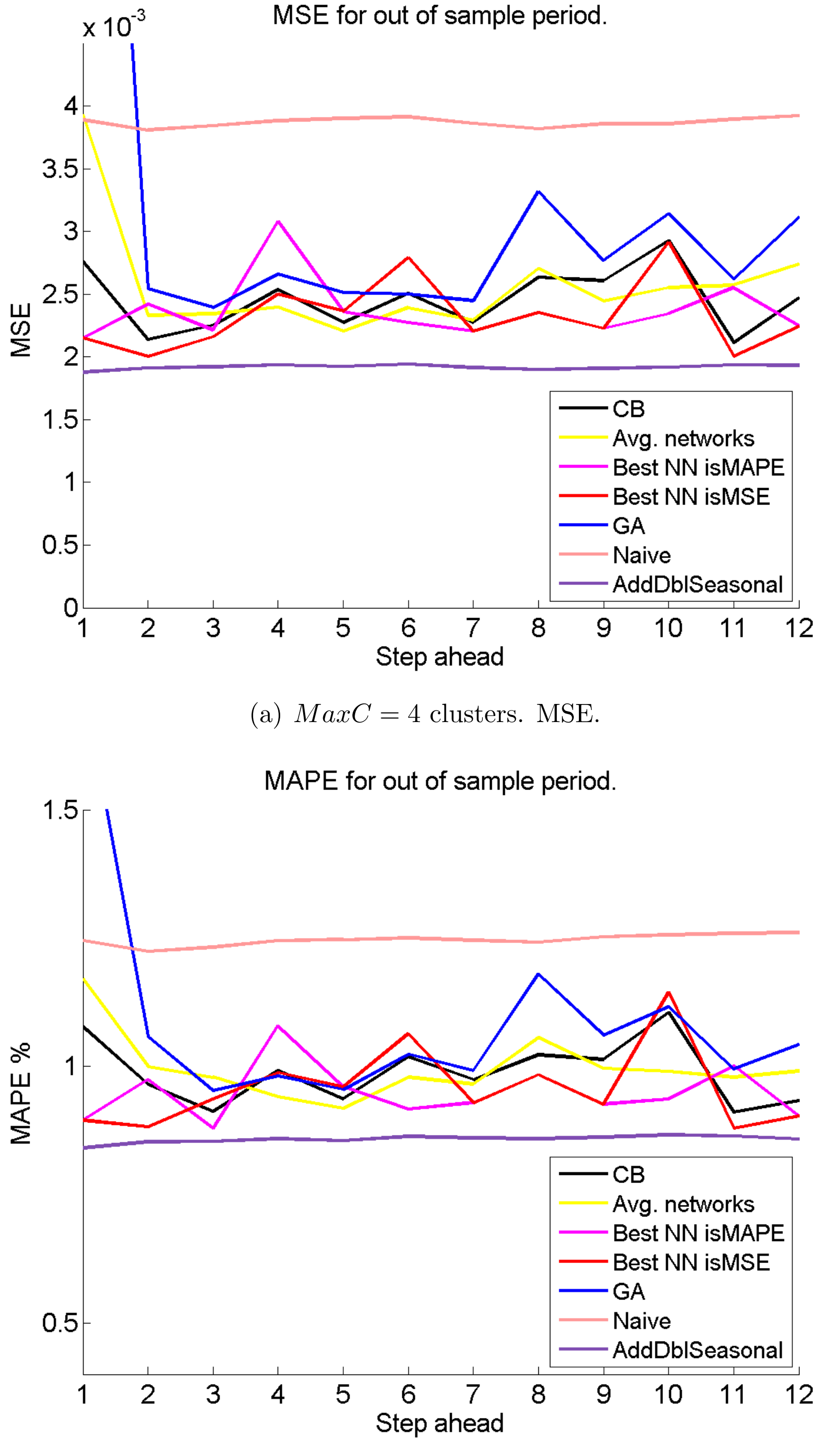
| h | NI | NU | h | NI | NU |
|---|---|---|---|---|---|
| 1 | 2 | 8 | 7 | 3 | 7 |
| 2 | 3 | 6 | 8 | 3 | 6 |
| 3 | 3 | 6 | 9 | 4 | 6 |
| 4 | 3 | 7 | 10 | 4 | 6 |
| 5 | 3 | 7 | 11 | 3 | 8 |
| 6 | 3 | 6 | 12 | 2 | 7 |

Figure 4.16 shows the MSE and MAPE error metrics for the out-of-sample period (only CB models with a maximum of 4 clusters are included and the full set of graphs are provided in Figure C.5 from Appendix C). Performance has a similar order of magnitude over the different forecasts horizons which might be due to the the regularity of the time series.

The AddDblSeasonal benchmark outperforms all models and combinations in all forecast horizons, which tallies with the fact that it comes from the family of models from which the time series was generated. Additionally, the detail provided in Figure 4.17 reveals that CB combinations generally have a better and more stable performance than GA. This suggests that GA combination approach might not be suitable for more complex regular time series behaviour, in this case, seasonality.

Table 4.13 summarises forecasting performance. At least one of the three cluster based models outperforms the average of NN in almost all forecast horizons, with CB8 being the model that most consistently improves over such benchmark. In general, the performance of structural combinations is irregular.

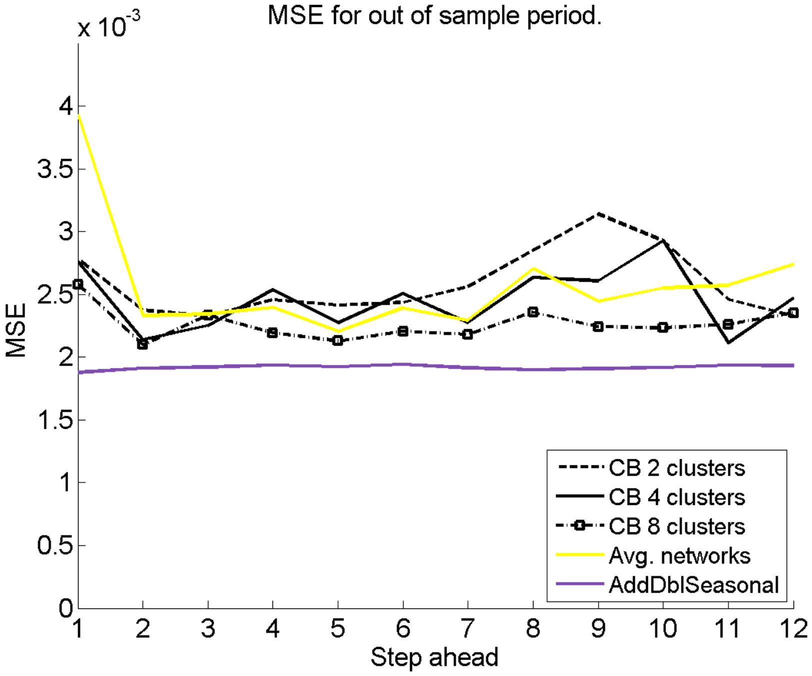
Figure 4.18 illustrates how the MSE has a mixed pattern with respect to the number of clusters, as it can be observed that performance curves have different shapes: convex, concave or almost straight. A similar patter was observed for MAPE metric (graphs are omitted). This could be due to the regularity of the series and the use of the direct forecast approach, which uses different NNs (and



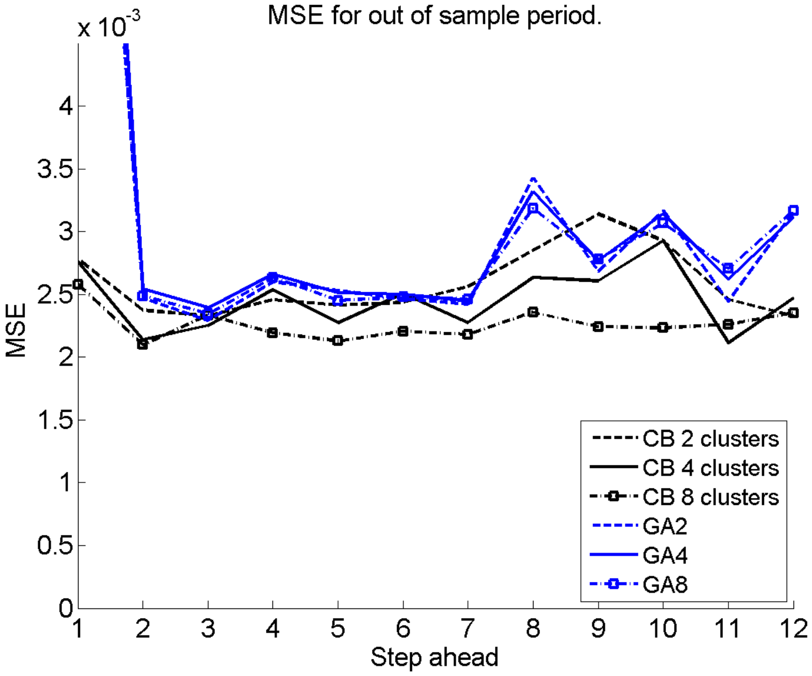
(a) $MaxC = 4$ clusters. MSE.

(b) $MaxC = 4$ clusters. MAPE.

Figure 4.16: Out-of-sample MSE and MAPE for Synthetic-2S series.

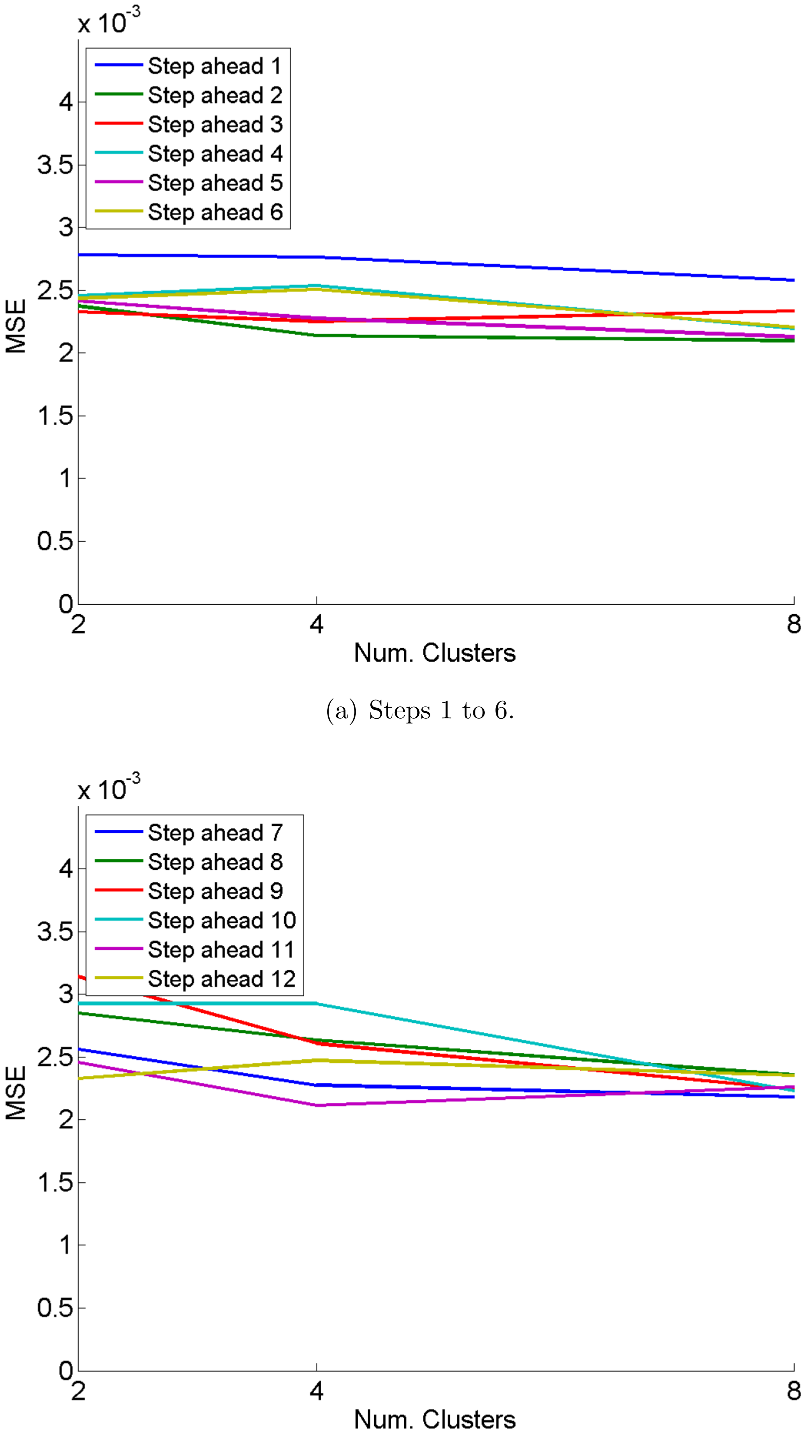


(a) Comparison with Avg. and additive Dbl. seasonal.



(b) Comparison with GA.

Figure 4.17: Out-of-sample MSE for Synthetic-2S series.



(a) Steps 1 to 6.

(b) Steps 7 to 12.

Figure 4.18: Out-of-sample MSE vs. number of clusters. Synthetic-2S series.

different ensembles) for each horizon.

Details of the best performing CB model configuration for some forecast horizons are given in Table 4.14. The weights applied to the forecasts given by different clusters, $\Phi$, are very similar for most model configurations and forecast horizons. This suggests homogeneity in forecasts obtained from clusters. However, the moderate percentage differences in performance observed with respect to the average in Table 4.13, for most horizons, do not suggest that the combination can be assimilated to an average. The $\alpha_j$ coefficients signal a generally stable, but diverse weighting of forecasts within clusters. In terms of cluster configuration (Table 4.15) there are CB4 models for horizons 1, 2 and 9 that have a clearer separation when compared to other models, but in general, the validity indexes do not suggest a strong separation between clusters.

The Ljung-Box tests showed that for most forecast horizons ($h \geq 3$) almost all models exhibit forecast errors that are serially correlated. However, Jarque-Bera tests and Lilliefors tests revealed that most models, including CB combinations, exhibit normally distributed errors for almost all forecast horizons. Finally, the assessment of uncertainty in forecast exhibit very narrow bands (Figure 4.19) which stems from the relatively good performance of NN ensembles fitted to specific forecast horizons for a very regular time series.

Table 4.13: Forecasting performance. Synthetic-2S series.

| Forecast horizon | Model | MSE | Rank | %Δ wrt Avg. | MAPE | | %Δ wrt Avg. |
|---|---|---|---|---|---|---|---|
| h=1 | CB2 | 0.002783 | 6 | -29.21% | 1.0241% | 5 | -12.49% |
| | CB4 | 0.002762 | 5 | -29.74% | 1.0775% | 6 | -7.92% |
| | CB8 | 0.002579 | 4 | -34.41% | 1.0131% | 4 | -13.43% |
| | Avg. Net. | 0.003931 | 8 | 0.00% | 1.1702% | 7 | 0.00% |
| | Best Net. isMAPE | 0.002149 | 2 | -45.34% | 0.8948% | 2 | -23.54% |
| | Best Net. isMSE | 0.002149 | 2 | -45.34% | 0.8948% | 2 | -23.54% |
| | GA2 | 0.009340 | 9 | 137.57% | 1.7058% | 9 | 45.77% |
| | GA4 | 0.010450 | 11 | 165.81% | 1.7613% | 11 | 50.51% |
| | GA8 | 0.009833 | 10 | 150.11% | 1.7088% | 10 | 46.03% |
| | Naive | 0.003890 | 7 | -1.05% | 1.2456% | 8 | 6.44% |
| | AddDblSeasonal | 0.001876 | 1 | -52.28% | 0.8408% | 1 | -28.15% |
| h=2 | CB2 | 0.002374 | 6 | 1.80% | 1.0110% | 7 | 1.13% |
| | CB4 | 0.002137 | 4 | -8.36% | 0.9650% | 4 | -3.47% |
| | CB8 | 0.002099 | 3 | -10.01% | 0.9355% | 3 | -6.42% |
| | Avg. Net. | 0.002332 | 5 | 0.00% | 0.9997% | 6 | 0.00% |
| | Best Net. isMAPE | 0.002421 | 7 | 3.83% | 0.9741% | 5 | -2.56% |
| | Best Net. isMSE | 0.002002 | 2 | -14.14% | 0.8825% | 2 | -11.72% |
| | GA2 | 0.002466 | 8 | 5.75% | 1.0410% | 9 | 4.14% |
| | GA4 | 0.002544 | 10 | 9.09% | 1.0585% | 10 | 5.89% |
| | GA8 | 0.002489 | 9 | 6.73% | 1.0407% | 8 | 4.11% |
| | Naive | 0.003809 | 11 | 63.32% | 1.2242% | 11 | 22.46% |
| | AddDblSeasonal | 0.001912 | 1 | -18.01% | 0.8530% | 1 | -14.67% |
| h=3 | CB2 | 0.002329 | 6 | -0.61% | 0.9306% | 4 | -4.94% |
| | CB4 | 0.002251 | 4 | -3.95% | 0.9124% | 3 | -6.80% |
| | CB8 | 0.002335 | 7 | -0.37% | 0.9680% | 9 | -1.12% |
| | Avg. Net. | 0.002343 | 9 | 0.00% | 0.9790% | 10 | 0.00% |
| | Best Net. isMAPE | 0.002215 | 3 | -5.47% | 0.8791% | 2 | -10.20% |
| | Best Net. isMSE | 0.002158 | 2 | -7.89% | 0.9367% | 6 | -4.32% |
| | GA2 | 0.002292 | 5 | -2.19% | 0.9360% | 5 | -4.39% |
| | GA4 | 0.002393 | 10 | 2.12% | 0.9527% | 7 | -2.68% |
| | GA8 | 0.002343 | 8 | -0.01% | 0.9660% | 8 | -1.32% |
| | Naive | 0.003843 | 11 | 64.01% | 1.2321% | 11 | 25.86% |
| | AddDblSeasonal | 0.001920 | 1 | -18.06% | 0.8537% | 1 | -12.80% |
| h=4 | CB2 | 0.002456 | 4 | 2.50% | 0.9658% | 5 | 2.66% |
| | CB4 | 0.002537 | 6 | 5.84% | 0.9923% | 9 | 5.48% |
| | CB8 | 0.002194 | 2 | -8.43% | 0.9048% | 2 | -3.83% |
| | Avg. Net. | 0.002397 | 3 | 0.00% | 0.9408% | 3 | 0.00% |
| | Best Net. isMAPE | 0.003080 | 10 | 28.50% | 1.0789% | 10 | 14.68% |
| | Best Net. isMSE | 0.002501 | 5 | 4.37% | 0.9876% | 8 | 4.97% |
| | GA2 | 0.002600 | 7 | 8.49% | 0.9522% | 4 | 1.21% |
| | GA4 | 0.002660 | 9 | 11.00% | 0.9812% | 7 | 4.30% |
| | GA8 | 0.002633 | 8 | 9.87% | 0.9733% | 6 | 3.46% |
| | Naive | 0.003883 | 11 | 62.02% | 1.2447% | 11 | 32.30% |
| | AddDblSeasonal | 0.001937 | 1 | -19.17% | 0.8587% | 1 | -8.72% |

$\%\Delta = 100(M_{model} - M_{Avg})/M_{Avg}$ with $M_i$ being the metric for model $i$.

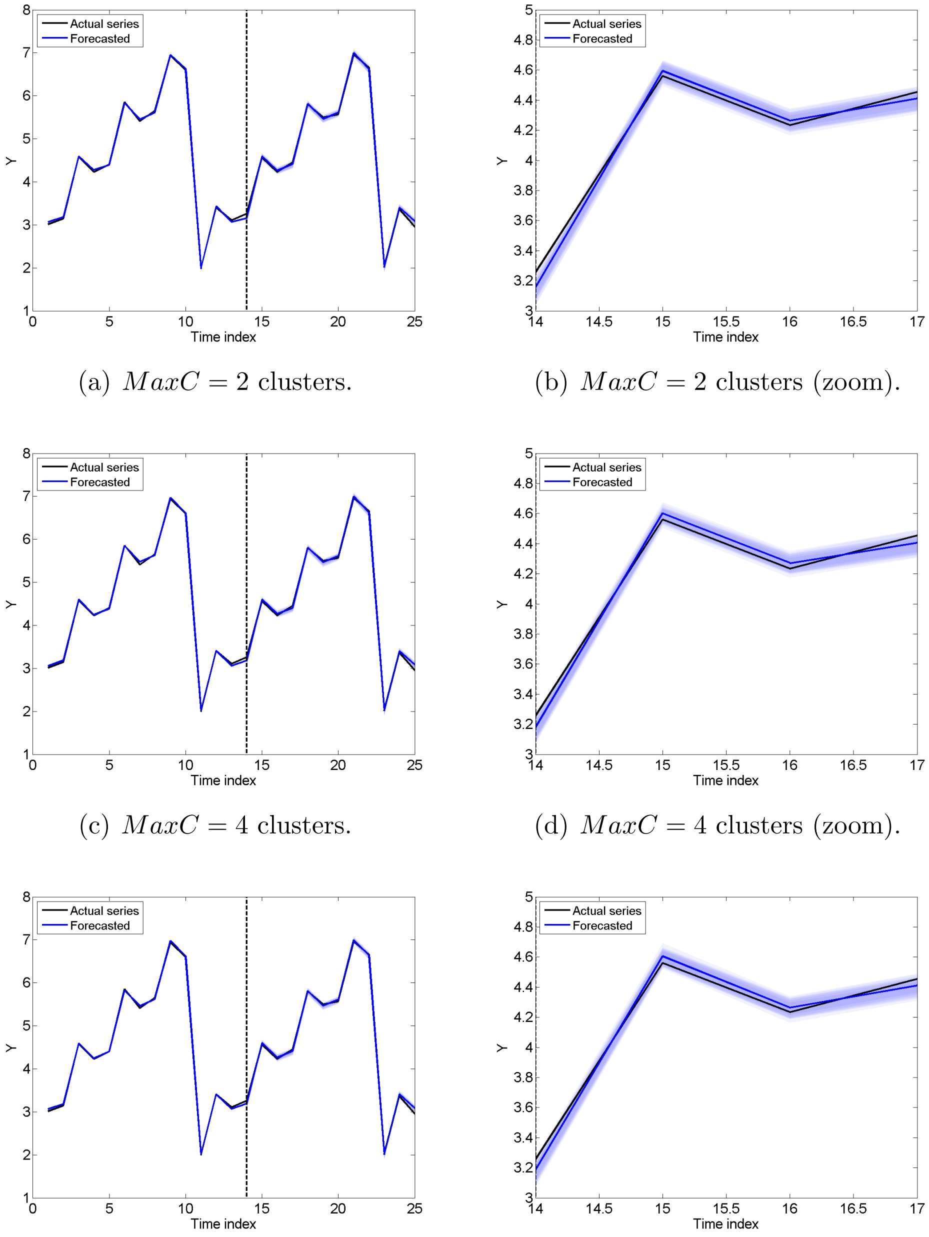
Negative % values indicate improvement over the average.

(Continued) Forecasting performance. Synthetic-2S series.

| Forecast horizon | Model | MSE | Rank | %Δ wrt Avg. | MAPE | | %Δ wrt Avg. |
|---|---|---|---|---|---|---|---|
| h=5 | CB2 | 0.002414 | 7 | 9.49% | 0.9519% | 6 | 3.58% |
| | CB4 | 0.002276 | 4 | 3.22% | 0.9366% | 4 | 1.92% |
| | CB8 | 0.002129 | 2 | -3.45% | 0.9177% | 2 | -0.13% |
| | Avg. Net. | 0.002205 | 3 | 0.00% | 0.9189% | 3 | 0.00% |
| | Best Net. isMAPE | 0.002362 | 5 | 7.13% | 0.9603% | 9 | 4.50% |
| | Best Net. isMSE | 0.002362 | 5 | 7.13% | 0.9603% | 9 | 4.50% |
| | GA2 | 0.002531 | 10 | 14.80% | 0.9546% | 7 | 3.88% |
| | GA4 | 0.002511 | 9 | 13.89% | 0.9549% | 8 | 3.91% |
| | GA8 | 0.002450 | 8 | 11.13% | 0.9463% | 5 | 2.98% |
| | Naive | 0.003901 | 11 | 76.92% | 1.2467% | 11 | 35.67% |
| | AddDblSeasonal | 0.001923 | 1 | -12.78% | 0.8550% | 1 | -6.96% |
| h=6 | CB2 | 0.002436 | 5 | 1.84% | 0.9619% | 4 | -1.75% |
| | CB4 | 0.002507 | 9 | 4.82% | 1.0182% | 8 | 4.01% |
| | CB8 | 0.002207 | 2 | -7.75% | 0.9438% | 3 | -3.59% |
| | Avg. Net. | 0.002392 | 4 | 0.00% | 0.9790% | 5 | 0.00% |
| | Best Net. isMAPE | 0.002274 | 3 | -4.94% | 0.9163% | 2 | -6.40% |
| | Best Net. isMSE | 0.002791 | 10 | 16.70% | 1.0636% | 10 | 8.64% |
| | GA2 | 0.002467 | 6 | 3.14% | 1.0172% | 6 | 3.90% |
| | GA4 | 0.002500 | 8 | 4.52% | 1.0238% | 9 | 4.58% |
| | GA8 | 0.002479 | 7 | 3.65% | 1.0179% | 7 | 3.98% |
| | Naive | 0.003913 | 11 | 63.59% | 1.2503% | 11 | 27.72% |
| | AddDblSeasonal | 0.001941 | 1 | -18.85% | 0.8633% | 1 | -11.81% |
| h=7 | CB2 | 0.002561 | 10 | 11.60% | 1.0107% | 10 | 4.71% |
| | CB4 | 0.002275 | 5 | -0.85% | 0.9741% | 6 | 0.92% |
| | CB8 | 0.002181 | 2 | -4.94% | 0.9473% | 4 | -1.86% |
| | Avg. Net. | 0.002295 | 6 | 0.00% | 0.9652% | 5 | 0.00% |
| | Best Net. isMAPE | 0.002207 | 3 | -3.82% | 0.9291% | 2 | -3.75% |
| | Best Net. isMSE | 0.002207 | 3 | -3.82% | 0.9291% | 2 | -3.75% |
| | GA2 | 0.002420 | 7 | 5.47% | 0.9819% | 7 | 1.73% |
| | GA4 | 0.002449 | 8 | 6.73% | 0.9918% | 8 | 2.75% |
| | GA8 | 0.002460 | 9 | 7.21% | 0.9990% | 9 | 3.50% |
| | Naive | 0.003861 | 11 | 68.27% | 1.2455% | 11 | 29.04% |
| | AddDblSeasonal | 0.001913 | 1 | -16.63% | 0.8601% | 1 | -10.89% |
| h=8 | CB2 | 0.002851 | 7 | 5.32% | 1.0386% | 6 | -1.62% |
| | CB4 | 0.002636 | 5 | -2.63% | 1.0224% | 5 | -3.15% |
| | CB8 | 0.002355 | 4 | -13.00% | 0.9960% | 4 | -5.66% |
| | Avg. Net. | 0.002707 | 6 | 0.00% | 1.0557% | 7 | 0.00% |
| | Best Net. isMAPE | 0.002352 | 2 | -13.09% | 0.9837% | 2 | -6.82% |
| | Best Net. isMSE | 0.002352 | 2 | -13.09% | 0.9837% | 2 | -6.82% |
| | GA2 | 0.003427 | 10 | 26.61% | 1.1917% | 10 | 12.88% |
| | GA4 | 0.003319 | 9 | 22.62% | 1.1801% | 9 | 11.78% |
| | GA8 | 0.003183 | 8 | 17.60% | 1.1456% | 8 | 8.52% |
| | Naive | 0.003819 | 11 | 41.09% | 1.2419% | 11 | 17.63% |
| | AddDblSeasonal | 0.001900 | 1 | -29.80% | 0.8587% | 1 | -18.66% |

(Continued) Forecasting performance. Synthetic-2S series.

| Forecast horizon | Model | MSE | Rank | %Δ wrt Avg. | MAPE | | %Δ wrt Avg. |
|---|---|---|---|---|---|---|---|
| h=9 | CB2 | 0.003141 | 10 | 28.41% | 1.1281% | 10 | 13.29% |
| | CB4 | 0.002606 | 6 | 6.57% | 1.0136% | 6 | 1.79% |
| | CB8 | 0.002242 | 4 | -8.32% | 0.9630% | 4 | -3.29% |
| | Avg. Net. | 0.002446 | 5 | 0.00% | 0.9958% | 5 | 0.00% |
| | Best Net. isMAPE | 0.002226 | 2 | -8.97% | 0.9262% | 2 | -6.99% |
| | Best Net. isMSE | 0.002226 | 2 | -8.97% | 0.9262% | 2 | -6.99% |
| | GA2 | 0.002684 | 7 | 9.75% | 1.0502% | 7 | 5.47% |
| | GA4 | 0.002767 | 8 | 13.14% | 1.0607% | 9 | 6.52% |
| | GA8 | 0.002780 | 9 | 13.67% | 1.0564% | 8 | 6.09% |
| | Naive | 0.003857 | 11 | 57.71% | 1.2526% | 11 | 25.80% |
| | AddDblSeasonal | 0.001907 | 1 | -22.02% | 0.8616% | 1 | -13.47% |
| h=10 | CB2 | 0.002927 | 7 | 14.71% | 1.0604% | 5 | 7.13% |
| | CB4 | 0.002925 | 6 | 14.66% | 1.1051% | 7 | 11.65% |
| | CB8 | 0.002233 | 2 | -12.49% | 0.9292% | 2 | -6.12% |
| | Avg. Net. | 0.002551 | 4 | 0.00% | 0.9898% | 4 | 0.00% |
| | Best Net. isMAPE | 0.002344 | 3 | -8.11% | 0.9358% | 3 | -5.46% |
| | Best Net. isMSE | 0.002914 | 5 | 14.21% | 1.1441% | 10 | 15.59% |
| | GA2 | 0.003166 | 10 | 24.10% | 1.1065% | 8 | 11.79% |
| | GA4 | 0.003140 | 9 | 23.08% | 1.1167% | 9 | 12.82% |
| | GA8 | 0.003066 | 8 | 20.18% | 1.0940% | 6 | 10.53% |
| | Naive | 0.003857 | 11 | 51.18% | 1.2561% | 11 | 26.91% |
| | AddDblSeasonal | 0.001917 | 1 | -24.86% | 0.8666% | 1 | -12.45% |
| h=11 | CB2 | 0.002461 | 6 | -4.28% | 0.9573% | 6 | -2.23% |
| | CB4 | 0.002113 | 3 | -17.80% | 0.9106% | 3 | -7.01% |
| | CB8 | 0.002261 | 4 | -12.05% | 0.9275% | 4 | -5.28% |
| | Avg. Net. | 0.002571 | 8 | 0.00% | 0.9792% | 7 | 0.00% |
| | Best Net. isMAPE | 0.002552 | 7 | -0.74% | 1.0010% | 9 | 2.23% |
| | Best Net. isMSE | 0.002007 | 2 | -21.94% | 0.8799% | 2 | -10.14% |
| | GA2 | 0.002441 | 5 | -5.05% | 0.9533% | 5 | -2.64% |
| | GA4 | 0.002621 | 9 | 1.96% | 0.9943% | 8 | 1.54% |
| | GA8 | 0.002705 | 10 | 5.22% | 1.0077% | 10 | 2.91% |
| | Naive | 0.003895 | 11 | 51.53% | 1.2593% | 11 | 28.60% |
| | AddDblSeasonal | 0.001935 | 1 | -24.73% | 0.8643% | 1 | -11.73% |
| h=12 | CB2 | 0.002329 | 4 | -14.87% | 0.9251% | 4 | -6.61% |
| | CB4 | 0.002472 | 6 | -9.65% | 0.9332% | 6 | -5.79% |
| | CB8 | 0.002354 | 5 | -13.97% | 0.9303% | 5 | -6.08% |
| | Avg. Net. | 0.002736 | 7 | 0.00% | 0.9906% | 7 | 0.00% |
| | Best Net. isMAPE | 0.002246 | 2 | -17.91% | 0.9028% | 2 | -8.86% |
| | Best Net. isMSE | 0.002246 | 2 | -17.91% | 0.9028% | 2 | -8.86% |
| | GA2 | 0.003151 | 9 | 15.18% | 1.0434% | 9 | 5.33% |
| | GA4 | 0.003117 | 8 | 13.94% | 1.0428% | 8 | 5.27% |
| | GA8 | 0.003169 | 10 | 15.84% | 1.0539% | 10 | 6.39% |
| | Naive | 0.003924 | 11 | 43.42% | 1.2610% | 11 | 27.29% |
| | AddDblSeasonal | 0.001934 | 1 | -29.31% | 0.8585% | 1 | -13.34% |



(a) $MaxC = 2$ clusters.

(b) $MaxC = 2$ clusters (zoom).

(c) $MaxC = 4$ clusters.

(d) $MaxC = 4$ clusters (zoom).

(e) $MaxC = 8$ clusters.

(f) $MaxC = 8$ clusters (zoom).

Figure 4.19: Forecast intervals for Synthetic-2S time series.

The graphs cover the period for $t - 12 \leq t \leq t + H$ where $t$ is the last observation of the in-sample period and $H = 12$ is the number of forecast horizons. The shades, from lighter to darker, correspond to $\alpha$ levels 0.95, 0.90, 0.85, 0.80, 0.75 and 0.60.

Table 4.14: Coefficients for structural combination of NN for Synthetic-2S series.

| h | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 1 | $\alpha_1$ | 0.0018 | 0.0580 | -0.0608 | 0.0313 | 0.0549 | 0.9100 | |
| | $\alpha_2$ | -0.0038 | 0.2494 | 0.0662 | 0.0575 | 0.6996 | -0.0788 | |
| | $\alpha_3$ | -0.0024 | -0.2469 | 0.2834 | 0.2910 | 0.1293 | 0.5361 | |
| | $\alpha_4$ | -0.0182 | 0.0701 | -0.0011 | 0.0608 | -0.2369 | 1.1035 | |
| | $\alpha_5$ | 0.0246 | -0.0460 | 0.1941 | -0.1541 | 0.1910 | 0.8026 | |
| | $\alpha_6$ | 0.0035 | -0.0660 | 0.0221 | -0.0749 | 0.0982 | 1.0139 | |
| | $\alpha_7$ | 0.0003 | -0.0259 | -0.0752 | 0.0842 | 0.0855 | 0.9247 | |
| | $\Phi$ | 0.1393 | 0.1420 | 0.1462 | 0.1447 | 0.1423 | 0.1447 | 0.1469 |
| 2 | $\alpha_1$ | -0.0045 | 0.0142 | 0.4543 | 0.2349 | 0.2906 | | |
| | $\alpha_2$ | 0.0028 | 0.7876 | -0.2586 | 0.0158 | -0.0250 | 0.4712 | |
| | $\alpha_3$ | 0.0031 | 0.0151 | 0.2054 | -0.1964 | 0.6569 | 0.3096 | |
| | $\alpha_4$ | 0.0030 | 0.3746 | -0.1562 | -0.0093 | 0.1326 | 0.6489 | |
| | $\alpha_5$ | -0.0025 | -0.0856 | 0.8669 | 0.0851 | 0.1283 | | |
| | $\Phi$ | 0.2056 | 0.1973 | 0.1983 | 0.2121 | 0.1945 | | |
| 6 | $\alpha_1$ | -0.0188 | 0.0031 | -0.5853 | 0.7243 | -0.1585 | 1.0087 | |
| | $\alpha_2$ | -0.0011 | 0.1214 | 0.0871 | 0.0260 | -0.0788 | 0.8334 | |
| | $\alpha_3$ | -0.0051 | -0.1999 | 0.5205 | 0.0710 | 0.0668 | 0.5308 | |
| | $\alpha_4$ | -0.0063 | -0.1643 | 0.2516 | 0.9033 | 0.0341 | -0.0355 | |
| | $\alpha_5$ | -0.0023 | 0.1968 | 0.3278 | 0.6160 | -0.1424 | -0.0096 | |
| | $\Phi$ | 0.1942 | 0.2006 | 0.2036 | 0.2050 | 0.2080 | | |

$MaxC = 8$ is the maximum number of clusters. $h$ denotes the forecast horizon, $\alpha_i$ are the coefficients applied to point-forecasts from models in cluster $i$ and $\Phi$ are the weights applied to the outputs from clusters.

Table 4.15: Cluster validity indexes. Series: Synthetic-2S.

| | Maximum number of clusters: 2 | | | |
|---|---|---|---|---|
| $h$ | PC | MPC | MDO | Final num. clusters |
| 1 | 1.0000 | NA | 1.0000 | 1 |
| 2 | 1.0000 | NA | 1.0000 | 1 |
| 3 | 1.0000 | NA | 1.0000 | 1 |
| 4 | 1.0000 | NA | 1.0000 | 1 |
| 5 | 0.5004 | 0.0008 | 0.0000 | 2 |
| 6 | 1.0000 | NA | 1.0000 | 1 |
| 7 | 1.0000 | NA | 1.0000 | 1 |
| 8 | 0.5007 | 0.0014 | 0.0000 | 2 |
| 9 | 1.0000 | NA | 1.0000 | 1 |
| 10 | 1.0000 | NA | 1.0000 | 1 |
| 11 | 1.0000 | NA | 1.0000 | 1 |
| 12 | 1.0000 | NA | 1.0000 | 1 |

| | Maximum number of clusters: 4 | | | |
|---|---|---|---|---|
| $h$ | PC | MPC | MDO | Final num. clusters |
| 1 | 0.5050 | 0.0100 | 0.3000 | 2 |
| 2 | 0.5041 | 0.0082 | 0.4000 | 2 |
| 3 | 0.2520 | 0.0027 | 0.0000 | 4 |
| 4 | 0.3340 | 0.0010 | 0.0000 | 3 |
| 5 | 0.3337 | 0.0005 | 0.0000 | 3 |
| 6 | 0.3344 | 0.0017 | 0.0000 | 3 |
| 7 | 0.3343 | 0.0015 | 0.0000 | 3 |
| 8 | 0.2502 | 0.0003 | 0.0000 | 4 |
| 9 | 0.5026 | 0.0052 | 0.2000 | 2 |
| 10 | 0.5015 | 0.0030 | 0.1000 | 2 |
| 11 | 0.5008 | 0.0017 | 0.0000 | 2 |
| 12 | 0.2506 | 0.0008 | 0.0000 | 4 |

| | Maximum number of clusters: 8 | | | |
|---|---|---|---|---|
| $h$ | PC | MPC | MDO | Final num. clusters |
| 1 | 0.1430 | 0.0001 | 0.0000 | 7 |
| 2 | 0.2004 | 0.0004 | 0.0000 | 5 |
| 3 | 0.2006 | 0.0007 | 0.0000 | 5 |
| 4 | 0.2507 | 0.0009 | 0.0000 | 4 |
| 5 | 0.3343 | 0.0015 | 0.0000 | 3 |
| 6 | 0.2002 | 0.0003 | 0.0000 | 5 |
| 7 | 0.1669 | 0.0003 | 0.0000 | 6 |
| 8 | 0.1429 | 0.0001 | 0.0000 | 7 |
| 9 | 0.1251 | 0.0001 | 0.0000 | 8 |
| 10 | 0.1251 | 0.0001 | 0.0000 | 8 |
| 11 | 0.1250 | 0.0000 | 0.0000 | 8 |
| 12 | 0.2505 | 0.0007 | 0.0000 | 4 |

$h$ denotes the forecast horizon, PC denotes the Partition Coefficient, MPC denotes the Modified Partition Coefficient, and MDO denotes the Membership Degree Optimum. Values closer to 1 are preferable.

## 4.6 Discussion of Findings from Synthetic Series

It was observed that for STAR2 and Synthetic-1S series, GA combinations have a more stable performance for different forecast horizons. This relative stability is relevant, because the architecture of models used in ensembles for different forecast horizons differ. Hence, there seems to be support for using GAs in structural combinations of NNs when the serial dependency involves a limited number of past lags.

For Synthetic-1S and Synthetic-2S time series, it is clear that CB models with more clusters forecast better. The more complex task of forecasting at longer horizons is reflected in the influence of more clusters (more models) and more parameters (for combination). In some cases, a preferable number of clusters is identified, as in steps 3 and 11 for Synthetic-2S, when the error metric curve (in Figure 4.18) shows a convex shape and a minimum. In other horizons (4 and 6), the performance curve is concave, but also shows a minimum.

When comparing the ability of CB and GA combinations to outperform the NN-based benchmarks, the following is observed: for the non-seasonal STAR2 series, GA combinations outperform the simple average in more occasions than CB combinations (see summary Table 4.16). For the seasonal and double-seasonal series (Synthetic-1S and Synthetic-2S), however, only CB performs well when compared to the simple average. The ability of CB and GA to outperform the models with best fit in the ensemble is similar in the cases of STAR2 and Synthetic-1S series; while for Synthetic-2S, CB combinations are clearly better.

When considering the overall ability of structural combinations (CB and GA) to outperform NN-based benchmarks, it is noticed that the benefit of building ensembles and combining structurally is clearer in the case of the non-seasonal series (with high noise), since the models with best fit and the simple average are more consistently outperformed by the structural combination. For the single-seasonal series, the structural combinations improve markedly over the models with best fit, but not over the simple average, while for the double-seasonal series, improvement over the best models and the average is similar (except for CB8). It seems that the use of separate models for a regular time series can create well performing individual models and, consequently, well performing average forecasts. In consequence, such benchmarks tend to be difficult to outperform with feed-forward NNs and ensembles.

Table 4.16: Number of forecast horizons for which CB and GA combinations outperform benchmarks.

| Metric | Series / Models | Reference benchmarks | | | |
|---|---|---|---|---|---|
| MSE | STAR2 Series | Avg. | Bst. IsMAE | Bst. IsMSE | Stat. |
| | CB2 | 3 | 2 | 3 | 3 |
| | GA2 | 5 | 3 | 4 | 5 |
| | CB4 | 3 | 2 | 3 | 2 |
| | GA4 | 4 | 3 | 4 | 5 |
| | CB8 | 2 | 3 | 3 | 2 |
| | GA8 | 4 | 4 | 4 | 5 |
| MSE | Synthetic-1S Series | Avg. | Bst. IsMAPE | Bst. IsMSE | Stat. |
| | CB2 | 6 | 10 | 9 | 3 |
| | GA2 | 3 | 11 | 10 | 4 |
| | CB4 | 5 | 9 | 9 | 4 |
| | GA4 | 5 | 10 | 10 | 4 |
| | CB8 | 8 | 12 | 11 | 4 |
| | GA8 | 5 | 11 | 11 | 4 |
| MSE | Synthetic-2S Series | Avg. | Bst. IsMAPE | Bst. IsMSE | Stat. |
| | CB2 | 4 | 3 | 2 | 0 |
| | GA2 | 2 | 2 | 1 | 0 |
| | CB4 | 7 | 4 | 2 | 0 |
| | GA4 | 0 | 1 | 1 | 0 |
| | CB8 | 12 | 7 | 5 | 0 |
| | GA8 | 1 | 1 | 1 | 0 |

| Metric | Series / Models | Reference benchmarks | | | |
|---|---|---|---|---|---|
| *MAE | STAR2 Series | Avg. | Bst. IsMAE | Bst. IsMSE | Stat. |
| | CB2 | 2 | 4 | 4 | 4 |
| | GA2 | 5 | 4 | 3 | 5 |
| | CB4 | 3 | 2 | 3 | 4 |
| | GA4 | 4 | 4 | 3 | 5 |
| | CB8 | 3 | 3 | 3 | 4 |
| | GA8 | 4 | 4 | 3 | 5 |
| MAPE | Synthetic-1S Series | Avg. | Bst. IsMAPE | Bst. IsMSE | Stat. |
| | CB2 | 5 | 10 | 9 | 2 |
| | GA2 | 5 | 10 | 10 | 3 |
| | CB4 | 4 | 9 | 9 | 3 |
| | GA4 | 6 | 10 | 10 | 4 |
| | CB8 | 7 | 12 | 11 | 4 |
| | GA8 | 4 | 10 | 10 | 4 |
| MAPE | Synthetic-2S Series | Avg. | Bst. IsMAPE | Bst. IsMSE | Stat. |
| | CB2 | 6 | 3 | 5 | 0 |
| | GA2 | 2 | 3 | 5 | 0 |
| | CB4 | 6 | 4 | 4 | 0 |
| | GA4 | 1 | 3 | 4 | 0 |
| | CB8 | 12 | 5 | 4 | 0 |
| | GA8 | 1 | 2 | 4 | 0 |

Avg. stands for the average of NN in the ensemble;
Bst. IsMAE, Bst. IsMAPE, Bst. IsMSE stand for the best NN in terms of in-sample MAE, MAPE or MSE in the ensemble;
Stat. stands for the corresponding statistical benchmark used.
* MAE preferred over MAPE when the time series has values close to zero.

## 4.7 Forecasting Wind Power Using Ensembles of NNs Based on Multivariate Time Series

Energy Forecasting Competition (Kaggle, 2012; Hong et al., 2014) provided hourly wind power scaled in the interval $[0, 1]$ and numerical weather forecast data (wind speed and direction) for 7 wind farms. The first wind farm was selected for the present study. The original arrangement in the competition allocated the period from 2009/07/01 to 2010/12/31 for training (fitting) and for the rest of the data-set missing periods of 48 hours length were defined, that the participants had to forecast (all data, however, is available so that forecasting performance can be inferred). No rolling forecast origin was used and only RMSE error metric was reported in the analyses of the competition (Hong et al., 2014). Here a data set-up was adopted that enabled the use of a rolling window over a sufficiently long period of data without missing values. The data were split between training (in-sample) and testing (out-of-sample): 66% and 33%, respectively. The subset thus selected covers the period from 2009/07/01 at 0:00 to 2010/12/31 at 12:00 clock and weather forecast for 48 hours ahead are available every 12 hours from 2009/07/01 to 2010/12/31.

The two-stages procedure of sensitivity analysis and fitting of models that is used in previous sections is followed here. The configuration parameters for the first stage are listed in Table 4.17. Inputs comprise 2 exogenous variables (the most recent forecast for wind speed and wind direction available for a specific forecast horizon), as well as lagged variable, i.e, lags 1 to 5 from the wind power series (a partial autocorrelogram shows that only the first 3 lags are important, but 5 are included in order to allow for a wider view concerning the sensitivity of error metrics to the number of inputs). As argued by Lee & Scholtes (2014), the use of different models for every forecast horizon is desirable given that the quality of wind forecasts differs depending on the horizon: as they are issued every 12 hours, the longer the time between the desired horizon and the last issued forecast, the worse is the quality of available information.

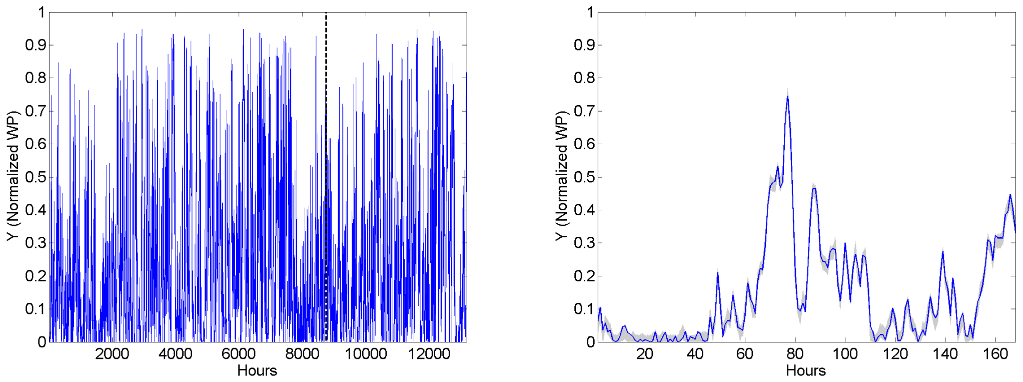
During the preliminary sensitivity analysis, variation was introduced into the original series in order to create replicas (100 for each factor combination), by adding noise uniformly distributed in the range $[-0.1\sigma_b, +0.1\sigma_b]$, where $\sigma_b$ is the standard deviation of the bootstrapped series. This permitted to mimic the conditions of the sensitivity analysis performed with synthetic time series in the previous chapter. Because the generating process was unknown in the case of wind power, the noise addition allowed to create small variations of the series, which play the role of different realisations of the process. The idea was inspired by Zhang (2007) and Brown et al. (2003) who studied the addition of noise to input data in neural networks. Their findings indicate that the effect on performance depends on the noise level added and its distribution. The noise level was chosen so that the general dynamics of the series is not substantially altered. The distribution accounts for a pessimistic case, since the likelihood of extreme values is the same as that of observations that are close to the mean.

The period between 2009/7/1 and 2010/06/30 (8760 observations) was used for model identification and training (in-sample period) and data from 2010/07/01 to 2010/12/31 (4416) was used for model evaluation (out-of-sample period). NMAPE and RMSE metrics were used to assess the performance of models during both the sensitivity analysis and the forecasting exercise. The former can be interpreted in terms of percentages and avoids divisions by zero, as there are times when a wind farm has no production. The latter is commonly used in the wind-power literature (see for example Giebel et al., 2003). The NMAPE is calculated as follows:

$$NMAPE = \frac{1}{n} \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{(1/n)\sum_{i=1}^{n} y_i} \tag{4.24}$$

Where $y_i$ is the $i$-th observation and $\hat{y}_i$ is the corresponding estimated value. RMSE was preferred over MSE because it is standard in the wind power forecasting literature.

The original series and replicas are depicted in Figure 4.20, where training (in-sample) and evaluation (out-of-sample) periods are separated by a dashed line.



(a) Wind power series.

(b) Replicas (first week).

Figure 4.20: Hourly electricity production and replicas for Kaggle wind farm 1, from 1 July 2009 to 31 December 2010.

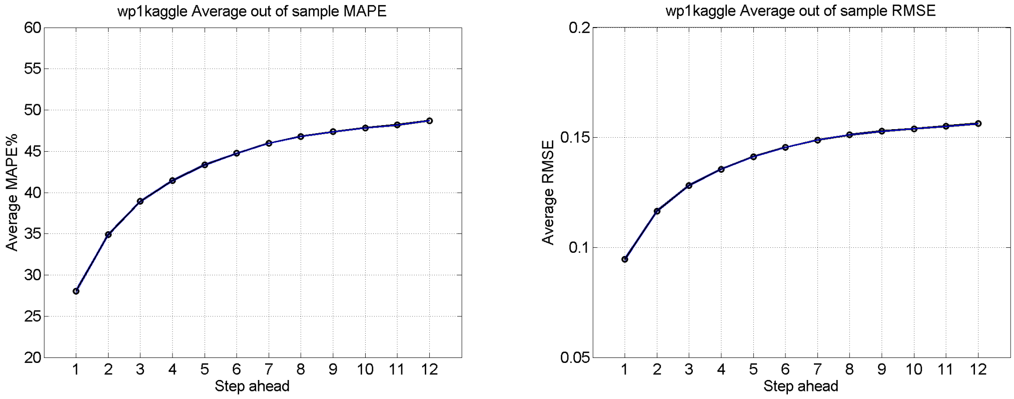
### 4.7.1 Preliminary Analysis and Specification of Individual Models for the Ensemble

Figure 4.21 shows the average out-of-sample NMAPE and RMSE of NN models for each forecast horizon. The latter metric is preferred over the MSE as it is traditionally used in the wind power industry. A very homogeneous behaviour is noticed, given the very narrow confidence bands, indicating little variability in forecast performance. Taking into account that the averages and confidence bands are calculated with data from different architectures, it can be inferred that the forecasting problem is insensitive to architectural decisions. Table 4.18 lists the architectures with the lowest average out-of-sample RMSE. Exogenous variables and the first lag of wind power are common among the specifications and complexity in terms of the number of neurons tends to be high.

Table 4.17: Configuration: Kaggle wind power data.

| **Factor** | **Symbol** | **Levels** |
|---|---|---|
| Number of inputs | NI | $\mathbf{1}, \ldots, \mathbf{7}$,<br>corresponding to the set<br>$[ws, wd, wp_{L1}, \ldots, wp_{L5}]$, where<br>where $ws$ and $wd$ are the most recent forecasts for wind speed and wind direction available at time $t$ for a specific horizon, $k$ (see Figure 4.22), and $wp_{L1}, \ldots, wp_{L5}$ are values of wind power at times $t, \ldots, t-4$ used to forecast $wp_{t+k}$, $k = 1, \ldots, 12$. |
| Number of hidden layers | NL | 1 |
| Number of hidden units | NU | $\mathbf{1}, \ldots, \mathbf{14}$ (1 to 2 times the number of levels for NI) |
| Activation function for hidden nodes | AF1 | Tangent Sigmoid |
| Activation function for the output node | AF2 | Linear |
| Initial values for the weights | W0 | Values in the range [-2 2] established by the Nguyen-Widrow algorithm (there is a degree of randomness) |
| Training algorithm | TA | Back-propagation with Levenberg-Marquardt optimisation. |
| Stopping criteria | SC | * The maximum number of epochs (repetitions) is reached: 300.<br>* The maximum amount of time is exceeded: $\infty$<br>* Performance is minimised to the goal: 0<br>* The performance gradient falls below $min_{grad}$ : $10^{-10}$<br>* $\mu$ exceeds $\mu_{max} = 10^3$<br>* Validation performance has increased more than $max_{fail}$ times since the last time it decreased (when using validation): 6 |
| Data normalisation | DN | Yes |
| Initial combination coefficient ($\mu$) | MU | 0.001 |
| Prune units | PU | **Yes, No** |
| Prune input variables | PI | No |
| Sample size | SS | 8760 |
| Data configuration for training, testing and validation (training + validation = in-sample period; testing=out-of-sample period) | DC | Conf. 1: $Ntr = 8760$, $Nva = 760$ $Nte = 4416$ |
| Extreme values treated | EV | No. |
| Sampling method | SM | **block, cross-validated** |
| Forecast approach | FA | Direct: a separate model for each forecast horizon |

In bold are the factors which vary.



(a) MAPE.

(b) RMSE.

Figure 4.21: Average NMAPE% and RMSE.

Table 4.18: Models with the lowest average out-of-sample RMSE.

| Forecast horizon | NI | NU | Average RMSE |
|---|---|---|---|
| 1 | 4 | 12 | 0.06866 |
| 2 | 4 | 14 | 0.09859 |
| 3 | 4 | 13 | 0.11445 |
| 4 | 3 | 12 | 0.12422 |
| 5 | 4 | 10 | 0.13204 |
| 6 | 3 | 11 | 0.13740 |

| Forecast horizon | NI | NU | Average RMSE |
|---|---|---|---|
| 7 | 3 | 10 | 0.14226 |
| 8 | 3 | 12 | 0.14550 |
| 9 | 3 | 8 | 0.14711 |
| 10 | 3 | 11 | 0.14937 |
| 11 | 3 | 7 | 0.15006 |
| 12 | 3 | 7 | 0.15163 |

When $NI = p$, the first $p$ variables of the set $[ws, wd, wp_{L1}, wp_{L2}]$ are used, where $ws$ and $wd$ are the most recent forecasts for wind speed and wind direction available at time $t$ for horizon $k$ and $wp_{L1}, wp_{L2}$ are values of wind power at times $t$ and $t-1$ used to forecast $wp_{t+k}$.

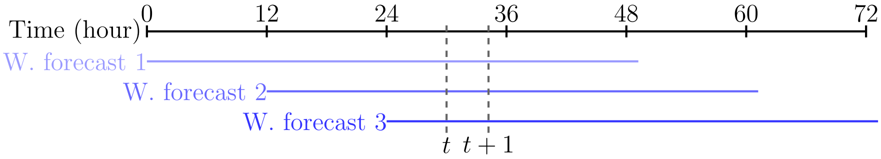


Figure 4.22: Weather forecasts usage.

Weather forecasts (wind speed and direction) are issued every 12 hours. To forecast wind power at time $t+1$ with origin at $t$, the most recent weather forecast is W. forecast 3.

The main effects graphs (Figure 4.23) show how NMAPE and RMSE (in both the in-sample and out-of-sample periods) are sensitive to NI, which is evident after $NI = 2$. The first part ($1 \leq NI \leq 2$) corresponds to the exogenous variables (wind speed and wind direction forecasts) and the second part ($3 \leq NI \leq 7$) corresponds to wind power lagged-values. Sensitivity with respect to the weather forecasts is surprisingly low, and the use of previous values has a marked effect, in particuar the first lag. The sensitivity of both metrics in both periods to NU is low, as judged by the main effects graphs, and pruning of synaptic connections (PU) slightly performance (the complete set of graphs, not shown here, overall supports these findings).

ANOVA and Kruskal-Wallis tests suggest that NI, NU and PU have significant effect on both in-sample and out-of-sample metrics. Jonckheere-Terpstra tests confirm such influence and further highlight that NI and NU tend to decrease the values of the metrics, while PR tends to increase them. In all, the tests confirm the general tendencies observed in the main-effects graphs.

The assessment of residual serial correlation reveals a general failure of the models to capture the dynamics of the series. For almost all forecast horizons, the tests show serial correlation in all trials made. Only the forecasts for one step ahead from a subset of NNs led to serially independent forecast errors. Such models have 3 or more inputs, which confirms the need for lagged values of wind power production. In models for the subsequent hour, these inputs are no longer influential in capturing the dynamics of the series.

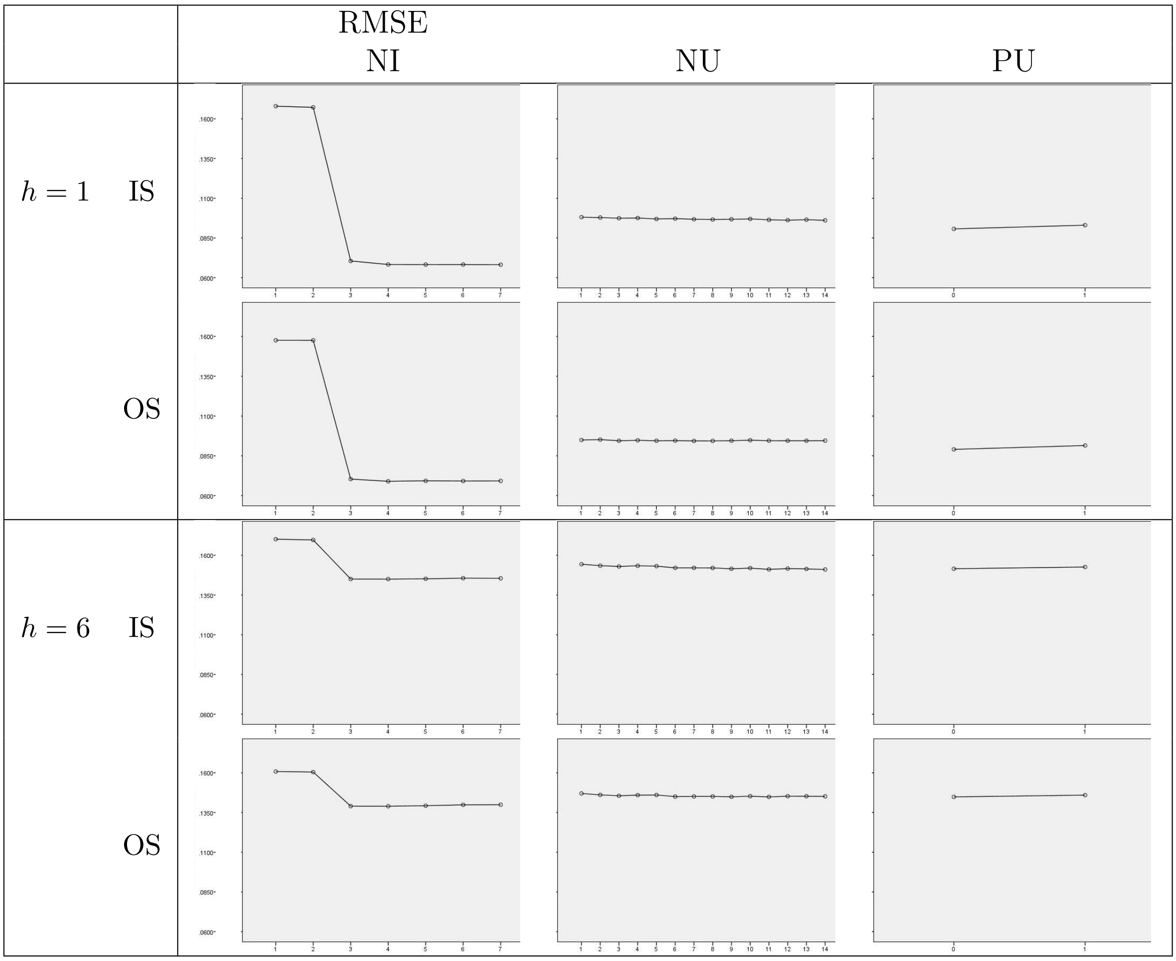
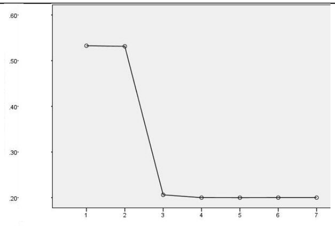
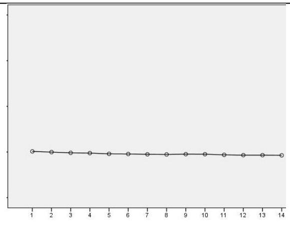
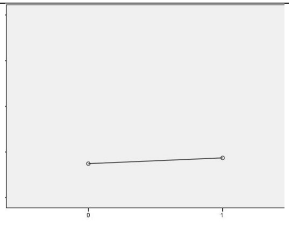
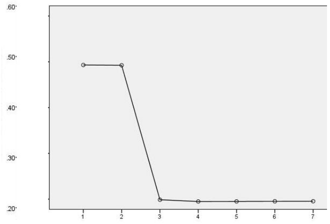
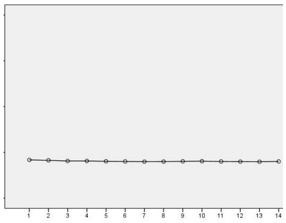
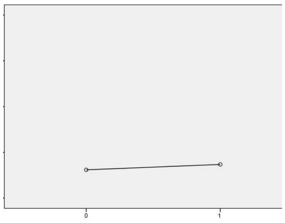
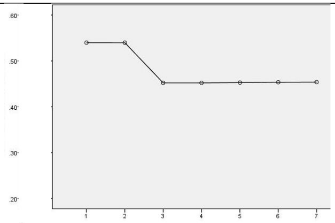
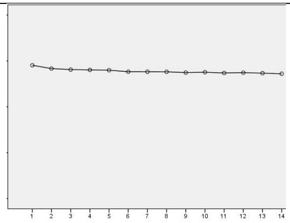
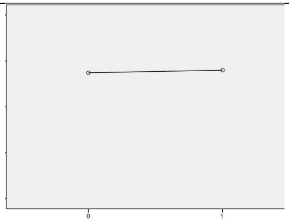
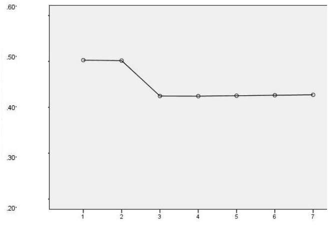
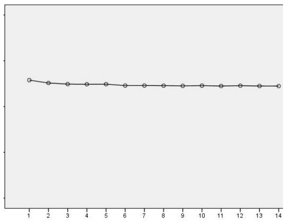
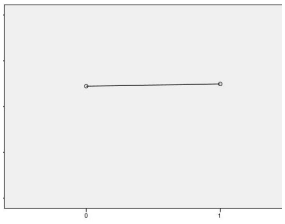


Figure 4.23: Main effects. Series: Kaggle wind power production series. IS stands for in-sample; OS stands for out-of-sample.

|  | NMAPE<br>NI | NU | PU |
|---|---|---|---|
| $h = 1$ IS |  |  |  |
| OS |  |  |  |
| $h = 6$ IS |  |  |  |
| OS |  |  |  |

(Continued) Main effects. Series: Kaggle wind power production series.

### 4.7.2 Combination of Models

The NN models selected are listed in Table 4.19. They are relatively similar in terms of the number of inputs and have a number of neurons such that $NU > NI$. This reflects the higher influence of the exogenous inputs and the need of a relatively complex forecasting mechanism.

Table 4.19: Selected models for Kaggle series.

| h | NI | NU |
|---|---|---|
| 1 | 4 | 9 |
| 2 | 4 | 13 |
| 3 | 4 | 13 |
| 4 | 3 | 12 |
| 5 | 4 | 10 |
| 6 | 3 | 10 |
| 7 | 5 | 9 |
| 8 | 3 | 12 |
| 9 | 3 | 8 |
| 10 | 3 | 11 |
| 11 | 3 | 7 |
| 12 | 3 | 7 |

When $NI = p$, the first $p$ variables of the set $[ws, wd, wp_{L1}, wp_{L2}, wp_{L3}]$ are used, where $ws$ and $wd$ are the most recent forecasts for wind speed and wind direction available at time $t$ for horizon $h$ and $wp_{L1}, wp_{L2}, wp_{L3}$ are values of wind power at times $t$, $t-1$ and $t-2$ used to forecast $wp_{t+h}$.

The analysis of performance for this series follows the general procedure that has already been applied to synthetic series. However, as previously mentioned, the RMSE metric was used instead of MSE, as it is traditionally used in the wind power industry and NMAPE metric is preferred over MAPE in order to avoid divisions by zero when there is no power production in the wind farm. An ARIMA model with exogenous variables (ARIMAX) was considered as statistical benchmark, and was selected using the routine *auto.arima* from *forecast* package in R (Hyndman & Khandakar, 2008; Hyndman, 2015). The model is the following:

$$\hat{y}_t = 0.0320ws + 0.3334y_{t-1} + 0.4916y_{t-2} - 0.1670\epsilon_{t-1} - 0.6504\epsilon_{t-2} - 0.1617\epsilon_{t-3} \quad (4.25)$$

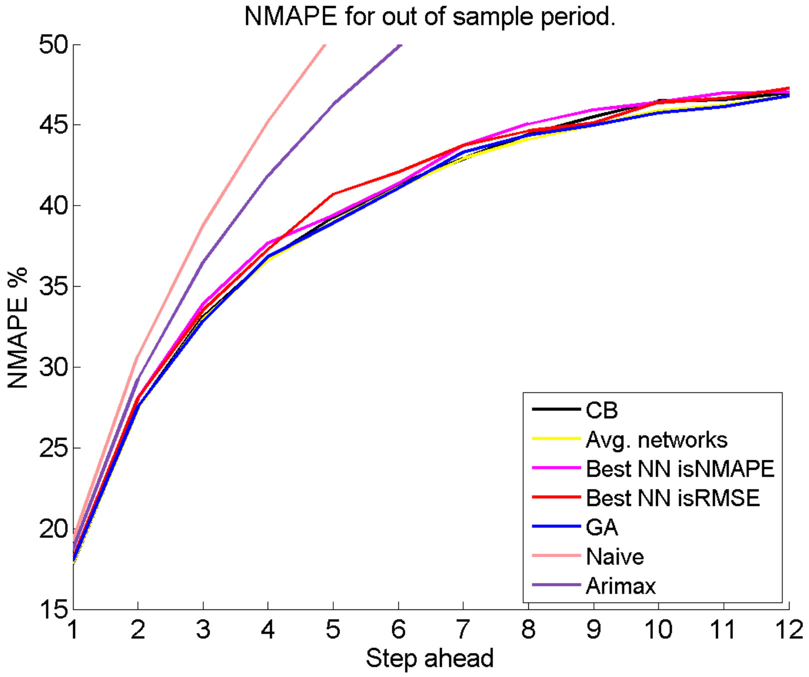
Where $y_t$ is the normalised wind power production at time $t$ and $ws$ is the most recent wind direction forecast available for time $t$ (the automatic specification routines in R found a zero coefficient for $wd$, which was confirmed by fitting other models directly).

Figure 4.24 shows the out-of-sample NMAPE and RMSE for the CB combination approach and Figure 4.25 focuses on the RMSE performance of key benchmarks (Figures C.7 and C.8 from Appendix C provide further detail). Behaviour of NMAPE, not shown, is similar. Numeric detail of performance and the ranking of models for every forecast horizon are presented in Table 4.20.

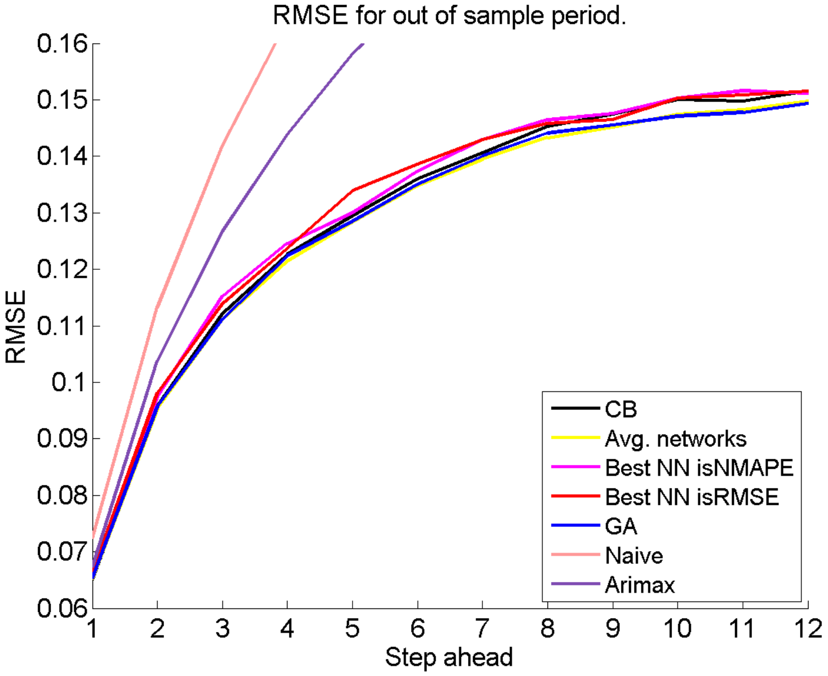
Taking into account the RMSE metric, all NN-based outperform the naive and statistical benchmark. Both CB and GA combinations outperform the individual NN with best in-sample metrics in most of the forecast horizons. However, CB combination outperforms the NN average only for the first step ahead and, interestingly, GA combinations outperformed the CB models and the simple average in several horizons.

For NMAPE, the behaviour is similar, with minor changes in the ranking of models, but results confirm the tendencies observed. GA combination show superiority with respect to CB, and both combination approaches easily outperform a single NN, naive and ARIMA benchmarks.

The behaviour of both metrics is very smooth in comparison with synthetic series. As it was shown before, models with the lowest out-of-sample RMSE and the selected models for the ensembles had similar structures in terms of inputs and neurons and the sensitivity analysis showed a relative insensitivity of models to such factors (flat regions). The smoothness in the metrics vs. forecast horizon graphs (Figure 4.25) might reflect the insensitivity of model performance to structural factors, which in turn might come from nature of the phenomenon (achievable performance with a given structure and input set is very homogeneous). Such insensitivity is also evident in Figure 4.26, where the number of clusters does not seem to affect the metrics.

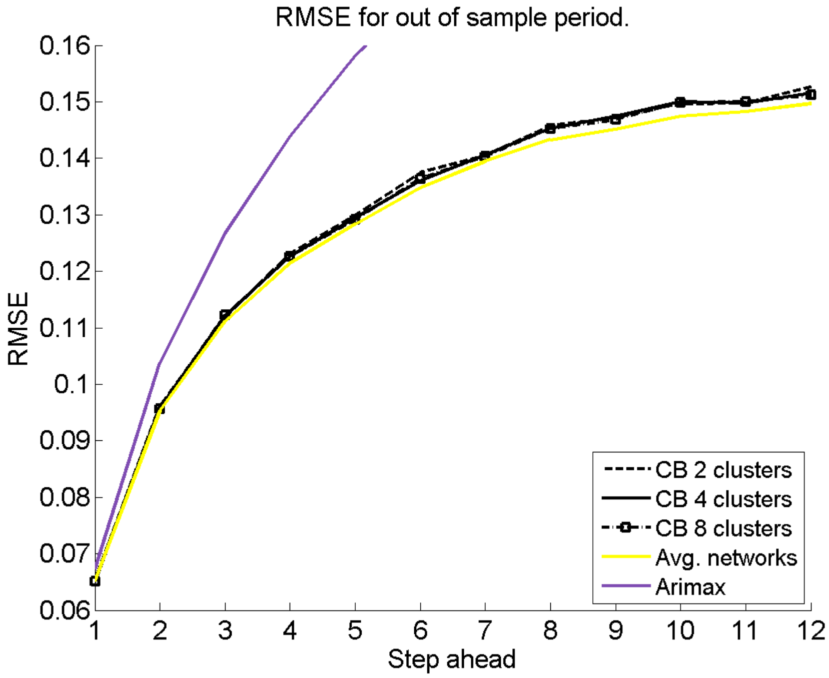


(a) 4 clusters. NMAPE.

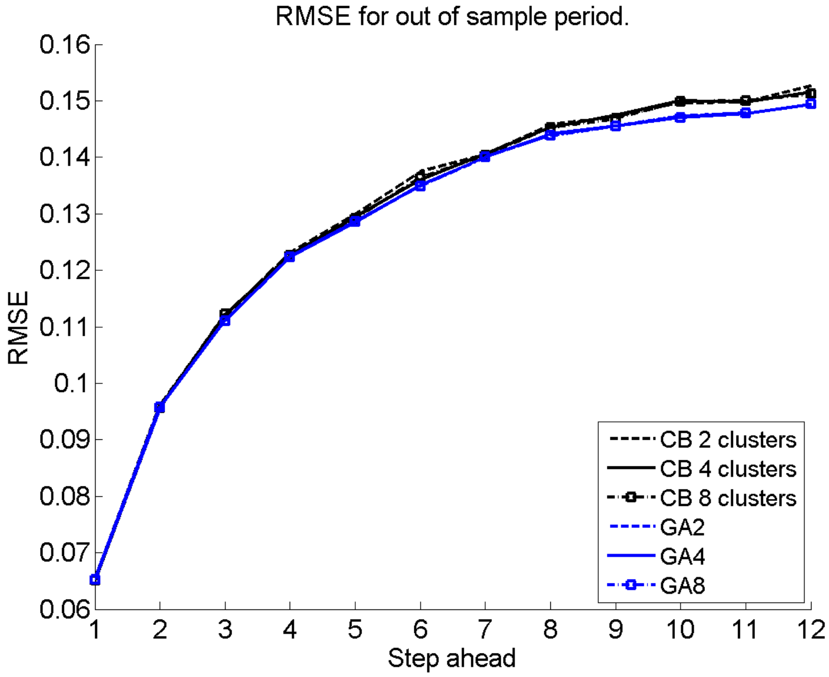


(b) 4 clusters. RMSE.

Figure 4.24: Out-of-sample NMAPE and RMSE for Kaggle wind power production series.



(a) Comparison with Avg. and ARIMAX. RMSE.



(b) Comparison with GA. RMSE.

Figure 4.25: Out-of-sample RMSE for Kaggle wind power production series.

Table 4.20: Forecasting performance. Kaggle wind power production series.

| F. horizon | Model | RMSE | Rank | %Δ wrt Avg. | NMAPE | Rank | %Δ wrt Avg. |
|---|---|---|---|---|---|---|---|
| h=1 | CB2 | 0.06526 | 6 | 0.11% | 17.8220% | 2 | -0.25% |
| | CB4 | 0.06510 | 2 | -0.13% | 17.7910% | 1 | -0.43% |
| | CB8 | 0.06509 | 1 | -0.14% | 17.8330% | 3 | -0.19% |
| | Avg. Net. | 0.06518 | 3 | 0.00% | 17.8670% | 4 | 0.00% |
| | Best Net. IsNMAPE | 0.06582 | 8 | 0.98% | 18.0270% | 8 | 0.90% |
| | Best Net. IsRMSE | 0.06614 | 9 | 1.47% | 18.1010% | 9 | 1.31% |
| | GA2 | 0.06523 | 4 | 0.07% | 17.8944% | 5 | 0.15% |
| | GA4 | 0.06526 | 6 | 0.11% | 17.9940% | 7 | 0.71% |
| | GA8 | 0.06525 | 5 | 0.10% | 17.9580% | 6 | 0.51% |
| | Naive | 0.07235 | 11 | 11.00% | 19.2395% | 11 | 7.68% |
| | ARIMAX | 0.06724 | 10 | 3.16% | 18.7348% | 10 | 4.86% |
| h=2 | CB2 | 0.09616 | 7 | 0.87% | 27.5890% | 6 | 0.51% |
| | CB4 | 0.09587 | 6 | 0.56% | 27.4870% | 3 | 0.13% |
| | CB8 | 0.09566 | 2 | 0.34% | 27.4570% | 2 | 0.03% |
| | Avg. Net. | 0.09533 | 1 | 0.00% | 27.4500% | 1 | 0.00% |
| | Best Net. IsNMAPE | 0.09743 | 8 | 2.20% | 28.0450% | 8 | 2.17% |
| | Best Net. IsRMSE | 0.09809 | 9 | 2.90% | 28.0620% | 9 | 2.23% |
| | GA2 | 0.09581 | 5 | 0.50% | 27.6183% | 7 | 0.61% |
| | GA4 | 0.09573 | 4 | 0.41% | 27.5592% | 4 | 0.40% |
| | GA8 | 0.09572 | 3 | 0.40% | 27.5826% | 5 | 0.48% |
| | Naive | 0.11323 | 11 | 18.77% | 30.6929% | 11 | 11.81% |
| h=3 | CB2 | 0.11190 | 5 | 0.69% | 33.0640% | 5 | 0.30% |
| | CB4 | 0.11214 | 6 | 0.91% | 33.1140% | 7 | 0.45% |
| | CB8 | 0.11218 | 7 | 0.94% | 33.1110% | 6 | 0.44% |
| | Avg. Net. | 0.11113 | 4 | 0.00% | 32.9650% | 4 | 0.00% |
| | Best Net. IsNMAPE | 0.11520 | 9 | 3.66% | 33.9090% | 9 | 2.86% |
| | Best Net. IsRMSE | 0.11383 | 8 | 2.43% | 33.5350% | 8 | 1.73% |
| | GA2 | 0.11109 | 2 | -0.04% | 32.8125% | 1 | -0.46% |
| | GA4 | 0.11111 | 3 | -0.02% | 32.8397% | 3 | -0.38% |
| | GA8 | 0.11108 | 1 | -0.05% | 32.8365% | 2 | -0.39% |
| | Naive | 0.14187 | 11 | 27.66% | 38.7617% | 11 | 17.58% |
| | ARIMAX | 0.12668 | 10 | 13.99% | 36.4649% | 10 | 10.62% |
| h=4 | CB2 | 0.12302 | 7 | 1.29% | 36.8270% | 6 | 0.57% |
| | CB4 | 0.12259 | 6 | 0.94% | 36.7100% | 3 | 0.25% |
| | CB8 | 0.12145 | 1 | 0.00% | 36.6180% | 1 | 0.00% |
| | Avg. Net. | 0.12145 | 1 | 0.00% | 36.6180% | 1 | 0.00% |
| | Best Net. IsNMAPE | 0.12462 | 9 | 2.61% | 37.6980% | 9 | 2.95% |
| | Best Net. IsRMSE | 0.12369 | 8 | 1.84% | 37.2800% | 8 | 1.81% |
| | GA2 | 0.12234 | 4 | 0.73% | 36.8116% | 4 | 0.53% |
| | GA4 | 0.12234 | 3 | 0.73% | 36.8234% | 5 | 0.56% |
| | GA8 | 0.12234 | 5 | 0.74% | 36.8339% | 7 | 0.59% |
| | Naive | 0.16401 | 11 | 35.04% | 45.1852% | 11 | 23.40% |
| | ARIMAX | 0.14387 | 10 | 18.46% | 41.8693% | 10 | 14.34% |

$\%\Delta = 100(M_{model} - M_{Avg})/M_{Avg}$ with $M_i$ being the metric for model $i$.
Negative % values indicate improvement over the average.

(Continued) Forecasting performance. Kaggle wind power production series.

| F. horizon | Model | RMSE | Rank | %Δ wrt Avg. | NMAPE | Rank | %Δ wrt Avg. |
|---|---|---|---|---|---|---|---|
| h=5 | CB2 | 0.12991 | 7 | 1.25% | 39.4140% | 8 | 1.10% |
| | CB4 | 0.12950 | 6 | 0.93% | 39.2780% | 6 | 0.75% |
| | CB8 | 0.12914 | 5 | 0.65% | 39.1470% | 5 | 0.42% |
| | Avg. Net. | 0.12831 | 1 | 0.00% | 38.9840% | 2 | 0.00% |
| | Best Net. IsNMAPE | 0.13000 | 8 | 1.32% | 39.4110% | 7 | 1.10% |
| | Best Net. IsRMSE | 0.13391 | 9 | 4.36% | 40.7060% | 9 | 4.42% |
| | GA2 | 0.12873 | 4 | 0.32% | 39.0871% | 4 | 0.26% |
| | GA4 | 0.12847 | 2 | 0.12% | 38.9201% | 1 | -0.16% |
| | GA8 | 0.12864 | 3 | 0.25% | 39.0475% | 3 | 0.16% |
| | Naive | 0.18232 | 11 | 42.10% | 50.6443% | 11 | 29.91% |
| | ARIMAX | 0.15812 | 10 | 23.23% | 46.2666% | 10 | 18.68% |
| h=6 | CB2 | 0.13742 | 8 | 1.97% | 41.5750% | 8 | 1.04% |
| | CB4 | 0.13604 | 5 | 0.95% | 41.3040% | 5 | 0.38% |
| | CB8 | 0.13640 | 6 | 1.22% | 41.5550% | 7 | 0.99% |
| | Avg. Net. | 0.13476 | 1 | 0.00% | 41.1470% | 4 | 0.00% |
| | Best Net. IsNMAPE | 0.13737 | 7 | 1.94% | 41.3530% | 6 | 0.50% |
| | Best Net. IsRMSE | 0.13853 | 9 | 2.80% | 42.0780% | 9 | 2.26% |
| | GA2 | 0.13491 | 3 | 0.11% | 41.0862% | 1 | -0.15% |
| | GA4 | 0.13500 | 4 | 0.18% | 41.0923% | 3 | -0.13% |
| | GA8 | 0.13486 | 2 | 0.07% | 41.0911% | 2 | -0.14% |
| | Naive | 0.19744 | 11 | 46.51% | 55.4741% | 11 | 34.82% |
| | ARIMAX | 0.16917 | 10 | 25.54% | 49.8369% | 10 | 21.12% |
| h=7 | CB2 | 0.14051 | 6 | 0.74% | 43.0370% | 4 | 0.28% |
| | CB4 | 0.14063 | 7 | 0.82% | 42.9110% | 1 | -0.01% |
| | CB8 | 0.14038 | 5 | 0.65% | 42.9930% | 3 | 0.18% |
| | Avg. Net. | 0.13948 | 1 | 0.00% | 42.9160% | 2 | 0.00% |
| | Best Net. IsNMAPE | 0.14296 | 8 | 2.49% | 43.7410% | 8 | 1.92% |
| | Best Net. IsRMSE | 0.14296 | 8 | 2.49% | 43.7410% | 8 | 1.92% |
| | GA2 | 0.14004 | 3 | 0.40% | 43.3054% | 6 | 0.91% |
| | GA4 | 0.14004 | 3 | 0.40% | 43.3054% | 6 | 0.91% |
| | GA8 | 0.14001 | 2 | 0.38% | 43.2899% | 5 | 0.87% |
| | Naive | 0.20975 | 11 | 50.38% | 59.5757% | 11 | 38.82% |
| | ARIMAX | 0.17890 | 10 | 28.26% | 53.2872% | 10 | 24.17% |
| h=8 | CB2 | 0.14570 | 7 | 1.68% | 44.7100% | 8 | 1.36% |
| | CB4 | 0.14526 | 5 | 1.37% | 44.4080% | 5 | 0.68% |
| | CB8 | 0.14528 | 6 | 1.39% | 44.5330% | 6 | 0.96% |
| | Avg. Net. | 0.14329 | 1 | 0.00% | 44.1080% | 1 | 0.00% |
| | Best Net. IsNMAPE | 0.14643 | 9 | 2.19% | 45.0560% | 9 | 2.15% |
| | Best Net. IsRMSE | 0.14577 | 8 | 1.73% | 44.6290% | 7 | 1.18% |
| | GA2 | 0.14390 | 3 | 0.43% | 44.2966% | 3 | 0.43% |
| | GA4 | 0.14407 | 4 | 0.55% | 44.3549% | 4 | 0.56% |
| | GA8 | 0.14381 | 2 | 0.36% | 44.2787% | 2 | 0.39% |
| | Naive | 0.21927 | 11 | 53.03% | 62.9785% | 11 | 42.78% |
| | ARIMAX | 0.18527 | 10 | 29.30% | 55.6429% | 10 | 26.15% |

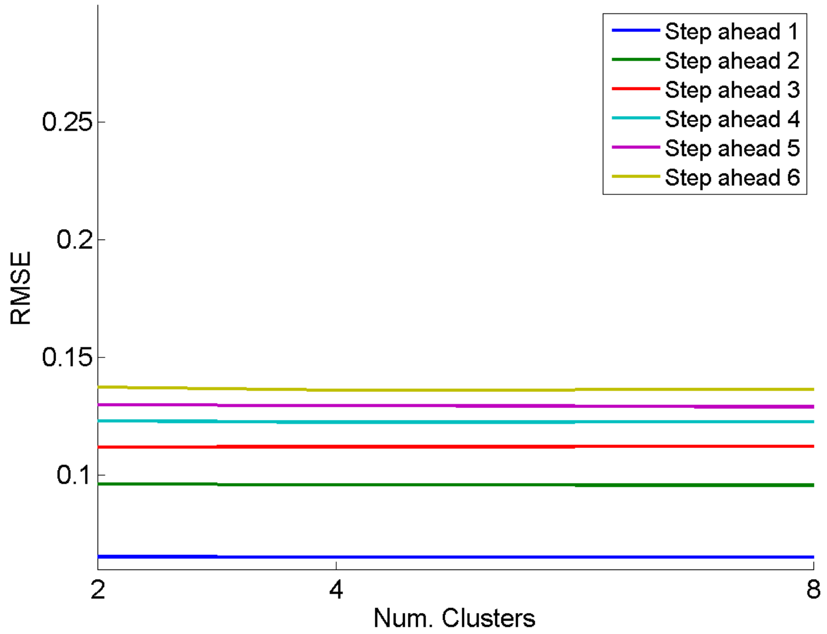
(Continued) Forecasting performance. Kaggle wind power production series.

| F. horizon | Model | RMSE | Rank | %Δ wrt Avg. | NMAPE | Rank | %Δ wrt Avg. |
|---|---|---|---|---|---|---|---|
| h=9 | CB2 | 0.14718 | 7 | 1.43% | 45.3890% | 7 | 0.90% |
| | CB4 | 0.14746 | 8 | 1.62% | 45.5140% | 8 | 1.18% |
| | CB8 | 0.14681 | 6 | 1.17% | 45.3600% | 6 | 0.83% |
| | Avg. Net. | 0.14511 | 1 | 0.00% | 44.9850% | 1 | 0.00% |
| | Best Net. IsNMAPE | 0.14755 | 9 | 1.68% | 45.9360% | 9 | 2.11% |
| | Best Net. IsRMSE | 0.14651 | 5 | 0.96% | 45.0980% | 5 | 0.25% |
| | GA2 | 0.14555 | 3 | 0.30% | 45.0052% | 2 | 0.04% |
| | GA4 | 0.14555 | 3 | 0.30% | 45.0052% | 2 | 0.04% |
| | GA8 | 0.14551 | 2 | 0.28% | 45.0276% | 4 | 0.09% |
| | Naive | 0.22745 | 11 | 56.74% | 65.7558% | 11 | 46.17% |
| | ARIMAX | 0.19231 | 10 | 32.53% | 57.7835% | 10 | 28.45% |
| h=10 | CB2 | 0.14969 | 5 | 1.51% | 45.8420% | 4 | -0.09% |
| | CB4 | 0.15000 | 7 | 1.72% | 46.4730% | 8 | 1.28% |
| | CB8 | 0.14989 | 6 | 1.65% | 46.4930% | 9 | 1.33% |
| | Avg. Net. | 0.14746 | 4 | 0.00% | 45.8850% | 5 | 0.00% |
| | Best Net. IsNMAPE | 0.15026 | 8 | 1.90% | 46.3970% | 6 | 1.12% |
| | Best Net. IsRMSE | 0.15026 | 8 | 1.90% | 46.3970% | 6 | 1.12% |
| | GA2 | 0.14729 | 3 | -0.12% | 45.6733% | 1 | -0.46% |
| | GA4 | 0.14710 | 2 | -0.24% | 45.7411% | 3 | -0.31% |
| | GA8 | 0.14703 | 1 | -0.29% | 45.7183% | 2 | -0.36% |
| | Naive | 0.23457 | 11 | 59.07% | 68.2057% | 11 | 48.64% |
| | ARIMAX | 0.19574 | 10 | 32.74% | 59.2290% | 10 | 29.08% |
| h=11 | CB2 | 0.14980 | 6 | 1.04% | 46.5880% | 6 | 0.79% |
| | CB4 | 0.14976 | 5 | 1.01% | 46.5480% | 5 | 0.70% |
| | CB8 | 0.15005 | 7 | 1.21% | 46.5890% | 7 | 0.79% |
| | Avg. Net. | 0.14826 | 4 | 0.00% | 46.2240% | 4 | 0.00% |
| | Best Net. IsNMAPE | 0.15158 | 9 | 2.24% | 46.9730% | 9 | 1.62% |
| | Best Net. IsRMSE | 0.15082 | 8 | 1.73% | 46.6610% | 8 | 0.95% |
| | GA2 | 0.14778 | 3 | -0.32% | 46.1489% | 3 | -0.16% |
| | GA4 | 0.14777 | 1 | -0.33% | 46.1334% | 1 | -0.20% |
| | GA8 | 0.14777 | 1 | -0.33% | 46.1334% | 1 | -0.20% |
| | Naive | 0.24092 | 11 | 62.50% | 70.0919% | 11 | 51.64% |
| h=12 | CB2 | 0.15264 | 9 | 1.98% | 47.2680% | 8 | 0.95% |
| | CB4 | 0.15151 | 8 | 1.22% | 46.9790% | 5 | 0.34% |
| | CB8 | 0.15119 | 6 | 1.01% | 47.0980% | 7 | 0.59% |
| | Avg. Net. | 0.14968 | 4 | 0.00% | 46.8220% | 4 | 0.00% |
| | Best Net. IsNMAPE | 0.15110 | 5 | 0.95% | 47.0290% | 6 | 0.44% |
| | Best Net. IsRMSE | 0.15150 | 7 | 1.22% | 47.2970% | 9 | 1.01% |
| | GA2 | 0.14944 | 1 | -0.16% | 46.7998% | 1 | -0.05% |
| | GA4 | 0.14944 | 1 | -0.16% | 46.7998% | 1 | -0.05% |
| | GA8 | 0.14944 | 1 | -0.16% | 46.7998% | 1 | -0.05% |
| | Naive | 0.24714 | 11 | 65.11% | 71.9670% | 11 | 53.70% |
| | ARIMAX | 0.20421 | 10 | 36.43% | 61.9097% | 10 | 32.22% |

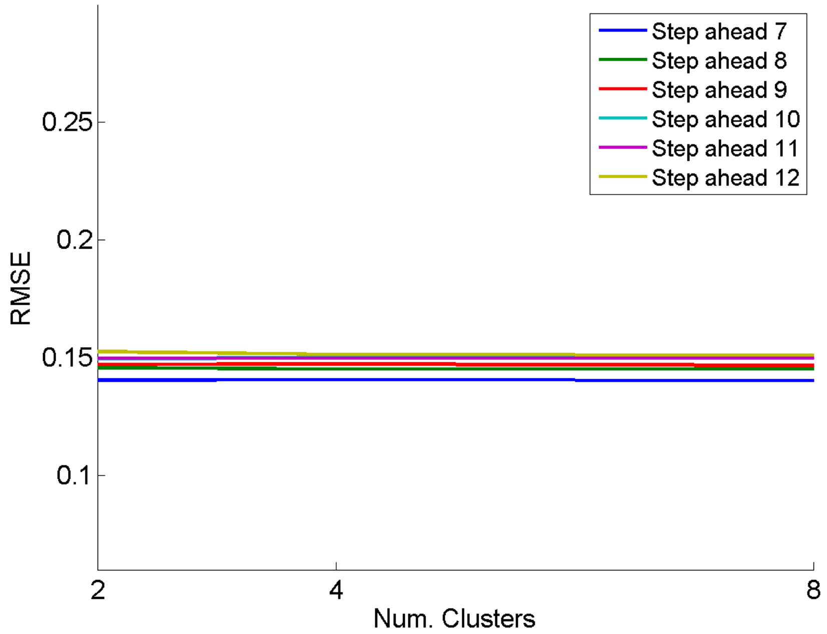
Not surprisingly, the average of NNs performs relatively well.

Table 4.21 summarises some CB models. Clusters are very homogeneous both in coefficient ranges and the number of model per cluster. This homogeneity is in accordance with previous observations. NN models for different forecast horizons have different structures but the use of such structure by CB combinations throughout the different horizons is rather similar, as it can also be observed in the prediction intervals (Figure 4.27).

The cluster validity indexes, suggest that the best structural differentiation in cluster configuration is found in CB2 models (Table 4.22). It is observed also that as the number of clusters increases, the clear separation between cluster decreases. However some CB8 models, with a high number of clusters, performed better than CB models with less clusters.

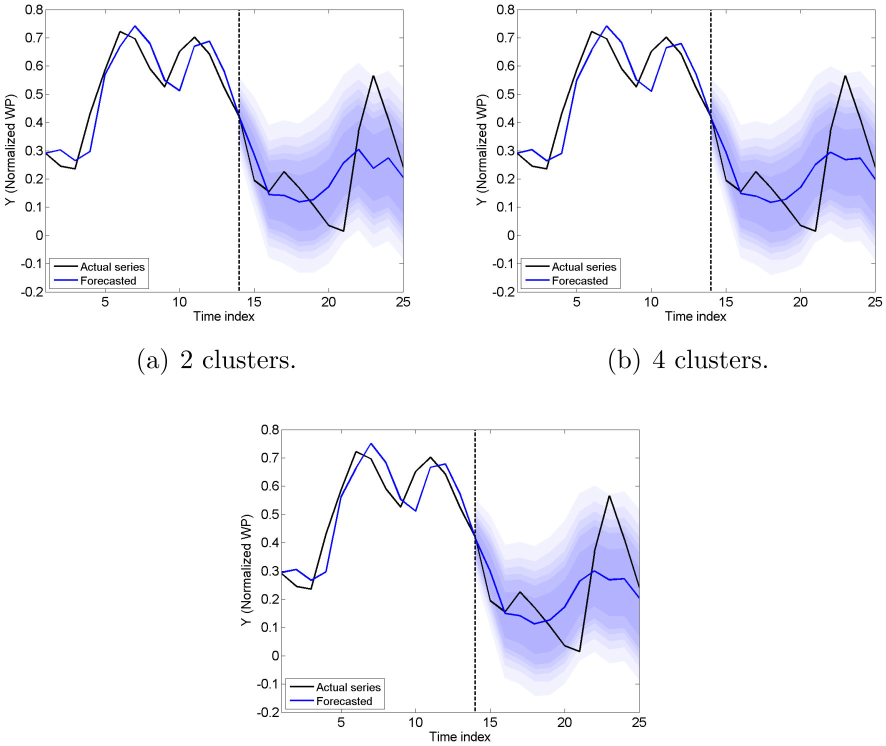


(a) Steps ahead 1 to 6. RMSE.



(b) Steps ahead 7 to 12. RMSE.

Figure 4.26: Out-of-sample RMSE vs. number of clusters. Kaggle wind power production series.



(a) 2 clusters.

(b) 4 clusters.

(c) 8 clusters.

Figure 4.27: Forecast intervals for Kaggle wind power time series.

The graphs cover the period for $t - 12 \leq t \leq t + H$ where $t$ is the last observation of the in-sample period and $H = 12$ is the number of forecast horizons. Therefore, the last 13 observations of the in-sample period are included along with the first 12 forecasts in the out-of-sample period. The horizontal axis indexes these hours as $1, \ldots, 25$. The shades, from lighter to darker, correspond to $\alpha$ levels 0.95, 0.90, 0.85, 0.80, 0.75 and 0.60.

Ljung-Box test on serial correlation for forecast errors (Table 4.24) shows that for the first forecast horizon, CB, GA and the average of NN seem to capture well the dynamics of the series. Yet, for higher horizons all models present serial correlated errors. Normality of the errors has been rejected for all models.

Table 4.21: Coefficients for structural combination of NN for Kaggle wind power production series.

| h | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 1 | $\alpha_1$ | -0.0010 | 0.0895 | 0.2424 | 0.1828 | 0.3799 | 0.1083 | |
| | $\alpha_2$ | -0.0004 | 0.4269 | 0.0263 | 0.0105 | 0.3656 | 0.1727 | |
| | $\alpha_3$ | -0.0008 | 0.0257 | 0.2009 | 0.4336 | 0.2005 | 0.1397 | |
| | $\alpha_4$ | -0.0002 | 0.2245 | -0.0448 | 0.2578 | 0.2553 | 0.3123 | |
| | $\Phi$ | 0.2720 | 0.1816 | 0.2705 | 0.2745 | | | |
| 2 | $\alpha_1$ | -0.0003 | 0.2133 | -0.2232 | 0.1395 | 0.5067 | 0.3655 | |
| | $\alpha_2$ | -0.0014 | 0.2501 | 0.4438 | 0.4358 | -0.1289 | | |
| | $\alpha_3$ | 0.0005 | -0.1831 | 0.2446 | 0.4291 | 0.0416 | 0.4695 | |
| | $\alpha_4$ | 0.0001 | 0.0511 | 0.2142 | -0.1937 | 0.5291 | 0.3993 | |
| | $\alpha_5$ | 0.0004 | 0.0953 | 0.3450 | -0.0108 | 0.2331 | 0.3372 | |
| | $\alpha_6$ | -0.0008 | 0.1824 | 0.4132 | 0.1223 | -0.0091 | 0.2946 | |
| | $\alpha_7$ | -0.0011 | 0.1116 | 0.1844 | 0.4762 | 0.2447 | -0.0104 | |
| | $\Phi$ | 0.1400 | 0.1494 | 0.1477 | 0.1491 | 0.1317 | 0.1462 | 0.1520 |
| 6 | $\alpha_1$ | -0.0031 | 0.0714 | -0.1468 | 0.3229 | 0.3042 | 0.4525 | |
| | $\alpha_2$ | 0.0012 | 0.1905 | -0.2177 | 0.1612 | 0.3663 | 0.5033 | |
| | $\alpha_3$ | -0.0063 | -0.1294 | 0.0917 | 0.5791 | 0.0665 | 0.3996 | |
| | $\alpha_4$ | -0.0023 | -0.0941 | 0.2461 | 0.1162 | 0.3234 | 0.4159 | |
| | $\alpha_5$ | -0.0039 | -0.3187 | 0.4621 | 0.1318 | 0.4293 | 0.3020 | |
| | $\alpha_6$ | -0.0016 | 0.0814 | 0.5232 | 0.3197 | -0.2932 | 0.3770 | |
| | $\Phi$ | 0.1688 | 0.1741 | 0.1693 | 0.1728 | 0.1701 | 0.1719 | |
| 12 | $\alpha_1$ | -0.0064 | -0.2765 | 0.4655 | 0.1362 | 0.4513 | 0.2572 | |
| | $\alpha_2$ | -0.0013 | -0.2529 | 0.2052 | 0.6912 | -0.2086 | 0.5819 | |
| | $\alpha_3$ | -0.0005 | 0.0209 | -0.1341 | 0.3927 | 0.3231 | 0.3993 | |
| | $\alpha_4$ | -0.0011 | -0.4656 | 0.3210 | 0.4589 | 0.6874 | | |
| | $\alpha_5$ | -0.0037 | 0.5154 | 0.4310 | -0.6665 | 0.4315 | 0.3067 | |
| | $\alpha_6$ | -0.0023 | 0.5206 | 0.2587 | 0.0191 | -0.2661 | 0.4741 | |
| | $\alpha_7$ | 0.0020 | -0.3528 | 0.7391 | 0.3773 | 0.2324 | | |
| | $\Phi$ | 0.1474 | 0.1474 | 0.1470 | 0.1404 | 0.1500 | 0.1435 | 0.1489 |

$MaxC = 8$ is the maximum number of clusters. For each selected horizon, $h$, $\alpha_i$ are the coefficients applied to point-forecasts from models in cluster $i$ and $\Phi$ are the weights applied to the outputs from clusters.

In general, the CB combination approach adopted here provided a very consistent performance, although with difficulties to outperform the average of NN forecasts. However, there is a gain in structural combinations, as evidenced by the good performance of GA structural combination.

Although results from other authors using the same data set are not available

Table 4.22: Cluster validity indexes. Series: Kaggle wind power production.

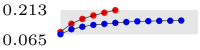
| | Maximum number of clusters: 2 | | | |
|---|---|---|---|---|
| *h* | PC | MPC | MDO | Final num. clusters |
| 1 | 1.0000 | NA | 1.0000 | 1 |
| 2 | 0.5087 | 0.0175 | 0.9000 | 2 |
| 3 | 0.5013 | 0.0026 | 0.1000 | 2 |
| 4 | 0.5008 | 0.0017 | 0.0000 | 2 |
| 5 | 0.5007 | 0.0014 | 0.0000 | 2 |
| 6 | 1.0000 | NA | 1.0000 | 1 |
| 7 | 0.5008 | 0.0015 | 0.0000 | 2 |
| 8 | 0.5007 | 0.0014 | 0.0000 | 2 |
| 9 | 0.5036 | 0.0072 | 0.3000 | 2 |
| 10 | 0.5027 | 0.0054 | 0.2000 | 2 |
| 11 | 0.5021 | 0.0042 | 0.3000 | 2 |
| 12 | 1.0000 | NA | 1.0000 | 1 |

| | Maximum number of clusters: 4 | | | |
|---|---|---|---|---|
| *h* | PC | MPC | MDO | Final num. clusters |
| 1 | 0.2509 | 0.0011 | 0.0000 | 4 |
| 2 | 0.3342 | 0.0014 | 0.0000 | 3 |
| 3 | 0.2521 | 0.0028 | 0.0000 | 4 |
| 4 | 0.2501 | 0.0002 | 0.0000 | 4 |
| 5 | 0.2505 | 0.0006 | 0.0000 | 4 |
| 6 | 0.5005 | 0.0011 | 0.0000 | 2 |
| 7 | 0.2521 | 0.0027 | 0.0000 | 4 |
| 8 | 0.3345 | 0.0018 | 0.0000 | 3 |
| 9 | 0.3352 | 0.0028 | 0.0000 | 3 |
| 10 | 0.2505 | 0.0006 | 0.0000 | 4 |
| 11 | 0.2508 | 0.0010 | 0.0000 | 4 |
| 12 | 0.3344 | 0.0016 | 0.0000 | 3 |

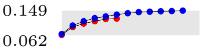
| | Maximum number of clusters: 8 | | | |
|---|---|---|---|---|
| *h* | PC | MPC | MDO | Final num. clusters |
| 1 | 0.2536 | 0.0048 | 0.0000 | 4 |
| 2 | 0.1431 | 0.0003 | 0.0000 | 7 |
| 3 | 0.1697 | 0.0036 | 0.0000 | 6 |
| 4 | 0.1251 | 0.0001 | 0.0000 | 8 |
| 5 | 0.1671 | 0.0005 | 0.0000 | 6 |
| 6 | 0.1668 | 0.0001 | 0.0000 | 6 |
| 7 | 0.1674 | 0.0009 | 0.0000 | 6 |
| 8 | 0.1251 | 0.0001 | 0.0000 | 8 |
| 9 | 0.1431 | 0.0003 | 0.0000 | 7 |
| 10 | 0.1431 | 0.0003 | 0.0000 | 7 |
| 11 | 0.1431 | 0.0003 | 0.0000 | 7 |
| 12 | 0.1431 | 0.0002 | 0.0000 | 7 |

*h* denotes the forecast horizon, PC denotes the Partition Coefficient, MPC denotes the Modified Partition Coefficient, and MDO denotes the Membership Degree Optimum. Values closer to 1 are preferable.

in great detail (Lee & Scholtes, 2014; Hong et al., 2014), comparisons can be made with similar studies by using the normalised RMSE. Wind power measurements, $p_i$, are frequently reported as $p_i/c$ where $c$ is the nominal capacity of the wind farm. This normalisation facilitates the comparison of results among different approaches[9]. The power measurements from the data-set used here were normalised in this way by their providers (Hong et al., 2014). The RMSEs obtained through NN structural combinations (considering CB, GA, the average forecast and the best models) for the normalised wind power is in the interval $[0.06509, 0.15264]$. In general, these figures are similar to the results obtained in recent wind power forecasting articles, as highlighted below.

Zhao et al. (2016) developed a forecasting model based on extreme learning machine and back-forecasting (in addition to the forward mechanisms commonly adopted in the literature). They normalised the RMSE by using the maximum power measurement instead of the nominal capacity, which produces a pessimistic error metric. Their normalised RMSE for steps 1 to 6 ranges from 0.079 to 0.21, approximately. The lowest value (for $h = 1$) is very close to the RMSE obtained with ensemble models here, but their maximum value (for $h = 6$) is worst than the RMSE obtained with GA4 for the same horizon: 

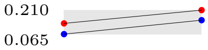
(in red are the RMSEs from Zhao et al., 2016, and in blue the RMSEs from GA4 combination).

Liang et al. (2016) proposed a framework to forecast wind power with support vector machines and historic error correction mechanisms. Their normalised RMSE, reported for steps 1 to 6, are similar to the RMSEs obtained here and marginally better (in terms of the send and third decimal place): 

(in red their RMSEs and in blue the RMSEs with GA4 combination).

Yan et al. (2016) proposed a model to forecast wind power based on Gaussian processes. They forecasted hourly wind power for up to 12 hours ahead and reported

[9] Other error measurements based on proportions, such as MAPE and NMAPE, are less standard.

normalised RMSE ranging from around 0.13 to 0.21, for their best model. An approximate comparison can be made only for $h = 1$ and $h = 12$:  (in red are the RMSEs from Yan et al., 2016, and in blue the RMSEs from GA4 combination).

Finally, Hong et al. (2014) collected the results from various participants in the Energy Forecasting Competition (Kaggle, 2012), who used the data set presented here plus data for other wind farms. Forecast accuracy was measured with a single RMSE, covering all intervals (of 48 hours) with missing power measurements that participants forecasted: there was no rolling window forecasting to produce performance figures for specific horizons. The reported RMSE ranges from 0.145 (for the best submission) to 0.18 (for the worst submission in the selection). As their focus was in two days, comparisons are not possible, but it can be seen that the error figure for the best submission is similar to the figures obtained for $h = 12$.

Table 4.23: Comparisons by forecast horizon. Series: Kaggle wind power.

| RMSE | | | | |
|---|---|---|---|---|
| | Avg. | Bst. IsNMAPE | Bst. IsRMSE | ARIMAX |
| CB2 | 0 | 10 | 10 | 12 |
| GA2 | 4 | 12 | 12 | 12 |
| CB4 | 1 | 11 | 10 | 12 |
| GA4 | 4 | 12 | 12 | 12 |
| CB8 | 1 | 11 | 11 | 12 |
| GA8 | 4 | 12 | 12 | 12 |
| NMAPE | | | | |
| | Avg. | Bst. IsNMAPE | Bst. IsRMSE | ARIMAX |
| CB2 | 2 | 9 | 10 | 12 |
| GA2 | 5 | 12 | 12 | 12 |
| CB4 | 2 | 11 | 10 | 12 |
| GA4 | 6 | 12 | 12 | 12 |
| CB8 | 1 | 9 | 10 | 12 |
| GA8 | 5 | 12 | 12 | 12 |

Number of forecast horizons for which CB and GA combinations outperform different benchmark models. Avg. stands for the average of NN in the ensemble; Bst. IsNMAPE stands for the best NN in terms of in-sample NMAPE in the ensemble; Bst. IsRMSE stands for the best NN in terms of in-sample RMSE in the ensemble.

Table 4.24: Ljung-Box test. Series: Kaggle wind power.

| | | Forecast horizon | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Model | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| 2C | CB | | * | * | * | * | * | * | * | * | * | * | * |
| | GA | | * | * | * | * | * | * | * | * | * | * | * |
| 4C | CB | * | * | * | * | * | * | * | * | * | * | * | * |
| | GA | | * | * | * | * | * | * | * | * | * | * | * |
| 8C | CB | | * | * | * | * | * | * | * | * | * | * | * |
| | GA | | * | * | * | * | * | * | * | * | * | * | * |
| | Avg. | | * | * | * | * | * | * | * | * | * | * | * |
| | Best Net. IsNMAPE | * | * | * | * | * | * | * | * | * | * | * | * |
| | Best Net. IsRMSE | * | * | * | * | * | * | * | * | * | * | * | * |
| | ARIMAX | * | * | * | * | * | * | * | * | * | * | * | * |

Ljung-Box test for serial correlation (with 95% confidence level) for Kaggle wind power production series. The rejection of the hypothesis of independent forecast errors is indicated with *.

## 4.8 Study with an Electricity Demand Time Series [10]

A time series of electricity demand was used to asses the performance of CB combination approach. The series contains hourly observations in Rio de Janeiro covering the period from Sunday 5 May 1996 to Saturday 30 November 1996 (Figure 4.28). It has been used by Taylor et al. (2006) to evaluate the performance of various univariate models, including a NN, which was implemented according to Darbellay & Slama (2000).

The direct approach of fitting different NNs for different forecast horizons, as in previous studies in this chapter, led to a performance markedly different from results obtained by Taylor et al. (2006). In their study, the authors fitted a NN with input lags 1, 2, 24, 25, 48, 72, 96, 120, 144, 168, 192, 216, 240, 264, 312, 336 and forecasted the differences in an iterative manner. Further experiments with this setting provided better results than the direct approach and therefore it

[10] Study to be presented at the Workshop on Data Mining for Energy Modelling, 12 December 2016, Barcelona, Spain. Rendon, J. and de Menezes, L.M. (2016) - "Structural combination of neural network models", DaMEMO 2016 Proceedings, IEEE Society Press.

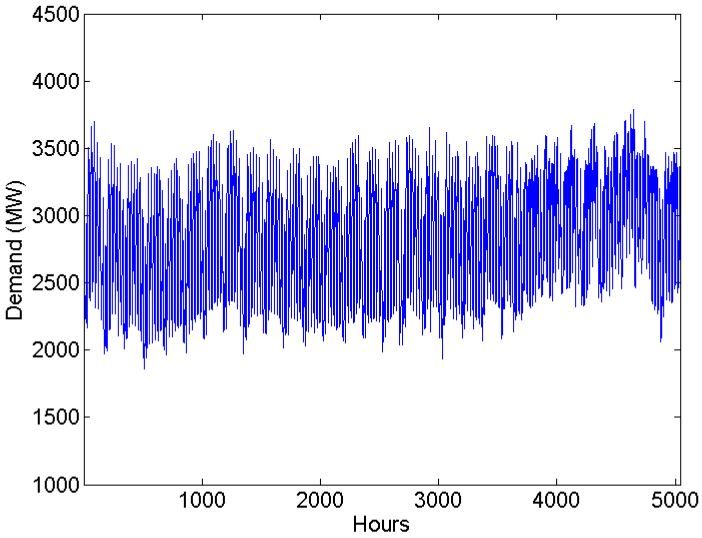


Figure 4.28: Hourly electricity demand in Rio de Janeiro for Sunday, 5 May 1996 to Saturday, 30 November 1996. Original Series.

was adopted for the present study. This requires an adaptation of the structural combination presented in Section 4.3.2.1, as follows.

### 4.8.1 Variant of the Model for an Iterative Forecast

When the individual NN models are used to issue forecasts in an iterative manner, the clustering procedure takes into account the in-sample one-step-ahead forecasts produced by NNs in order to calculate the loss function minimised during the optimisation.

Each network produces forecasts $\hat{y}_{t+1}, \ldots, \hat{y}_{t+h}$ based on a set of inputs (e.g lagged values of the series), $x_t = \{y_t, y_{t-1}, \ldots, y_{t-p}\}$, and the previously foreasted values. Therefore, $\hat{y}_{t+h} = f(x_t; \hat{y}_{t+1}, \ldots, \hat{y}_{t+h-1})$ The combined forecast output for $t + h$ is calculated based on a combination of the forecasts made by the clustered

NNs for such horizon.

$$\hat{y}_{t+h} = \sum_{i=1}^{n} \phi_i \hat{y}_{C_i,t+h} \tag{4.26}$$

Where $\hat{y}_{C_i,t+h}$ is the output from cluster $i$ for step $t+h$:

$$\hat{y}_{C_i,t+h} = \alpha_{i,0} + \alpha_{i,1}\hat{y}_{i,1,t+h} + \alpha_{i,2}\hat{y}_{i,2,t+h} + \ldots + \alpha_{i,L}\hat{y}_{i,L,t+h} \tag{4.27}$$

$\hat{y}_{i,1,t+h}, \hat{y}_{i,2,t+h}, \ldots$ are the forecasts for $t+h$ produced by models selected within cluster $i$.

The coefficients $\phi_k$ are calculated as an average of the normalised weights of vectors (models in vectorial form):

$$u_i(v) = e^{-\frac{D_i^2(v)}{\sum_{j=1}^{n} D_j^2(v)}} \tag{4.28}$$

$$w_i(v) = \frac{u_i(v)}{\sum_{j=1}^{n} u_j(v)} \tag{4.29}$$

$$\phi_k = \frac{\sum_{m \in C_k} w_m(v_{C_k})}{N_k} \tag{4.30}$$

where $C_k$ denotes cluster $k$, $v_{C_k}$ is the centre of such cluster and $N_k$ is the number of models in it.

### 4.8.2 Preliminary Analysis and Specification of Individual Models for the Ensemble Based on Iterative Forecasts

As in the study involving the wind power series, the generating process of the demand series is unknown. Therefore, variation in the series was created, during the sensitivity analysis, by adding noise to the original series. It is distributed as $N(0, 0.1\sigma_b)$, where $\sigma_b$ is the standard deviation of the bootstrapped series. The magnitude of such standard deviation, when compared to the interquartile range of

Table 4.25: Factor configuration.

| **Factor** | **Symbol** | **Levels** |
|---|---|---|
| Number of inputs | NI | $\mathbf{1}, \ldots, \mathbf{16}$, being 16 the number of lags. |
| Number of hidden layers | NL | 1 |
| Number of hidden units | NU | $\mathbf{2}, \mathbf{6}, \mathbf{10}$ |
| Activation function for hidden nodes | AF1 | Tangent Sigmoid |
| Activation function for the output node | AF2 | Linear |
| Initial values for the weights | W0 | Values in the range [-2 2] established by the Nguyen-Widrow algorithm (there is a degree of randomness) |
| Training algorithm | TA | Backpropagation with Levenberg-Marquardt optimisation. |
| Stopping criteria | SC | * The maximum number of epochs (repetitions) is reached: 4000.<br>* The maximum amount of time is exceeded: $\infty$<br>* Performance is minimised to the goal: 0<br>* The performance gradient falls below $min_{grad}$ : $10^{-10}$<br>* $\mu$ exceeds $\mu_{max} = 10^3$<br>* Validation performance has increased more than $max_{fail}$ times since the last time it decreased (when using validation): 6 |
| Data normalisation | DN | Yes |
| Combination coefficient ($\mu$) | MU | 0.001 |
| Prune units | PU | **Yes, No** |
| Prune input variables | PI | No |
| Sample size | SS | 5040 |
| Data configuration for training, validation and testing (training + validation = in-sample data; testing = out-of-sample data) | DC | Conf. 1: $Ntr$ = 3024, $Nva$ = 336 $Nte$ = 1680 |
| Extreme values treated | EV | No. |
| Sampling method | SM | **block, cross-validated** |
| Forecast approach | FA | Iterative |

In bold are the factors which vary in the study.

the time series, falls into the low level identified in the research by Barrow et al. (2010)[11].

In their study, Darbellay & Slama (2000), when using NN to forecast demand time series, indicated that it was unnecessary to use more than 10 hidden units and, therefore, they selected models with such factor ranging from 6 to 10. In a sensitivity analysis for seasonal time series, Crone & Dhawan (2007), by studying the average errors over a set of series, found that a good and robust performance could be achieved by using 3, 5, 6 or 9 hidden units. These findings inspired the selection of levels for NU (number of hidden units) in the preliminary analysis. Pruning of hidden layer weights was not adopted as it was not found helpful in improving performance for seasonal and double-seasonal series, as seen in Chapter 3.

Figure 4.29 depicts the original series with the replicas generated and Figure 4.30 the average out-of-sample MAPE of NN models for each forecast horizon. Table 4.26 shows the performance for the architecture with the lowest average out-of-sample MAPE, which is clearly simple in terms of neurons but complex in terms of inputs.

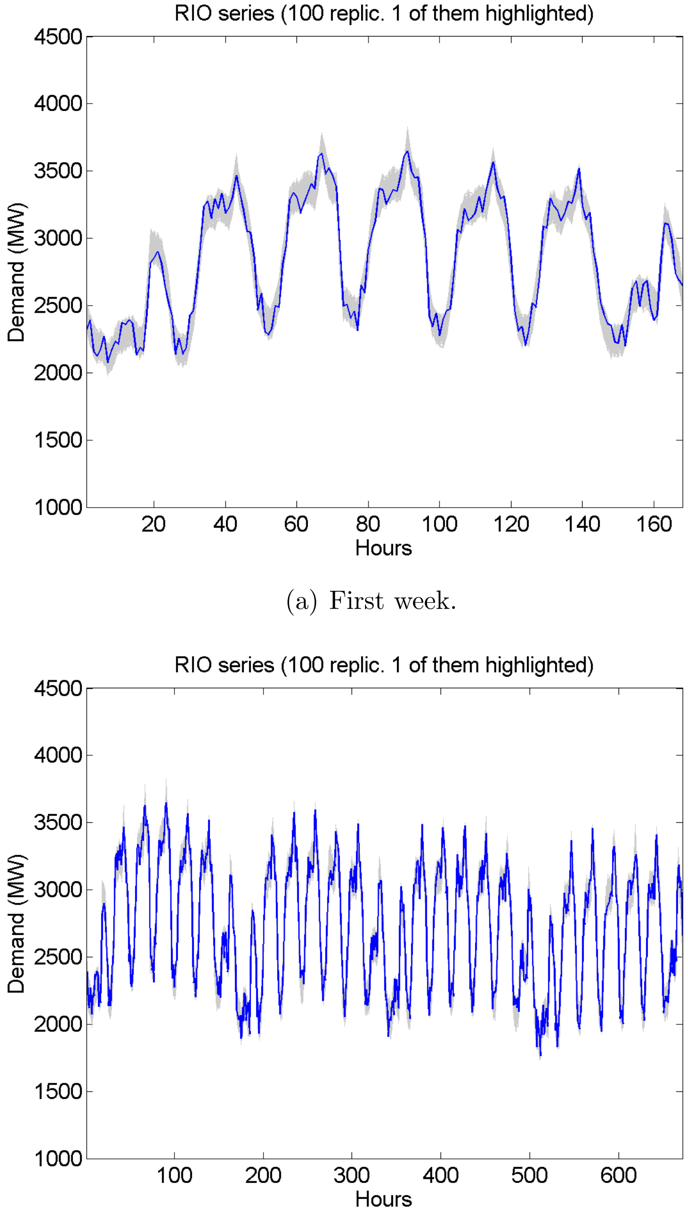
Table 4.26: Model with the lowest average out-of-sample MAPE.

| Forecast horizon | Average MAPE% |
|---|---|
| 1 | 2.4123 |
| 2 | 2.7485 |
| 3 | 3.0102 |
| 4 | 3.2032 |
| 5 | 3.3729 |
| 6 | 3.5216 |
| 7 | 3.6589 |
| 8 | 3.7688 |
| 9 | 3.8778 |
| 10 | 3.9876 |
| 11 | 4.0980 |
| 12 | 4.1812 |

Configuration: 16 inputs and 2 hidden units.

The main effects graphs (in Figure 4.31) show a clear sensitivity of both error metrics to the number of inputs (NI) and insensitivity to the number of neurons

[11] In Chapter 3 it was discussed how the authors considered three different levels of noise when developing ensembles of NNs for synthetic data.



(a) First week.

(b) First four weeks.

Figure 4.29: Replications of Rio de Janeiro electricity demand time series with added noise. One of the replications is highlighted in blue.

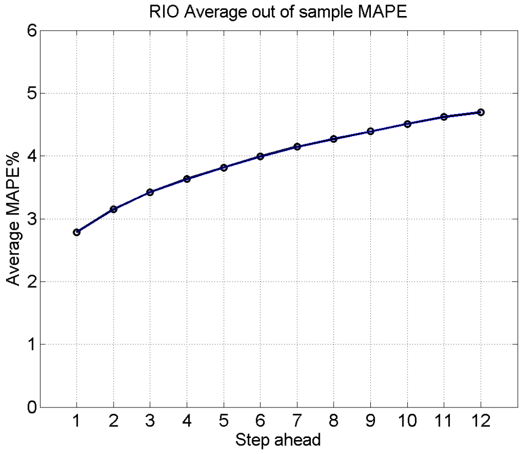


Figure 4.30: Average MAPE%.

(NU). All selected inputs are relevant, which is visible in the almost continuously decreasing error. ANOVA, Kruskal-Wallis and Jonckheere-Terpstra tests show a significant effect of NI over both error metrics in all horizons with improvements in fit and forecast accuracy as more inputs are used. Factor NU, on the other hand, has a significant effect only in some of the last horizons, and to the detriment of performance. The assessment of serial correlation revealed a generalised failure of different model configurations to fully capture the dymacis of the series.

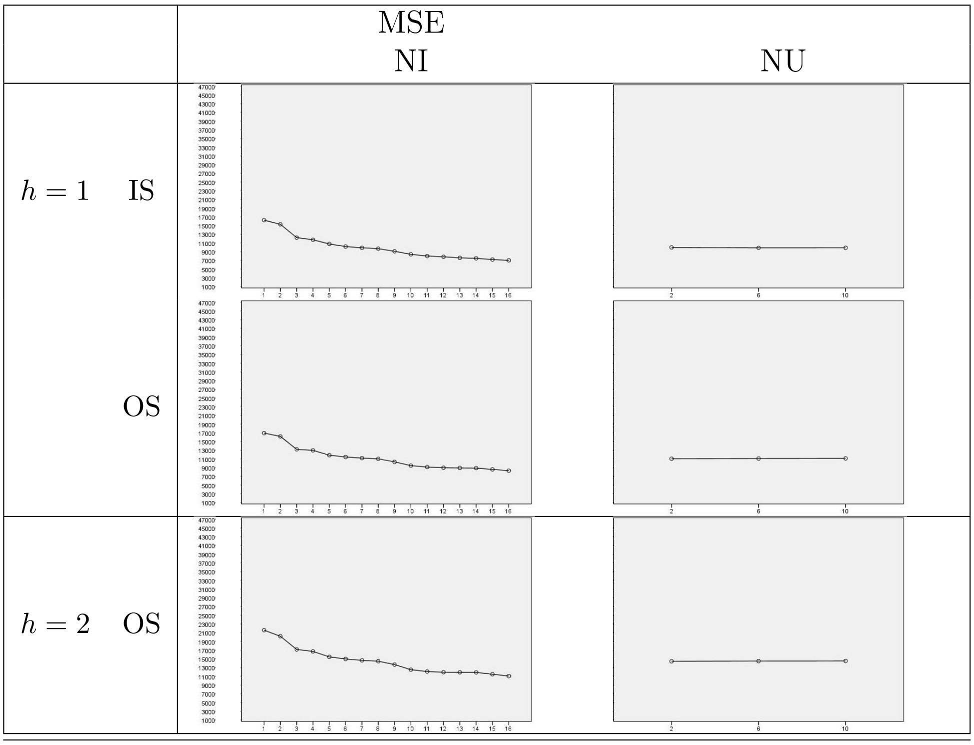
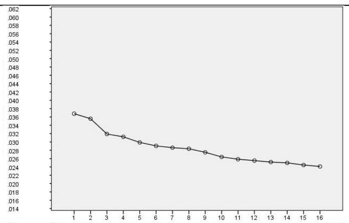
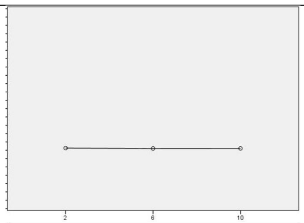
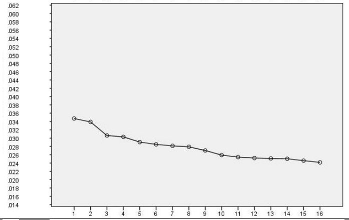
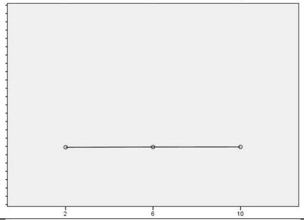
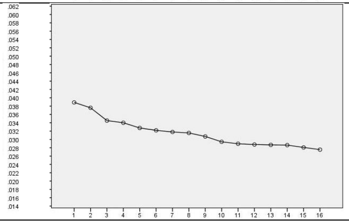
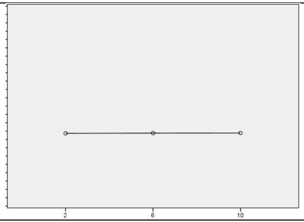


Figure 4.31: Main effects (sub-sample). Series: RIO. IS stands for in-sample; OS stands for out-of-sample.

| | | MSE | |
|---|---|---|---|
| | | NI | NU |
| $h = 1$ | IS |  |  |
| | OS |  |  |
| $h = 2$ | OS |  |  |

(Continued) Main effects (sub-sample). Series: RIO. IS stands for in-sample; OS stands for out-of-sample.

### 4.8.3 Combination of Models

The preliminary analysis suggested that an NN architecture with 16 inputs (corresponding to all lags considered) and 2 neurons would be the best choice. Such architecture is used here with the clustering procedure explained in Section 4.8.1 to test the structural combination of forecasting models. Due to the existence of extreme values in the out-of-sample performance of NNs, the ensemble to perform the combination was built with the over-produce and choose approach (Mendes-Moreira et al., 2012): 150 NNs where generated and 50 selected. The forecast performance for $h = 12$ was assessed with a rolling window in the in-sample period and the best models were used to conduct the structural combination.

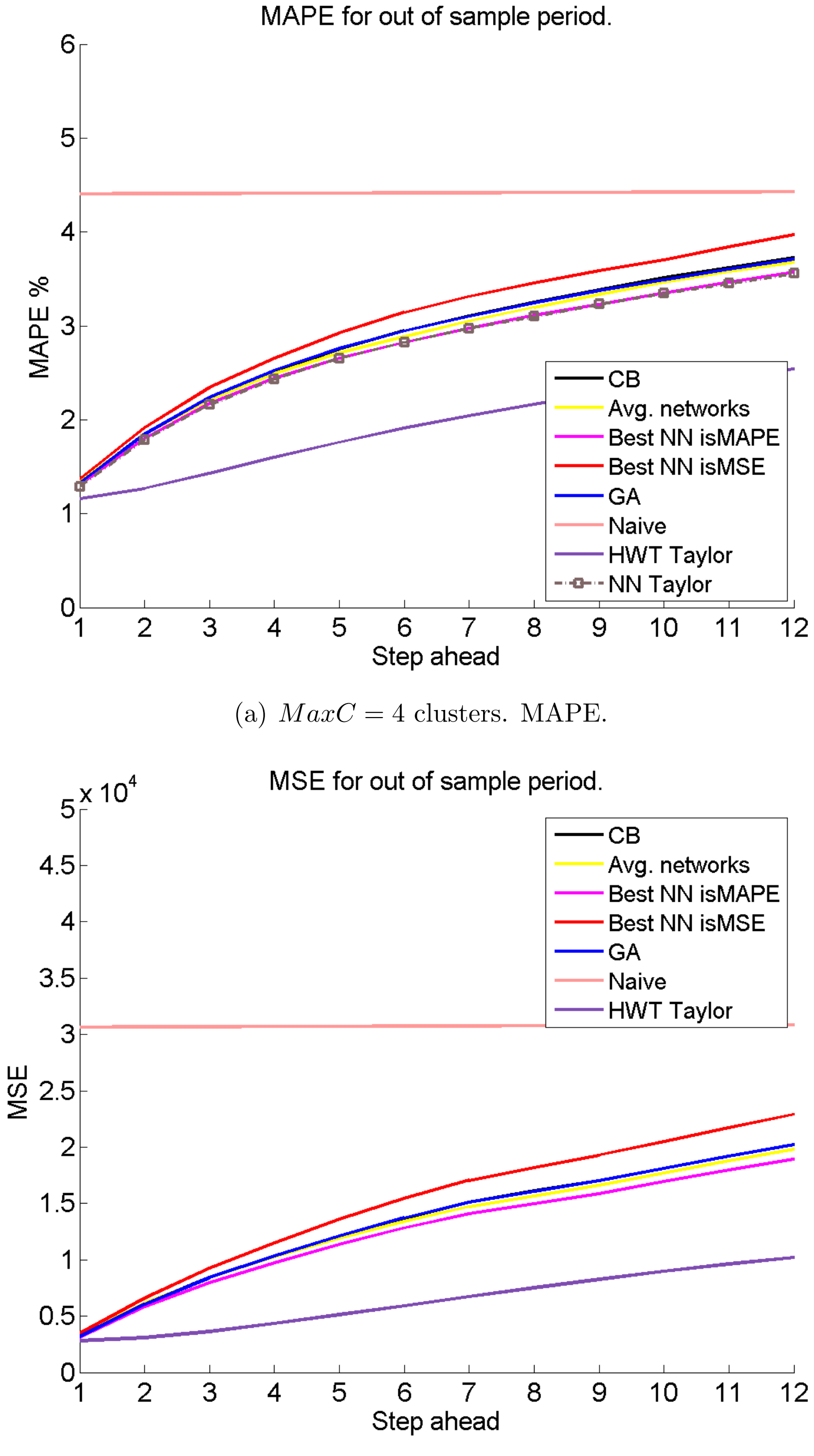
Results are compared with those obtained by Taylor et al. (2006) with a Holt-Winters-Taylor (HWT) exponential smoothing method and a NN. Comparisons are made by using MSE and MAPE error metrics as outlined in section 4.4.1. However, only the MAPE figures are available for the NN used by the authors.

Figures 4.32 and 4.33, along with Figures C.9 and C.10 from Appendix C, show the out-of-sample MAPE and MSE for the different CB models and the selected benchmarks and Table 4.27 provides a ranking of models for every forecast horizon, with the first position corresponding to the lowest metric value, and a percentage of error difference with respect to the forecast average of all the NNs in the ensemble.

The ranking of models is homogeneous with the HWT model performing best, followed by the NN in Taylor et al. (2006), the best NN in the ensemble, the CB models, the GA benchmark and finally the average forecast from the ensemble. For $h \geq 3$ the average outperforms the GA. All NN models and NN-based combinations have a very similar performance (which is also visible in the forecast intervals in Figure 4.35). The base NN of the ensemble seems to be relatively well specified and this, along with the initial filtering of models, possibly makes it harder for combinations, that include divers models, to defeat the best models.

Model structural differentiation of clusters is low (Table 4.31), which coincides with a better performance of CB2 that was reduced to a single-cluster configuration. In this case, model structural diversity seems low. On the other hand, coefficients in Table 4.29 show that CB combinations, for higher number of clusters, have heterogeneous weighting of cluster outputs. This suggests that little structural differentiation might still be accompanied by diversity in forecasts.

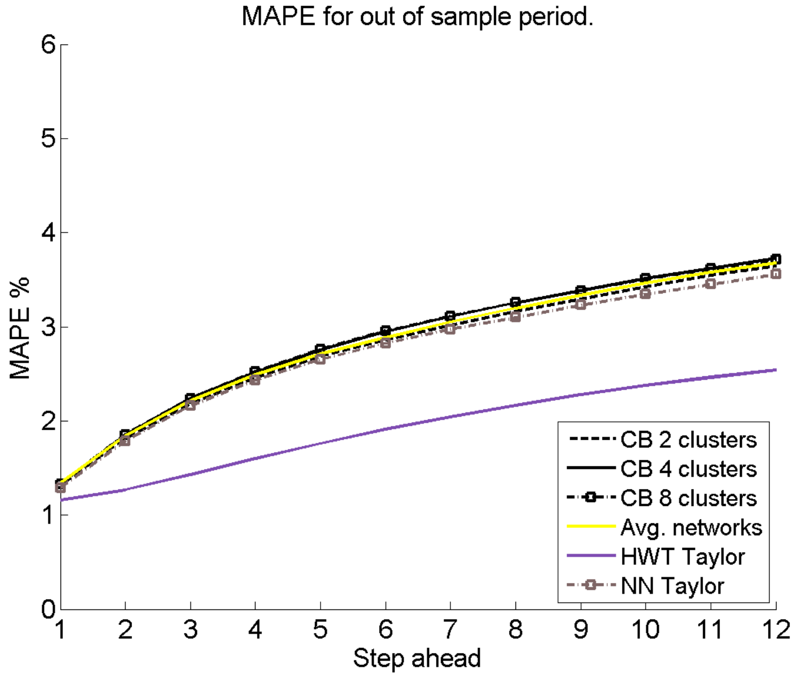
The assessment of serial correlation in forecast errors showed that all NN based models left some dynamics of the series unexplained. The Lilliefors test on normality supports this finding, but Jarque-Bera test (Tables 4.28) revealed normality in errors from GA models for some horizons.



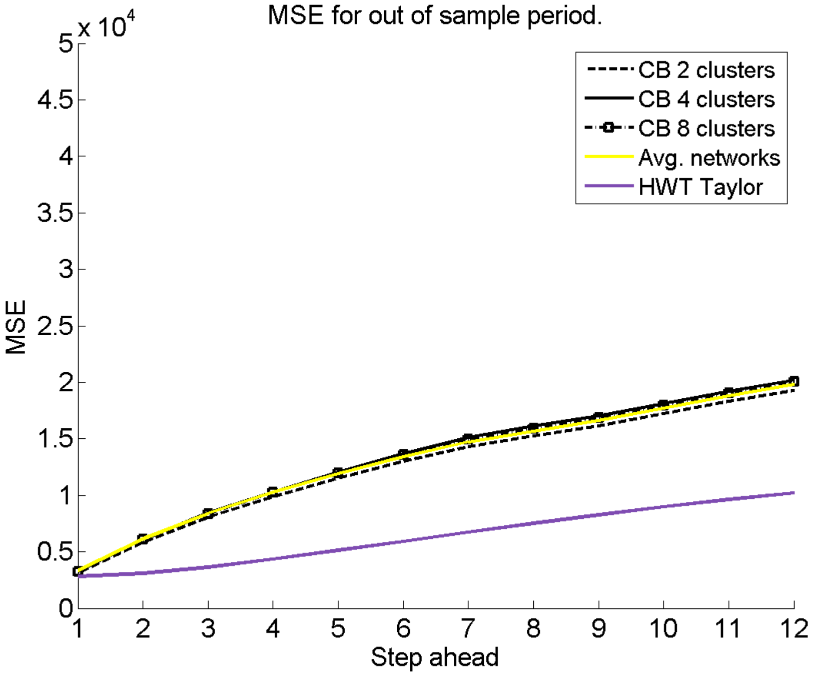
(a) $MaxC = 4$ clusters. MAPE.

(b) $MaxC = 4$ clusters. MSE.

Figure 4.32: Out-of-sample MAPE and MSE for Rio de Janeiro electricity demand series.

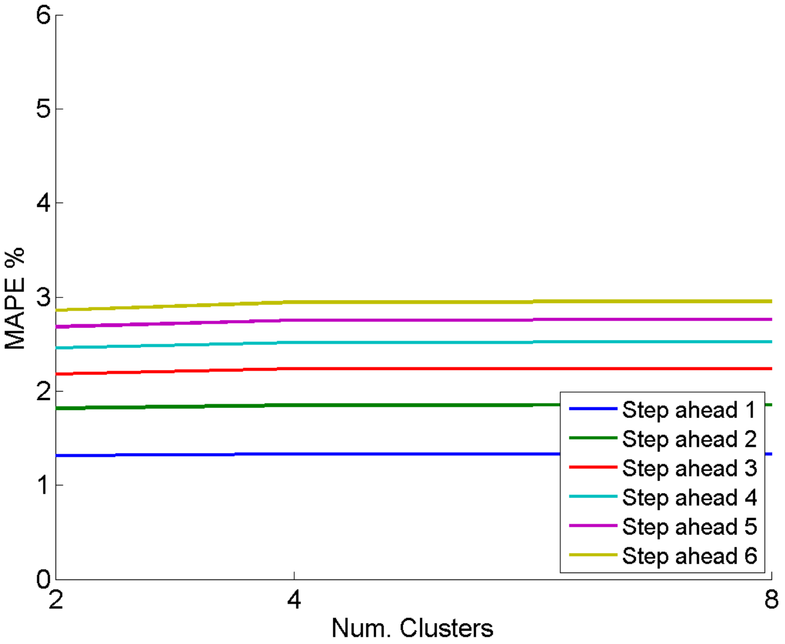


(a) Comparison with Avg. and HWT. MAPE.

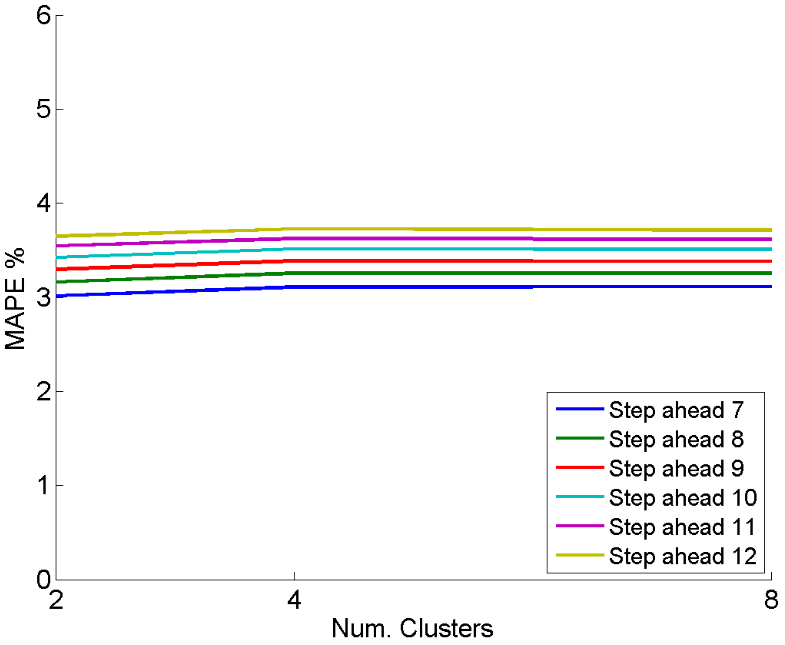


(b) Comparison with Avg. and HWT. MSE.

Figure 4.33: Out-of-sample MAPE and MSE for Rio de Janeiro electricity demand series.

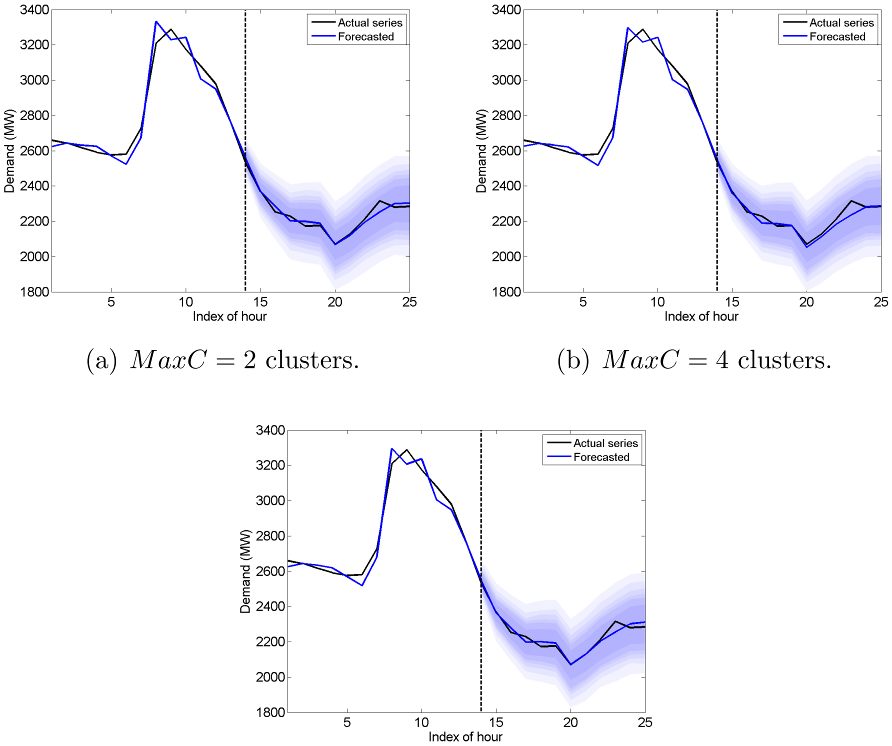


(a) Steps ahead 1 to 6. MAPE.



(b) Steps ahead 7 to 12. MAPE.

Figure 4.34: Out-of-sample MAPE vs. number of clusters. Rio de Janeiro electricity demand series.



(a) $MaxC = 2$ clusters.

(b) $MaxC = 4$ clusters.

(c) $MaxC = 8$ clusters.

Figure 4.35: Forecast intervals for Rio de Janeiro electricity demand time series.

The graphs cover the period for $t - 12 \leq t \leq t + H$ where $t$ is the last observation of the in-sample period and $H = 12$ is the number of forecast horizons. The shades, from lighter to darker, correspond to $\alpha$ levels 0.95, 0.90, 0.85, 0.80, 0.75 and 0.60.

Table 4.27: Forecasting performance. Rio de Janeiro electricity demand series.

| Forecast horizon | Model | MSE | Rank | %Δ wrt Avg. | MAPE | | %Δ wrt Avg. |
|---|---|---|---|---|---|---|---|
| H=1 | CB2 | 3151.72883 | 3 | -5.74% | 1.31262% | 4 | -1.66% |
| | CB4 | 3256.80153 | 7 | -2.60% | 1.33027% | 8 | -0.34% |
| | CB8 | 3267.40143 | 8 | -2.28% | 1.33040% | 9 | -0.33% |
| | Avg. Net. | 3343.58294 | 9 | 0.00% | 1.33479% | 10 | 0.00% |
| | Best Net. isMAPE | 3142.80143 | 2 | -6.00% | 1.30781% | 3 | -2.02% |
| | Best Net. isMSE | 3495.32677 | 10 | 4.54% | 1.36547% | 11 | 2.30% |
| | GA2 | 3230.69771 | 6 | -3.38% | 1.32558% | 7 | -0.69% |
| | GA4 | 3219.28209 | 5 | -3.72% | 1.32446% | 6 | -0.77% |
| | GA8 | 3210.09148 | 4 | -3.99% | 1.32207% | 5 | -0.95% |
| | Naive | 30638.85387 | 11 | 816.35% | 4.40468% | 12 | 229.99% |
| | HWT Taylor | 2817.24485 | 1 | -15.74% | 1.15789% | 1 | -13.25% |
| | NN Taylor | | | | 1.29025% | 2 | -3.34% |
| H=2 | CB2 | 5867.33684 | 3 | -4.65% | 1.81488% | 4 | -1.47% |
| | CB4 | 6089.60336 | 8 | -1.04% | 1.85059% | 9 | 0.47% |
| | CB8 | 6087.55396 | 7 | -1.07% | 1.85177% | 10 | 0.54% |
| | Avg. Net. | 6153.53015 | 9 | 0.00% | 1.84190% | 6 | 0.00% |
| | Best Net. isMAPE | 5831.06369 | 2 | -5.24% | 1.80243% | 3 | -2.14% |
| | Best Net. isMSE | 6624.70839 | 10 | 7.66% | 1.91473% | 11 | 3.95% |
| | GA2 | 6058.89821 | 5 | -1.54% | 1.84427% | 7 | 0.13% |
| | GA4 | 6061.83480 | 6 | -1.49% | 1.84660% | 8 | 0.26% |
| | GA8 | 6028.68023 | 4 | -2.03% | 1.84117% | 5 | -0.04% |
| | Naive | 30656.98541 | 11 | 398.20% | 4.40698% | 12 | 139.26% |
| | HWT Taylor | 3091.48104 | 1 | -49.76% | 1.26844% | 1 | -31.13% |
| | NN Taylor | | | | 1.78818% | 2 | -2.92% |
| H=3 | CB2 | 8040.95714 | 3 | -4.09% | 2.18189% | 4 | -1.51% |
| | CB4 | 8396.57159 | 8 | 0.15% | 2.23722% | 8 | 0.99% |
| | CB8 | 8371.82008 | 5 | -0.14% | 2.23874% | 9 | 1.06% |
| | Avg. Net. | 8383.71246 | 7 | 0.00% | 2.21531% | 5 | 0.00% |
| | Best Net. isMAPE | 7956.83617 | 2 | -5.09% | 2.17319% | 3 | -1.90% |
| | Best Net. isMSE | 9267.36009 | 10 | 10.54% | 2.34494% | 11 | 5.85% |
| | GA2 | 8376.06017 | 6 | -0.09% | 2.23404% | 7 | 0.85% |
| | GA4 | 8398.63640 | 9 | 0.18% | 2.23882% | 10 | 1.06% |
| | GA8 | 8335.41415 | 4 | -0.58% | 2.23011% | 6 | 0.67% |
| | Naive | 30675.25477 | 11 | 265.89% | 4.40958% | 12 | 99.05% |
| | HWT Taylor | 3630.40116 | 1 | -56.70% | 1.42916% | 1 | -35.49% |
| | NN Taylor | | | | 2.16067% | 2 | -2.47% |
| H=4 | CB2 | 9845.11384 | 3 | -3.73% | 2.45944% | 4 | -1.26% |
| | CB4 | 10286.07025 | 7 | 0.58% | 2.51625% | 7 | 1.02% |
| | CB8 | 10236.74480 | 6 | 0.10% | 2.52335% | 9 | 1.30% |
| | Avg. Net. | 10226.47264 | 4 | 0.00% | 2.49087% | 5 | 0.00% |
| | Best Net. isMAPE | 9736.42147 | 2 | -4.79% | 2.43960% | 3 | -2.06% |
| | Best Net. isMSE | 11511.47718 | 10 | 12.57% | 2.65476% | 11 | 6.58% |
| | GA2 | 10289.58501 | 8 | 0.62% | 2.51911% | 8 | 1.13% |
| | GA4 | 10330.84008 | 9 | 1.02% | 2.52640% | 10 | 1.43% |
| | GA8 | 10236.34174 | 5 | 0.10% | 2.51421% | 6 | 0.94% |
| | Naive | 30693.20304 | 11 | 200.13% | 4.41157% | 12 | 77.11% |
| | HWT Taylor | 4328.80658 | 1 | -57.67% | 1.59826% | 1 | -35.84% |
| | NN Taylor | | | | 2.43381% | 2 | -2.29% |

$\%\Delta = 100(M_{model} - M_{Avg})/M_{Avg}$ with $M_i$ being the metric for model $i$.
Negative % values indicate improvement over the average.

(Continued) Forecasting performance. Rio de Janeiro electricity demand series.

| Forecast horizon | Model | MSE | Rank | %Δ wrt Avg. | MAPE | | %Δ wrt Avg. |
|---|---|---|---|---|---|---|---|
| H=5 | CB2 | 11493.72511 | 3 | -3.46% | 2.68185% | 4 | -1.08% |
| | CB4 | 12060.69103 | 8 | 1.30% | 2.75439% | 8 | 1.59% |
| | CB8 | 11985.89748 | 5 | 0.68% | 2.75890% | 10 | 1.76% |
| | Avg. Net. | 11905.34801 | 4 | 0.00% | 2.71126% | 5 | 0.00% |
| | Best Net. isMAPE | 11358.45054 | 2 | -4.59% | 2.65482% | 3 | -2.08% |
| | Best Net. isMSE | 13605.60877 | 10 | 14.28% | 2.92296% | 11 | 7.81% |
| | GA2 | 12052.80270 | 7 | 1.24% | 2.74992% | 7 | 1.43% |
| | GA4 | 12115.85212 | 9 | 1.77% | 2.75801% | 9 | 1.72% |
| | GA8 | 11991.14679 | 6 | 0.72% | 2.74248% | 6 | 1.15% |
| | Naive | 30710.78550 | 11 | 157.96% | 4.41327% | 12 | 62.78% |
| | HWT Taylor | 5110.58695 | 1 | -57.07% | 1.75986% | 1 | -35.09% |
| | NN Taylor | | | | 2.65238% | 2 | -2.17% |
| H=6 | CB2 | 12991.43108 | 3 | -3.13% | 2.86016% | 4 | -0.86% |
| | CB4 | 13690.93309 | 8 | 2.08% | 2.94683% | 9 | 2.15% |
| | CB8 | 13592.09276 | 6 | 1.35% | 2.95443% | 10 | 2.41% |
| | Avg. Net. | 13411.42718 | 4 | 0.00% | 2.88488% | 5 | 0.00% |
| | Best Net. isMAPE | 12819.48827 | 2 | -4.41% | 2.82364% | 3 | -2.12% |
| | Best Net. isMSE | 15441.99157 | 10 | 15.14% | 3.14089% | 11 | 8.87% |
| | GA2 | 13634.73476 | 7 | 1.67% | 2.93101% | 7 | 1.60% |
| | GA4 | 13721.59864 | 9 | 2.31% | 2.94280% | 8 | 2.01% |
| | GA8 | 13574.41805 | 5 | 1.22% | 2.92283% | 6 | 1.32% |
| | Naive | 30728.61821 | 11 | 129.12% | 4.41511% | 12 | 53.04% |
| | HWT Taylor | 5926.55778 | 1 | -55.81% | 1.91042% | 1 | -33.78% |
| | NN Taylor | | | | 2.82060% | 2 | -2.23% |
| H=7 | CB2 | 14303.88902 | 3 | -2.99% | 3.01547% | 4 | -1.12% |
| | CB4 | 15098.75493 | 8 | 2.41% | 3.10725% | 9 | 1.89% |
| | CB8 | 14954.56448 | 6 | 1.43% | 3.11105% | 10 | 2.01% |
| | Avg. Net. | 14744.02128 | 4 | 0.00% | 3.04962% | 5 | 0.00% |
| | Best Net. isMAPE | 14091.42249 | 2 | -4.43% | 2.97216% | 2 | -2.54% |
| | Best Net. isMSE | 17044.50125 | 10 | 15.60% | 3.31318% | 11 | 8.64% |
| | GA2 | 15014.04330 | 7 | 1.83% | 3.08845% | 7 | 1.27% |
| | GA4 | 15118.43194 | 9 | 2.54% | 3.10015% | 8 | 1.66% |
| | GA8 | 14953.95279 | 5 | 1.42% | 3.08110% | 6 | 1.03% |
| | Naive | 30746.87366 | 11 | 108.54% | 4.41739% | 12 | 44.85% |
| | HWT Taylor | 6739.33119 | 1 | -54.29% | 2.04649% | 1 | -32.89% |
| | NN Taylor | | | | 2.97216% | 3 | -2.54% |
| H=8 | CB2 | 15266.51951 | 3 | -2.75% | 3.16174% | 4 | -1.11% |
| | CB4 | 16112.92750 | 9 | 2.64% | 3.25405% | 9 | 1.78% |
| | CB8 | 15937.52525 | 6 | 1.52% | 3.25406% | 10 | 1.78% |
| | Avg. Net. | 15698.52881 | 4 | 0.00% | 3.19725% | 5 | 0.00% |
| | Best Net. isMAPE | 15015.14508 | 2 | -4.35% | 3.11168% | 3 | -2.68% |
| | Best Net. isMSE | 18194.51474 | 10 | 15.90% | 3.45620% | 11 | 8.10% |
| | GA2 | 15986.47283 | 7 | 1.83% | 3.23222% | 7 | 1.09% |
| | GA4 | 16106.10352 | 8 | 2.60% | 3.24477% | 8 | 1.49% |
| | GA8 | 15927.52640 | 5 | 1.46% | 3.22417% | 6 | 0.84% |
| | Naive | 30764.93574 | 11 | 95.97% | 4.41936% | 12 | 38.22% |
| | HWT Taylor | 7526.12956 | 1 | -52.06% | 2.16884% | 1 | -32.17% |
| | NN Taylor | | | | 3.09950% | 2 | -3.06% |

(Continued) Forecasting performance. Rio de Janeiro electricity demand series.

| Forecast horizon | Model | MSE | Rank | %Δ wrt Avg. | MAPE | | %Δ wrt Avg. |
|---|---|---|---|---|---|---|---|
| H=9 | CB2 | 16165.81701 | 3 | -2.66% | 3.29473% | 4 | -1.11% |
| | CB4 | 17055.78937 | 9 | 2.70% | 3.38578% | 10 | 1.63% |
| | CB8 | 16857.29223 | 6 | 1.51% | 3.38345% | 9 | 1.56% |
| | Avg. Net. | 16607.27474 | 4 | 0.00% | 3.33156% | 5 | 0.00% |
| | Best Net. isMAPE | 15897.35335 | 2 | -4.27% | 3.23064% | 3 | -3.03% |
| | Best Net. isMSE | 19292.73843 | 10 | 16.17% | 3.58460% | 11 | 7.60% |
| | GA2 | 16907.45179 | 7 | 1.81% | 3.35955% | 7 | 0.84% |
| | GA4 | 17041.17000 | 8 | 2.61% | 3.37243% | 8 | 1.23% |
| | GA8 | 16852.17106 | 5 | 1.47% | 3.35486% | 6 | 0.70% |
| | Naive | 30783.20126 | 11 | 85.36% | 4.42158% | 12 | 32.72% |
| | HWT Taylor | 8271.15086 | 1 | -50.20% | 2.27873% | 1 | -31.60% |
| | NN Taylor | | | | 3.23024% | 2 | -3.04% |
| H=10 | CB2 | 17243.68031 | 3 | -2.70% | 3.42375% | 4 | -1.26% |
| | CB4 | 18142.41535 | 9 | 2.37% | 3.51324% | 10 | 1.32% |
| | CB8 | 17952.29223 | 6 | 1.30% | 3.50719% | 9 | 1.15% |
| | Avg. Net. | 17722.05909 | 4 | 0.00% | 3.46744% | 5 | 0.00% |
| | Best Net. isMAPE | 16947.28391 | 2 | -4.37% | 3.35471% | 3 | -3.25% |
| | Best Net. isMSE | 20497.39285 | 10 | 15.66% | 3.70631% | 11 | 6.89% |
| | GA2 | 17987.56095 | 7 | 1.50% | 3.48498% | 7 | 0.51% |
| | GA4 | 18126.04733 | 8 | 2.28% | 3.49400% | 8 | 0.77% |
| | GA8 | 17934.69572 | 5 | 1.20% | 3.47839% | 6 | 0.32% |
| | Naive | 30801.62089 | 11 | 73.80% | 4.42418% | 12 | 27.59% |
| | HWT Taylor | 8966.56642 | 1 | -49.40% | 2.37657% | 1 | -31.46% |
| | NN Taylor | | | | 3.34397% | 2 | -3.56% |
| H=11 | CB2 | 18309.99328 | 3 | -2.60% | 3.54540% | 4 | -0.97% |
| | CB4 | 19197.49977 | 8 | 2.12% | 3.62145% | 10 | 1.15% |
| | CB8 | 19012.02353 | 6 | 1.13% | 3.61687% | 9 | 1.03% |
| | Avg. Net. | 18798.66147 | 4 | 0.00% | 3.58011% | 5 | 0.00% |
| | Best Net. isMAPE | 18001.01188 | 2 | -4.24% | 3.46887% | 3 | -3.11% |
| | Best Net. isMSE | 21725.13402 | 10 | 15.57% | 3.84041% | 11 | 7.27% |
| | GA2 | 19055.25248 | 7 | 1.36% | 3.59477% | 7 | 0.41% |
| | GA4 | 19201.97084 | 9 | 2.15% | 3.60622% | 8 | 0.73% |
| | GA8 | 19000.10683 | 5 | 1.07% | 3.58912% | 6 | 0.25% |
| | Naive | 30819.82545 | 11 | 63.95% | 4.42631% | 12 | 23.64% |
| | HWT Taylor | 9605.91503 | 1 | -48.90% | 2.46244% | 1 | -31.22% |
| | NN Taylor | | | | 3.45048% | 2 | -3.62% |
| H=12 | CB2 | 19304.36839 | 3 | -2.47% | 3.64766% | 4 | -0.92% |
| | CB4 | 20196.13402 | 8 | 2.04% | 3.72599% | 10 | 1.21% |
| | CB8 | 20030.49655 | 6 | 1.20% | 3.71444% | 9 | 0.89% |
| | Avg. Net. | 19793.29945 | 4 | 0.00% | 3.68158% | 5 | 0.00% |
| | Best Net. isMAPE | 18959.65752 | 2 | -4.21% | 3.57599% | 3 | -2.87% |
| | Best Net. isMSE | 22918.82751 | 10 | 15.79% | 3.97378% | 11 | 7.94% |
| | GA2 | 20045.48860 | 7 | 1.27% | 3.69610% | 7 | 0.39% |
| | GA4 | 20199.70261 | 9 | 2.05% | 3.70961% | 8 | 0.76% |
| | GA8 | 19979.80420 | 5 | 0.94% | 3.69010% | 6 | 0.23% |
| | Naive | 30836.20581 | 11 | 55.79% | 4.42741% | 12 | 20.26% |
| | HWT Taylor | 10192.92250 | 1 | -48.50% | 2.53909% | 1 | -31.03% |
| | NN Taylor | | | | 3.55712% | 2 | -3.38% |

Table 4.28: Jarque-Bera normality test for Rio de Janeiro electricity demand series.

| | | Forecast horizon | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Model | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| 2C | CB | * | * | * | * | * | * | * | * | * | * | * | |
| | GA | | * | * | * | | * | * | * | * | * | * | |
| 4C | CB | * | * | * | * | * | * | * | * | * | * | * | |
| | GA | | * | * | * | | * | * | * | * | * | * | |
| 8C | CB | * | * | * | * | * | * | * | * | * | * | * | * |
| | GA | | * | * | * | | * | * | * | * | * | * | |
| | Avg. | * | * | * | * | * | * | * | * | * | * | * | |
| | Best net isMAPE | | * | * | * | | * | * | * | * | * | * | |
| | Best net isMSE | * | * | * | * | * | * | * | * | * | * | | |

The rejection of the hypothesis of normally distributed forecast errors (with 95% confidence level) is indicated with *.

Table 4.29: Coefficients for structural combination of NN for Rio de Janeiro electricity demand series.

| | | | | | | |
|---|---|---|---|---|---|---|
| CB2 | | | | | | |
| $\alpha_1$ | 24.9383 | 0.38 | 0.4028 | 0.0896 | -0.0942 | 0.2126 |
| $\Phi$ | 1 | | | | | |
| CB4 | | | | | | |
| $\alpha_1$ | 25.3822 | 0.7965 | -1.7627 | 1.1277 | 0.2929 | -0.4866 |
| $\alpha_2$ | 26.3815 | -1.0451 | 3.2846 | 0.5185 | 1.8219 | 1.4281 |
| $\alpha_3$ | 28.9069 | 0.5246 | -2.3253 | 0.5201 | 0.8754 | 0.2391 |
| $\Phi$ | 0.3984 | 0.1786 | 0.4233 | | | |
| CB8 | | | | | | |
| $\alpha_1$ | 24.9104 | 1.4617 | -5.6481 | 2.2431 | 0.3346 | 0.1906 |
| $\alpha_2$ | 26.1478 | 1.7667 | 6.8272 | 0.5522 | 3.0498 | 1.7813 |
| $\alpha_3$ | 27.6467 | -2.071 | 1.087 | 0.9352 | 0.0621 | 0.3737 |
| $\alpha_4$ | 21.5636 | -2.3524 | 0.62 | -0.3881 | 1.4783 | -1.4435 |
| $\alpha_5$ | 29.9801 | -6.0529 | 1.2764 | 2.8328 | 0.9112 | 1.7857 |
| $\Phi$ | 0.2199 | 0.1086 | 0.2297 | 0.2243 | 0.2169 | |

$\alpha_i$ are the coefficients applied to point-forecasts from models in cluster $i$ and $\Phi$ are the weights applied to the outputs from clusters.

Table 4.30: Comparisons by forecast horizon. Series: Rio de Janeiro electricity demand.

| MSE | | | | |
|---|---|---|---|---|
| | Avg. | Bst. IsMAPE | Bst. IsMSE | HWT |
| CB2 | 12 | 0 | 12 | 0 |
| GA2 | 3 | 0 | 12 | 0 |
| CB4 | 2 | 0 | 12 | 0 |
| GA4 | 2 | 0 | 12 | 0 |
| CB8 | 3 | 0 | 12 | 0 |
| GA8 | 3 | 0 | 12 | 0 |

| MAPE | | | | |
|---|---|---|---|---|
| | Avg. | Bst. IsMAPE | Bst. IsMSE | HWT |
| CB2 | 12 | 0 | 12 | 0 |
| GA2 | 1 | 0 | 12 | 0 |
| CB4 | 1 | 0 | 12 | 0 |
| GA4 | 1 | 0 | 12 | 0 |
| CB8 | 1 | 0 | 12 | 0 |
| GA8 | 2 | 0 | 12 | 0 |

Number of forecast horizons for which CB and GA combinations outperform different benchmark models. Avg. stands for the average of NN in the ensemble; Bst. IsMAPE stands for the best NN in terms of in-sample MAPE in the ensemble; Bst. IsMSE stands for the best NN in terms of in-sample MSE in the ensemble.

Table 4.31: Cluster validity indexes. Series: Rio de Janeiro electricity demand.

| Max. num clusters | PC | MPC | MDO | Final. num clusters |
|---|---|---|---|---|
| 2 | 1.0000 | NA | 1.0000 | 1 |
| 4 | 0.3669 | 0.0504 | 0.0000 | 3 |
| 8 | 0.2092 | 0.0115 | 0.0000 | 5 |

*h* denotes the forecast horizon, PC denotes the Partition Coefficient, MPC denotes the Modified Partition Coefficient and MDO denotes the Membership Degree Optimum. Values closer to 1 are preferable.

## 4.9 Discussion of Findings from Real Time Series

For the wind power and electricity demand series, the NN-based models have similar performance and the preliminary analysis showed more evident sensitivity of the error metrics to the number of inputs than to the number of neurons. However, the statistical tests showed the benefit in using more neurons for the wind power series. This can be due to the complex nature of the time series, which involves non-linearities and weather predictors.

The best performing NNs for the electricity demand series are not only simple in terms neurons, but also the best performing CB models are the simplest. Only CB2, with a maximum of 2 clusters, was able to consistently outperform the average forecast from the NNs. The need for structural simplicity in the case of the electricity demand series is manifested both at single model level and at the ensemble level.

A difference between the two studies is the modelling effort that was required prior to the fitting of the models. The electricity demand time series required a sensible selection of inputs in the early stages of the study, so as to capture the regularities in the series (in accord with Crone & Kourentzes, 2010). However, this also implies that the statistical benchmarks are very well suited to the data, and thus we observed difficulty of NN models and ensembles in outperforming the well specified models. It thus appears that faced with regular data, it pays off to invest time and effort in the selection of inputs and use a well-specified model that address these regularities to forecast the series.

Structural combination using GA is performing well in forecasting the wind power series, and is able to outperform benchmarks (ARIMAX, best NN and average of NN forecasts) in several forecast horizons. It outperforms the simple average in up to 6 forecast horizons and clearly shows improvements over CB combination. As it was mentioned in Section 4.3.3, GA combination can be viewed as a structurally informed average. This might be one of the reasons for its good performance, as in general the average is known to be a robust benchmark.

## 4.10 Discussion of Findings in the Context of Ensembles

Timmermann (2006) argued, from a theoretical perspective, that unless one can find ex ante a particular forecasting model producing smaller forecast errors than its competitors, forecast combinations offer diversification gains that make them more attractive than relying on a forecast from a single model. Variety comes from different forecasters or models. In the present study, model diversity (proxied with structural descriptors) is explicitly included and used to inform forecast combinations. Diversification in models was explored by generating ensembles of models trained with randomised input-output data patterns.

One of the main characteristics of ensembles of NNs, as mentioned in Chapter 2, is how the steps of generation pruning and combination (Figure 2.1) are followed. Some authors (such as Hansen & Salamon, 1990; Drezga, 1999; Siwek et al., 2009) adopt a sequential approach, where one stage feeds into the next. Others follow more dynamic approaches, where the stages are interrelated. Fore example, Liu & Yao (1999), Liu et al. (2000) and Zhou et al. (2002) created ensembles with evolutionary algorithms, in which case the dynamics of the ensemble building process are interlinked. In the present study, the most dynamic part of the approach is located in the combination stage. Pruning is part of the normal working of the combining algorithms: both CB and GA combination routines, by leaving out a proportion of the models in the generated pool, perform pruning. Although there is an intimate relation between pruning and combining, as the latter informs dynamically the former, the general outline puts the present research more on the ground of sequential approaches.

Both the dynamic and sequential approaches mentioned above make use of forecasts produced by models. In the present research, departing from that approach, the information used to combine forecast is the parameter set of NNs. This allows to include information about the structure of the model into the combining stage (a procedure inspired by Bakker & Heskes, 2003).

The contribution of the present study is in the context of sequential ensemble generation and structurally informed forecast combinations. The approaches followed to incorporate structural parameters are a clustering-based algorithm and a genetic-based algorithm. In this way, the inclusion of model structure via clustering, as suggested by Bakker & Heskes (2003), is explored. The adoption of a clustering algorithm and a structurally informed benchmark (based on genetic algorithms), that resembles an average, permits comparisons between combination mechanisms of the same orientation, but of different complexity. Therefore, both in terms of the use of clustering and the use of genetic algorithms, this research makes a contribution. The study of different synthetic and real time series (with a multivariate case) and the inclusion of different levels in the number of clusters, permitted a realistic assessment of the performance of the proposed forecasting combinations.

A key point in the study is the rationale for the inclusion of structure in combining forecasts. The motivation is the use of characteristics of models, and not merely their single output forecasts. However, the interpretation of what constitutes internal characteristics could differ from the approach adopted here. Larger structural representations or measures based on structural components rather than components themselves could be used. Additionally, the study of relationships between internal components and outputs in models, suggested by Garson (1991) and Goh (1995), could be informative. These are natural extensions of the present research.

Regarding the calculation of forecast intervals, there is considerable room for research in the future. The Delta method relies heavily on the difference between the estimated weights of the network $\hat{w}$ and the optimal $w*$ (see Khosravi et al., 2011). As the CB method finds clusters in the weight space (w), it would be interesting to see if this method can be adapted to take into account the distance between models in such space. The best model in a cluster could be used as a reference model, assumed to have weights closer to the true values, and could therefore be used to inform the calculations in the Delta method within the context of ensembles. The MVE method estimates the mean and variance of the target variable by using separate NNs. An extension of this method could be attempted for the case in which several networks are used to estimate the mean. The bootstrap method relies in the production of forecasts by different models obtained through the re-sampling of training (in-sample) data. The different forecasts thus produced are used to calculate the mean and variance needed for the intervals. Therefore, it could be adapted easily for the ensembles, despite of its main drawback of being more complex than others to implement (Khosravi et al., 2011). It could be adapted to take into account the models included in the clusters rather than all models in the ensemble and, perhaps, to include cluster configuration into the calculations. The empirical method by Lee & Scholtes (2014), used here, has the limitation of not considering the possibility of a time-varying error distribution. This issue could be solved by using more sophisticated methods that incorporate, for example, exponential weighting schemes, such as those discussed by Taylor (2007), in order to make the quantile estimation adaptive (Lee & Scholtes, 2014).

The study of ensembles with the structural combination proposed here expands on the set of methods used in the literature on electricity demand forecasting. Specifically for the Rio de Janeiro electricity demand, results revealed a considerable similarity in performance between the single NN specified by Taylor et al. (2006) and the ensembles in the present research, and also showed the clear superiority of the exponential smoothing benchmark (a HWT model). This contributes to the evidence that to forecast this double-seasonal time series, NNs are not the best choice, thus the findings of this study add to the discussion about the suitability of NNs for real world seasonal time series (Crone et al., 2011; Zhang & Qi, 2005). This suggests that the structural combinations of exponential smoothing models should be investigated, as they may be more appropriate in the case of seasonal time series data. Their suitability for regular time series and their ability to easily adapt to new information makes them attractive for building and combining ensembles. Moreover, the findings motivate the investigation of structural combinations of forecasting models that specifically address seasonality and have a good track record on performance.

Considering the wind power forecasting application, the situation is similar to the electricity demand forecasting study, as the structural combination is infrequent. Relevant research in forecasting wind power include Wang & Hu (2015), Giebel et al. (2003), Sanchez (2008), Salcedo-Sanz et al. (2009), Li et al. (2011). The approaches include very complex modelling, mixing different elements from statistics and computing intelligence, but the combinations mostly focus on output forecasts with no information about internals of the models. In the context of this dissertation, this application illustrates the potential for adopting out proposed approaches in multivariate time series forecasting, something that can be further explored in future research.

The results for the seasonal and double-seasonal series highlight a limitation of the ensembles used in the present research. They are composed of models of the same nature and the same basic specification. From the perspective of model diversity (Timmermann, 2006; Bunn, 1975), there is a gain in diversity that might not be exploited. Our results show how NNs ensembles combined structurally can be competitive against the average forecast, but are outperformed by the statistical benchmarks. The use of a single family of models limits the gains in variety. If combinations of NNs and HWT models were made, there could be a bigger gain than in combining NNs or HWT alone. That is why, apart from studying structural combinations of statistical and other standard forecasting models, future research may structurally combine ensembles of models of different nature. This could be achieved by constructing bundles of models, such as $B_i =< ARIMA, NN >$, and proposing a structural representation for $B_i$.

The quality of clusters as measured by several indexes suggests that there are clusters found by the CB algorithm that are not well differentiated or separated. Exploring the inclusion of the forecasts and the structure of models as mentioned above (Garson, 1991; Goh, 1995) seems to be a promising research avenue, as this could lead to better differentiated clusters. This scheme can be combined with the exploration of models of different nature, so that both diversity in structure and diversity in forecasts are exploited.

When the indexes do not give sufficient information to take a decision on the number of clusters, a rule of thumb can be used: if significant forecast improvement by a cluster-based combination with respect to benchmarks is observed in the first horizons, then the cluster-based combination is preferred. If insignificant improvement is obtained for the first horizons through a well-performing benchmark or a GA combination, such model is preferred. In all cases, the number of clusters should be small, due to the homogeneity of the belongingness of models to clusters as the number of these is increased and taking into account the tendency of well performing combinations to have a few clusters.

## 4.11 Conclusions and Research Agenda

This chapter has presented a novel forecasting model combination approach that involves the creation of ensembles of NNs and the combination of a subset of them based on parameters from their structure. The first implementation of the proposed combination approach is based on clustering algorithms, which groups together models that share a measure of similarity (in this case a measure of distance in the parameter space of models). The second implementation uses genetic algorithms to select models by using reference points (analogous to cluster centres) in the parameter space and can be seen as a structurally informed average of forecasts. Different levels in the number of clusters were used and synthetic and real data series of different complexity and nature were selected to assess the combination scheme.

Structural combination with genetic algorithms (GA) outperforms the simple average more easily than cluster based (CB) combination for non-seasonal time series (STAR2 and wind power production[12]), whereas for the seasonal series (Synthetic-1S, Synthetic-2S and electricity demand) the CB tend to do better in relation to the simple average. CB and GA easily outperform the best NNs in the ensembles in the non-seasonal synthetic series and wind power series. For the seasonal time series, Synthetic-1S, Synthetic-2S and electricity demand, on the other hand, there is no marked superiority of structural combinations over individual models. In spite of this, CB shows better performance than GA with respect to the best models.

GA combinations not only perform well compared to CB for non-seasonal series, but also show a smoother performance pattern. This is a desirable feature in light of the structural differences in NNs for different different horizons. These findings suggest that different forms of structural combination can be explored with the aim of finding the best combination approach for a given application.

CB and GA structural combinations were outperformed by the chosen statistical benchmark in the cases of the single-seasonal and double-seasonal synthetic series, and the double-seasonal real series. Exponential smoothing models are better equipped to capture the regularities in these time series than NNs. Consequently, the potential structural combinations of these models should be investigated, as they

[12] Additional experiments to structurally combine NNs for BL1 time series, which has an even higher noise than STAR2, support these findings. Characteristics of this series can be found in Section 3.3.

may be more attractive in the case of seasonal time series data.

Nonetheless, for non-seasonal series (synthetic and real) the NNs and the structurally combined ensembles showed a clear advantage. In the case of Kaggle wind power series, the superiority of CB models to the statistical benchmark is likely to come from the high complexity and non-linearity of the forecasting problem, for which NNs based models are more robust. It is important, however, that future studies with wind power data consider the existence and influence of diurnal cycles as this could further clarify the conditions under which NNs combined structurally are able to better forecast data of this kind. Based on these findings, future research could investigate the behaviour of a structural combination approach when models of different nature are combined. If, for example, instead of using single unit models (statistical or computational intelligence model), bundles of the form $B_i =< ARIMA, NN >$ are formed, the structural combination of such bundles could potentially improve performance, by better exploiting diversity, when dealing with complex forecasting problems.

A direct multi-step-ahead forecasting approach, with separate ensembles for different forecast horizons, was used in some applications, whereas a single ensemble, in an iterative approach, producing forecasts for all horizons, was used in one application, due to its characteristics. Findings suggest that an extension of the present research could investigate the way in which the chosen multi-step-ahead forecast approach affects the performance of structural combination.

This understanding of the effect of the forecasting approach could be supported by more sophisticated forms of forecast interval calculations. The empirical method by Lee & Scholtes (2014), used here, may not be entirely suitable for time-varying error distributions. This issue could be solved by using methods that incorporate, for example, exponential weighting schemes, such as those discussed by Taylor (2007) in order to make the quantile estimation adaptive (Lee & Scholtes, 2014).

There is a basic difference between CB and GA structural combinations: the first is deterministic and the second is random. As the first performs better with seasonal series and the second worked well with non-seasonal series, the question arises of how the nature of structural clustering underlying the combination is related to the regularity in the data. This question is left for future research.

In general, a natural extension of the present study is to explore other interpretations of what constitutes internal characteristics of a model. The study can be extended to explore the conditions under which CB and GA combinations work better. One factor that was limited in the present study is the maximum number of clusters allowed for the combination mechanism. A more general study of the effect of changes in this factor is needed.

Other particular improvements on the CB combination mechanism can include the introduction of a performance criterion, so that the models selected in each cluster are filtered to contain only the best around the centroid. Additional selection inside clusters can be achieved by discarding individual models with very small $\alpha$ coefficients in the regressions performed for each cluster. Experiments with different functional forms for the final forecast and different optimisation mechanisms, for both CB and GA, can also be tried. For example, the optimisation for CB combinations could obtain $\alpha$ and $\phi$ coefficients in one single step rather than two.

Future research can also explore switching of models in the context of ensembles of NNs built for iterative forecasting and the adoption of bootstrapping to analyse the effect of a bagging-like approach in the structural combinations. Additionally, in a context of big data and business analytics, an automated connection between the databases produced during the sensitivity analysis of individual models and the CB or GA structural combinations would be desirable. Such databases can be used to automatically identify candidates for ensembles, which may improve the quality of inputs.

Overall, the exploration of structural forecasting model combination, and its implementation in two forms, opens the possibility to investigate new types of ensembles, which are based on statistical models.

# Chapter 5

# Structural Combination of Seasonal Exponential Smoothing Models

## 5.1 Abstract

This chapter extends the structural combination of NNs in order to develop structural combinations of forecasting models that are suitable for specific types of seasonal time series. Two exponential smoothing models are considered: the single-seasonal multiplicative Holt-Winters model (MHW) and the double-seasonal multiplicative Holt-Winters-Taylor model (MHWT). Structural diversity in models is promoted either by adding normally distributed noise to the series or by swapping blocks of data prior to training (fitting). In the first part of the chapter, the representation, creation and combination of models are discussed. Subsequently, three empirical studies are described, which were conducted in order to evaluate the behaviour of the combination procedure. The first application aims to forecast daily peak electricity demand for the next 7 days. A second study forecasts hourly electricity demand for the next 24 hours. Finally, the third study forecasts half-hourly electricity demand in England and Wales for the next 24 hours. Results show that structural combinations can outperform competitive benchmarks. Moreover, when the method used to add diversity to the original series is considered, structurally combined ensembles of MHW and MHWT seem to be better suited to forecast the daily peak demand of electricity and the half-hourly demand en England and Wales when noise is added to the series; by contrast, for the double-seasonal hourly electricity demand, block swapping provided better results when combining the models.

## 5.2 Introduction

Ensembles originated in climate forecasting, where physical models with different initial conditions provide different predictions (i.e. Murphy et al., 2004; Parker, 2010). With univariate models, there is no modelling of the phenomena, but a hypothesis of how historical values and previous forecast errors relate to the future values of the series (Pankratz, 2009, p. 8). In both cases there are initial conditions, but of different nature. In models for climate prediction, these conditions are initial values of state variables in physical systems. In univariate time series forecasting models, they are the initial values in optimisation algorithms. The exploration of ensembles for univariate time series forecasting by changing these algorithmic conditions has been widely done in the research in NNs, as summarised in Table 2.1, but is rare in cases of traditional statistical and forecasting models. Combinations using ensemble forecasts provided by climate agencies have been done, for example, by Taylor & Buizza (2003) and the building of ensembles containing statistical models (ARMA) was done by Matijaš et al. (2013), who pioneered the notion of exploring variety in statistical models. However, the creation of pools of standard forecasting models, with diversity induction mechanisms has, to the best of our knowledge, received minimum attention in the time series forecasting literature. In general, by using the structure when combining models, this research departs from the sole use of point forecasts made by models or experts (see Clemen, 1989; Diebold & Lopez, 1996; De Menezes et al., 2000; Timmermann, 2006; Newbold & Harvey, 2007).

The combination mechanism proposed in this chapter generates, in a first stage, models with diversity induction mechanisms. This task relates to the choice of initial parameter values, with which a modeller is usually confronted when selecting a single model. Data variation techniques, e.g. noise addition or swapping data blocks, have been used to promote model diversity and less needed with NNs due to their volatility.

Holt-Winters models are widely used in business (Hyndman et al., 2008) and Holt-Winters-Taylor models have been successfully applied to load forecasting (e.g. Taylor, 2010). Such models were devised to address the dynamics of seasonal and double-seasonal time series, respectively, and tend to be robust, due to the way that they can adapt to changes in the data pattern. NNs, in contrast, as highlighted in previous chapters, are general purpose models that tend to be unstable and require several design decisions prior to their use. Therefore, the HW and HWT models permit an assessment of the structural combination approach that is proposed in this dissertation with models that are less volatile than NNs, but also adaptive to the dynamics of the time series. Three specific models are chosen for this study.

The multiplicative form of the HW model is used to forecast daily peak electricity demand. Peak demand is the maximum amount of power that must be delivered (Willis, 2000, p. 40). This forecasting task is one of the basic operations undertaken by the transmission or distribution operator when scheduling generation for the next day (Haida & Muto, 1994; Iizaka et al., 2002; Amjady, 2001). It is also required by operators of dispatching centres in order to schedule maintenance or conduct adequacy assessments (Amjady, 2001). Different studies have addressed peak demand forecasting, for example, by using regression methods (Joe H. Chow, 2004), clustering based on classification of load curves (Goia et al., 2010), NNs (Wang & Cao, 2006) and ARIMA-based models (Amjady, 2001). The structural combination adopted here expands the set of approaches that could be undertaken in practice. Peak electricity demand data from the Rio de Janeiro time series, which was used in an application of structural combination of NNs in the previous chapter, is chosen for assessing the combination of MHW models.

The second study focuses on the multiplicative double-seasonal Holt-Winters-Taylor. This model was proposed by Taylor (2003) and was used by Taylor et al.

(2006) when comparing the performance of various univariate models. It is structurally combined to forecast hourly electricity demand from Rio de Janeiro.

Finally, the third study, considers the same double-seasonal Holt-Winters-Taylor model, but with different seasonal cycle lengths. The model was structurally combined to forecast electricity demand in England and Wales.

The following sections describe the base models, the way in which they were prepared to form part of ensembles, and the combination approach.

## 5.3 The Base Models

The multiplicative Holt-Winters model used for the first application, with an autocorrelation error correction term added to the standard Holt-Winters model presented, for example, by Hyndman et al. (2008), is:

$$\begin{aligned}
l_t &= \alpha \frac{y_t}{w_{t-S_1}} + (1-\alpha)(l_{t-1} + b_{t-1}) && (5.1) \\
b_t &= \gamma(l_t - l_{t-1}) + (1-\gamma)b_{t-1} \\
w_t &= \omega \frac{y_t}{l_{t-1} + b_{t-1}} + (1-\omega)w_{t-S_1} \\
\hat{y}_t(k) &= (l_t + b_t k)w_{t-S_1+k} + \phi^k(y_t - (l_{t-1} + b_{t-1})w_{t-S_1}) && (5.2)
\end{aligned}$$

where $\alpha$, $\gamma$ and $\omega$ are smoothing parameters; $w_t$ is the seasonal index, $b_t$ represents the trend, $l_t$ the level; $S_1$ is the season length, $\hat{y}_t(k)$ is the $k$ step-ahead forecast from forecast origin $t$ and $\phi$ is the parameter of the autocorrelation error correction. Forecasts are produced up to $S_1$ steps ahead.

The multiplicative form of the Holt-Winters-Taylor exponential smoothing model by Taylor (2003), that is used here for the second and third applications (forecasting electricity demand for Rio de Janeiro and England and Wales), has the following formulation:

$$\begin{aligned}
S_t &= \alpha\left(\frac{y_t}{D_{t-S_1}W_{t-S_2}}\right) + (1-\alpha)(S_{t-1}+T_{t-1}) \\
T_t &= \gamma(S_t - S_{t-1}) + (1-\gamma)T_{t-1} \\
D_t &= \delta\frac{y_t}{S_t W_{t-S_2}} + (1-\delta)D_{t-S_1} \\
W_t &= \omega\frac{y_t}{S_t D_{t-S_1}} + (1-\omega)W_{t-S_2}
\end{aligned} \tag{5.3}$$

$$\hat{y}_t(k) = (S_t + kT_t)D_{t-S_1+k}W_{t-S_2+k} + \phi^k\left(y_t - (S_{t-1}+T_{t-1})D_{t-S_1}W_{t-S_2}\right)$$

$S_t$ and $T_t$, are the smoothed level and trend. $D_t$ and $W_t$ are the seasonal indices for the intraday and intraweek seasonal cycles, respectively; $S_1$ and $S_2$ are the intraday and intraweek season lengths, respectively; $\alpha$, $\gamma$, $\delta$, $\omega$ are the smoothing parameters; $\hat{y}_t(k)$ is the $k$ step-ahead forecast made from forecast origin $t$ and $\phi$ is the parameter of the autocorrelation error correction. Forecasts are produced up to $S_1$ steps ahead.

The models were implemented in Matlab® and were used as building blocks to construct ensembles and their structural combination. The level, trend and seasonal components are estimated by averaging the early observations through moving average filters.

## 5.4 Methodology

Structural combination of ensembles of models was conducted following the general procedure described in Section 4.8.1 from Chapter 4 with some variations, as seen in Figure 5.1, illustrated below.

HW and HWT models are characterised by having a low number of parameters, when compared to NNs, and they tend to be more stable. For a given time series,

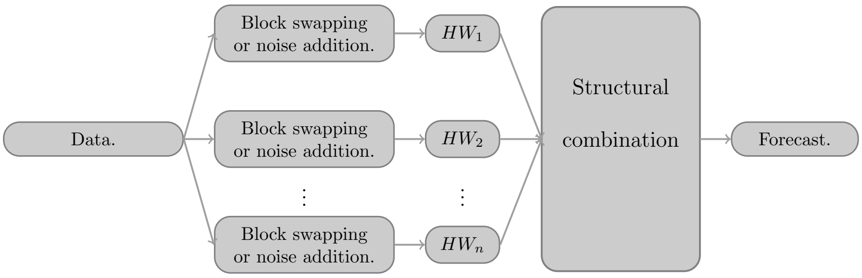

Data.
Block swapping or noise addition.
Block swapping or noise addition.
Block swapping or noise addition.
$HW_1$
$HW_2$
$HW_n$
Structural combination
Forecast.

Figure 5.1: Model scheme.

parameters of HW and HWT models in different trials easily converge to very similar values. In order to perform the structural combinations, model diversity was promoted, by using different mechanisms. Model parameter variations were induced by fitting models to replicas of the original time series (data) that were obtained via two mechanisms. These are the second stage in Figure 5.1 and are described in the next subsection.

In general in this study several time series are forecasted with exponential smoothing models and forecast combinations are found through the structural approach proposed. Results are produced for different levels of noise added or block swapping performed on the series and three different levels of the maximum number of clusters allowed in the combination procedure. Results are compared against the naïve benchmark, the simple average of forecasts produced by NN, the base best model (a model with the best in-sample performance obtained with the original time series in 100 training operations) and standard benchmarks from other studies (see sub-section 5.4.3). It is expected to have a picture of the benefit of structurally combining forecasting models for these specific time series.

### 5.4.1 Block Swapping and Noise Addition

Noise addition (Zhang, 2007) and block bootstrapping (Jing, 1997) have been used in different forms to create ensembles of NNs. Zhang (2007) explained the rationale for the addition of noise to time series data in the following way:

> “in almost all practical time-series problems, it is impossible to make more than one observation at any given time. Thus, although it may be possible to increase the sample size by varying the length of the observed time series, there will only be a single observation on the underlying random variable at time $t$. Nevertheless we may regard the observed time series as just one of a set of an infinite number of time series that might have been observed from the underlying process. [...] The jittered ensemble method is based on the idea that at each time point, many possible observations could be made. Thus, for each realized time series, if we can create many ”noisy” or jittered time series that aim to mimic the behavior of the data generating process, we will have multiple samples and each of these time series can be viewed as a possible realization of the DGP. These jittered time series can then be used to enhance neural network training and model building by effectively forming an ensemble of neural networks built on different samples from the same DGP[1]” (Zhang, 2007, p. 5332)

The same author mentions the moving blocks bootstrap as an alternative to the addition of noise:

> “0ne common resampling method for time series analysis is the moving blocks bootstrap where blocks of consecutive observations are randomly drawn [...]. The basic idea in the moving blocks bootstrap is to form $b$ blocks of data $z_t = (y_t, \ldots, y_{t+k-1})$ of length $k$ from the original time series $(y_1, y_2, \ldots, y_T)$, where $b = T - k + 1$. The sampling with replacement from blocks $(z_1, z_2, \ldots, z_b)$ yields resamples $(z_1^*, z_2^*, \ldots, z_l^*)$” (Zhang, 2007, p. 5333)

[1] DGP stands for data generating process.

The first approach has been used by Zhang (2007) as an alternative to bootstrapping (Efron & Tibshirani, 1994). Both methods aim at improving the generalisation capabilities of ensembles of NNs. The present study adopts both alternatives in the construction of ensembles of exponential smoothing models, with the aim of producing model diversity while attempting to preserve the data process and, finally, improve forecasting accuracy.

Considering Zhang (2007) and Brown et al. (2003) have found that in neural networks the effect of noise addition on performance depends on the noise level and its distribution, in this study initially a normally distributed noise $N(0, k\sigma_b)$ was added to the time series, where $\sigma_b$ was the standard deviation of a bootstrapped replica, $S'$, of the original time series, $S$, and $k$ was a constant that allowed to use a fraction of the standard deviation and thus create different levels of noise[2]. However, it was observed that the use of multiples samples of the standard deviation of $S$ decreased the correlation between in-sample errors for the ensembles and even promoted negative correlation more frequently than the use of a single sample of the standard deviation. Furthermore, with a single sample of the standard deviation, large proportions of the in-sample errors and out-of-sample residuals converged to the same values. The in-sample fit and out-of-sample performance of models under both approaches is similar, but more variety is observed in the second case, which was therefore applied to the ensembles: if $n$ is the length of $S$, then $n$ bootstrapped versions of $S$ are created leading to $n$ samples of the standard deviation, $\vec{\sigma_b}$, which are then used to add noise to $S$ in the form of $\vec{N}(0, k\vec{\sigma_b})$.

The block swapping applied in the present study is a simplified version of the moving block bootstrap in Zhang (2007). This modification aims at reducing a potentially negative effect of the moving block bootstrap on the short-term dependencies in the series. Instead of building entire series from blocks of data taken

[2] Bootstrapping is performed by using the *bootstrp* Matlab routine (see Matlab, 2017).

from the original series, randomly selected pairs of data blocks are swapped in the in-sample period (see Figure 5.2). The block size is equal to the longest seasonal cycle and the number swaps is kept low, in order to guarantee that the structure of the series is fairly preserved[3]. Based on our observations, it appears to be advisable to use small numbers of block swaps in order to produce the variety sought for in the models and yet preserve the general dynamics of the series.

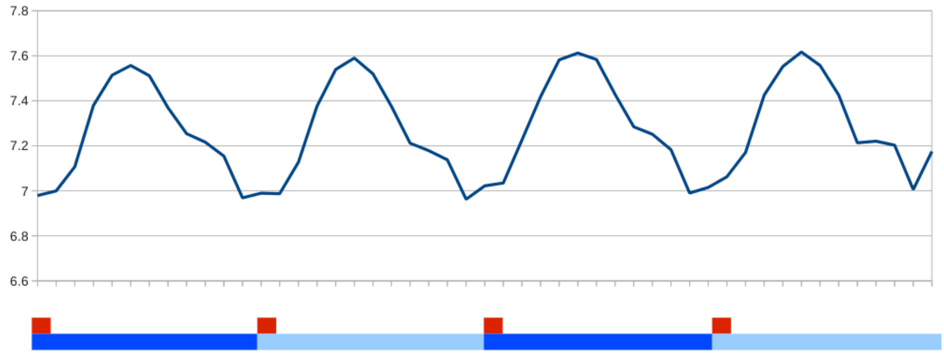


Figure 5.2: Illustration of block swapping. The in-sample period is partitioned in blocks (blue colour) using reference points (red colour). Randomly selected pairs of blocs are interchanged. This is performed $n$ times, with replacement.

These variation inducing mechanisms are used to create pools of MHW and MHWT models, depending on the type of time series, to be combined. Each model is fitted to a different replica of the original series. For a given time series, the diversity induction mechanisms are applied separately: one set of experiments with combinations is done where diversity in models in the pool is induced by fitting individual models to replicas of the time series generated through the swapping of a number of blocks (SW); the second set of experiments uses pools where individual models were fitted to replicas of the time series generated with noise addition. Indi-

[3] For the peak electricity demand and hourly demand time series there were 20 blocks of data in the training in-sample period and the number of swaps, for the different levels, was 2, 4, and 6 (comprising the proportions 0.1, 0.2 and 0.3). For the half-hourly electricity demand series from England and Wales there were 35 blocks of data and the number of swaps, for the different levels, was 3, 7 and 10.

vidual models were estimated by minimising the one-step-ahead root mean squared errors, as in most recent applications in forecasting (e.g. Arora, 2013).

### 5.4.2 Experimental Setup

The maximum number of clusters (2, 4, and 8), the size of the ensembles (50 models) and the number of models selected per cluster (5) were kept as in the previous study with NNs (Chapter 4). This allows for a maximum of 80% of the models in the pool to be included in the combined forecasts.

Three levels of block swapping or noise addition were adopted. For the block swapping method, the adopted levels are $0.1I$, $0.2I$ and $0.3I$, where $I =$ In-sample length $/S_2$ and $S_2$ is the length of the longest cycle in the series. For the case of noise addition, levels $\vec{\sigma_1} = 0.1\vec{\sigma_b}$, $\vec{\sigma_2} = 0.2\vec{\sigma_b}$ and $\vec{\sigma_3} = 0.3\vec{\sigma_b}$ were used, $\vec{\sigma_b}$ being the standard deviation of the bootstrapped original series. A normally distributed noise, $\vec{N}(0, \vec{\sigma_i})$, $i = 1, 2, 3$, was then generated and added to the series, thus leading to three different levels of uncertainty (variation).

### 5.4.3 Analysis Procedure

Results for all application are analysed in two groups, depending on whether variety in the series was induced by noise addition or block swapping. Forecasting performance is assessed by using the Mean Squared Error (MSE) and Mean Absolute Percentage Error (MAPE). For the time horizon, $h$, they are calculated as follows:

$$MSE_h = \frac{1}{N-h-IS+1} \sum_{Fo=IS}^{N-h} (x_{Fo+h} - \hat{x}_{Fo+h})^2 \tag{5.4}$$

$$MAPE_h = \frac{1}{N-h-IS+1} \sum_{Fo=IS}^{N-h} \left| \frac{x_{Fo+h} - \hat{x}_{Fo+h}}{x_{Fo+h}} \right| \tag{5.5}$$

Where $N$ is the length of the time series, $IS$ is the index of the last in-sample observation, $Fo$ is the forecast origin, $x_i$ is the observed value and $\hat{x_i}$ is the forecasted value. Comparisons are made against the following benchmarks:

- Naïve benchmark: the forecast for time period $t$ and lead time $k$ is $\hat{y}_t(k) = y_{t+k-S}$, where $S$ is the longest seasonal cycle.

- The best model in terms of in-sample MSE (denoted as *Best isMSE*): a model is selected from the ensemble, having the lowest in-sample MSE from all the models.

- The best model in terms of the in-sample MAPE (denoted as *Best isMAPE*): a model is selected from the same ensemble, having the lowest in-sample MAPE from all the models.

- Average of point forecasts of all models in the ensemble (denoted as *Avg. models*).

- Base best model (denoted as *Base MHW* or *Base MHWT*): An instance of the base model fitted with the original series (without noise addition or block swapping). The model is selected as having the lowest in sample RMSE metric within 100 trials that had different random starting points.

For the peak electricity demand series two other benchmarks were added. The first is a seasonal ARIMA (SARIMA) model, fitted through the *auto.arima* routine from the *forecast* R package (Hyndman & Khandakar, 2008; Hyndman, 2015). The second is a single NN with iteratively produced forecasts. For the hourly electricity demand, a model evaluated with the parameters that are reported in Taylor et al. (2006) is included. This model is referred to as MHWT PT (to denote that **p**arameters are taken from **T**aylor et al., 2006). For the half-hourly electricity demand in England and Wales, a double seasonal model implementation from R *forecast* package was included.

## 5.5 Structural Combinations of MHW Models to Forecast Daily Peak Electricity Demand

In this investigation, the multiplicative form of the HW model is adopted to forecast the daily peak demand in Rio de Janeiro, using data from Sunday 5 May 1996 to Saturday 30 November 1996 (Taylor et al., 2006). The maximum demand from each day was extracted from the hourly time series, leading to a sample with 210 daily observations. The first 20 weeks of data (140 observations) were used for training (fitting) and the remaining 10 weeks (70 observations) were used for evaluating the accuracy of forecasts up to 7 days ahead. The time series is depicted in Figure 5.3, where seasonality and time-varying volatility are observed.

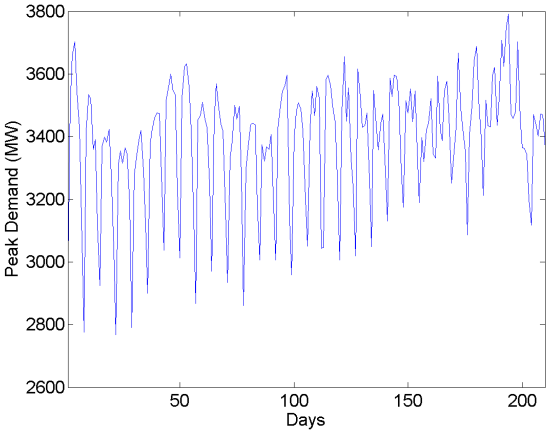


Figure 5.3: Daily peak electricity demand in Rio de Janeiro from Sunday, 5 May 1996 to Saturday, 30 November 1996. Original Series.

The seasonal ARIMA (SARIMA) model that is used as benchmark was obtained through automated routines available in the *forecast* R package, by using a logarithmic transformation of the time series. The specification is the following:

$$(1-\phi_1 B-\phi_2 B^2)(1-\Phi_1 B^7)(1-B^7)(y_t-\mu)=(1+\theta_1 B)(1+\Theta_1 B^7+\Theta_2 B^{14})e_t$$

where $\phi_1 = 1.2609, \phi_2 = -0.3271, \Phi_1 = -0.7837,$

$\theta_1 = -0.7529, \Theta_1 = 0.0012, \Theta_2 = -0.5595$ and $\mu = 0.0001$

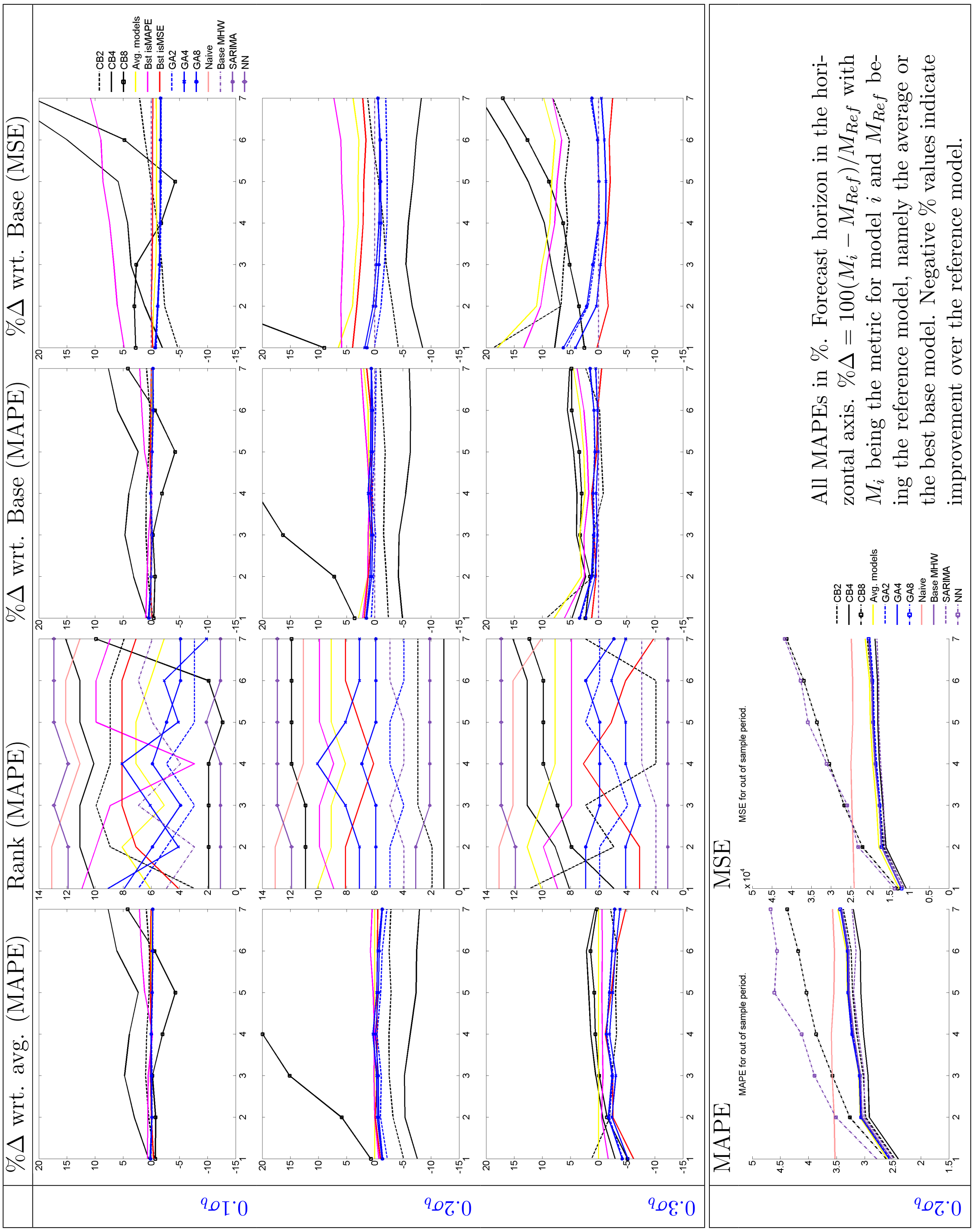
$B$ denotes the backward shift operator. Additionally, a feed-forward NN was fitted with the series differenced as $d_t = (1-B)(1-B^7)y_t$ and using lags 1, 2, 6, 7, 8, 13, 14, 15 as inputs. An exploration of a range of configurations, with models having from 2 to $Num.\ Inputs + 1 = 9$ neurons, led to a NN with 4 hidden units. This NN was used to iteratively forecast 7 steps (days) ahead.

### 5.5.1 Results when Model Variation was Introduced through Noise Addition to Generate MHW Combinations

Figures 5.4 and 5.5 summarise the results based on experiments where variation in the time series was introduced via noise addition to the original series. They provide a sub-set of out-of-sample MAPE and MSE for the different CB models and the selected benchmarks. Significant volatility can be observed in CB combinations (bottom of Figure 5.4), whereas GA combinations are stable, with similar performance to the base best model.

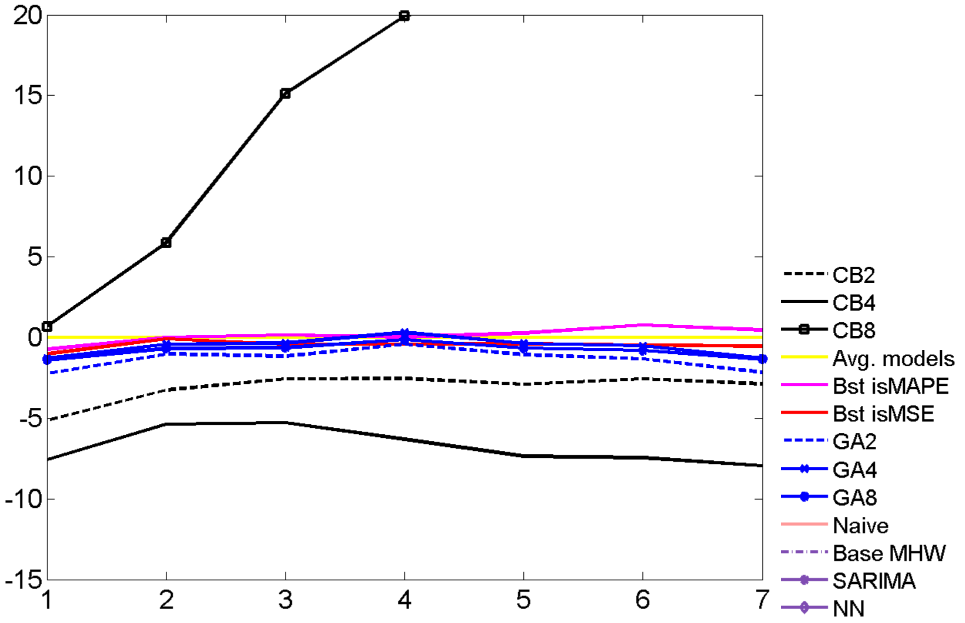
Figure 5.4 (top) and 5.5 also summarise forecasting performance, for each forecast horizon with rankings and percentage differences with respect to the average forecast and the base best model. The percentage of difference is negative when the model has a smaller error metric than the reference model (average or base model) and positive when the model has a higher error. This difference is calculated only for models derived from the ensemble: CB combinations, GA combinations, best model in terms of in-sample MAPE and best model in terms of in-sample MSE.

Again, the CB model is volatile. In terms of MAPE, for lowest level of noise, a CB combination with a maximum of 8 clusters provides remarkable improvement

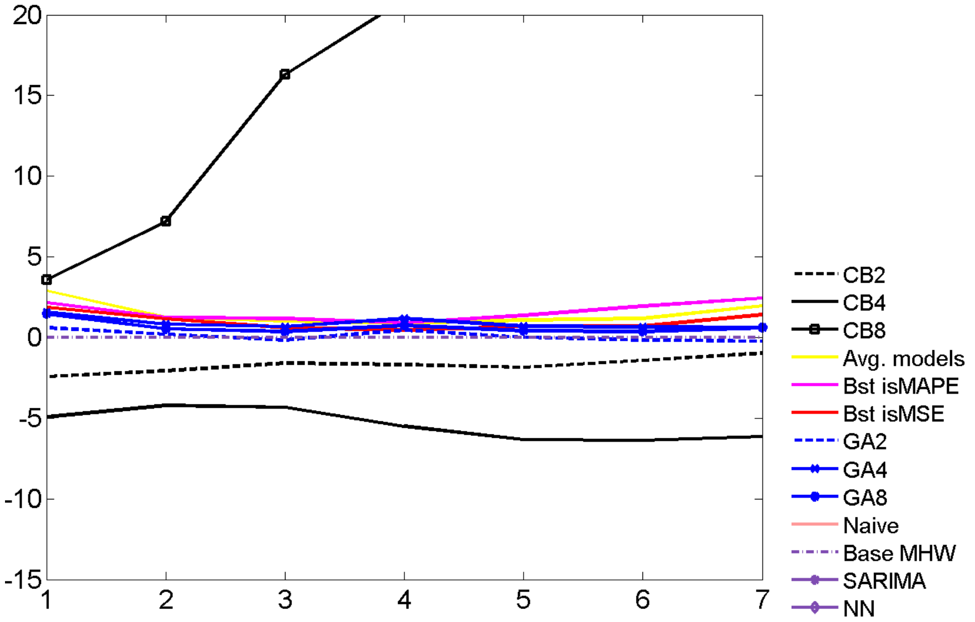


All MAPEs in %. Forecast horizon in the horizontal axis. $\%\Delta = 100(M_i - M_{Ref})/M_{Ref}$ with $M_i$ being the metric for model $i$ and $M_{Ref}$ being the reference model, namely the average or the best base model. Negative % values indicate improvement over the reference model.

Figure 5.4: Summary for Rio de Janeiro peak electricity demand under noise addition.



(a) %Δ wrt. avg. (MAPE).



(b) %Δ wrt. Base model (MAPE).

Figure 5.5: %Δ difference in metric vs. forecast horizon for ensembles generated with noise at $0.2\sigma_b$.

with respect to the average and, additionally, outperforms most benchmarks (the SARIMA model is closely followed and outperformed in one forecast horizon). For the level 2 (medium) of noise, CB8 model performs very poorly, but CB4 improves over the average and all other benchmarks. For level noise 3, CB2 is the only CB model that was able to improve over the average. In terms of MSE (results not shown), the patterns of improvement over the average and the model rankings change, mostly for the first (lowest) level of noise. However, the relevance of CB4 for the level of noise 2 remains.

In general, GA combinations are much more stable, consistently providing marginal improvements over the average (the higher the level of noise, the higher the gain). However, they are unable to outperform the SARIMA benchmark. Improvement over the base best model (Figure 5.4) is observed clearly in CB models for levels of noise 1 and 2. GA combinations, on the other hand, have a very close performance to the base best model, without outperforming it consistently. CB4, the best performing in this set of models (produced with noise addition), is reduced to a single cluster model during the pruning stage of the clustering algorithm. Finally, looking at the MAPE and MSE rankings of models, a more stable performance is observed from the medium level (2) of noise, which also produced the best performing CB model.

Table 5.1 shows sample models (see Equation 5.1) from the three different pools created with different noise levels. Some variety in the coefficients is noticeable, which is one of the desirable features for the structural combination. Table 5.2 shows the set of parameters for a CB model obtained with a pool of MHW models fitted with noisy time series. The first set of parameters ($\alpha_{\mathbf{j}}$) are the coefficients to be applied to point-forecasts from the selected models for every cluster, and $\Phi$ defines the weights applied to the forecasts extracted from each cluster. The weighting of models within clusters and the weighting in the overall forecasts from clusters is diverse, which confirms that CB combinations are making use of model diversity.

Jarque-Bera and Lilliefors tests for normality and Ljung-Box test for serial correlation were conducted on the forecast errors. Evidence of normality and serial independence is common at the first level of noise. However, with greater variation (noise), there is serial correlation in forecast errors for horizons $h > 1$ in all models derived from the ensembles.

Table 5.1: Sample MHW models with noise addition.

| With noise at level 1: $N(0, 0.1\sigma_b)$ | | | |
|---|---|---|---|
| $\alpha$ | $\gamma$ | $\omega$ | $\phi$ |
| 0.1705 | 0.0130 | 0.1513 | 0.3113 |
| 0.1294 | 0.0000 | 0.1583 | 0.3951 |
| 0.1486 | 0.0000 | 0.1692 | 0.3724 |
| With noise at level 2: $N(0, 0.2\sigma_b)$ | | | |
| $\alpha$ | $\gamma$ | $\omega$ | $\phi$ |
| 0.1256 | 0.0001 | 0.1556 | 0.3587 |
| 0.1888 | 0.0235 | 0.1239 | 0.2111 |
| 0.1389 | 0.0000 | 0.1547 | 0.3437 |
| With noise at level 3: $N(0, 0.3\sigma_b)$ | | | |
| $\alpha$ | $\gamma$ | $\omega$ | $\phi$ |
| 0.1246 | 0.0084 | 0.1329 | 0.3223 |
| 0.1158 | 0.0323 | 0.0631 | 0.3044 |
| 0.1943 | 0.0333 | 0.1099 | 0.1215 |

Each row corresponds to a model.

Table 5.2: Coefficients for sample CB combination of MHW models.

| CB8 | | | | | | |
|---|---|---|---|---|---|---|
| $\alpha_1$ | 233.5037 | -4.2509 | -12.1411 | 4.5640 | 1.5756 | 20.7354 |
| $\alpha_2$ | 243.2322 | -23.6656 | 3.2236 | -9.5765 | 0.7404 | |
| $\alpha_3$ | 253.8552 | 33.0597 | -12.1856 | -32.0101 | 17.4262 | 9.9941 |
| $\alpha_4$ | 245.9087 | -22.1712 | 1.7946 | -9.1367 | 3.9267 | 33.6397 |
| $\Phi$ | 0.2541 | 0.2582 | 0.2308 | 0.2568 | | |

Noise $\sim N(0, 0.1\sigma_b)$.

### 5.5.2 Results when Model Variation was Introduced through Block Swapping to Generate MHW Combinations

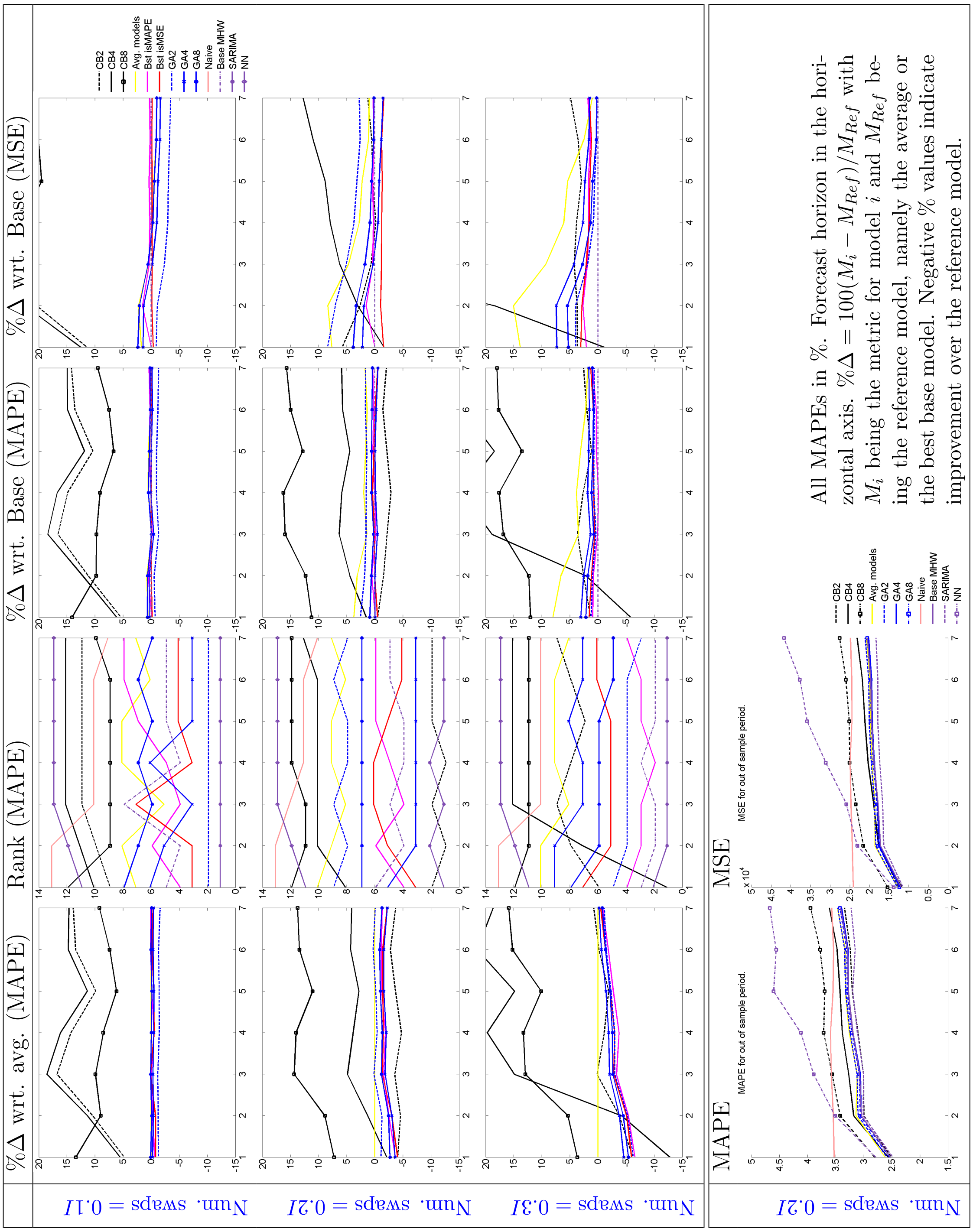
The number of in-sample observations for the electricity peak demand series is 140 and $S_1 = 7$. The number of block swaps ($I =$ Train length$/S_1$) is $0.1I = 2$, $0.2I = 4$ and $0.3I = 6$ for the three levels used in this variant of ensembles.

Figures 5.6 and 5.7 summarise combinations where block swapping was applied to the original time series. Deterioration of performance in CB combinations is significant (bottom left of Figure 5.6), when comparing with the noise addition results that were described above, whereas GA combinations tend to be stable and perform better.

Figures also show the ranking of models and the percentage differences of performance with respect to the average of forecasts of models in the ensemble and the base best model. When focusing on the rankings, the graphs show the stability in performance of GA models since their error metrics are located in a clearly distinguishable range, whereas performance for CB varies widely. In the graphs that summarise the improvement with respect to the average and the base best model, the higher volatility of CB with respect to more stable GA can be fully appreciated.

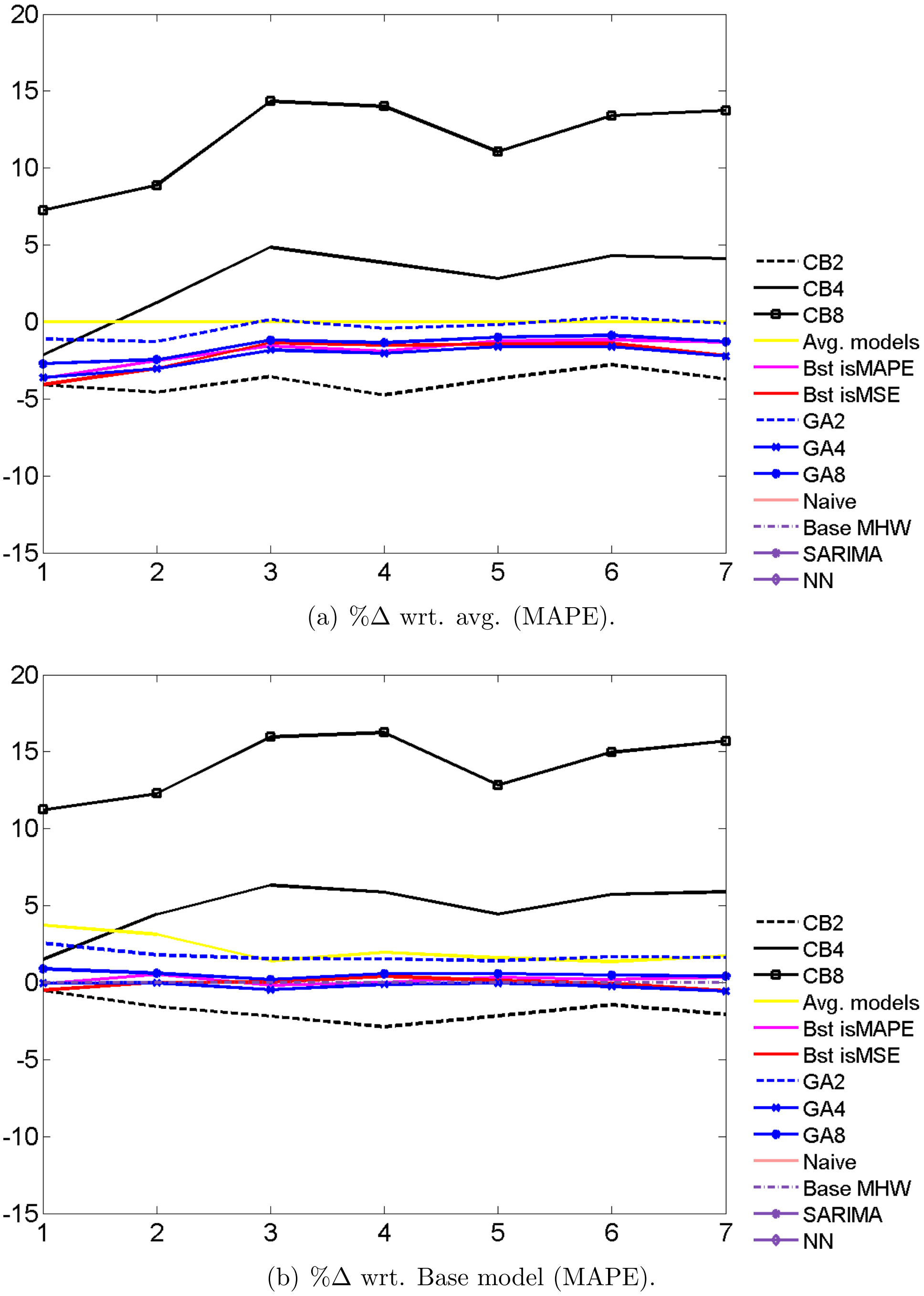
In general, CB combinations remain volatile and GA continue to be stable. GA2 consistently outperforms the average in the first level of block swapping and CB2 does it in the second level. For the highest level of block swapping, both models outperform the average in most of the forecast horizons, but the other combinations and best models also offer advantages with respect to the average. GA2 (for the first level of block swapping) and CB2 (for the second level) also improve over the base best model, but GA is more consistent. Overall, GA combinations outperform CB combinations.

Table 5.3 provides details of selected MHW models generated with block swapping and Table 5.4 shows parameters for a sample CB model obtained from a pool of



All MAPEs in %. Forecast horizon in the horizontal axis. $\%\Delta = 100(M_i - M_{Ref})/M_{Ref}$ with $M_i$ being the metric for model $i$ and $M_{Ref}$ being the reference model, namely the average or the best base model. Negative % values indicate improvement over the reference model.

Figure 5.6: Summary for Rio de Janeiro peak electricity demand under block swapping.



(a) %Δ wrt. avg. (MAPE).

(b) %Δ wrt. Base model (MAPE).

Figure 5.7: %Δ difference in metric vs. forecast horizon for ensembles generated with Num. swaps $= 0.2I$.

such models (the maximum number of clusters was two for this model and the pruning routine left only one). Variation is noticed in the $\alpha$ parameters, which signals a diverse contribution of models to the final forecast. This relatively well-performing model was produced with the medium level of block swapping in an ensemble with more stable ranking of models (in terms of MAPE).

In the post-estimation tests that were conducted there is evidence of normality in forecast errors, but serial correlation persists in most forecast horizons. However, when comparing with results when variation was introduced via noise addition, the models with block swapping exhibit less serial correlation for $h = 2$.

Table 5.3: Sample MHW models with block swapping.

| With block swapping at level 1 | | | |
|---|---|---|---|
| $\alpha$ | $\gamma$ | $\omega$ | $\phi$ |
| 0.1266 | 0.0000 | 0.1523 | 0.4042 |
| 0.0970 | 0.0000 | 0.1690 | 0.4405 |
| 0.1581 | 0.0000 | 0.1298 | 0.3774 |
| With block swapping at level 2 | | | |
| $\alpha$ | $\gamma$ | $\omega$ | $\phi$ |
| 0.0990 | 0.0007 | 0.0000 | 0.4761 |
| 0.1250 | 0.0000 | 0.1585 | 0.4651 |
| 0.1709 | 0.0000 | 0.1184 | 0.3290 |
| With block swapping at level 3 | | | |
| $\alpha$ | $\gamma$ | $\omega$ | $\phi$ |
| 0.1253 | 0.0000 | 0.0231 | 0.4736 |
| 0.1165 | 0.0000 | 0.1705 | 0.4595 |
| 0.0408 | 0.0320 | 0.1451 | 0.6324 |

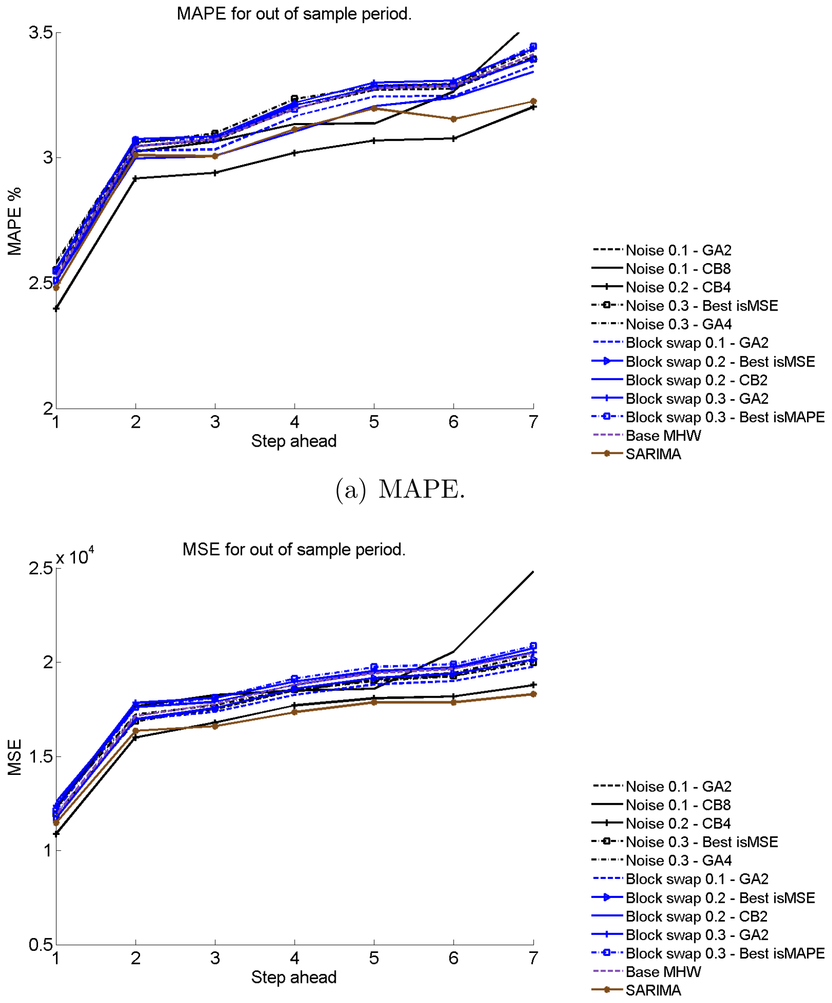
Each row corresponds to a model.

Table 5.4: Coefficients for sample CB combination of MHW models.

| CB2 | | | | | | |
|---|---|---|---|---|---|---|
| $\alpha_1$ | 252.8867 | 0.4311 | 9.0562 | -10.8901 | 6.7905 | -4.4615 |
| $\Phi$ | 1 | | | | | |

Num. swaps = $0.2I$, where $I$ = Train length/$S_1$.

Considering the results obtained for the peak electricity demand, GA combinations perform well in both noise addition and block swapping, but most notably for the latter approach. By contrast, CB models do better with noise addition than with block swapping, but are volatile with respect to performance. Considering the medium level of noise, such combination, with a maximum of 4 clusters, is capable of improvements over the base and average models (once outperforming all benchmarks): improvement of this CB4 with respect to base model ranges from 4.93% to 6.14% in MAPE and between 7.39% and 8.54% for MSE. Improvement over the SARIMA model ranges between 0.71% and 3.33% in terms of MAPE for all forecast horizons, although in terms of MSE, the improvement is only present in the first two forecast horizons: 5.11% and 2.10%. In most configurations (of noise and maximum number of clusters) they perform worse than the benchmarks in several forecast horizons. Figure 5.8 shows a selection of the best models with their performance in terms of MSE and MAPE. Performance curves for models under noise addition are sparse compared to the performance curves of models with block swapping. In general, improvements over the base best model, the average and the SARIMA benchmark suggest that MHW models can be structurally combined and this combination can be competitive against established benchmarks.



(a) MAPE.

(b) MSE.

Figure 5.8: Best ensemble-based models for Rio de Janeiro peak electricity demand.

## 5.6 Structural Combinations of MHWT Models to Forecast Rio de Janeiro Electricity Demand

We refer to the double-seasonal time series in Section 4.8, Chapter 4, to forecast electricity demand. It is now used to investigate combinations of multiplicative Holt-Winters-Taylor (MHWT) models. Hourly observations from Sunday 5 May 1996 to Saturday 30 November 1996 (see Figure 5.9) are considered. Base model and data partitioning follow the procedure by Taylor et al. (2006). The MHWT was used to generate ensembles using the first 20 weeks of data (equivalent to 3360 hourly observations) for training (fitting) and the remaining 10 weeks (equivalent to 1680 observations) for evaluation of the accuracy of forecasts up to 24 hours ahead.

The results of structural combinations are compared against the benchmarks described in Section 5.4.3, as well as a MHWT model evaluated with the parameters that are reported in Taylor et al. (2006): $\alpha = 0.01$, $\gamma = 0.00$, $\delta = 0.09$, $\omega = 0.15$ and $\phi = 0.88$ (see Equation 5.3). This model is referred to as MHWT PT (to denote that **p**arameters are taken from **T**aylor et al., 2006).

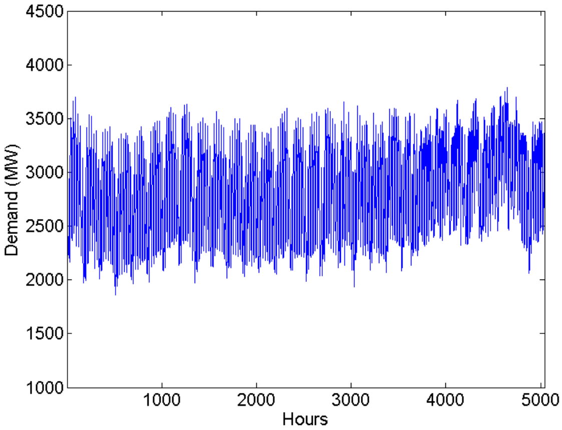


Figure 5.9: Hourly electricity demand in Rio de Janeiro for Sunday, 5 May 1996 to Saturday, 30 November 1996. Original Series.

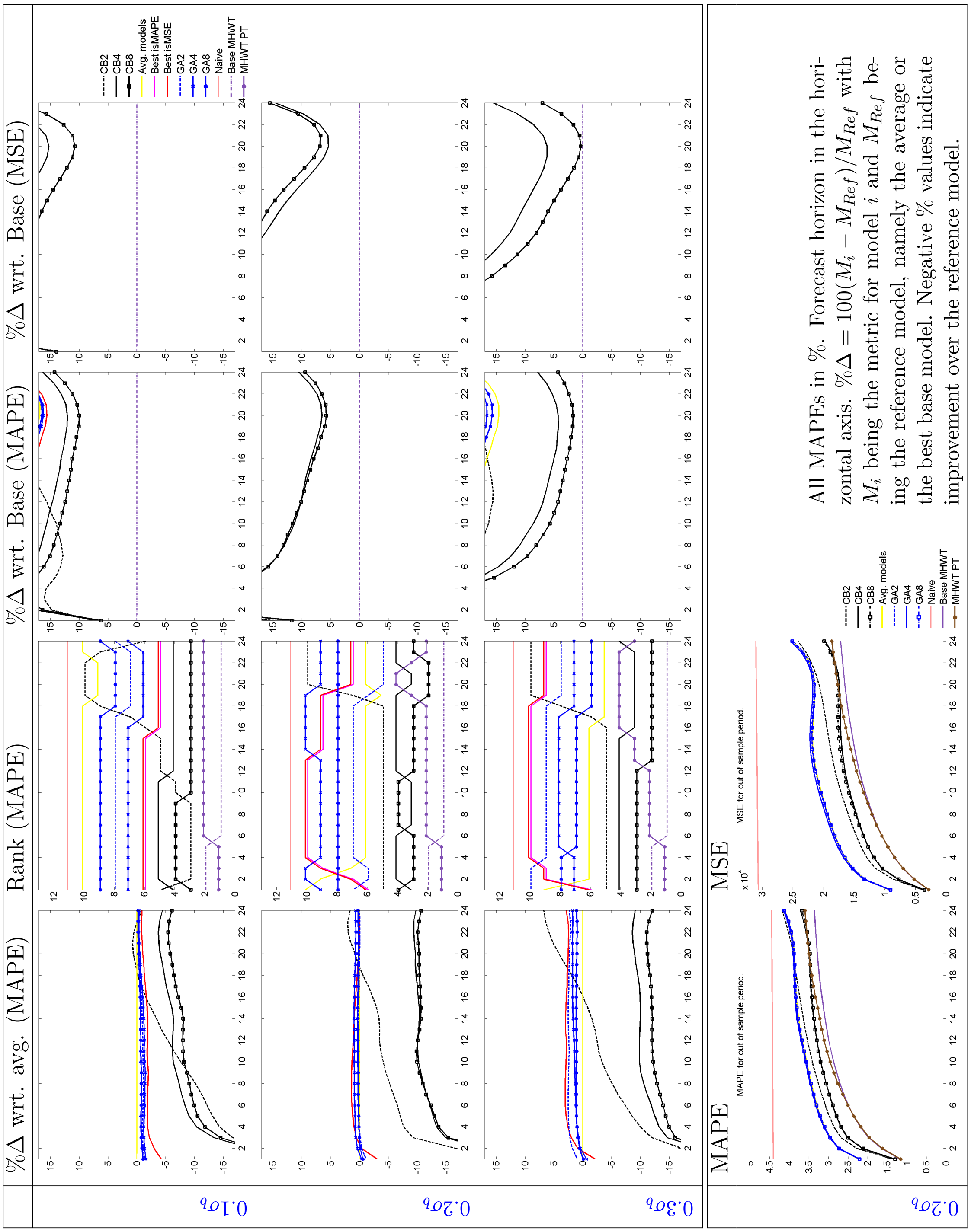
### 5.6.1 Results when Model Variation was Introduced through Noise Addition to Generate MHWT Combinations

Figures 5.10 and 5.11 summarise forecasting performance of the combinations that were based on models that were estimated using the disturbed original time series (through noise addition). Performance of CB combinations, GA and the simple average of the individual forecasts in the ensemble deteriorates with this training strategy, and this is evident by the superior performance of the base best model and MHWT PT. However, the rankings and percentage differences with respect to the average show how CB combinations are able to consistently improve over the simple average during the first 16 hours. Although all models do not compare well to the base best model, Figure 5.10 shows how CB combinations perform comparatively better than most benchmarks. The individual models with best fit metrics do not perform well out of sample. Overall, from the rankings, a relatively homogeneous ordering of models emerges, with clear abrupt changes in the case of CB models.

Contrary to the case of the single seasonal peak load series when noise was added, CB models outperform the average more consistently than GA combinations. The latter tend to perform poorly. As the level of noise is increased, GA combinations' performance further deteriorates.

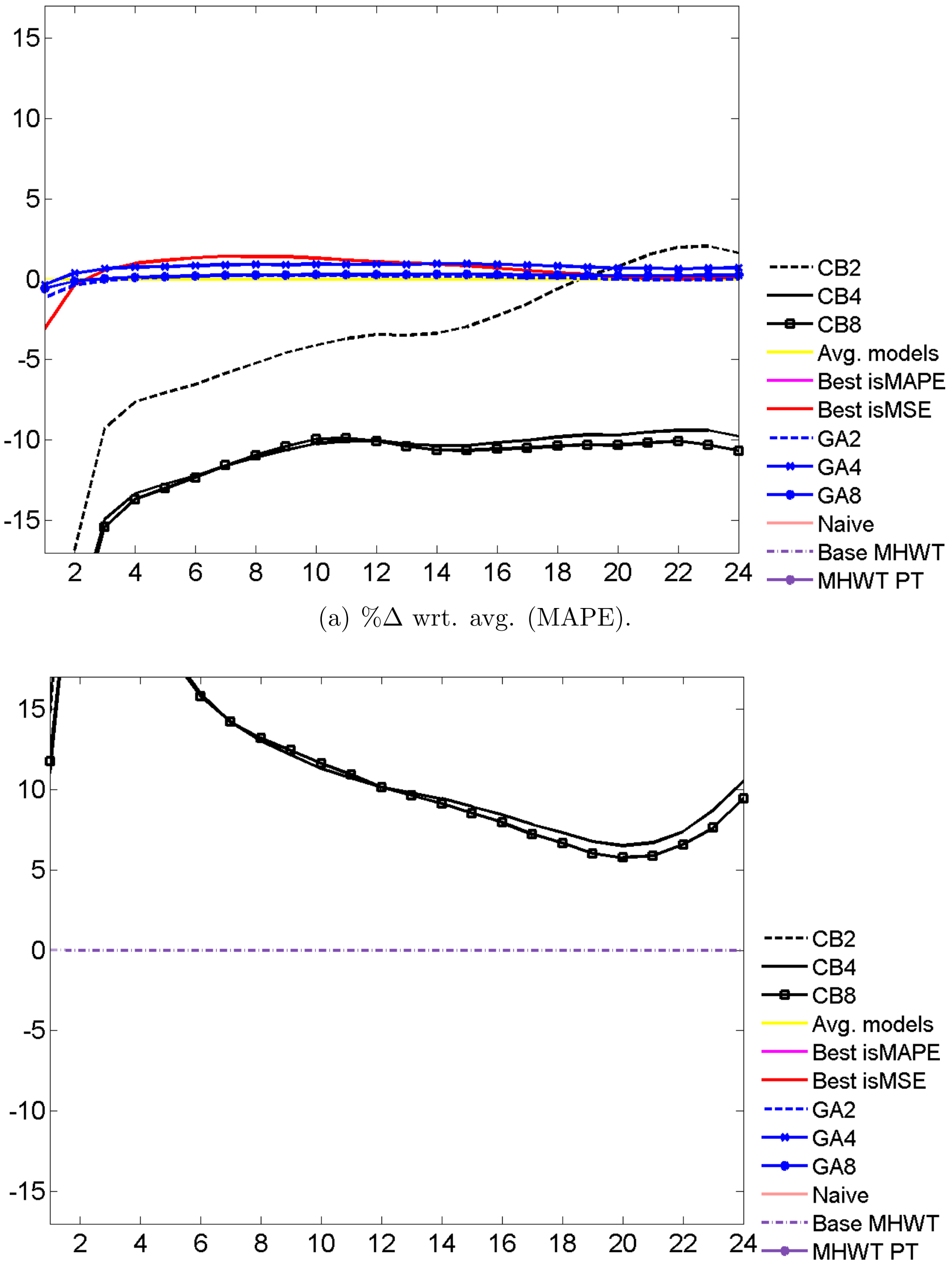
Sample models from the ensembles are detailed in Table 5.5. Parameters values for $\gamma$ and $\omega$ are similar to the estimates obtained by Taylor et al. (2006), whereas differences are also observed: the MHWT model estimates seem to be less sensitive to the addition of noise when capturing the dynamics of the trend and the weekly cycle ($\omega$).

The best performing model obtained with the noise addition approach (CB8) is specified in Table 5.6. Variation in the coefficients for forecasts in every cluster is observed, but also a relatively homogeneous weighting of forecasts produced by clusters. This means that variety in forecasts is exploited mostly at the level of clus-



All MAPEs in %. Forecast horizon in the horizontal axis. $\%\Delta = 100(M_i - M_{Ref})/M_{Ref}$ with $M_i$ being the metric for model $i$ and $M_{Ref}$ being the reference model, namely the average or the best base model. Negative % values indicate improvement over the reference model.

Figure 5.10: Summary for Rio de Janeiro hourly demand under noise addition.



(a) %Δ wrt. avg. (MAPE).

(b) %Δ wrt. Base model (MAPE).

Figure 5.11: %Δ difference in metric vs. forecast horizon for ensembles generated with noise at $0.2\sigma_b$.

ters and less at the global level. According to the Jarque-Bera test, most ensemble models for several forecast horizons ($2 \leq h \leq 12$), had normally distributed forecast errors, when the level of noise is low (Table 5.7). For the medium level of noise, the pattern changes (Table 5.8), as only CB2 model led to normally distributed errors in several forecast horizons. For the high level of noise, normality of the errors was rejected. These results are inconclusive as the Lilliefors test rejected normality, when considering all the models and levels of noise. The Ljung-Box indicated serially correlated forecast errors for all models in all forecast horizons, thus highlighting that they failed to capture the dynamic of the time series.

Table 5.5: Sample MHWT models with noise addition.

| From the pool with noise at level 1: $N(0, 0.1\sigma_b)$ | | | | |
|---|---|---|---|---|
| $\alpha$ | $\gamma$ | $\delta$ | $\omega$ | $\phi$ |
| 0.1475 | 0.0000 | 0.0399 | 0.1504 | 0.3922 |
| 0.1520 | 0.0000 | 0.0370 | 0.1517 | 0.3475 |
| 0.1360 | 0.0000 | 0.0449 | 0.1355 | 0.3997 |
| From the pool with noise at level 2: $N(0, 0.2\sigma_b)$ | | | | |
| $\alpha$ | $\gamma$ | $\delta$ | $\omega$ | $\phi$ |
| 0.1393 | 0.0000 | 0.0160 | 0.1574 | 0.1222 |
| 0.1103 | 0.0000 | 0.0219 | 0.1417 | 0.1897 |
| 0.1171 | 0.0000 | 0.0245 | 0.1394 | 0.1714 |
| From the pool with noise at level 3: $N(0, 0.3\sigma_b)$ | | | | |
| $\alpha$ | $\gamma$ | $\delta$ | $\omega$ | $\phi$ |
| 0.0545 | 0.0001 | 0.0053 | 0.1651 | 0.0864 |
| 0.0759 | 0.0000 | 0.0000 | 0.1616 | 0.0993 |
| 0.0791 | 0.0000 | 0.0152 | 0.1387 | 0.1051 |

Each row corresponds to a model.

Table 5.6: Coefficients for sample CB combination of MHWT models.

| | | | | | | |
|---|---|---|---|---|---|---|
| CB8 | | | | | | |
| $\alpha_1$ | 7.7504 | 9.6292 | -48.8083 | 52.4999 | 0.0024 | 5.6734 |
| $\alpha_2$ | 13.6929 | -11.4271 | 24.4374 | -0.7998 | -24.9317 | 25.1327 |
| $\alpha_3$ | -0.5409 | 16.2968 | 2.0425 | -18.6673 | 24.8590 | -21.4949 |
| $\alpha_4$ | 9.0731 | -14.4804 | 5.1019 | 47.3863 | -29.6399 | -12.2696 |
| $\alpha_5$ | -0.9714 | -22.5574 | -1.0564 | 4.7719 | 9.1180 | -15.7031 |
| $\alpha_6$ | 27.7027 | 13.4725 | 8.3640 | -20.6679 | | |
| $\Phi$ | 0.1669 | 0.1673 | 0.1687 | 0.1668 | 0.1692 | 0.1644 |

Noise $\sim N(0, 0.1\sigma_b)$.

Table 5.7: Jarque-Bera test of forecast errors for HWT combinations (low noise).

| | | Forecast horizon | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 |
| 2C | CB | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | | |
| | GA | * | * | * | | | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * |
| 4C | CB | * | * | | * | | | | | | | | | * | * | * | * | * | * | * | * | * | * | * | * |
| | GA | * | * | | | | * | * | * | | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * |
| 8C | CB | * | * | * | * | * | * | | | | | | | * | * | * | * | * | * | * | * | * | * | * | |
| | GA | * | * | * | * | | * | | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * |
| | Avg. | * | * | * | | | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * |
| | Bst IsMAPE | * | | * | * | | | * | * | | | | * | * | * | * | * | * | * | * | * | * | * | * | * |
| | Bst IsMSE | * | | * | * | | | * | * | | | | * | * | * | * | * | * | * | * | * | * | * | * | * |

Noise $\sim N(0, 0.1\sigma_b)$.
The rejection of the hypothesis of normally distributed errors (with 95% confidence level) is indicated with *.

Table 5.8: Jarque-Bera test of forecast errors for HWT combinations (medium noise).

| | | Forecast horizon | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 |
| 2C | CB | * | * | | | | | | | | | | | | | | | | | * | * | * | * | * | |
| | GA | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * |
| 4C | CB | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * |
| | GA | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * |
| 8C | CB | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * |
| | GA | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * |
| | Avg. | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * |
| | Bst IsMAPE | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * |
| | Bst IsMSE | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * |

Noise $\sim N(0, 0.2\sigma_b)$.
The rejection of the hypothesis of normally distributed errors (with 95% confidence level) is indicated with *.

### 5.6.2 Results when Model Variation was Introduced through Block Swapping to Generate MHWT Combinations

Figures 5.12 and 5.13 provide graphs of the out-of-sample MAPE and MSE of ensemble models as well as the combinations created through block swapping. A very similar performance is noticed. Rankings and percentage differences with respect to the average highlight how difficult it is for combinations in this setting to outperform the average of forecasts. Only CB2 has performed better than the simple average in most of horizons for the first 2 levels of block swapping under MAPE and GA8 for the third level. When considering MSE, the comparative performance of CB2 is similar for the first two levels of block swapping, but in the highest level, the simple average outperforms all models. All improvements with respect to the average forecast are marginal, as it can be seen in the percentage differences given in the same graphs.

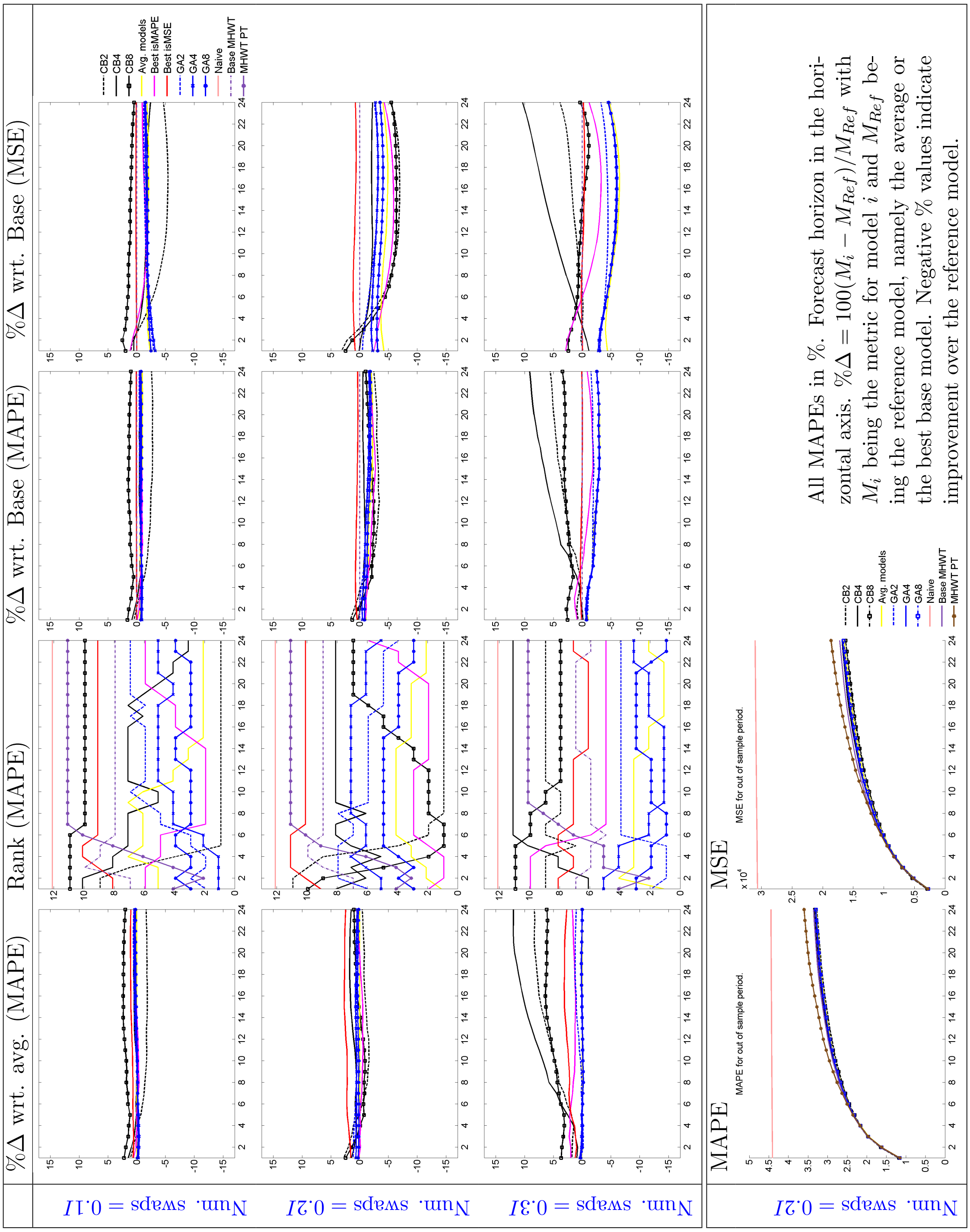
Consistent improvement over the base best model is observed for GA combinations and the average (Figure 5.12). CB combinations improve on such benchmark in an irregular way: CB2 in first level of block swapping and all CB combinations in the second level, for $h \geq 3$; for the last level, improvement is scant and only present under MSE. The models with the lowest in-sample metrics tend to perform well, which suggests that relatively well-performing ensembles were created.

Sample MHWT models from the ensembles created with block swapping are detailed in Table 5.9. Parameters are similar to the values obtained by Taylor et al. (2006). Models appear to capture the time series dynamics, which was not the case when the ensembles were generated using noise addition to the original series to create diversity.

Parameters from the best performing model (CB2) are reported in Table 5.10. Greater stability in model parameters is noticed in this setting, when comparing with the noise addition. A heterogeneous contribution of models inside clusters to their outputs and from clusters to the final forecast is also noticed in the variations within $\alpha_j$ and $\Phi$ parameters.

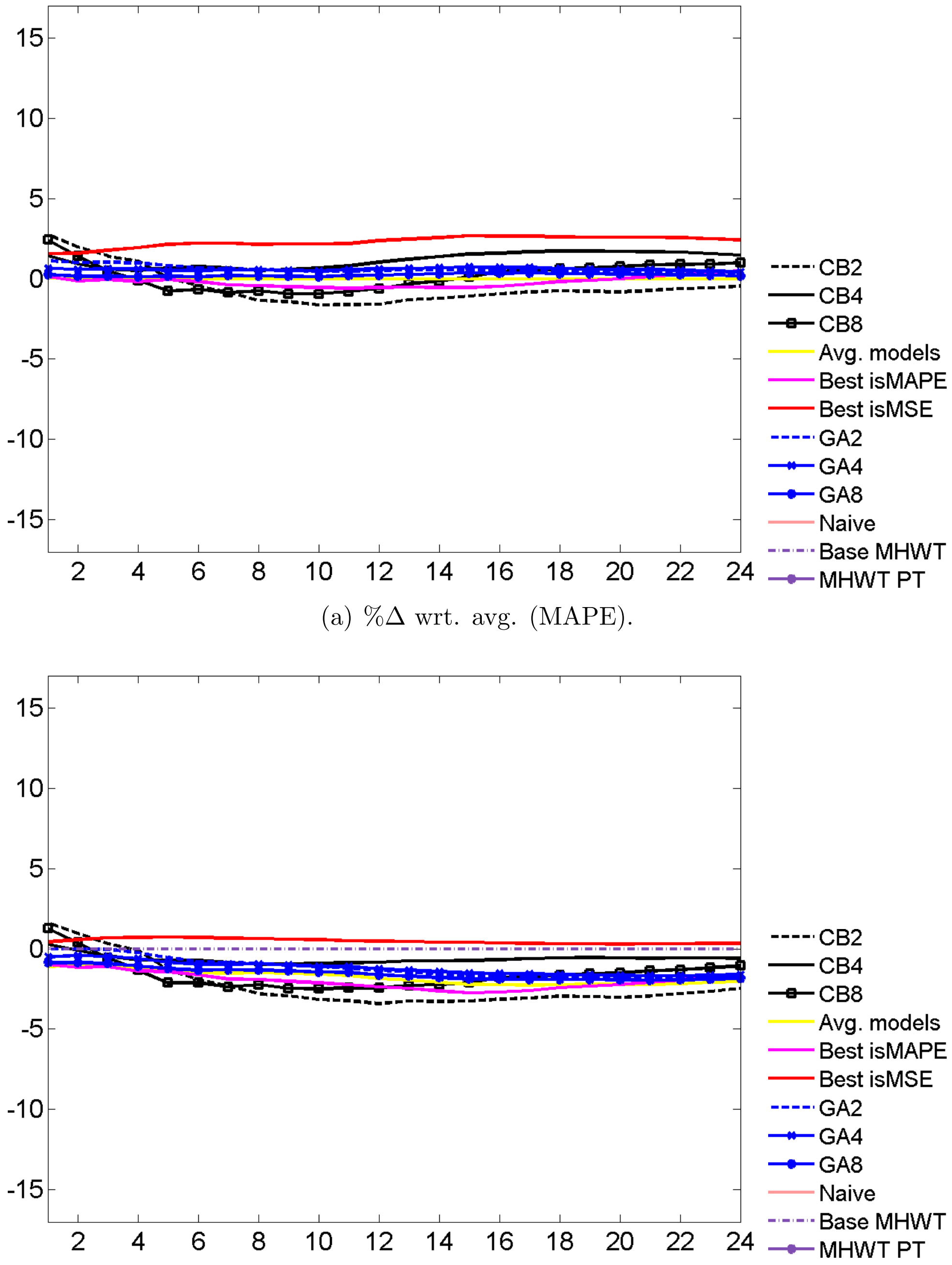
The Jarque-Bera and Lilliefors tests rejected the hypothesis normally distributed forecast errors. Additionally, the Ljung-Box supported serial correlation in forecast errors for all models and forecast horizons. Hence, the full dynamics of the time series are not captured, though MAPEs are significantly lower than 5% on all forecast horizons and about 1% in the case of one-step-ahead forecasts.

Results indicate that for this double-seasonal time series, the best performance is achieved with block swapping, as shown in Figure 5.14. All models based on block swapping perform similarly to the simple average, which is known to be robust to serial correlation in forecast errors and changes in the data pattern. Nevertheless, CB2 and GA4 provide marginal improvements with respect to the best average (from the ensemble with block swapping at level 3). Additionally, they outperform the base best model, and thus are promising alternatives for this task.



All MAPEs in %. Forecast horizon in the horizontal axis. $\%\Delta = 100(M_i - M_{Ref})/M_{Ref}$ with $M_i$ being the metric for model $i$ and $M_{Ref}$ being the reference model, namely the average or the best base model. Negative % values indicate improvement over the reference model.

Figure 5.12: Summary for Rio de Janeiro hourly demand under block swapping.



(a) %Δ wrt. avg. (MAPE).

(b) %Δ wrt. Base model (MAPE).

Figure 5.13: %Δ difference in metric vs. forecast horizon for ensembles generated with Num. swaps = $0.2I$.

Table 5.9: Sample MHWT models with block swapping.

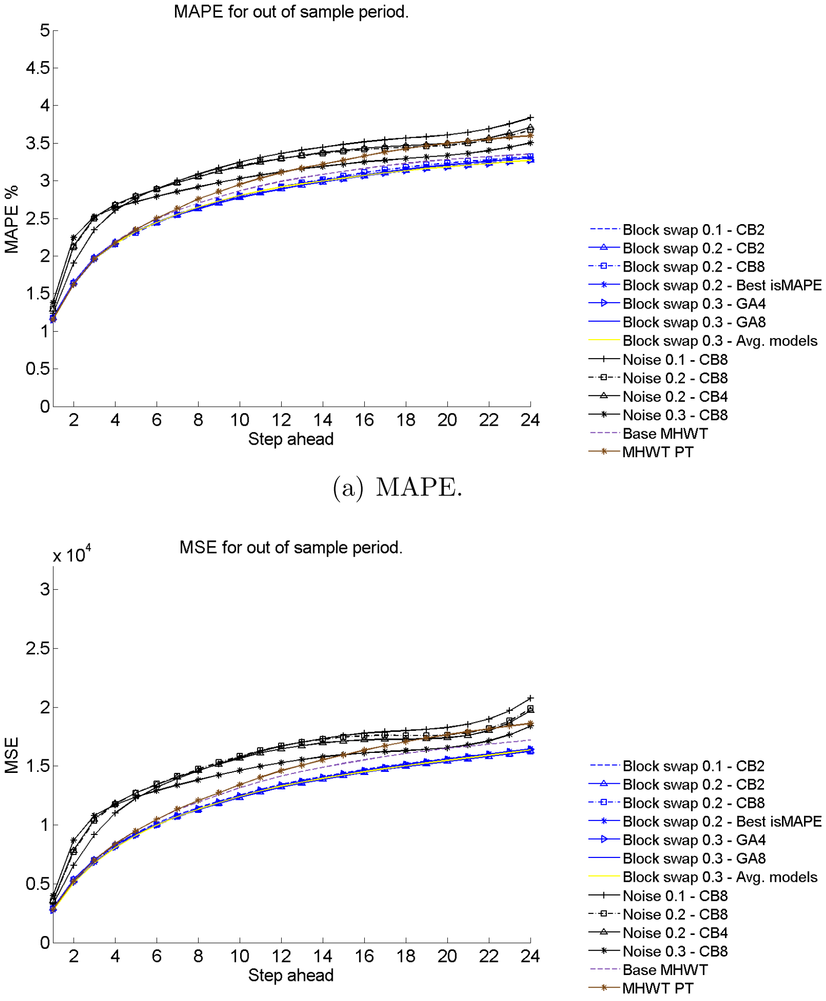
| From the pool with block swapping at level 1 (Num. swaps = $0.1I$) | | | | |
|---|---|---|---|---|
| $\alpha$ | $\gamma$ | $\delta$ | $\omega$ | $\phi$ |
| 0.0088 | 0.0013 | 0.0916 | 0.1167 | 0.8534 |
| 0.0141 | 0.0012 | 0.0841 | 0.1775 | 0.8404 |
| 0.0096 | 0.0013 | 0.0955 | 0.1357 | 0.8638 |
| From the pool with block swapping at level 2 (Num. swaps = $0.2I$) | | | | |
| $\alpha$ | $\gamma$ | $\delta$ | $\omega$ | $\phi$ |
| 0.0090 | 0.0017 | 0.0868 | 0.1462 | 0.8576 |
| 0.0084 | 0.0016 | 0.1041 | 0.1550 | 0.8655 |
| 0.0092 | 0.0013 | 0.1090 | 0.1325 | 0.8700 |
| From the pool with block swapping at level 3 (Num. swaps = $0.3I$) | | | | |
| $\alpha$ | $\gamma$ | $\delta$ | $\omega$ | $\phi$ |
| 0.0011 | 0.0095 | 0.2271 | 0.1270 | 0.8844 |
| 0.0081 | 0.0011 | 0.0804 | 0.1318 | 0.8666 |
| 0.0182 | 0.0007 | 0.0811 | 0.1425 | 0.8440 |

Each row corresponds to a model.

Table 5.10: Coefficients for sample CB combination of MHWT models.

| | | | | | | |
|---|---|---|---|---|---|---|
| CB2 | | | | | | |
| $\alpha_1$ | 9.6908 | -0.4484 | 0.5746 | -3.7500 | 1.7764 | 4.2474 |
| $\alpha_2$ | 9.6010 | -3.6880 | 1.7389 | -2.2772 | 1.3248 | 2.6450 |
| $\Phi$ | 0.4718 | 0.5285 | | | | |

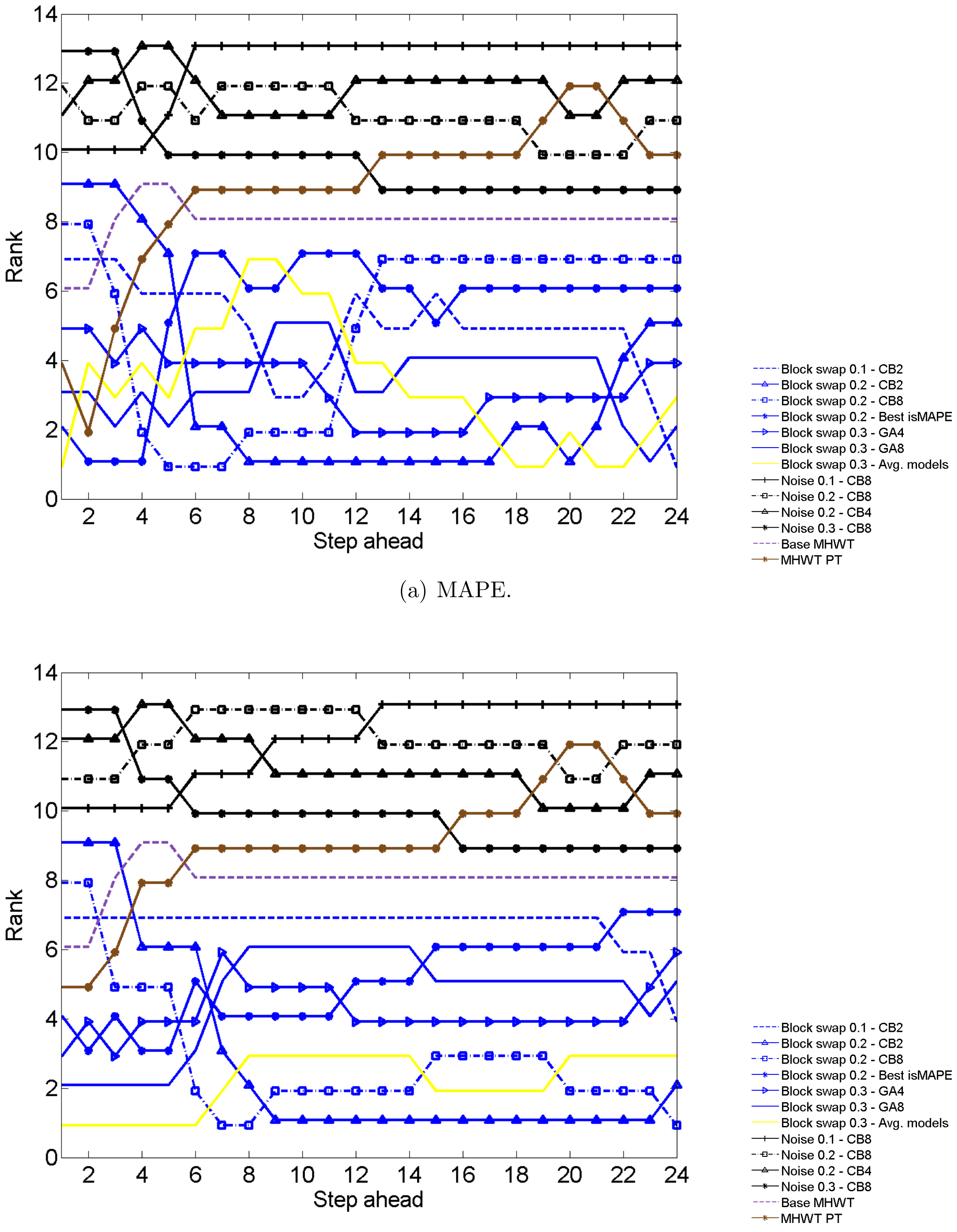
Block swapping at level 2.



(a) MAPE.

(b) MSE.

Figure 5.14: Best ensemble-based models for Rio de Janeiro electricity demand.



(a) MAPE.

(b) MSE.

Figure 5.15: Rank of best ensemble-based models for Rio de Janeiro demand series.

## 5.7 Structural Combinations of MHWT Models to Forecast England and Wales Electricity Demand

Combinations of the multiplicative Holt-Winters-Taylor (MHWT) were used to forecast electricity demand from England and Wales. Half hourly observations for the year 2016 (from Friday, 1 January 2016, to Saturday, 31 December 2016) were split into a training period consisting of 35 weeks and an evaluation period of 17.3 weeks (121 days) to test accuracy of forecasts 24 ahours ahead.

Adjustments were made on 27 March and 30 October, when the clock went forward or backward respectively, in order to have 48 observations for all days. On the first date, when the clock went forward one hour, the resulting missing data points were linearly interpolated. On the second date, data for the repeated observations (as the clock went backwards) were averaged. Two additional missing points were linearly interpolated. Observations corresponding to public holidays (according to the Bank of England) and Christmas were smoothed prior to fitting and evaluating forecasting models and combinations. This was done by replacing demand on each special day by the mean of the demand in the corresponding periods of the two adjacent weeks. The time series, comprised of 17568 observations, is depicted in Figure 5.16.

An implementation of the double-seasonal model in Equation 5.3, called *dshw*, from the R *forecast* package (Hyndman, 2017), was included as benchmark. After fitting the model to the data, the resulting parameters were: $\alpha = 0.9892$, $\gamma = 0.0000$, $\delta = 0.2507$, $\omega = 0.0001$, $\phi = 0.2618$. This model is referred to as Db. seasonal R.

As in the previous study, results are arranged in two main groups: the first reports on the generation and combination of ensembles where noise was added to the series and the second covers the ensembles and combination where the series

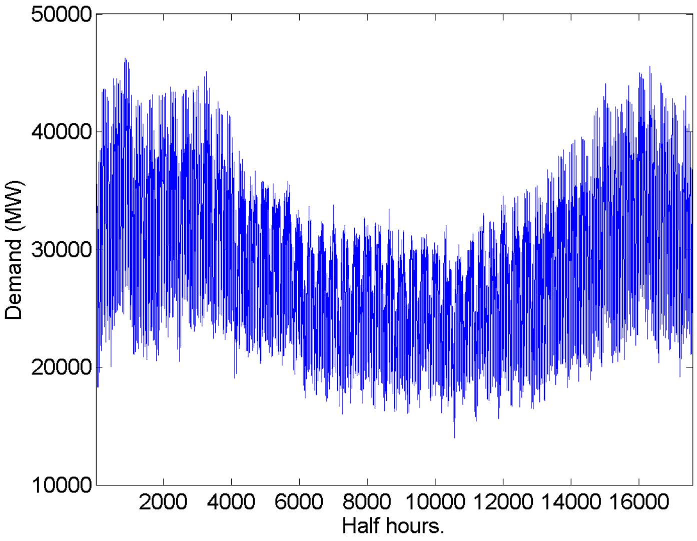


Figure 5.16: Hourly electricity demand in England and Wales for Friday, 1 January 2016 to Saturday, 31 December 2016.

was subject to block swapping.

### 5.7.1 Results when Model Variation was Introduced through Noise Addition to Generate MHWT Combinations

Figures 5.17 and 5.18 summarise the forecasting performance for the different CB models and benchmarks. The first includes performance rankings in terms of MAPE for every forecast horizon, the percentage of difference in error metric with respect to the average of forecasts in the ensemble, and the percentage differences in error metric of ensemble models, with respect to the base best model.

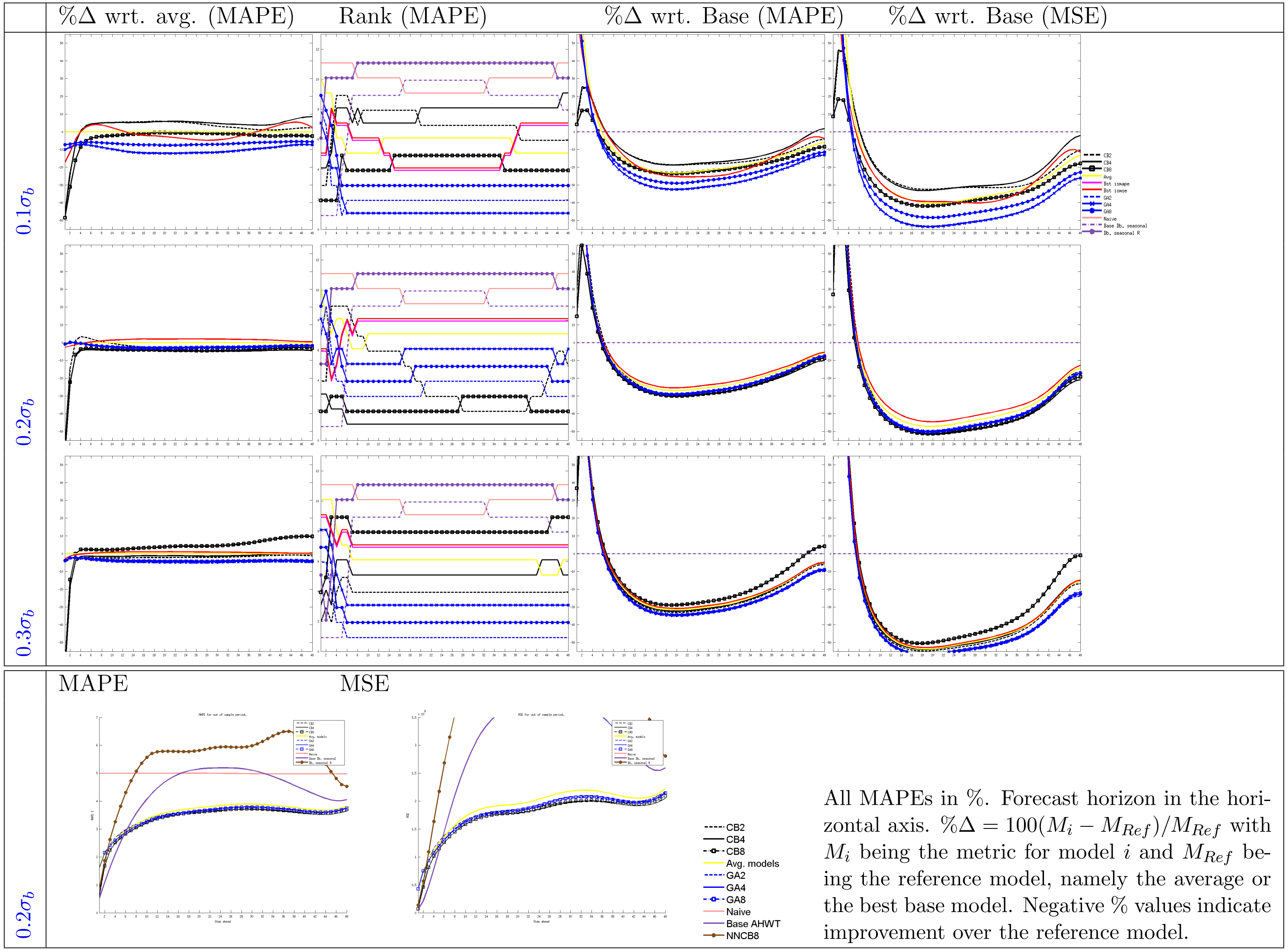
Noise addition produces a well-performing set of combinations, specially with genetic algorithms. In the low level of noise, all three combinations, GA2, GA3 and GA8, outperformed the other benchmarks in most horizons. In the middle level of noise, CB4 and CB8 are the best, while in the high level of noise GA combinations perform better again. Additionally, the CB combinations, when performing well, are better at forecasting the first four horizons than the GA combinations.

Following a similar pattern, GA combinations are better at outperforming the average of the ensembles and the base model in cases of low and high levels of noise, whereas at the middle level, the CB combinations dominate. The improvement of combinations over the basemodel are markedly higher than improvments over the average. It is clear that the combinations are better than the base model for horizons $h \geq 4$.

Coefficients of sample MHWT models are provided in Table 5.11. The level coefficient ($\alpha$) and intra-day cycle coefficient ($\delta$) seem to be affected by the level of noise, while the remaining parameters are not. As in the case of hourly electricity demand in Rio de Janeiro, the $\phi$ coefficients tend to be low.

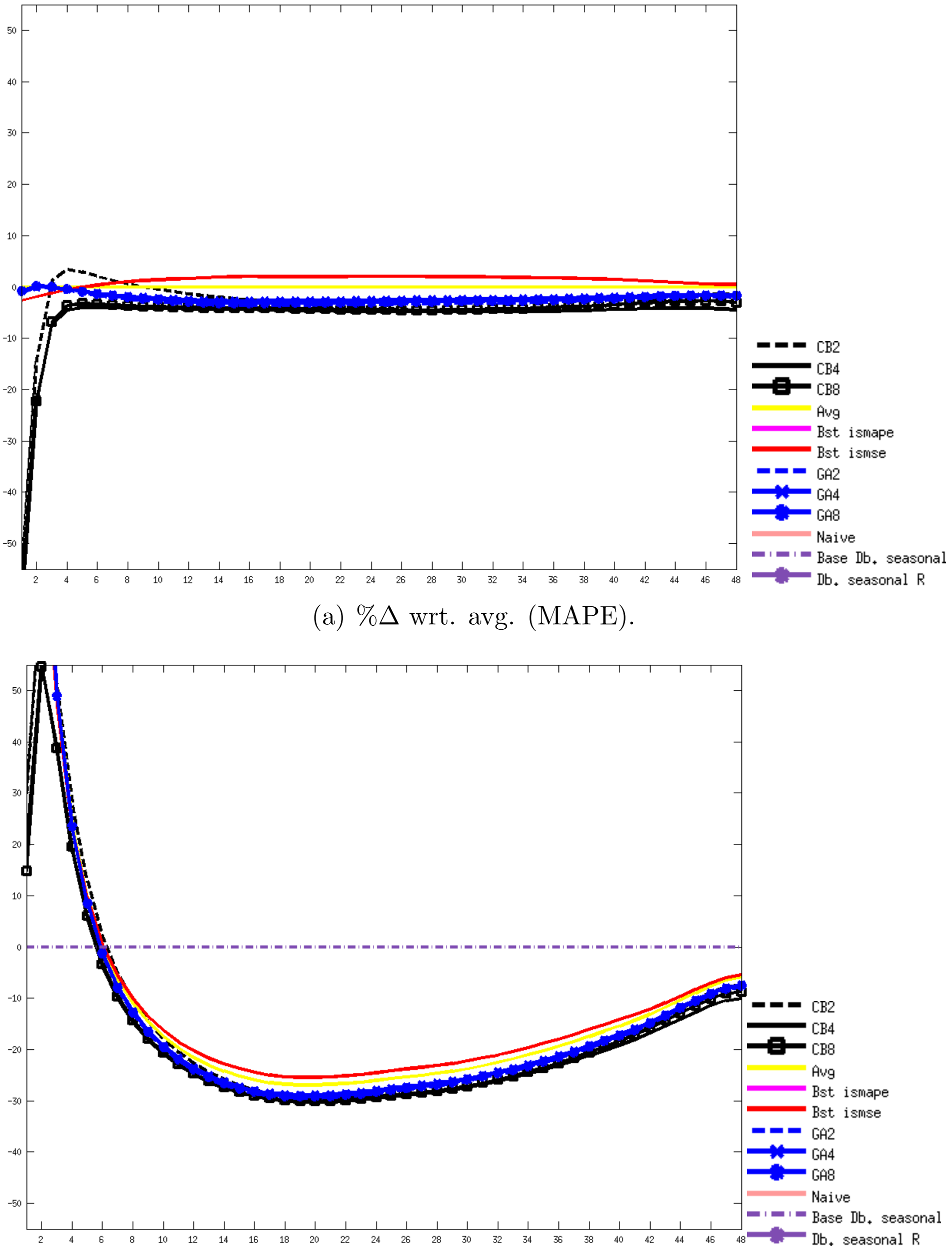
The estimated coefficients of the best performing CB combination for the noise addition setting, produced at a medium level, are reported in Table 5.12. Coefficients for combinations within clusters are diverse, while the final combination is performed with relatively homogeneous combination parameters. Diversity is therefore, playing a more relevant role within clusters than between clusters.

Jarque-Bera and Lilliefors test for normallity indicate that for all horizons and levels of noise the combinations and individual models produced non-normal and serially correlated residuals.



All MAPEs in %. Forecast horizon in the horizontal axis. $\%\Delta = 100(M_i - M_{Ref})/M_{Ref}$ with $M_i$ being the metric for model $i$ and $M_{Ref}$ being the reference model, namely the average or the best base model. Negative % values indicate improvement over the reference model.

Figure 5.17: Summary for England and Wales half-hourly demand under noise addition.



(a) %Δ wrt. avg. (MAPE).

(b) %Δ wrt. Base model (MAPE).

Figure 5.18: %Δ difference in metric vs. forecast horizon for ensembles generated with noise at $0.2\sigma_b$.

Table 5.11: Sample of double seasonal models with noise addition.

| With noise at level 1: $N(0, 0.1\sigma_b)$ | | | | |
|---|---|---|---|---|
| $\alpha$ | $\gamma$ | $\delta$ | $\omega$ | $\phi$ |
| 0.2031 | 0.0001 | 0.1344 | 0.1337 | 0.3535 |
| 0.2110 | 0.0001 | 0.1426 | 0.1320 | 0.3300 |
| 0.1997 | 0.0001 | 0.1343 | 0.1382 | 0.3643 |
| With noise at level 2: $N(0, 0.2\sigma_b)$ | | | | |
| $\alpha$ | $\gamma$ | $\delta$ | $\omega$ | $\phi$ |
| 0.1321 | 0.0002 | 0.0827 | 0.1170 | 0.1687 |
| 0.1306 | 0.0002 | 0.0844 | 0.1165 | 0.1487 |
| 0.1346 | 0.0002 | 0.0850 | 0.1228 | 0.1655 |
| With noise at level 3: $N(0, 0.3\sigma_b)$ | | | | |
| $\alpha$ | $\gamma$ | $\delta$ | $\omega$ | $\phi$ |
| 0.0692 | 0.0003 | 0.0492 | 0.1077 | 0.1346 |
| 0.0670 | 0.0005 | 0.0506 | 0.1088 | 0.1091 |
| 0.0705 | 0.0004 | 0.0505 | 0.1129 | 0.1296 |

Each row corresponds to a model.

Table 5.12: Coefficients for sample CB combination of MHWT models.

| CB4 | | | | | | |
|---|---|---|---|---|---|---|
| $\alpha_1$ | -34.7265 | 51.0860 | -21.1629 | -14.8134 | -4.3644 | -8.5177 |
| $\alpha_2$ | 58.7348 | 76.6740 | 1.8792 | -20.3662 | 45.4008 | -70.936 |
| $\alpha_3$ | -184.4589 | -46.5433 | 31.1391 | -7.9123 | -3.9978 | -3.3043 |
| $\Phi$ | 0.3457 | 0.3200 | 0.3337 | | | |

Noise $\sim N(0, 0.2\sigma_b)$.

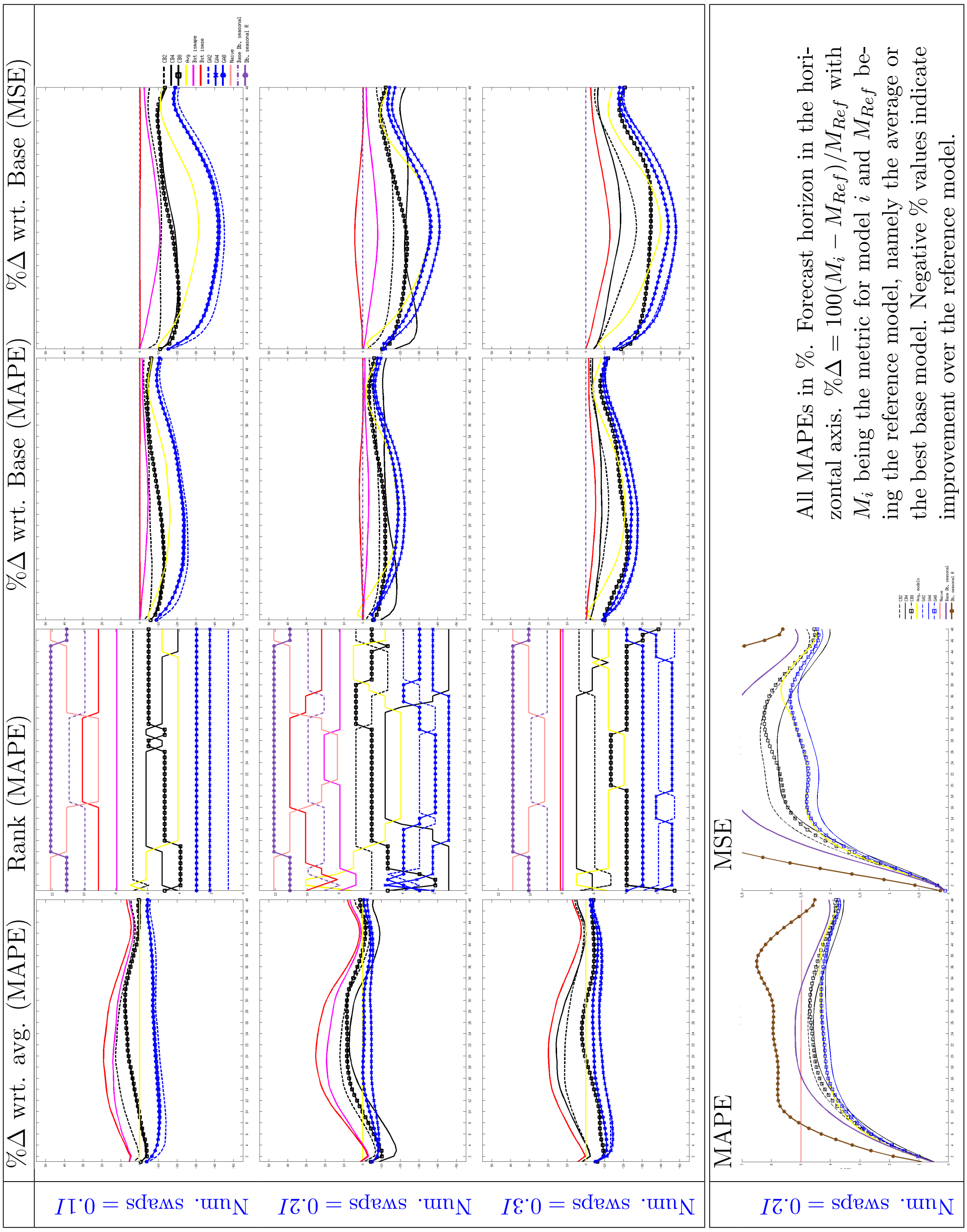
### 5.7.2 Results when Model Variation was Introduced through Block Swapping to Generate MHWT Combinations

Figures 5.19 and 5.20 show a sub-set of the out-of-sample MAPE and MSE for the different CB models and benchmarks. Rankings and percentage differences of performance metrics with respect to the average of the ensemble and the best base model are also included.

In all the levels of block swapping, GA combinations outperform CB. Improvement over the average is achieved consistently by GA combinations and improvement over the base model is observed in both CB and GA combinations. Sample individual models' coefficients obtained for the block-swapping scheme are listed in Table 5.13. The error term's coefficients ($\phi$) has higher values than in the case of noise addition. This might be due to limitations of the block-swapping approach when there is a seasonal cycle longer than the seasonal cycles considered in the forecasting model. This is the case of England and Wales data, where there is an annual cycle, which is longer than the data block size being swapped (equal to the number of observations in a week, that is, 336).

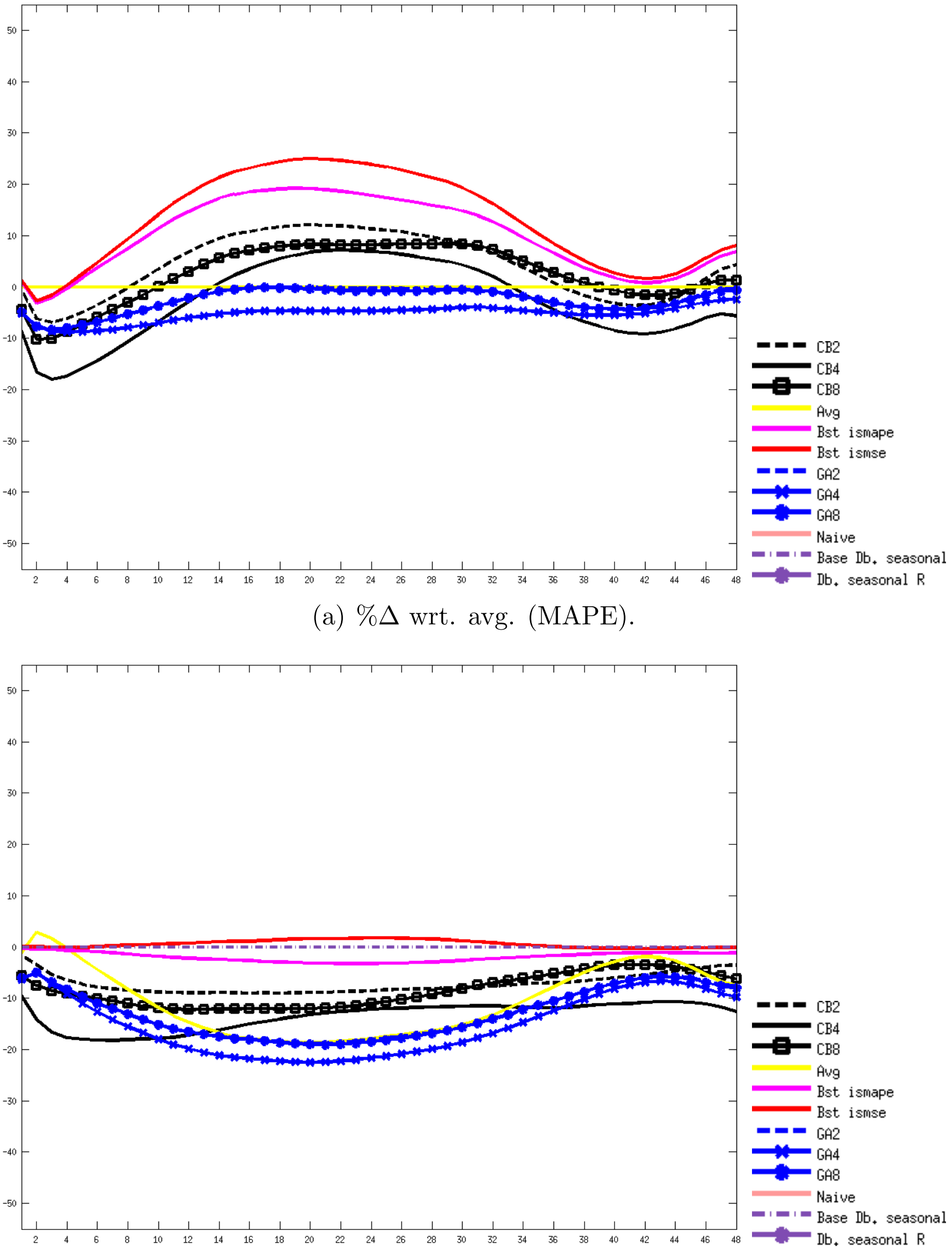
The best performing cluster-based combination is CB4, specified in Table 5.14. As in the case of noise addition, within-cluster diversity and between cluster homogeneity is noticed. In general, all models presented non-normality and serial correlation in forecast errors.

When comparing the two approaches for diversity generation in ensembles, in the case of England and Wales electricity demand, there is a clear benefit from using the noise addition over the use of block swapping, specially for longer forecast horizions (see Figure 5.21 for a selection of the best combinations). A ranking of models is provided in Figure 5.22. In both approaches the structural combinaions outperformed the average and the double seasonal R benchmark.



All MAPEs in %. Forecast horizon in the horizontal axis. $\%\Delta = 100(M_i - M_{Ref})/M_{Ref}$ with $M_i$ being the metric for model $i$ and $M_{Ref}$ being the reference model, namely the average or the best base model. Negative % values indicate improvement over the reference model.

Figure 5.19: Summary for England and Wales half-hourly demand under block swapping.



(a) %Δ wrt. avg. (MAPE).

(b) %Δ wrt. Base model (MAPE).

Figure 5.20: %Δ difference in metric vs. forecast horizon for ensembles generated with Num. swaps = $0.2I$.

Table 5.13: Sample double seasonal models with block swapping.

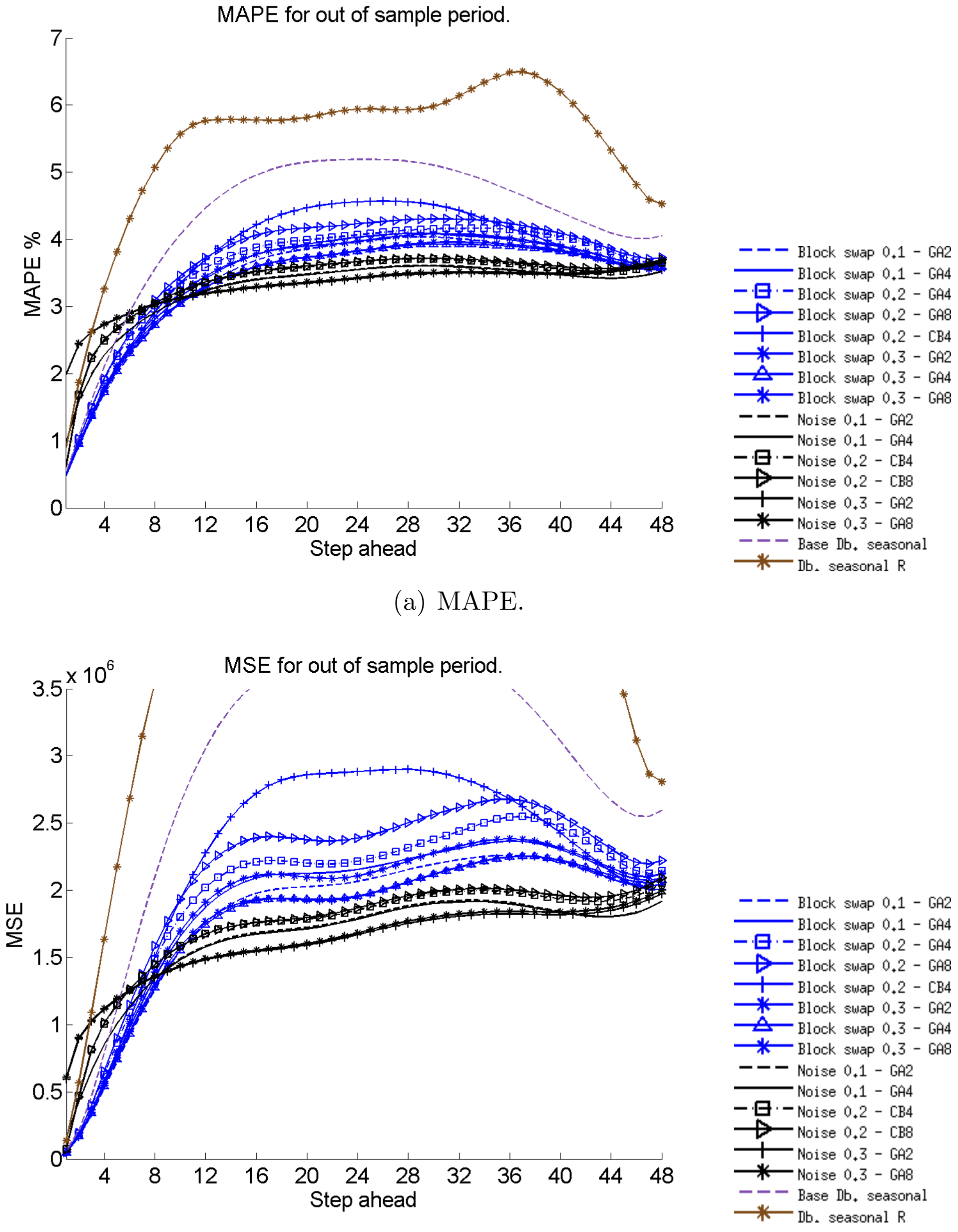
| With block swapping at level 1 | | | | |
|---|---|---|---|---|
| $\alpha$ | $\gamma$ | $\delta$ | $\omega$ | $\phi$ |
| 0.0000 | 0.7563 | 0.2866 | 0.1162 | 0.9896 |
| 0.4286 | 0.0000 | 0.4323 | 0.1235 | 0.8786 |
| 0.4030 | 0.0000 | 0.4140 | 0.2429 | 0.8904 |
| With block swapping at level 2 | | | | |
| $\alpha$ | $\gamma$ | $\delta$ | $\omega$ | $\phi$ |
| 0.3981 | 0.0000 | 0.3854 | 0.1738 | 0.8843 |
| 0.4114 | 0.0000 | 0.4009 | 0.1773 | 0.8827 |
| 0.9616 | 0.0000 | 0.2089 | 0.9952 | 0.6969 |
| With block swapping at level 3 | | | | |
| $\alpha$ | $\gamma$ | $\delta$ | $\omega$ | $\phi$ |
| 0.9620 | 0.0000 | 0.0000 | 1.0000 | 0.6211 |
| 0.9209 | 0.0000 | 0.0300 | 1.0000 | 0.6745 |
| 0.4404 | 0.0000 | 0.3155 | 0.0836 | 0.8561 |

Each row corresponds to a model.

Table 5.14: Coefficients for sample CB combination of MHWT models.

| CB4 | | | | | | |
|---|---|---|---|---|---|---|
| $\alpha_1$ | 27.0983 | -2.1562 | 4.6571 | 1.7779 | 4.9938 | 4.6172 |
| $\alpha_2$ | 31.7028 | -4.4105 | -12.4740 | 2.6930 | -5.7148 | 1.3574 |
| $\alpha_3$ | 9.6241 | -8.2948 | 10.4048 | 2.0044 | | |
| $\alpha_4$ | 31.1695 | -3.1384 | 8.0233 | -4.3081 | | |
| $\Phi$ | 0.2657 | 0.2117 | 0.2637 | 0.2599 | | |

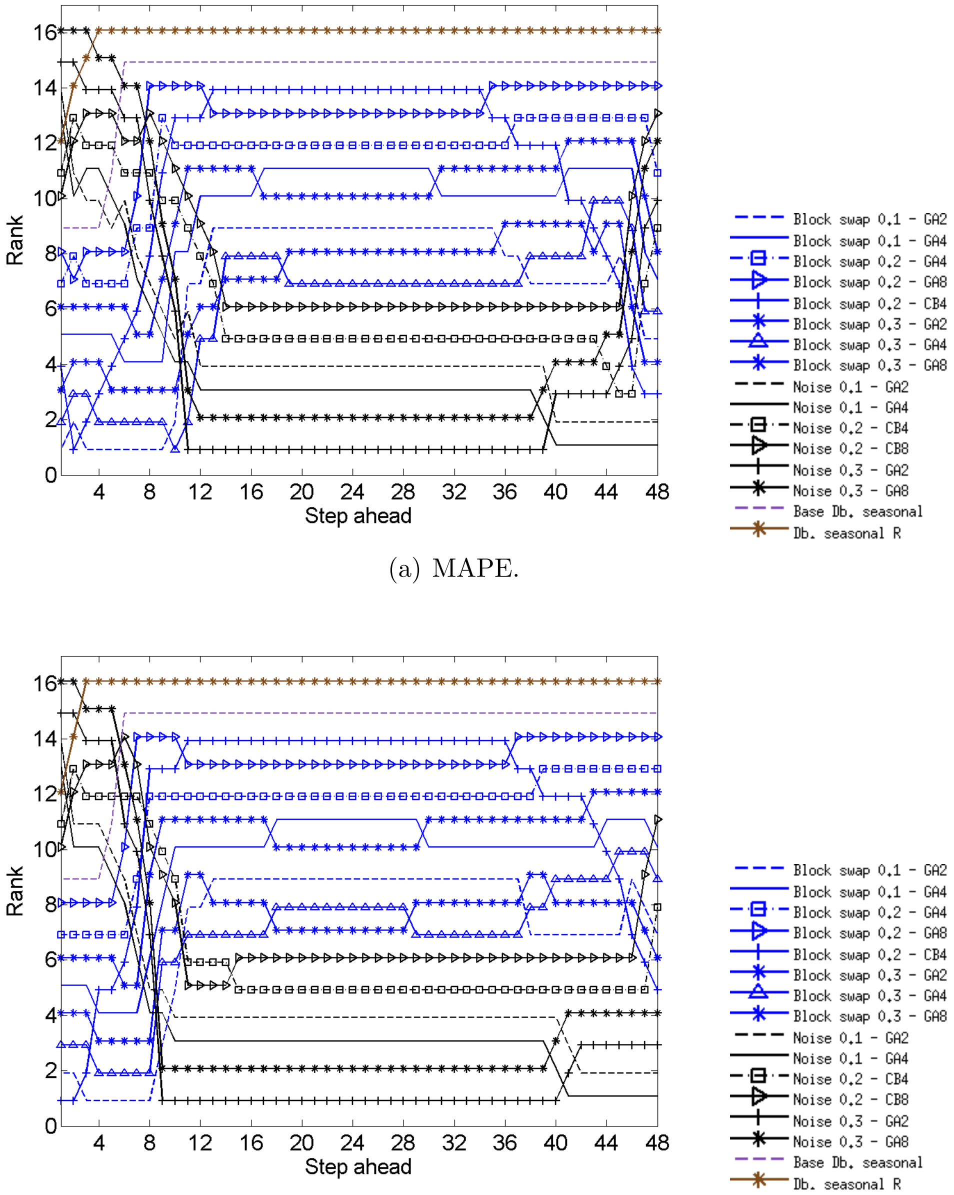
Num. swaps = $0.2I$, where $I$ = Train length$/S_1$.



(a) MAPE.

(b) MSE.

Figure 5.21: Best ensemble-based models for England and Wales demand series.



(a) MAPE.

(b) MSE.

Figure 5.22: Rank of best ensemble-based models for England and Wales demand series.

## 5.8 Discussion

The advantage of non-linear combinations with NNs over linear combination schemes has been highlighted in the literature (Donaldson & Kamstra, 1996). Yet, in the previous chapter, inherent limitations in NNs and a superiority of exponential smoothing models for seasonal and double-seasonal time series were found. This motivated the present study, where the focus is on combinations of seasonal exponential smoothing models.

It could be argued that ensemble forecasting in climate prediction (Murphy et al., 2004; Parker, 2010) is a practical way to tackle the lack of knowledge about nature by using numerous trials of different models. In univariate time series forecasting, the production of ensembles, as done here, is similar. Lack of knowledge about the workings of a phenomena is stated from the beginning by focusing on temporal dependencies. In this study, replicas of the time series are created with the aim of obtaining as much information as possible from the data in order to better approximate the temporal dependence. This idea relates to the benefit in exploiting diversification gains, as argued by Timmermann (2006). That is, different models created with different initial conditions could be combined to obtain performance gains from model diversity.

Model diversity was partly explored with a procedure inspired by bootstrapping. Although bootstrapping has been more commonly used to estimate properties of an estimator (such as the variance) and aid the construction of measures of uncertainty, defined in terms of, for example, bias, variance or confidence intervals (Efron & Tibshirani, 1994), it has also been used to train different instances of forecasting models to build ensembles (Zhang, 2007). In both cases, bootstrapping generates different data sets based on a sample, but the purpose is different. In the estimation of uncertainty, the purpose is to vary data after it is available in order to calculate a property, and when building ensembles it is used to vary the data before it is used to fit a model, so that ensembles created with resampled data series can produce a better combined forecast (Zhang, 2007). In both cases, the underlying assumption is that the data at hand is a single realisation of a generating process and that several realisations can be simulated via resampling. In turn, the underlying assumption in fitting models with different data sets derived from the original one is that there is a true model and, some of the generated model instances when combined are better able to approximate the true model than a single one. The diversity (members of the ensembles) generated in search for such model benefits the final combination of forecasts (as argued before in reference to Timmermann, 2006). However, as the application of bootstrapping might affect the autocorrelation structure of the data, block bootstrapping has been studied (see Lahiri, 2013, p. 23) and adapted here (block swapping), with the aim of mantaining the seasonality present in the data.

The other mechanism for generating diversity in models, the addition of noise, also relies on the assumption of a true generating process around which variations are created (this time through noise). This mechanism is robust with respect to serial correlation in the data, but the forecast accuracy might be affected, depending on the level of noise that is added to the time series.

This study makes a contribution to combining univariate time series models by using their structure, in contrast to the use of point forecasts (Clemen, 1989; Diebold & Lopez, 1996; De Menezes et al., 2000; Timmermann, 2006; Newbold & Harvey, 2007), specifically by building ensembles of exponential smoothing models. Additionally, variation induction mechanisms were explored and used to generate model diversity, which is uncommon in time series forecasting with established forecasting models.

When using Holt-Winters or Holt-Winters-Taylor models, the modeller is usually confronted with the choice of starting model parameter values in order to find the best possible performance in a model. Model diversity induction is thus an alternative way to search for improved performance, since starting points are then randomly generated and models are trained with slightly different replicas of the time series.

The results showed that when noise was added to the original series, some good performing structural combinations were created, such as CB in the case of the daily single-seasonal time series (peak electricity demand) and GA in the case of double-seasonal multiplicative models for the England and Wales time series. Forecast averages from the pools of multiplicative double-seasonal models and GA structural combinations with stable and good performance were obtained with block swapping for the hourly electricity demand in Rio de Janeiro.

Even when the structural combinations do not provide competitive results under noise addition, the improvement over the simple average of forecast is noticeable. Hence, faced with varied forecasts, the structural combination has the potential to deliver a reasonable performance. In the case of the multiplicative double-seasonal time series for Rio de Janeiro, results suggest that noise addition might have altered the dynamic of the series, making it more difficult for models to capture it. It is possible that the error correction mechanism that was introduced by Taylor et al. (2006) explains the greater forecasting performance of this benchmark for the Rio de Janeiro time series. It appears that there has been a change in the out-of-sample data pattern, that might have been captured by the strong auto-regressive component in their model.

The better performance of noise addition over block swapping for the England and Wales time series might indicate a negative impact of block-swapping in time series with longer cycles than those considered in the swapping of data blocks.

The effort in producing ensembles with these variations might be compensated by identifying a few well-performing model combinations that provide relatively good improvement over the best model and competing benchmarks. With the availability of parallel computing and high speed multi-core computers, the implementation of these ensembles with running times only slightly higher to the training of a single model is within practical means. The selection of a structural combination can be made by using a rule of thumb: if significant forecast improvement by a cluster-based combination with respect to the base best model and other benchmarks is observed in the first horizons, then the cluster-based combination is preferred. If insignificant improvement is obtained for the first horizons, either the base best model, a well-performing benchmark or a GA combination is preferred.

The cluster validity measures (material available upon request), reveal little differentiation between clusters, which might stem from the homogeneity in model specification and the tendency of models to converge to similar parameters, despite the use of mechanisms to promote diversity. For Rio de Janeiro electricity demand, the combination with block swapping at level 2, which showed good forecasting accuracy compared to others (2 clusters), scored better in terms of cluster differentiation. These two conditions, however, do not match for other cases. This assessment comprises combinations with more than one final cluster.

The results obtained can be viewed from the perspective of a learning process, interpreted as a link between a problem space and a solution space (Kasabov, 1996, p. 332). For different problems (data) there are different mappings (forecasting algorithms) that lead to a solution. This interpretation was exploited by Matijaš et al. (2013), when ranking statistical forecasting models, and is common when training pools of neural networks. In this research the fitting of a forecasting model was used instead of a learning algorithm (as in neural networks), and variation was introduced into the problem by altering the data. These variations led to different mappings (of the same kind, namely Holt-Winters or Holt-Winters-Taylor models). Subsequently, such mappings were combined, structurally. When using cluster-based combinations (which showed volatility), the forecasting performance could undergo favourable jumps when noise was added to the data and therefore improve markedly over benchmarks. This could be interpreted as a jump in the search of a problem-solution mapping. When data were diversified through block swapping, the mappings (or fitted models) provided a more stable performance, closer to the average and the base model. The first situation is specially observable in CB combination, when applied to the single-seasonal time series (peak demand in Rio de Janeiro). The second situation is observed in various CB and GA combinations in both hourly and half-hourly electricity demand time series.

Both approaches to generate diversity, block swapping and noise addition, have limitations. Block swapping might be too simplistic in handling the temporal dependencies in the data as long term dynamics might not be preserved. Noise addition, on the other hand, might modify the data regularity patterns.

Resuming the argument on diversity gains obtained in forecast combinations (Timmermann, 2006), as in the previous study, there is a limit for the diversity gain that can be obtained: the ensembles are composed of models of the same nature and therefore only diversity coming from structural variations of models of the same family is obtained. This diversity can be extended by combining ensembles of models of different families. Or, simpler, models of the same family but trained with different algorithms (e.g. that minimised alternative loss functions) or with different training configurations.

In Chapter 4, it was suggested that ensembles of NNs trained to forecast iteratively could be modified as to resemble a strategy of switching between different forecasts at different periods (Granger, 1993; Deutsch et al., 1994; Taylor & Majithia, 2000). This could be implemented, at the level of models, if a different clustering instance (of the same ensemble) was performed for every forecast horizon. Additionally, during the training of the ensemble, the fit for different forecast horizons could be assessed, and this information could be included in the structural representation of each model.

## 5.9 Conclusions and Research Agenda

Model ensembles originated in climate modelling due to the existence of different sources of uncertainty (Parker, 2010). It then has been applied since the work by Hansen & Salamon (1990) to neural networks, with the general idea of creating diverse models under different conditions. This approach is extended here to the Holt-Winters, Multiplicative Holt-Winters-Taylor and Additive Holt-Winters-Taylor statistical models. For such models, optimal parameters (structural descriptors) tend to be homogeneous. Therefore, diversity was promoted by adding noise or swapping blocks of data. The resulting ensembles are used to perform structural combinations proposed in the previous chapter: one approach with a clustering-based algorithm (CB) and other with genetic algorithms (GA).

Three applications were conducted. The first focuses on peak electricity demand from Rio de Janeiro, the second uses hourly electricity demand from Rio de Janeiro as well and the third application forecasts half-hourly electricty demand from England and Wales. Performance comparisons were made against the average of point forecasts in the ensemble, the base best model (building block in each ensemble) and other competitive benchmarks. Table 5.15 summarises forecasting performance of structural combinations with respect to the average and the base best model.

The average forecast from the ensemble is a robust benchmark, but it was observed that under noisy conditions, improvements over it are not associated with improvements over competitive benchmarks. The base best model gives another reference point that allows the modeller to further assess the benefit of building ensembles. As this model has been trained with the original time series and is subject to a selection process, comparisons against it can give a clearer idea of how

Table 5.15: Summary of findings for structural combinations.

| | | Rio peak | | Rio hourly | | England & Wales half-hourly | |
|---|---|---|---|---|---|---|---|
| | | CB | GA | CB | GA | CB | GA |
| Improvement wrt. Average | Noise addition | ✓✓✓ | ✓✓✓ | ✓✓✓ | ✓ | ✓✓✓ | ✓✓✓ |
| | Block swapping | ✓✓ | ✓✓✓ | ✓✓ | ✓ | ✓ | ✓✓✓ |
| Improvement wrt. base best model | Noise addition | ✓✓ | | | | ✓✓✓ | ✓✓✓ |
| | Block swapping | | ✓✓ | ✓✓ | ✓✓✓ | ✓✓✓ | ✓✓✓ |

Summary of findings for CB and GA combinations. Improvement of performance over the respective benchmark is marked with tick symbols, one for every level of noise or block swapping for which the improvement was observed. Each mark stands for improvement in most of horizons.

competitive the combinations are.

For the single-seasonal daily time series series (peak electricity demand), improvements over both the average and the base best model were observed, but the strategy of noise addition worked better than block swapping in producing competitive structural combinations. Results suggest that CB combinations are better at exploiting model variations coming from noisy data in order to improve performance on this series. GA combinations, on the other hand, seem to be well suited to exploit model variations through block swapping. Additionally, it was found that CB combinations are volatile while GA are not. Having found a CB combination which outperformed other benchmarks under the scheme of noise addition (at a middle level) opens the question of a possible interaction between the type of time series (single seasonal with a changing pattern), the type of combination (based on clustering and exhibiting volatility) and the level of noise in the data. This is a topic for further research.

For the first multiplicative double-seasonal time series (hourly electricity demand in Rio de Janeiro), improvement over the average forecast in the ensembles produced with noise addition was more common in CB combinations than in GA. However, none of the combinations or models with best in-sample metrics improved over the best base model. When using block swapping, the performance of the ensembles improved markedly and produced competitive averages, with further improvements being observed in GA and CB combinations. However, while the averages and the GA consistently outperformed the base best model, CB combinations were less consistent. The later were volatile and tended to performed poorly for the first forecast horizons. A question remains in relation to the fact that time series in the studies were split in a similar fashion. As sample size could influence forecasting performance and its stability, future research could consider the sensitivity to different partitioning of the time series. In doing so, it would address the question of how much forecast history is needed in such a computer intensive approach.

For the second multiplicative time series (hourly electricity demand in England and Wales), improvement over the average was easier for GA combinations than for CB under both noise addition and block swapping. However, improvement over the base model (and the double-seasonal R benchmark) was present in both approaches for all levels, specially for the noise addtion scheme (30% reductions in MAPE over the base model are reached for some forecast horizons). However, combinations performed poorly against the base model in the first forecast horizons. Overall, less volatility in performance was noticed for this time series, which might be due to the use of more data than in the previous applications.

Regarding the models, other forms of structural combination, besides clustering and genetic algorithms, can be explored. Other features can be included in the structural description of models, such as a measure of the evolution of the error during the training period. Additionally, a strategy similar to the switching of forecasts could be implemented by using a different clustering instance (of the same ensemble) for every forecast horizon. Regarding the variety induction mechanisms, more sophisticated ways of introducing noise or performing block swapping can be devised to further study the effect of these strategies on the performance of forecasting ensembles.

In general, this study permitted the extension of the structural forecast combination approach from the neural networks to exponential smoothing models from the Holt-Winters and Holt-Winters-Taylor family and, in this way, combinations were better equipped to deal with seasonality. Results obtained suggest a potential improvement in forecast accuracy when the structural combination is applied, but also highlight how robust exponential smoothing models can be.

In the context of business analytics, with the contemporary abundance of data, the combination models proposed here could be implemented by using different data sources. Available information can be used to train the individual forecasting models in the ensembles and also to determine the kind of data variations to be introduced. In general, the process of pooling models and combining them by using their structure fits in the tendency to search for knowledge in large data repositories: if a model is used to *learn* from the data, several can be used for the same purpose to then find a combination based on a proxy of the learned knowledge, such as the structure.

# Chapter 6

## Summary and Directions for Future Research

In this dissertation, a forecast combination approach that uses the structure of forecasting models was proposed, thus departing from the general approach of combining forecasts. The research was inspired by the existence of sophisticated models that tend to exhibit some form of intelligent behaviour and the possibility of using their structure to inform forecasts.

Ensembles of Neural networks (NN) were initially investigated. NNs were selected as they have a clear structural representation and are widely applied in forecasting. Prior to using feed-forward NNs in combining forecasts, a sensitivity analysis was conducted in order to investigate their behaviour when forecasting time series of diverse complexity. This study permitted a more objective selection of models to construct ensembles. Subsequently, the structural combination of feed-forward NNs was conducted. Ensembles were created and represented in their parameter space and were then combined according to two algorithms. The first proposal finds groups of models that are close together in their parameter space, so that forecasts are produced based on selected models from each identified cluster. The second proposal is based on genetic algorithms. It finds reference points in the parameter space and selects models around it, to then average their respective forecasts. Finally, to complement the analysis of the NN structural combination, a study applied the proposed procedures to two forecasting models that specifically address the type of seasonality in the time series and that are adaptive to changes in the data pattern: the multiplicative Holt-Winters and the multiplicative Holt-Winters-

Taylor models. The following sections summarise the findings for the three studies. Finally, this chapter concludes by suggesting potential lines of future research that emerged from this dissertation.

## 6.1 A Sensitivity Analysis of the Performance of Feed-Forward Neural Networks

The first study analysed the sensitivity of NN fit and forecasting performance, according to different forecast error metrics, to several NN design factors. Non-seasonal, single-seasonal and double-seasonal synthetic time series were considered. The factors studied were the number of inputs, the number of neurons, the sample size and the pruning of weights in the hidden layer.

Volatility in performance was commonly found when inspecting the plots of average MSE with confidence bands. More general tendencies, with less volatility, were observed through cross-validated variants of the experiments. Additionally, robust non-parametric tests were used to further assess the influence of factors over the MSE and MAE error metrics.

Considering previous literature, Zhang et al. (2001) reported a significant impact of input nodes (NI in this study) on MSE and MdAPE, for both training (in-sample) and test (out-of-sample) sets across different sample sizes for one-step-ahead forecasts. The experiments conducted in the present study (by using MSE and MAE) produced results which support their findings for the in-sample period and similar time series. Additionally, the direction of influence found here is almost always negative in error metric for the first horizon (improved fit). For other forecast horizons, the findings are mixed, and emphasise the volatile performance of neural networks for univariate time series forecasting and the risk of over-fitting the time series as models become more complex: generally an increase in the number of inputs improves model fit, but not necessarily forecasting performance. In the out-of-sample periods, significant effects of the number of inputs were found, but the influence of the number of inputs is mixed and forecasting performance may deteriorate.

Zhang et al. (2001) observed an error metric curve, with respect to NI, that decreased until achieving a minimum for a certain number of inputs and growing afterwards. This pattern was observed in one series (STAR2) from the non-seasonal set, but for the rest, results were mixed. It is noteworthy that while previous literature focused on one-step-ahead forecasts, the study addressed multiple steps ahead.

The number of neurons was found to be a significant factor for all forecast horizons and series, in terms of MSE and MAE during the training (in-sample) period, thus improving fit. In the out-of-sample period, the effect is significant for most forecast horizons and series. The effect is mixed for non-seasonal time series and there are indications of over-fitting, but for the single-seasonal and double-seasonal time series, there is evidence of improvement in accuracy when the number of neurons is increased.

Balestrassi et al. (2009) reported that the number of neurons was significant for all the series considered in their study (SAR, BL1, BL2, TAR, NAR1, NMA, STAR1 and STAR2). The number of inputs was fixed in their univariate time series approach. The sample size was also found significant, which are in line with the findings obtained in the present study. Yet the simple pruning of NNs weights conducted here led to mixed results, with a common tendency to worsen the forecast accuracy.

A plausible explanation for these mixed findings, in terms of significance of factors, specially for higher forecast horizons, is the effect of high levels of noise on the data generating processes. Zhang et al. (2001) used generating processes (adopted here for non-seasonal series) which add noise equivalent to a high proportion of the interquartile range. However, in this research, when performing experiments with configurations used by Barrow et al. (2010), it was found that high levels of noise in the data generating process impact the forecast accuracy of NNs, and the impact can be amplified for longer forecast horizons. STAR2 series appears to have been less affected by the high level of noise than other non-seasonal series, thus leading to more stable and better model performance.

In general, it was observed that the time series that were better captured by the feed-forward NNs used in this study are stable and are characterised by a regular pattern and strong serial dependence (supporting findings by Crone et al., 2011). For other series, the consideration of multi-step-ahead forecasts revealed that NNs are capable most of the times to capture the main dependencies present in the generating processes.

The present study found ordinary least-square regression unsuitable to assess the effects of design factors over error metrics because the model assumptions did not hold. In our experience, one should use non-parametric statistical tests as the data are subject to serial dependence and non-normality.

NNs needed low complexity for non-seasonal series. For the single-seasonal, NNs with relatively high number of inputs and neurons showed better forecast accuracy and better behaved errors (in the autocorrelation maps). For the double-seasonal time series, the benefit of complexity in terms of forecast accuracy seems to reach a limit after 6 inputs and 12 hidden units after which over-fitting appears, but forecast errors are better behaved with more complex models and the main effects graphics support the benefit of having more complex models as well. This importance of inputs (past lags), specially for the single-seasonal and double-seasonal series has implications for the combining algorithm, which will be considered in the next section.

The research question formulated in this first study was *How can a sensitivity analysis, based on design of experiments, be used to aid the selection of NN models* *for a forecasting ensemble?* The sensitivity analysis can aid the model selection by and objectively assessing the average behaviour of different error metrics within the NN model constrains imposed by the ensemble design. The model constraints are translated into design factors (and respective levels) in the study, such as the number of hidden units or the number or inputs.

Overall, model selection for the ensembles would comprise several steps: establishing the characteristics of the ensemble (big or small models, same or different structure, etc.), designing the experiments, creating the performance data-base, using it to visualise performance and conduct statistical tests and, finally, selecting models based on performance and residual behaviour. Given the volatility of NNs, ideally the modeller would make use of several metrics.

## 6.2 Structural Combination of Neural Network Forecasting Models

The next stage of the investigation was the proposal of a structural forecast combination method and the assessment of its performance. Based on previous research, the synaptic weight space of NN was used to perform clustering of models. This permitted the development of forecast combinations based on structural characteristics. An algorithm based on recursive partitioning was used so that from each region or cluster found, a subset of models to be combined is selected. Additionally, an implementation based on genetic algorithms (GA) was proposed as a simpler alternative structural combination. It relies on using reference points in the parameter space of models, from which models are selected, based on their distance to reference, and their forecasts are averaged.

Both synthetic and real time series were used to assess the performance of the models. The synthetic time series were STAR2, Synthetic-1S and Synthetic-2S. A real time series application focused on electricity demand and a multivariate time series applications forecasted wind power production. Different numbers of clusters were considered. Results from the initial study on sensitivity analysis were used to select base models for the STAR2, Synthetic-1S and Synthetic-2S ensembles. Additionally, sensitivity analyses were conducted to select models for the electricity demand and the wind power production series.

Structural combination with genetic algorithms (GA) outperformed the average more easily than cluster based (CB) combination for non-seasonal time series (STAR2 and wind power production) whereas for the seasonal series (Synthetic-1S, Synthetic-2S and electricity demand) the CB tended to do better in outperforming such benchmark. CB and GA easily outperformed the best NNs in the ensembles in the non-seasonal synthetic series and wind power series. For Synthetic-1S, Synthetic-2S and electricity demand, on the other hand, there was no marked superiority of structural combinations over such individual models. Nonetheless, CB showed better performance than GA with respect to the best models.

GA combinations showed a smoother performance pattern when compared to CB in most applications. This is an interesting feature in light of the structural differences in NNs for different horizons: for all the series, except the electricity demand, a separate ensemble was used for each horizon and the base specification for each ensemble was usually different. The GA benchmark is an average version of CB and it thus may replicate the robustness of the simple average that h.as been observed in the forecasting literature. These findings suggest that different forms of structural combination can be explored for different forecast horizons and that simpler forms are competitive.

CB and GA structural combinations were outperformed by the best benchmarks in the cases of the single-seasonal and double-seasonal series (synthetic and real). Exponential smoothing models are better equipped to adapt to changes in the regularities in these series than NNs. However, for non-seasonal series (synthetic and real) the NNs and the structurally combined ensembles showed a clear advantage.

Nonetheless, for non-seasonal series (synthetic and real) the NNs and the structurally combined ensembles showed a clear advantage. In the case of Kaggle wind power series, the superiority of CB models to the statistical benchmark is likely to come from the high complexity and non-linearity of the forecasting problem, for which NNs based models are more robust. This study suggests potential gains in multivariate time series forecasting, which should be addressed in future research.

For the electricity demand series, the specification of inputs was crucial (coinciding with the importance of inputs in the sensitivity analysis). The base specification had many inputs and a few hidden units. Because the internal parameters of NNs includes weights associated to inputs, the structural combination of models for this application has a component heavily associated to inputs. In cases like this, when a neural network has many inputs and a few hidden units, the clustering of models in their parameter space becomes more similar to a clustering of the weighted inputs.

In summary, the research questions of *How can the structure of neural networks combined?* and *How do the proposed models perform in forecasting?* are answered by empirically examining two variants of structural combination under different conditions provided by several time series. The use of internal parameters of models permitted to include the structure of models in the combination of forecasts, and the implementation of two forms of structural combination allowed to distinguish the conditions under which such approach had potential to improve forecast accuracy. CB tended to work better with seasonal and double-seasonal time series and GA with non-seasonal time series.

## 6.3 Structural Combination of Seasonal Exponential Smoothing Models

The last study of this dissertation extended the structural combination of NNs in order to address seasonality, which is a common feature of time series in the energy sector (e.g. demand, prices of electricity and gas), and assess combinations of statistical models. Two seasonal exponential smoothing models are explored, namely: the Holt-Winters and Holt-Winters-Taylor multiplicative models. The first was used to forecast a single-seasonal time series of peak electricity demand in Rio de Janeiro. The second was used to forecast hourly electricity demand in Rio de Janeiro and half-hourly electricity demand in England and Wales. Structural diversity in models was promoted via modelling replicas that were generated by adding normally distributed noise to the series or by swapping blocks of data previous to the training stage.

For the multiplicative single-seasonal time series (peak electricity demand), improvements over both the average and the base model were observed. Results suggest that cluster-based (CB) combinations are better at exploiting model variations coming from noisy data in order to improve performance on this series. Combinations based on genetic algorithms (GA), on the other hand, seem to be well suited to exploit model variations through block swapping. Additionally, it was found that CB combinations are volatile while GA are not.

For the multiplicative double-seasonal time series applied to the hourly Rio de Janeiro electricity demand, improvement over the average forecast in the ensembles produced with noise addition was found (more clearly in CB than in GA combinations). However, as judged by the sample parameters obtained in this setting, the dynamics of the series is time-varying, which could be the cause of the poorer performance observed when comparing with the base best model: none of the combinations improved over this benchmark. When using block swapping, the performance of the ensembles improved markedly and produced competitive averages and further improvements with GA and CB combinations. However, while the averages and the GA consistently outperformed the base best model, CB combinations were less consistent. The latter were volatile and tended to performed poorly for the first forecast horizons.

In the case of the second multiplicative time series (hourly electricity demand in England and Wales), improvement over the average was easier for GA combinations than for CB under both noise addition and block swapping. However, both approaches outperformed the base model in most forecast horizons. Overall, less volatility in performance was noticed for this time series, which might be due to the use of more data than in the previous applications.

To summarise, the last research question formulated in the dissertation was *How can the structural combination approach be extended from NN to other forecasting models?* This extension has to take into account the fact that exponential smoothing models usually have less parameters than NNs and the training tends to produce very similar optimal parameters. Consequently, there was a need to promote model diversity to generate combinations, and solutions based on computational intelligence literature were considered. Structural model diversity is promoted in statistical models by fitting them to replicas of the original time series. Two mechanisms were explored to create the replicas: noise addition or data block swapping. The obtained models are structurally combined to finally obtain the forecast. A detailed assessment of forecast performance under different levels of noise addition or block swapping is then undertaken.

In general, the results obtained suggested a potential improvement in forecast accuracy when the structural combination is applied. CB combinations performed better for the peak single-seasonal electricity demand and the double-seasonal electricity demand in Rio de Janeiro. GA performed well in all time series, were more stable, and outperformed CB when forecasting the double-seasonal electricity demand in England and Wales. Additionally, it was found that the robustness of MHWT models can be better exploited when forecasting regular time series if ensembles are generated through block swapping. When the time series is less regular, or when there is a longer cycle not considered in the models or in the data swapping scheme, the noise-addition mechanism is preferable.

## 6.4 Implications for the Literature and Future Research

Timmermann (2006) argued, from a theoretical perspective, that unless one can find ex ante a particular forecasting model producing smaller forecast errors than its competitors, forecast combination offers diversification gains that make it attractive to combine individual models rather than relying on a forecast from a single model. The present research made use of diversification in models to create ensembles and incorporated an aspect of such diversity into the combination of forecasts from those ensembles. The models have the same specification but can differ in parameter values, that is, in their structural descriptors. Bakker & Heskes (2003) suggested the use of clustering to summarise ensembles of neural networks(NN). Matijaš et al. (2013) ranked models by exploiting the interpretation of the learning process as a link between a problem space and a solution space (Kasabov, 1996, p. 332). Alamaniotis et al. (2012) built ensembles of Kernel-based Gaussian processes by using a linear (multi-objective) problem for which a solution was sought with genetic algorithms (GA). Different error measures constituted the vector of objectives. In this series of research there are attempts to enrich the context that informs the combinations. The present research contributes to this literature by building ensembles of neural networks, Holt-Winters (HW) and Holt-Winters-Taylor (HWT) models in their parameter space. NNs are easily diverse, given that they are volatile (Mendes-Moreira et al., 2012), but HW and HWT models tend to have very similar optimal parameters. To exploit diversity gains, the NNs were trained with randomised input-out patterns, but HW and HWT models, given their relative homogeneity in final parameters, were fitted with replicas of the original time series. Results confirm that there are gains in promoting model diversity. This relates to the diversification mentioned by Timmermann (2006). Here, however, the model diversity is explicitly included (proxied by structural parameters) and used to inform forecast combinations. Therefore, diversity is explored in a context where there is access to the internal (structural) characteristics of the forecasting models.

One of the main characteristics of ensembles of NNs as mentioned in Chapter 2 is how the steps of generation, pruning and combination are followed. Some authors (such as Hansen & Salamon, 1990; Drezga, 1999; Siwek et al., 2009) adopt a sequential approach, where one stage feeds into the next. Others follow more dynamic approaches, where the stages are interrelated. Fore example, Liu & Yao (1999), Liu et al. (2000) and Zhou et al. (2002) created ensembles with evolutionary algorithms, in which case the dynamics of the ensemble building process is interlinked. In the present study, the most dynamic part of the approach is located in the stage of combination. Although there is an intimate relation between pruning and combining as the latter informs dynamically the former, the general outline puts the present research more on the ground of sequential approaches.

Both the more dynamic approaches and the sequential ones mentioned above, make use of forecasts produced by models. In the present research, departing from that approach, the information used to combine forecast is the parameter set of NNs. This allows to include information about the structure of the model into the combining stage (a procedure inspired by Bakker & Heskes, 2003).

Therefore, the contribution of the present study is located in the context of sequential ensemble generation and structurally informed forecast combinations. The approaches followed to incorporate structural parameters are a clustering-based algorithm and genetic-based algorithm. In this way, the inclusion of model structure via clustering, as suggested by Bakker & Heskes (2003), is explored. The adoption of a clustering algorithm and a structurally informed benchmark (based on genetic algorithms), that resembles an average, permits to make comparisons of combination mechanisms of the same orientation but different complexity. Additionally, the general idea of the combination, by including internal characteristics of objects, naturally suggests the use of genetic algorithms, as these can represent objects whose components are evolved. Here, instead of using genetic algorithms to evolve the weights assigned to NNs outputs, as Zhou et al. (2002), we evolve a configuration of centres in the parameter space from which models are selected and averaged. Therefore, both in terms of the use of clustering and the use of genetic algorithms, the structural extension in forecasting makes a contribution. The study of different synthetic and real time series (with a multivariate case) and the inclusion of different levels in the number of clusters, permitted to have a realistic view of the performance of the proposed forecasting combinations.

A key point in this dissertation is the reasoning behind the inclusion of structure in combining forecasts. The motivation is the use of characteristics of models and not merely their single output forecasts. The approach and results contribute in offering a large empirical study about one possible approach for the inclusion of structural information in forecast combination. However, the interpretation of what constitutes internal characteristics could be different from the approach adopted here. Additionally, the study of relationships between internal components and outputs in models, suggested by Garson (1991) and Goh (1995), could be productive. Both ideas are natural extensions of the present research.

An implication of this research for the ensemble literature is the potential benefit in including structural information in the combination of forecasts. As the likelihood of the appearance of increasingly intelligent models is high, this orientation in combining forecasting models could become gradually more productive. This research shows how variety can be exploited within a set of models of the same type. The same approaches for variety generation, specially in HW and HWT models, could be explored with other forms of forecast combinations (structural or not) to extend the present research. The complexity of models would create a challenging number of features to perform structural combinations. A proper balance between dimension reduction and the use of a sufficiently rich structural representation would be needed to structurally combine ensembles within practical computing times.

There is a basic difference between CB and GA structural combinations: the first is deterministic and the second is random. As the first tended to perform better with more regular time series and the second worked well with less regular data, the question arises of how the nature of structural clustering behind the combination is related to the presence of regularity in the data. The appropriateness of a deterministic structural combination for a regular series when compared to a random one, and the opposite situation (appropriateness of a random structural combination for less regular data) is an issue that could be further investigated.

The multi-step ahead approach with NNs affects both training (fitting) and forecasting. Other approaches should be explored for different applications. Here the direct approach was used for synthetic series where a separate NN is used for each forecast horizon. This approach allows for networks to specialise in a specific forecast horizon (Gouriveau & Zerhouni, 2012) and, consequently, for a division of the training task between multiple machines. The iterative approach was used for the electricity demand application given that the direct approach led to a performance markedly different from results obtained by Taylor et al. (2006). However, a systematic exploration of multi-step-ahead forecasts for structurally combined ensembles allow to compare different approaches for the same data. If such comparison is made within a sensitivity analysis scheme, as in this dissertation, the computational cost would be higher than in the present research. Therefore, it could concentrate on a smaller number of series.

Another point for future research, in relation to training, involves the use of a bootstrap strategy to create the individual models that participate in the clustering algorithm. The adoption of this strategy, instead of the randomisation of the in-sample data set as done in the present research, would move the approach in the direction of bagging [1]. The GA combination can also benefit from such modification and these are potential research avenues for future research.

Simplified forms of design of experiments (DOE) can be devised, in order to speed up the exploration of models and the application of statistical tests. This could facilitate automation. The judgement exercised by the modeller in applying the heuristics suggested here could be automated to then feed the ensemble mechanisms without human intervention. The practicality of this would depend on finding a proper balance between the computational complexity of the experiments (combination of factors and levels) and the availability of parallel computing.

Further research on forms of structural combination with low computational cost, which could be inspired by genetic algorithms, could potentially be productive, given the good results obtained with such variant. They would have the advantages of speed and simplicity, which would contribute as an enhancement of the combinations proposed.

The limits imposed on the number of models per cluster in the combination algorithms allowed to use at least 20% of the total of models and at most 80% in

[1] Breiman (1996) defines bagging predictors as *"a method for generating multiple versions of a predictor and using these to get an aggregated predictor. The aggregation averages over the versions when predicting a numerical outcome and does a plurality vote when predicting a class. The multiple versions are formed by making bootstrap replicates of the learning set and using these as new learning sets"*.

the forecast combination. This choice permitted to follow findings by Zhou et al. (2002) which suggest that it is better to ensemble many available NNs but not all. The same proportions were used when combining exponential smoothing models. In the literature, however, there are dynamic options that could suggest modifications to the combinations routines. For example, Chen & Yao (2007) use an evolutionary algorithm whose selection mechanism produces a set of models in the last stage of the ensemble building process; Yu et al. (2008) establishes the number of models to combine by minimising the conditional generalised variance and Kourentzes et al. (2014) choose a varying number, depending on the forecast median or mode. Given that a structural combination with genetic algorithms has provided competitive results with low computational cost, a mechanism could be added to such approach to automatically determine the final population size, that is the number of clusters, and the number of models to be selected from each of them.

There is no conclusive evidence about the superiority of the combination of ensembles generated with noise addition over the ensembles combined when block swapping was performed. There could be an interaction between the type of time series (single-seasonal or double-seasonal), the type of combination (based on clustering or genetic algorithms), the length of the time series, the existence of longer cycles than those considered in individual models or the swapping of data blocks, and the level of noise or block swapping in the data. However, it was noticed, firstly, that noise addition can be superior in the case of single-seasonal and not very regular time series. Secondly, noise addition was also superior when a double-seasonal time series has dynamics not captured by individual models or when the length of data blocks to swap is shorter than other cycles present in the data. And thirdly, block swapping can perform better for more regular double-seasonal time series. This, plus a wider exploration of levels and approaches for noise addition and block swapping, are topics for further research.

The results obtained highlight a limitation of the ensembles used in the present research. They are composed of models of the same nature and the same basic specification. From the perspective of model diversity (Timmermann, 2006; Bunn, 1975), there is a gain that might not be exploited. Our results show that NNs ensembles combined structurally can be favourably compared to the average forecast, but are outperformed by the statistical benchmarks. Additionally, the study of HW and HWT model ensembles show how the base best model is difficult to outperform. The use of a single family of models might be limiting variety gains. If, for example, combinations of NNs and HWT models were made, there could be a bigger gain than in combining NNs or HWT alone. That is why, it would be interesting to structurally combine ensembles of models of different nature. One way of accomplishing that would be to combine bundles $M_i =< A, B >$, where each $M_i$ is composed of a model of type $A$ and a model of type $B$, such that $M_i$ has internal characteristics that could be useful for a structural combination. Combinations of pairs of models in the parameter space defined by $M$ can be done. Optionally, combinations could be performed first in the parameter space of models $A$ , then on the space of models $B$ to then perform a final combination.

The quality of clusters as measured by several indexes suggests that there are clusters found by the CB algorithm that are not well differentiated or separated. Exploring the inclusion of the forecasts and the structure of models, based on suggestions by (Garson, 1991; Goh, 1995), seems to be a promising research avenue, as this could lead to better differentiated clusters. This scheme can be combined with the exploration of models of different nature, as mentioned earlier, so that both diversity in structure and diversity in forecasts are exploited.

Applications of the proposed combinations to high frequency data can be explored. In the case of NNs, the computing time can be lowered, if needed, by using an iterative forecasting approach, which leads to a single pool of models to run the combinations, instead of a direct approach, which leads to a number of pools equal to the number of forecast horizons.

## 6.5 Main Contributions of this Research

The first contribution of this research is the preliminary sensitivity analysis, performed before combining neural networks, in order to analyse the influence of different design factors on several error forecasting metrics. It permitted to expand on previous research regarding design of experiments, by adding detail, time series and a more robust statistical analysis of the influence of the chosen factors on the performance of NNs. The results of this analysis allowed to have a more objective selection of specifications for the obtention of structural combinations.

The second contribution of this research is the incorporation of model structures in forecasting combinations. It is done both with neural networks and several models of the Holt-Winters and Holt-Winters-Taylor families, thus providing a view of the capacity of such combinations to improve forecast performance. The study of seasonal time series constitute a contribution, as the capacity of neural networks to forecast such series is contended. In general, the studies permitted to explore the conditions under which the proposed combinations tend to work better and the possible routes for research. Some conditions of promising performance were found: Structural combination of neural networks (NN) with genetic algorithms (GA) seem to work better than cluster based (CB) combinations for non-seasonal time series, whereas, for seasonal series, the CB tend to do better. When combining exponential smoothing models for a single seasonal series of peak electricity demand, results suggest that cluster-based (CB) combinations are better at exploiting model variations coming from noisy data in order to improve performance. Combinations based on genetic algorithms (GA), on the other hand, seem to be well suited to exploit model variations through block swapping. The structural combination of exponential smoothing models to forecast a multiplicative double-seasonal series of electricity demand in Rio de Janeiro, produced better results under block swapping with CB combinations. Finally, the results obtained with a double-seasonal time series of electricity demand from England and Wales suggest that noise addition is better suited to exploit model diversity in the structural combinations when the data have dynamics that might be affected by the block swapping scheme. Under noise addition, GA combinations performed better than CB. In all, the simpler structural combination with genetic algorithms tended to be more stable than the cluster-based structural combination.

# Bibliography

Abdallatif, M., Schramm, S., & Gö tze, J. (2016). Application of fuzzy c-means for proactive clustering of electrical power systems. In *2016 IEEE 16th International Conference on Data Mining Workshops*, (pp. 382–389). IEEE.

Abdel-Aal, R. E. (2005). Improving electric load forecasts using network committees. *Electric Power Systems Research*, *74*(1), 83–94.

Adeodato, P. J., Arnaud, A. L., Vasconcelos, G. C., Cunha, R. C., & Monteiro, D. S. (2011). MLP ensembles improve long term prediction accuracy over single networks. *International Journal of Forecasting*, *27*(3), 661–671.

Adya, M. & Collopy, F. (1998). How effective are neural networks at forecasting and prediction? a review and evaluation. *Journal of Forecasting*, *17*(5-6), 481–495.

Alamaniotis, M., Ikonomopoulos, A., & Tsoukalas, L. H. (2012). Evolutionary multiobjective optimization of kernel-based very-short-term load forecasting. *IEEE Transactions on Power Systems*, *27*(3), 1477–1484.

Alessandrini, S., Delle Monache, L., Sperati, S., & Cervone, G. (2015). An analog ensemble for short-term probabilistic solar power forecast. *Applied Energy*, *157*, 95–110.

Alvarez-Diaz, M. & Alvarez, A. (2005). Genetic multi-model composite forecast for non-linear prediction of exchange rates. *Empirical Economics*, *30*(3), 643–663.

Amjady, N. (2001). Short-term hourly load forecasting using time-series modeling with peak load estimation capability. *IEEE Transactions on Power Systems*, *16*(4), 798–805.

Anders, U. & Korn, O. (1999). Model selection in neural networks. *Neural networks : the official journal of the International Neural Network Society*, *12*(2), 309–323.

Armstrong, J. S. (2001). *Principles of forecasting: a handbook for researchers and practitioners*, volume 30. Springer.

Arora, S. (2013). *Time Series Forecasting with Applications in Macroeconomics and Energy*. PhD thesis, Somerville College, University of Oxford.

Atiya, A. F., El-Shoura, S. M., Shaheen, S. I., & El-Sherif, M. S. (1999). A comparison between neural-network forecasting techniques–case study: river flow forecasting. *IEEE transactions on neural networks*, *10*(2), 402–9.

Bakker, B. & Heskes, T. (2003). Clustering ensembles of neural network models. *Neural networks : the official journal of the International Neural Network Society*, *16*(2), 261–269.

Balestrassi, P., Popova, E., a.P. Paiva, & Marangon Lima, J. (2009). Design of experiments on neural network's training for nonlinear time series forecasting. *Neurocomputing*, *72*(4-6), 1160–1178.

Balkin, S. D. & Ord, J. (2000). Automatic neural network modeling for univariate time series. *International Journal of Forecasting*, *16*(4), 509–515.

Bao, Y., Xiong, T., & Hu, Z. (2014). Multi-step-ahead time series prediction using multiple-output support vector regression. *Neurocomputing*, *129*, 482–493.

Barrow, D., Crone, S., & Kourentzes, N. (2010). An evaluation of neural network ensembles and model selection for time series prediction. In *Neural Networks (IJCNN), The 2010 International Joint Conference on*, (pp. 1–8).

Basak, D., Pal, S., & Patranabis, D. C. (2007). Support vector regression. *Neural Information Processing-Letters and Reviews*, *11*(10), 203–224.

Bates, J. M. & Granger, C. W. (1969). The combination of forecasts. *Or*, 451–468.

Bishop, C. (1995). *Neural Networks for Pattern Recognition.* Advanced Texts in Econometrics. Clarendon Press.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140.

Breiman, L. (2001). Random forests. *Machine Learning*, (45), 5–32.

Breiman, L. et al. (1996). Heuristics of instability and stabilization in model selection. *The annals of statistics*, *24*(6), 2350–2383.

Brown, W. M., Gedeon, T. D., & Groves, D. I. (2003). Use of noise to augment training data: a neural network method of mineral–potential mapping in regions of limited known deposit examples. *Natural Resources Research*, *12*(2), 141–152.

Bunn, D. W. (1975). A bayesian approach to the linear combination of forecasts. *Operational Research Quarterly*, 325–329.

Burger, E. M. & Moura, S. J. (2015). Gated ensemble learning method for demand-side electricity load forecasting. *Energy and Buildings*, *109*, 23–34.

Chandra, A. & Yao, X. (2006). Ensemble learning using multi-objective evolutionary algorithms. *Journal of Mathematical Modelling and Algorithms*, *5*(4), 417–445.

Che, J. (2015). Optimal sub-models selection algorithm for combination forecasting model. *Neurocomputing*, *151*(P1), 364–375.

Chen, H. & Yao, X. (2007). Evolutionary random neural ensembles based on negative correlation learning. (pp. 1468–1474). Ieee.

Chen, H. & Yao, X. (2009). Regularized negative correlation learning for neural network ensembles. *Neural Networks, IEEE Transactions on*, *20*(12), 1962–1979.

Chen, P.-A., Chang, L.-C., & Chang, F.-J. (2013). Reinforced recurrent neural networks for multi-step-ahead flood forecasts. *Journal of Hydrology*, *497*, 71–79.

Cheng, C.-T., Xie, J.-X., Chau, K.-W., & Layeghifard, M. (2008). A new indirect multi-step-ahead prediction model for a long-term hydrologic prediction. *Journal of Hydrology*, *361*(1-2), 118–130.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, *5*(4), 559 – 583.

Cooper, J. P. & Nelson, C. R. (1975). The ex ante prediction performance of the st. louis and frb-mit-penn econometric models and some results on composite predictors. *Journal of Money, Credit and Banking*, 1–32.

Cover, T. M. & Thomas, J. A. (2006). *Elements of information theory*. John Wiley & Sons.

Crone, S. F. & Dhawan, R. (2007). Forecasting seasonal time series with neural networks: A sensitivity analysis of architecture parameters. (pp. 2099–2104). Ieee.

Crone, S. F., Hibon, M., & Nikolopoulos, K. (2011). Advances in forecasting with neural networks? Empirical evidence from the NN3 competition on time series prediction. *International Journal of Forecasting*, *27*(3), 635–660.

Crone, S. F. & Kourentzes, N. (2010). Feature selection for time series prediction – A combined filter and wrapper approach for neural networks. *Neurocomputing*, *73*(10-12), 1923–1936.

Daneshi, H. & Daneshi, A. (2008). Real time load forecast in power system. In *Electric Utility Deregulation and Restructuring and Power Technologies, 2008. DRPT 2008. Third International Conference on*, number April, (pp. 689–695).

Darbellay, G. A. & Slama, M. (2000). Forecasting the short-term demand for electricity: Do neural networks stand a better chance? *International Journal of Forecasting*, *16*(1), 71–83.

De Menezes, L., Wbunn, D., & Taylor, J. W. (2000). Review of guidelines for the use of combined forecasts. *European Journal of Operational Research*, *120*(1), 190–204.

Deutsch, M., Granger, C. W., & Teräsvirta, T. (1994). The combination of forecasts using changing weights. *International Journal of Forecasting*, *10*(1), 47–57.

Diebold, F. X. & Lopez, J. A. (1996). 8 forecast evaluation and combination. In G. Maddala & C. Rao (Eds.), *Statistical Methods in Finance*, volume 14 of *Handbook of Statistics* (pp. 241 – 268). Elsevier.

Diebold, F. X. & Pauly, P. (1987). Structural change and the combination of forecasts. *Journal of Forecasting*, *6*(1), 21–40.

Dietterich, T. G. (1997). Machine-learning research. *AI magazine*, *18*(4), 97.

Donaldson, R. G. & Kamstra, M. (1996). Forecast combining with neural networks. *Journal of Forecasting*, *15*(1), 49–61.

Drezga, I. (1999). Short-term load forecasting with local ANN predictors. *IEEE Transactions on Power Systems*, *14*(3), 844–850.

Efron, B. & Tibshirani, R. (1994). *An Introduction to the Bootstrap*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis.

Fair, R. C. & Shiller, R. J. (1990). Comparing information in forecasts from econometric models. *The American Economic Review*, 375–389.

Fan, S., Chen, L., & Lee, W.-j. (2009). Short-term load forecasting using comprehensive combination based on multimeteorological information. *IEEE Transactions on Industry Applications*, *45*(4), 1460–1466.

Fay, D. & Ringwood, J. V. (2010). On the influence of weather forecast errors in short-term load forecasting models. *IEEE Transactions on Power Systems*, *25*(3), 1751–1758.

Field, A. (2009). *Discovering Statistics Using SPSS*. ISM (London, England). SAGE Publications.

Fiordaliso, a. (1998). A nonlinear forecasts combination method based on Takagi-Sugeno fuzzy systems. *International Journal of Forecasting*, *14*(3), 367–379.

Fisher, P. G. & Wallis, K. F. (1990). The historical tracking performance of uk macroeconometric models 1978–1985. *Economic Modelling*, *7*(2), 179–197.

Freund, Y., Schapire, R. E., et al. (1996). Experiments with a new boosting algorithm. In *ICML*, volume 96, (pp. 148–156).

Friedman, J. H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics*, *19*(1), 1–67.

Garson, G. D. (1991). Interpreting neural-network connection weights. *AI Expert*, *6*(4), 46–51.

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/-variance dilemma. *Neural computation*, *4*(1), 1–58.

Giebel, G., Landberg, L., Kariniotakis, G., & Brownsword, R. (2003). State-of-the-art on methods and software tools for short-term prediction of wind energy production. In *Proc. of the 2003 European Wind Energy Association Conference*, (pp. 16–19).

Goh, A. (1995). Back-propagation neural networks for modeling complex systems. *Artificial Intelligence in Engineering*, *9*(3), 143 – 151.

Goh, S., Chen, M., Popović, D., Aihara, K., Obradovic, D., & Mandic, D. (2006). Complex-valued forecasting of wind profile. *Renewable Energy*, *31*(11), 1733–1750.

Goia, A., May, C., & Fusai, G. (2010). Functional clustering and linear regression for peak load forecasting. *International Journal of Forecasting*, *26*(4), 700–711.

Gouriveau, R. & Zerhouni, N. (2012). Connexionist-systems-based long term prediction approaches for prognostics. *IEEE Transactions on Reliability*, *61*(4), 909–920.

Granger, C. (1993). *Modelling Nonlinear Economic Relationships*. Oxford University Press.

Granger, C. W. & Ramanathan, R. (1984). Improved methods of combining forecasts. *Journal of Forecasting*, *3*(2), 197–204.

Gunter, S. I. & Aksu, C. (1989). N-step combinations of forecasts. *Journal of Forecasting*, *8*(3), 253–267.

Haida, T. & Muto, S. (1994). Regression based peak load forecasting using a transformation technique. *IEEE Transactions on Power Systems*, *9*(4), 1788–1794.

Hansen, L. K. & Salamon, P. (1990). Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *12*(10), 993–1001.

Harvey, D. & Newbold, P. (2000). Tests for multiple forecast encompassing. *Journal of Applied Econometrics*, *15*(5), 471–482.

Hassan, S., Khosravi, A., & Jaafar, J. (2015). Examining performance of aggregation algorithms for neural network-based electricity demand forecasting. *International Journal of Electrical Power & Energy Systems*, *64*, 1098–1105.

Haykin, S. (1999). *Neural Networks: A Comprehensive Foundation.* International edition. Prentice Hall.

Hinton, G. (2011). Deep belief nets. In *Encyclopedia of Machine Learning* (pp. 267–269). Springer.

Hippert, H., Pedreira, C., & Souza, R. (2001). Neural networks for short-term load forecasting: a review and evaluation. *IEEE Transactions on Power Systems*, *16*(1), 44–55.

Hong, T., Pinson, P., & Fan, S. (2014). Global energy forecasting competition 2012. *International Journal of Forecasting*, *30*(2), 357–363.

Huang, G., Huang, G.-B., Song, S., & You, K. (2015). Trends in extreme learning machines: A review. *Neural Networks*, *61*, 32–48.

Hyndman, R., Koehler, A., Ord, J., & Snyder, R. (2008). *Forecasting with Exponential Smoothing: The State Space Approach.* Springer Series in Statistics. Springer Berlin Heidelberg.

Hyndman, R. J. (2015). *forecast: Forecasting functions for time series and linear models.* R package version 6.1.

Hyndman, R. J. (2017). *Double-Seasonal Holt-Winters Forecasting.* R package version 6.1.

Hyndman, R. J. & Khandakar, Y. (2008). Automatic time series forecasting: the forecast package for R. *Journal of Statistical Software*, *26*(3), 1–22.

Iizaka, T., Matsui, T., & Fukuyama, Y. (2002). A novel daily peak load forecasting method using analyzable structured neural network. In *Transmission and Distribution Conference and Exhibition 2002: Asia Pacific. IEEE/PES*, volume 1, (pp. 394–399 vol.1).

Islam, M. M., Yao, X., & Murase, K. (2003). A constructive algorithm for training cooperative neural network ensembles. *IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council*, *14*(4), 820–34.

Jacobs, R. & Jordan, M. I. (1991). A competitive modular connectionist architecture. In L. R.P., J. Moody, & D. Touretzky (Eds.), *Advances in Neural Information Processing Systems 3*, volume 2 (pp. 767–773). Morgan-Kaufmann.

Janczura, J., Trück, S., Weron, R., & Wolff, R. C. (2013). Identifying spikes and seasonal components in electricity spot price data: A guide to robust modeling. *Energy Economics*, *38*, 96–110.

Jang, J., Sun, C., & Mizutani, E. (1997). *Neuro-fuzzy and soft computing: a computational approach to learning and machine intelligence*. Prentice Hall.

Jing, B.-Y. (1997). On the relative performance of the block bootstrap for dependent data. *Communications in Statistics-Theory and Methods*, *26*(6), 1313–1328.

Joe H. Chow, Felix F. Wu, J. A. M. (2004). *Applied Mathematics for Restructured Electric Power Systems: Optimization, Control, and Computational Intelligence* (1 ed.). Power Electronics and Power Systems. Springer.

Kaggle (2012). Global energy forecasting competition - wind forecasting. `https://www.kaggle.com/c/GEF2012-wind-forecasting`. Viewed in September 2015.

Kasabov, N. (1996). *Foundations of Neural Networks, Fuzzy Systems, and Knowledge Engineering*. A Bradford book. MIT Press.

Kaur, A., Pedro, H. T. C., & Coimbra, C. F. M. (2014). Ensemble re-forecasting methods for enhanced power load prediction. *Energy Conversion and Management*, *80*, 582–590.

Khadem, S. & Dillon, T. (2012). Optimization of neural network configurations for short-term traffic flow forecasting using orthogonal design. In *2012 IEEE Congress on Evolutionary Computation*, (pp. 1–7). Ieee.

Khosravi, A., Nahavandi, S., Creighton, D., & Atiya, A. F. (2011). Comprehensive review of neural network-based prediction intervals and new advances. *IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council*, *22*(9), 1341–56.

Khotanzad, A., Afkhami-Rohani, R., & Maratukulam, D. (1998). ANNSTLF - Artificial Neural Network Short-Term Load Forecaster - Generation Three. *IEEE Transactions on Power Systems*, *13*(4), 1413–1422.

Khotanzad, A., Elragal, H., & Lu, T. L. (2000). Combination of artificial neural-network forecasters for prediction of natural gas consumption. *IEEE transactions on neural networks*, *11*(2), 464–73.

Kourentzes, N., Barrow, D. K., & Crone, S. F. (2014). Neural network ensemble operators for time series forecasting. *Expert Systems with Applications*, *41*(9), 4235–4244.

Krasnopolsky, V. M. (2007). Reducing uncertainties in neural network Jacobians and improving accuracy of neural network emulations with NN ensemble approaches. *Neural networks : the official journal of the International Neural Network Society*, *20*(4), 454–61.

Lahiri, S. (2013). *Resampling Methods for Dependent Data*. Springer Series in Statistics. Springer New York.

Lee, K. L. & Billings, S. a. (2003). A new direct approach of computing multi-step ahead predictions for non-linear models. *International Journal of Control*, *76*(8), 810–822.

Lee, Y. S. & Scholtes, S. (2014). Empirical prediction intervals revisited. *International Journal of Forecasting*, *30*(2), 217–234.

Leith, C. E. (1974). Theoretical skill of Monte-Carlo forecasts. *Monthly Weather Review*, *102*, 409–418.

Lemke, C. & Gabrys, B. (2010). Meta-learning for time series forecasting and forecast combination. *Neurocomputing*, *73*(10-12), 2006–2016.

LeSage, J. P. & Magura, M. (1992). A mixture-model approach to combining forecasts. *Journal of Business & Economic Statistics*, *10*(4), 445–452.

Li, G., Shi, J., & Zhou, J. (2011). Bayesian adaptive combination of short-term wind speed forecasts from neural network models. *Renewable Energy*, *36*(1), 352 – 359.

Liang, Z., Liang, J., Wang, C., Dong, X., & Miao, X. (2016). Short-term wind power combined forecasting based on error forecast correction. *Energy Conversion and Management*, *119*, 215–226.

Liu, Y. & Yao, X. (1999). Simultaneous training of negatively correlated neural networks in an ensemble. *IEEE transactions on systems, man, and cybernetics. Part B, Cybernetics : a publication of the IEEE Systems, Man, and Cybernetics Society*, *29*(6), 716–25.

Liu, Y., Yao, X., & Higuchi, T. (2000). Evolutionary ensembles with negative correlation learning. *IEEE Transactions on Evolutionary Computation*, *4*(4), 380–387.

Lorenz, E. N. (1965). Deterministic nonperiodic flow. *Journal of the Athmospheric Sciences*, *20*(3), 130–141.

Lu, B. L. & Ito, M. (1999). Task decomposition and module combination based on class relations: a modular neural network for pattern classification. *IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council*, *10*(5), 1244–56.

Maines, L. a. (1996). An experimental examination of subjective forecast combination. *International Journal of Forecasting*, *12*(2), 223–233.

Makridakis, S. & Winkler, R. L. (1983). Averages of forecasts: Some empirical results. *Management Science*, *29*(9), 987–996.

Maqsood, I., Khan, M., & Abraham, A. (2004). An ensemble of neural networks for weather forecasting. *Neural Computing and Applications*, *13*(2), 112–122.

Martí, I., Kariniotakis, G., Pinson, P., Sanchez, I., Nielsen, T. S., Madsen, H., Giebel, G., & Usaola, J. (2006). Evaluation of advanced wind power forecasting models - Results of the Anemos project. In *Proceedings of European wind energy conference.*

Matijaš, M., a.K. Suykens, J., & Krajcar, S. (2013). Load forecasting using a multivariate meta-learning system. *Expert Systems with Applications*, *40*(11), 1–11.

Matlab (2017). Bootstrp. `https://uk.mathworks.com/help/stats/bootstrp.html`. Viewed in May 2017.

Medeiros, M. C., Teräsvirta, T., & Rech, G. (2006). Building neural network models for time series: a statistical approach. *Journal of Forecasting*, *25*(1), 49–75.

Mendes-Moreira, J., Soares, C., Jorge, A. M., & Sousa, J. F. D. (2012). Ensemble approaches for regression: A survey. *ACM Computing Surveys (CSUR)*, *45*(1), 10.

Montgomery, D. C. (2008). *Design and analysis of experiments.* John Wiley & Sons.

Montgomery, G. J. & Drake, K. C. (1991). Abductive reasoning networks. *Neurocomputing*, *2*(3), 97–104.

Murphy, J. M., Sexton, David M. H. Barnett, D. N., Jones, G. S., Webb, M. J., Collins, M., & Stainforth, D. A. (2004). Quantification of modelling uncertainties in a large ensemble of climate change simulations. *Nature*, *430*, 768 – 772.

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

Newbold, P. & Harvey, D. I. (2007). Forecast combination and encompassing. In M. P. Clements & D. F. Hendry (Eds.), *A Companion to Economic Forecasting* (pp. 268–283). Blackwell Publishing Ltd.

Nguyen, D. & Widrow, B. (1990). Improving the learning speed of 2-layer neural networks by choosing initial values of the adaptive weights. In *Neural Networks, 1990., 1990 IJCNN International Joint Conference on*, (pp. 21–26). IEEE.

Ord, K. & Fildes, R. (2012). *Principles of Business Forecasting.* Cengage Learning.

Pai, P.-F. & Hong, W.-C. (2005). Forecasting regional electricity load based on recurrent support vector machines with genetic algorithms. *Electric Power Systems Research*, *74*(3), 417 – 425.

Palit, A. K. & Popovic, D. (2000). Nonlinear combination of forecasts using artificial neural network, fuzzy logic and neuro-fuzzy approaches. In *Fuzzy Systems, 2000. FUZZ IEEE 2000. The Ninth IEEE International Conference on*, volume 2, (pp. 566–571). IEEE.

Pan, R. (2010). *Holt-Winters Exponential Smoothing*. John Wiley & Sons, Inc.

Pankratz, A. (2009). *Forecasting with Univariate Box - Jenkins Models: Concepts and Cases*. Wiley Series in Probability and Statistics. Wiley.

Pao, Y.-H. & Takefuji, Y. (1992). Functional-link net computing: theory, system architecture, and functionalities. *Computer*, *25*(5), 76–79.

Parker, W. S. (2010). Predicting weather and climate: Uncertainty, ensembles and probability. *Studies In History and Philosophy of Science Part B: Studies In History and Philosophy of Modern Physics*, *41*(3), 263–272.

Qiu, X., Zhang, L., Ren, Y., & Suganthan, P. N. (2014). Ensemble deep learning for regression and time series forecasting.

Ren, Y., Suganthan, P., & Srikanth, N. (2015). Ensemble methods for wind and solar power forecasting – A state-of-the-art review. *Renewable and Sustainable Energy Reviews*, *50*, 82–91.

Salcedo-Sanz, S., Perezbellido, A., Ortizgarcia, E., Portillafigueras, A., Prieto, L., & Correoso, F. (2009). Accurate short-term wind speed prediction by exploiting diversity in input data using banks of artificial neural networks. *Neurocomputing*, *72*(4-6), 1336–1341.

Sanchez, I. (2006). Short-term prediction of wind energy production. *International Journal of Forecasting*, *22*, 43 – 56.

Sanchez, I. (2008). Adaptive combination of forecasts with application to wind energy. *International Journal of Forecasting*, *24*(4), 679–693.

Siwek, K., Osowski, S., & Szupiluk, R. (2009). Ensemble neural network approach for accurate load forecasting in a power system. *International Journal of Applied Mathematics and Computer Science*, *19*(2), 303–315.

Stephenson, D. B., Coelho, C. A. S., Doblas-Reyes, F. J., & Balsameda, M. (2005). Forecast assimilation: a unified framework for the combination of multi-model weather and climate predictions. *Tellus A*, *57*(3), 253–264.

Stock, J. H. & Watson, M. W. (2004). Combination forecasts of output growth in a seven-country data set. *Journal of Forecasting*, *23*(6), 405–430.

Taguchi, G. & Yokoyama, Y. (1993). *Taguchi methods: design of experiments*. TAGUCHI METHODS SERIES. ASI Press.

Taieb, S. B., Sorjamaa, A., & Bontempi, G. (2010). Multiple-output modeling for multi-step-ahead time series forecasting. *Neurocomputing*, *73*(10), 1950–1957.

Taylor, J., McSharry, P., & Buizza, R. (2009). Wind power density forecasting using ensemble predictions and time series models. *Energy Conversion, IEEE Transactions on*, *24*(3), 775 –782.

Taylor, J. W. (2003). Short-term electricity demand forecasting using double seasonal exponential smoothing. *Journal of the Operational Research Society*, *54*(8), 799–805.

Taylor, J. W. (2007). Forecasting daily supermarket sales using exponentially weighted quantile regression. *European Journal of Operational Research*, *178*(1), 154–167.

Taylor, J. W. (2010). Triple seasonal methods for short-term electricity demand forecasting. *European Journal of Operational Research*, *204*(1), 139–152.

Taylor, J. W. & Buizza, R. (2002). Neural network load forecasting with weather ensemble predictions. *IEEE Transactions on Power Systems*, *17*(3), 626–632.

Taylor, J. W. & Buizza, R. (2003). Using weather ensemble predictions in electricity demand forecasting. *International Journal of Forecasting*, *19*(1), 57–70.

Taylor, J. W., de Menezes, L. M., & McSharry, P. E. (2006). A comparison of univariate methods for forecasting electricity demand up to a day ahead. *International Journal of Forecasting*, *22*(1), 1–16.

Taylor, J. W. & Majithia, S. (2000). Using combined forecasts with changing weights for electricity demand profiling. *The Journal of the Operational Research Society*, *51*(1), pp. 72–82.

Teräsvirta, 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*(4), 755–774.

Terui, N. & van Dijk, H. K. (2002). Combined forecasts from linear and nonlinear time series models. *International Journal of Forecasting*, *18*(3), 421–438.

Timmermann, A. (2006). Chapter 4 forecast combinations. volume 1 of *Handbook of Economic Forecasting* (pp. 135 – 196). Elsevier.

Wang, J. & Hu, J. (2015). A robust combination approach for short-term wind speed forecasting and analysis – Combination of the ARIMA (Autoregressive Integrated Moving Average), ELM (Extreme Learning Machine), SVM (Support Vector Machine) and LSSVM (Least Square SVM) forecasts usi. *Energy*, *93*, 41–56.

Wang, Z. & Cao, Y. (2006). Mutual information and non-fixed ANNs for daily peak load forecasting. *2006 IEEE PES Power Systems Conference and Exposition, PSCE 2006 - Proceedings*, (1), 1523–1527.

Webby, R. & O'Connor, M. (1996). Judgemental and statistical time series forecasting: a review of the literature. *International Journal of Forecasting*, *12*(1), 91–118.

Willis, H. (2000). *Distributed Power Generation: Planning and Evaluation.* Power Engineering (Willis). Taylor & Francis.

Wu, K.-l. & Yang, M.-s. (2005). A cluster validity index for fuzzy clustering. *Pattern Recognition Letters*, *26*, 1275–1291.

Xiong, L., Shamseldin, A. Y., & O'Connor, K. M. (2001). A non-linear combination of the forecasts of rainfall-runoff models by the first-order takagi–sugeno fuzzy system. *Journal of Hydrology*, *245*(1–4), 196 – 217.

Yan, J., Li, K., Bai, E., Yang, Z., & Foley, A. (2016). Time series wind power forecasting based on variant Gaussian Process and TLBO. *Neurocomputing*, *189*, 135–144.

Yan, W. (2012). Toward automatic time-series forecasting using neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, *23*(7), 1028–1039.

Yang, Y. (2004). Combining forecasting procedures: some theoretical results. *Econometric Theory*, *20*(01), 176–222.

Yao, X. & Islam, M. (2008). Evolving artificial neural network ensembles. *Computational Intelligence Magazine, IEEE*, *3*(1), 31–42.

Yu, H. & Wilamowski, B. M. (2010). Levenberg-Marquardt Training 12.1. In *Industrial Electronics Handbook, 2nd Edition* (pp. 1–16). CRC Press.

Yu, L., Lai, K. K., & Wang, S. (2008). Multistage RBF neural network ensemble learning for exchange rates forecasting. *Neurocomputing*, *71*(16-18), 3295–3302.

Yu, L., Wang, S., & Lai, K. (2005). A novel nonlinear ensemble forecasting model incorporating GLAR and ANN for foreign exchange rates. *Computers & Operations Research*, *32*(10), 2523–2541.

Yu, L., Wang, S., & Lai, K. K. (2008). Forecasting crude oil price with an EMD-based neural network ensemble learning paradigm. *Energy Economics*, *30*(5), 2623–2635.

Zhang, D., Ji, M., Yang, J., Zhang, Y., & Xie, F. (2014). A novel cluster validity index for fuzzy clustering based on bipartite modularity. *Fuzzy Sets and Systems*, *253*(c), 122–137.

Zhang, G. & Qi, M. (2005). Neural network forecasting for seasonal and trend time series. *European Journal of Operational Research*, *160*(2), 501–514.

Zhang, G., Wu, Y., & Liu, Y. (2014). An advanced wind speed multi-step ahead forecasting approach with characteristic component analysis. *Journal of Renewable and Sustainable Energy*, *6*(5), 053139.

Zhang, G. P. (2007). A neural network ensemble method with jittered training data for time series forecasting. *Information Sciences*, *177*(23), 5329–5346.

Zhang, G. P., Patuwo, B. E., & Hu, M. Y. (2001). A simulation study of artificial neural networks for nonlinear time-series forecasting. *Computers & Operations Research*, *28*, 381–396.

Zhang, G. P., Patuwo, E. B., & Michael Y., H. (1998). Forecasting with artificial neural networks: The state of the art. *International Journal of Forecasting*, *14*(1), 35–62.

Zhao, Y., Ye, L., Li, Z., Song, X., Lang, Y., & Su, J. (2016). A novel bidirectional mechanism based on time series model for wind power forecasting. *Applied Energy*, *177*, 793–803.

Zhou, Z.-H., Wu, J., & Tang, W. (2002). Ensembling neural networks: Many could be better than all. *Artificial Intelligence*, *137*(1-2), 239–263.

# Appendix A

## Multi-Step-Ahead Forecasts with Neural Networks

Forecasting with statistical models has been done normally for several horizons (Ord & Fildes, 2012). On the contrary, the NN literature has tended to focus on a single one. This focus is partly driven by the greater complexity of the multi-step ahead forecasting problem, as highlighted by the different approaches that have been proposed. Here a brief summary of those approaches is made in order to inform design decisions for the present research. A review and classification of works according to the approach for multi-step ahead forecast is given by Gouriveau & Zerhouni (2012).

According to the authors, simple approaches to multi-step ahead forecast include the parallel one, which uses one network with multiple outputs, one for every step. It is practical as all forecasts are obtained with a single model, but it suffers from the presence of serious rounding errors. Another simple approach is the iterative, which uses one model with one output. The model is fed iteratively with previous estimated values ($\hat{x}_{t+1}$) to produce forecasts at the following steps ($\hat{x}_{t+2}, \ldots$). It is easy to implement but suffers from error propagation.

More complex methods include the direct approach (Taieb et al., 2010). It fits one network for every step ahead with all models using the same data. Specialisation on a single horizon is the main advantage of this approach. However, complex dependencies between variables is not taken into account. DirRec (or cascade) has one model fitted for every horizon and the forecast for period t+1 is used as input for a network producing the forecast for period t+2. Although this method is easy to implement it also suffers from error propagation. The MIMO approach works as the parallel one, but with several networks, each one of them producing forecasts for several horizons until all horizons are covered. Finally, MISMO (Multiple-Input Several Multiple-Outputs) consists of several MIMO with a parameter $s$ that determines the output number of all MIMO.

These approaches are adopted with variations in the literature. Sample works are cited to illustrate.

Drezga (1999) and Matijaš et al. (2013) adopted the iterative approach for hourly electricity load forecasting while Siwek et al. (2009) uses the parallel approach. However, authors do not include comparisons with other multi-step ahead approaches.

Atiya et al. (1999) used several NNs forecasting approaches to conduct river flow forecasting. When comparing the direct, iterative and cascade methods, better results were reported with the direct one, although they clarified that the comparative ability of the different approaches is usually problem dependent. Lee & Billings (2003) also found evidence in favour of the direct approach when compared with the iterative one, in the case of non-linear time series, and proposed a modification of the former in order to reduce the mean squared prediction errors that come from the existence of autocorrelation in the prediction errors.

Goh et al. (2006) proposed to forecast several steps ahead the vector composed of wind speed and wind direction, with a recurrent NN architecture based on a cascade scheme. Although performance was satisfactory, no comparison is made with other multi-step ahead forecasting schemes.

Cheng et al. (2008) used NNs with time delays and splines obtaining better performance with respect to other time delays networks. Multi-step ahead forecasts were produced with the iterative approach. The improvement in performance obtained with the proposed method came from sophisticated input treatment and model training but not from a fundamental difference in the form of producing multi-step ahead forecasts.

Yan (2012) proposed an approach for automatic time-series forecasting with NN which included a direct approach for multi-step ahead forecasting with good results in the NN3 competition, for its reduced data set category. The method for producing multi-step ahead forecasts is not compared against others, but it is clear that a simple approach was helpful in his attempt to automate the forecasting process.

Chen et al. (2013) developed a cascade version of a training algorithm for recurrent NN. They reported superior results in multi-step ahead forecasts when comparing with the a training algorithm limited to one-step ahead forecasts. This research shows how the training algorithm and the scheme for multi-step ahead forecasts tend to be interrelated. A given training algorithm may limit the options to produce multi-step ahead forecasts and a given strategy to multi-step ahead forecasting can require modification in the training algorithm. This is also clear from the research by Bao et al. (2014). They implemented MIMO strategy with support vector regression and obtained lower forecast errors when comparing with the direct and iterative approaches. Their strategy had to be carefully chosen taking into account the type of network. Zhang et al. (2014) also used the MIMO strategy with extreme learning machine and obtained superior results when comparing with the iterated and direct approaches.

In general the superiority of a given approach to produce multi-step ahead forecasts is undecided and dependent on the problem, but it can be seen that the tendency is to use recurrent networks and adapt iterative, cascade or MIMO approaches depending on the architecture characteristics and training algorithms. The direct approach is well positioned with the potential of producing smaller independent models at the expense of loosing interaction between steps ahead. It is a practical solution with moderate complexity. The parallel and iterative approaches are also practical but tend to be less used used with simple networks probably due to the propagation of error. In the case of electricity load forecasting, the iterative approach remains in use throughout the years.

# Appendix B

## Selected Model Configurations for Sample Series

The base procedure proposed in section 3.4.5 is used to select models to ensemble for several series. Such series are chosen as to have different levels of complexity. The procedure is used to favour compact models because this conveniently limits the running times for the combination routines to be used in the following chapter. The model database and results gathered can also be used to select models with different criteria so that performance on ensembles constructed with such models improves.

Given that the generating processes for the non-seasonal series are based on information from a few lags and taking into account the homogeneity of the metrics behaviour for higher forecast horizons, the selection made here is restricted to the first 6 horizons in the selected non-seasonal series. Table B.1 shows the selected models for STAR2, Synthetic-1S and Synthetic-2S series.

For non-seasonal series no benefit was observed in adding complexity (high number of inputs or neurons) but relatively large models were found appropriate for the seasonal series. The most practical way to use the results from the design of experiments to identify those configurations is going to the main effects graphs and locating the levels of the NI and NU factors for the out-of-sample period where the error metric has reached a relatively low value and further increases in the factors do not lower the metric considerably. This information is contrasted, by the proposed heuristic, with tests and serial correlation maps in order to reduce the selection to parsimonious configurations, if possible.

Table B.1: Selected NN models.

| h | NI | NU |
|---|---|---|
| 1 | 2 | 3 |
| 2 | 2 | 1 |
| 3 | 3 | 5 |
| 4 | 2 | 5 |
| 5 | 2 | 1 |
| 6 | 2 | 3 |

STAR2 series.

| h | NI | NU |
|---|---|---|
| 1 | 7 | 9 |
| 2 | 7 | 9 |
| 3 | 6 | 10 |
| 4 | 7 | 9 |
| 5 | 6 | 9 |
| 6 | 7 | 8 |
| 7 | 6 | 9 |
| 8 | 4 | 7 |
| 9 | 3 | 8 |
| 10 | 3 | 8 |
| 11 | 2 | 7 |
| 12 | 1 | 8 |

Synthetic-1S series.

| h | NI | NU |
|---|---|---|
| 1 | 2 | 8 |
| 2 | 3 | 6 |
| 3 | 3 | 6 |
| 4 | 3 | 7 |
| 5 | 3 | 7 |
| 6 | 3 | 6 |
| 7 | 3 | 7 |
| 8 | 3 | 6 |
| 9 | 4 | 6 |
| 10 | 4 | 6 |
| 11 | 3 | 8 |
| 12 | 2 | 7 |

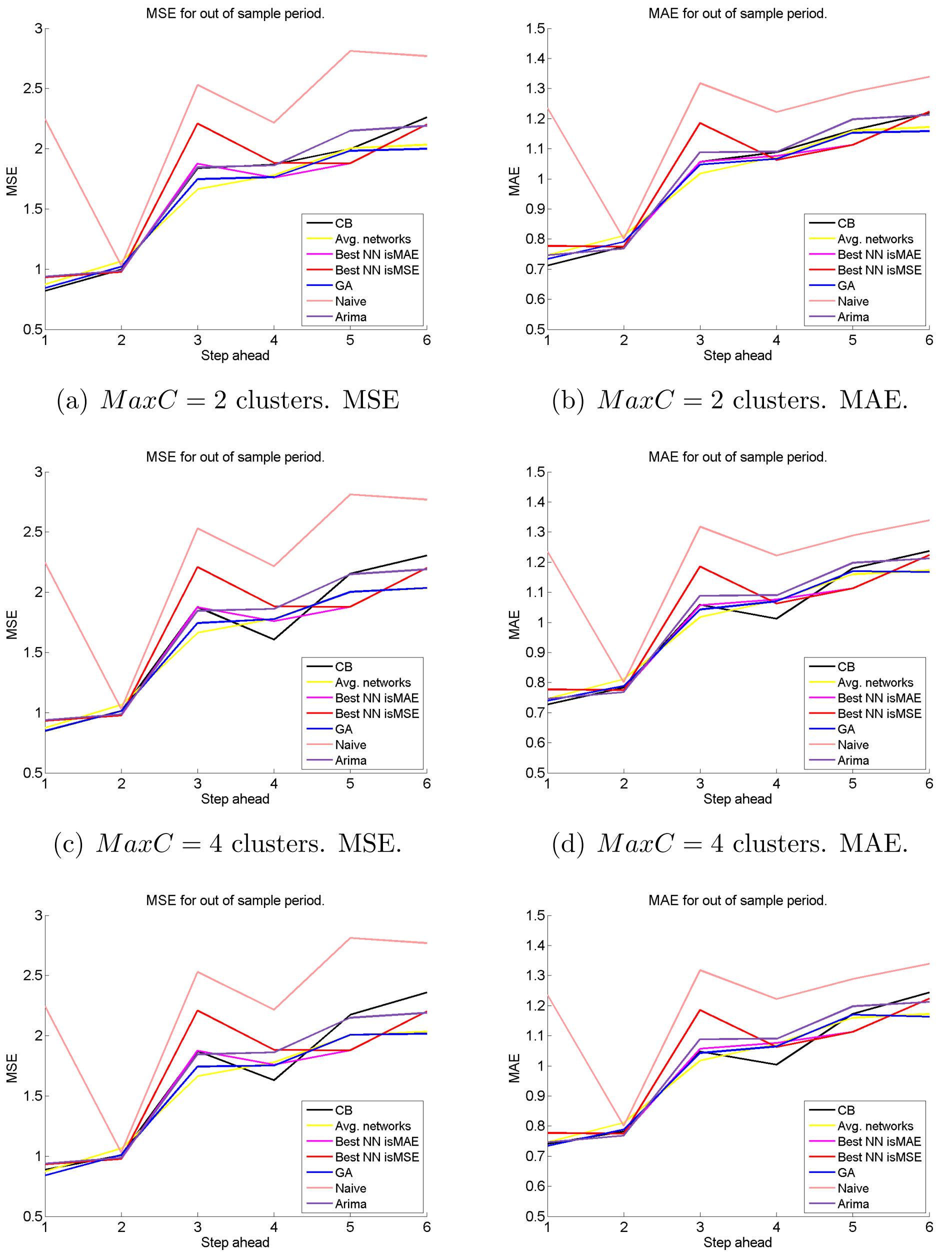
Synthetic-2S series.

# Appendix C

## Additional Material for the Chapter on Structural Combination of Neural Network Forecasting Models

### C.1 STAR2 Time Series

This section contains results obtained with structural combinations for the STAR2 time series.



(a) $MaxC = 2$ clusters. MSE

(b) $MaxC = 2$ clusters. MAE.

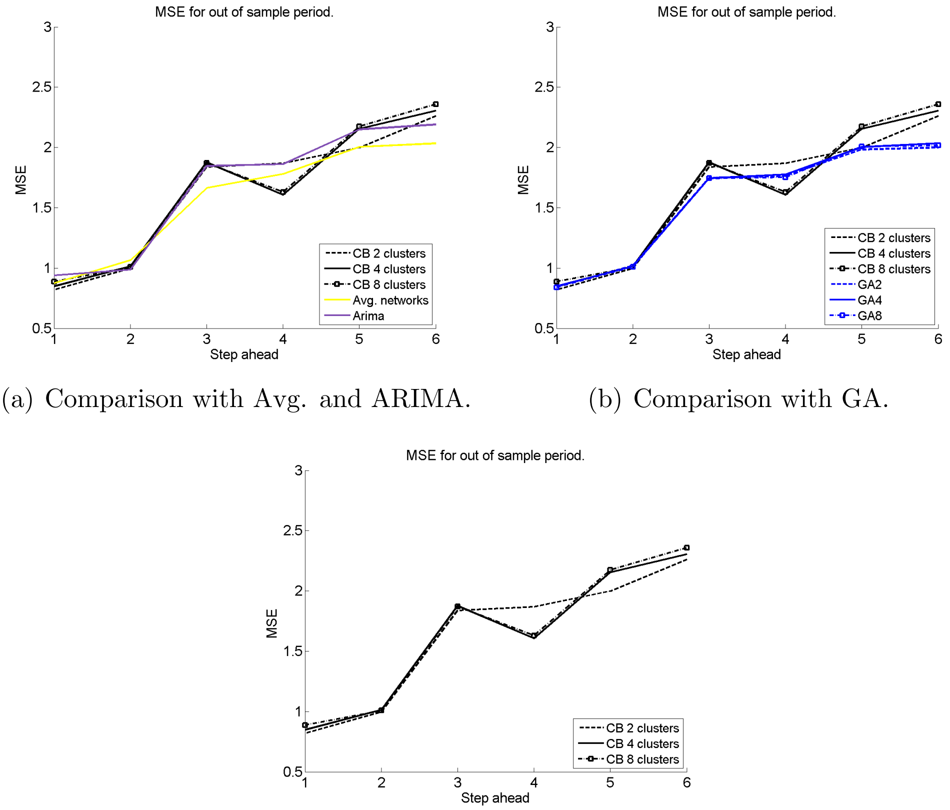
(c) $MaxC = 4$ clusters. MSE.

(d) $MaxC = 4$ clusters. MAE.

(e) $MaxC = 8$ clusters. MSE

(f) $MaxC = 8$ clusters. MAE.

Figure C.1: Out-of-sample MSE and MAE for STAR2 series.



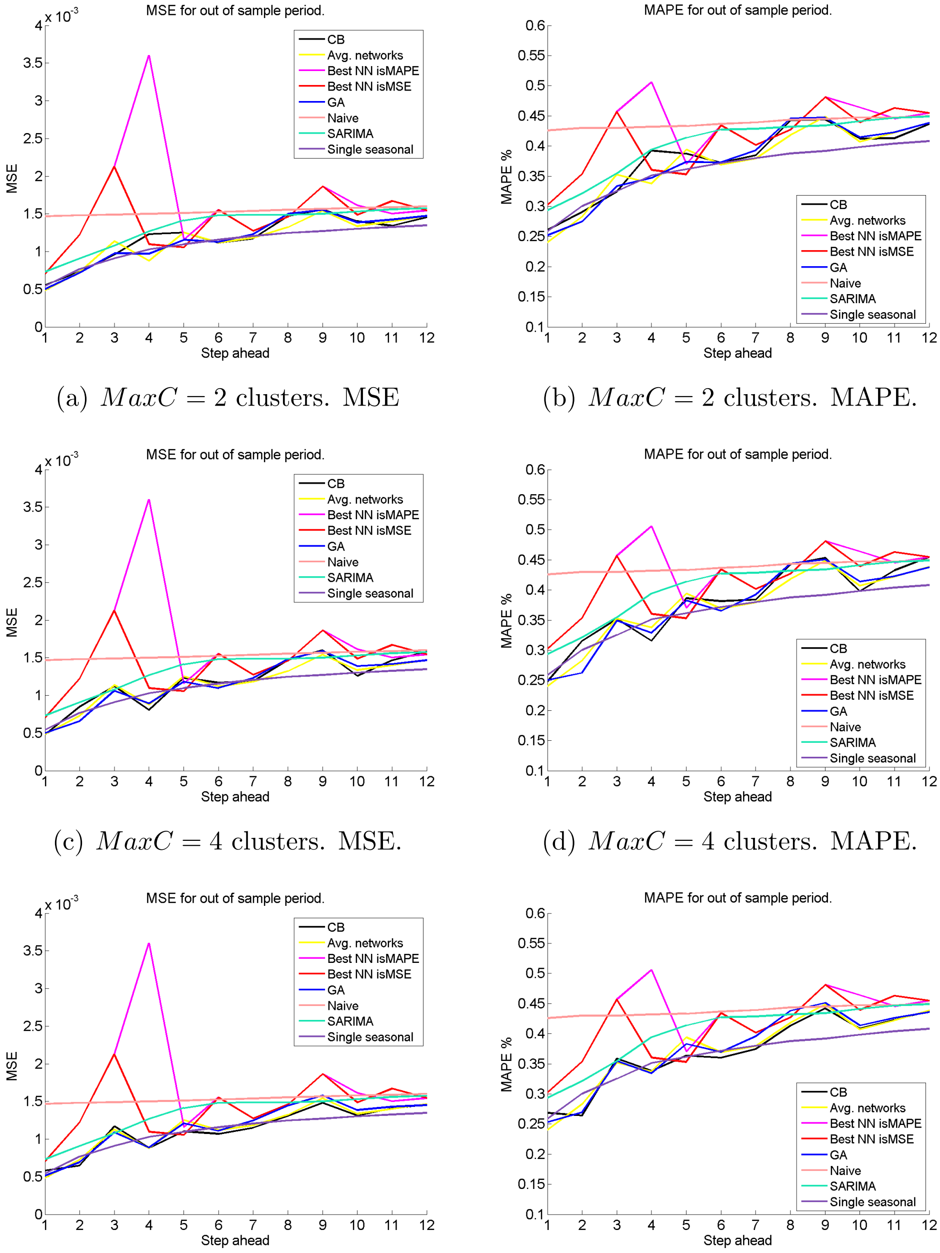
(a) Comparison with Avg. and ARIMA.

(b) Comparison with GA.

(c) All CB.

Figure C.2: Out-of-sample MSE for STAR2 series.

## C.2 Synthetic-1S Time Series



(a) $MaxC = 2$ clusters. MSE

(b) $MaxC = 2$ clusters. MAPE.

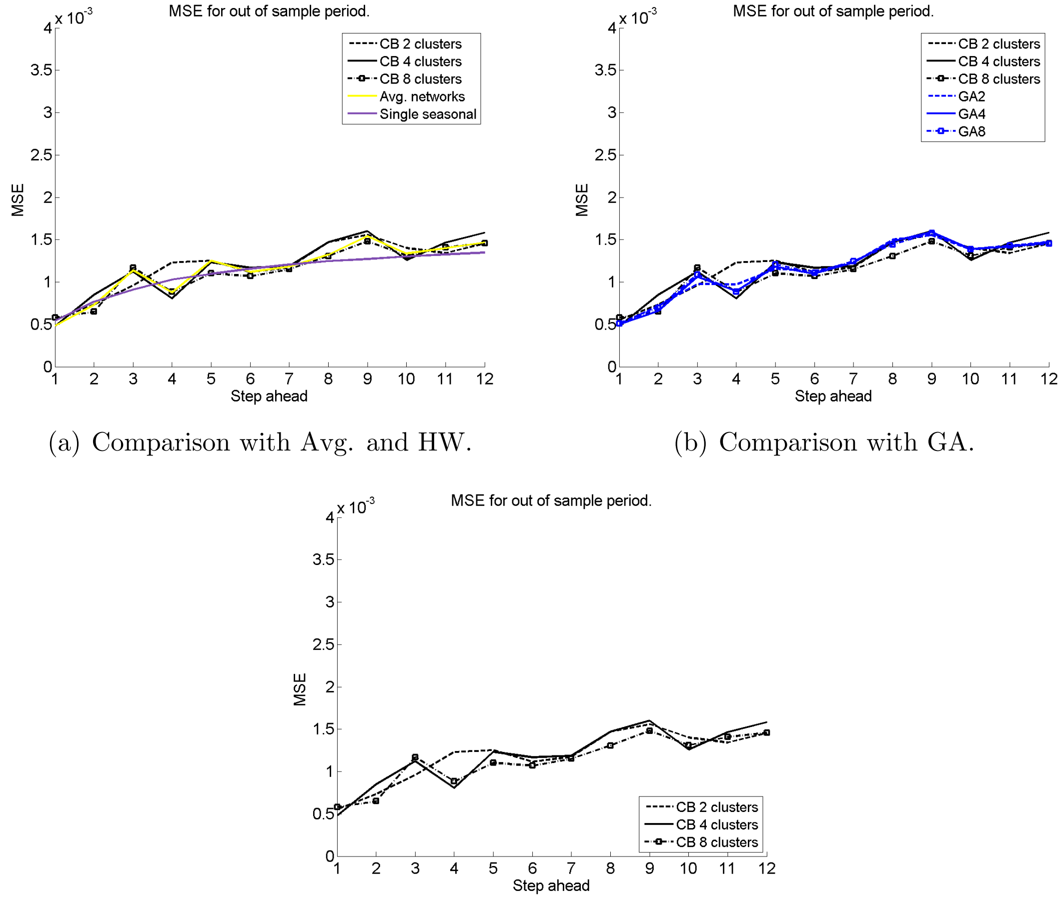
(c) $MaxC = 4$ clusters. MSE.

(d) $MaxC = 4$ clusters. MAPE.

(e) $MaxC = 8$ clusters. MSE

(f) $MaxC = 8$ clusters. MAPE.

Figure C.3: Out-of-sample MSE and MAPE for Synthetic-1S series.



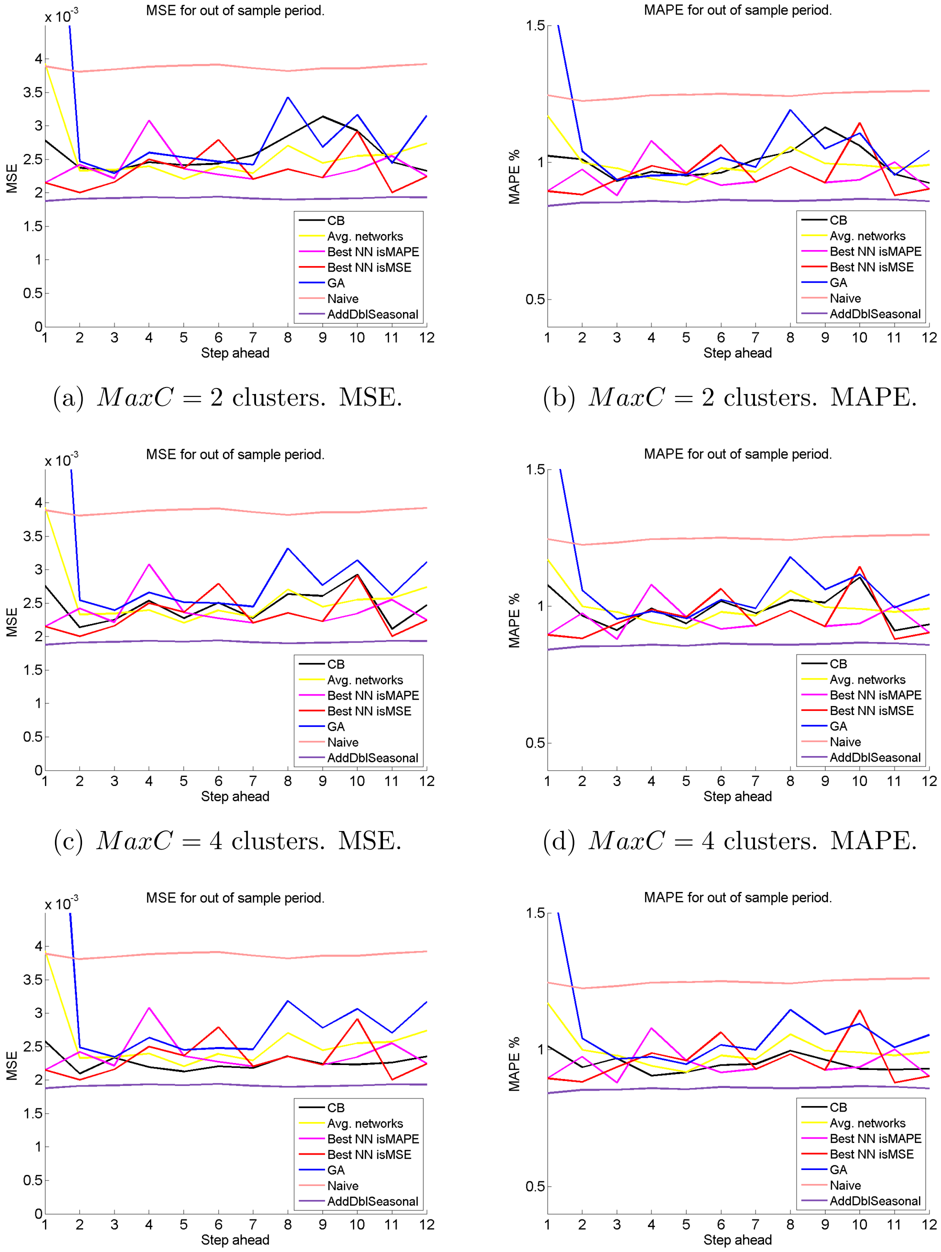
(a) Comparison with Avg. and HW.

(b) Comparison with GA.

(c) All CB.

Figure C.4: Out-of-sample MSE for Synthetic-1S series.

## C.3 Synthetic-2S Series



(a) $MaxC = 2$ clusters. MSE.

(b) $MaxC = 2$ clusters. MAPE.

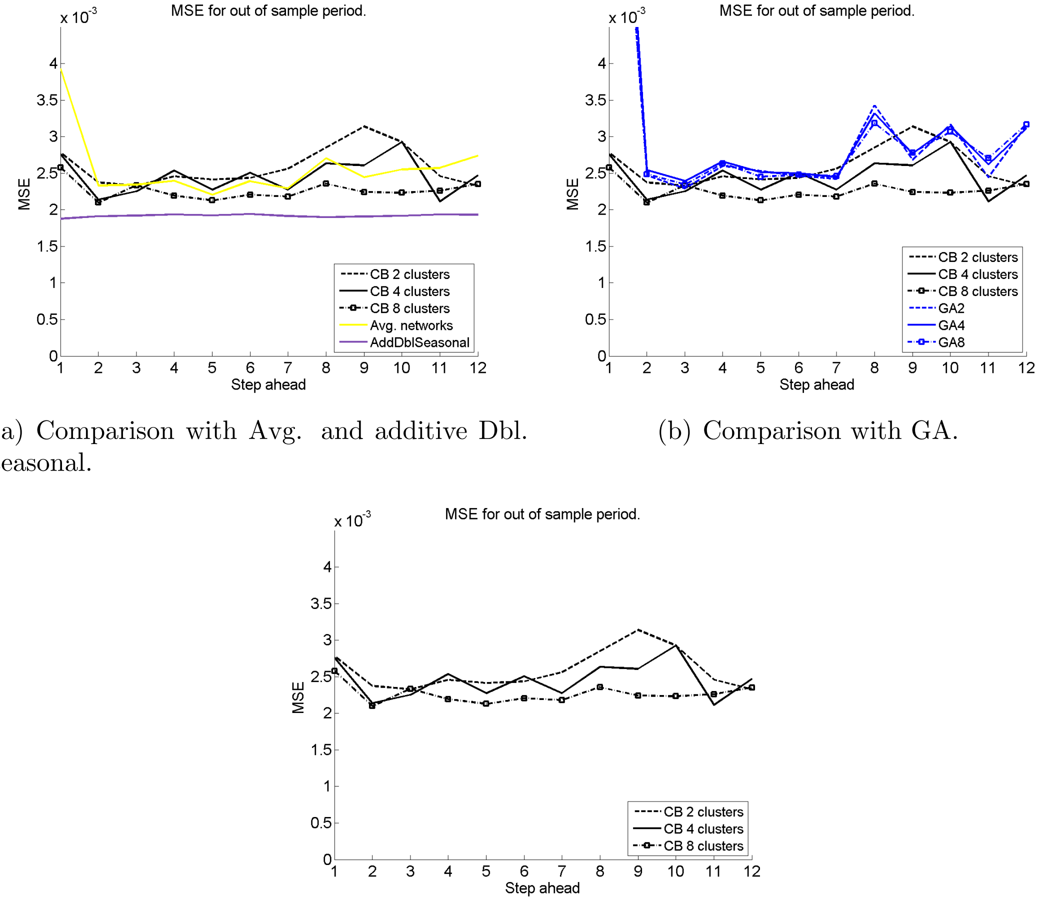
(c) $MaxC = 4$ clusters. MSE.

(d) $MaxC = 4$ clusters. MAPE.

(e) $MaxC = 8$ clusters. MSE.

(f) $MaxC = 8$ clusters. MAPE.

Figure C.5: Out-of-sample MSE and MAPE for Synthetic-2S series.



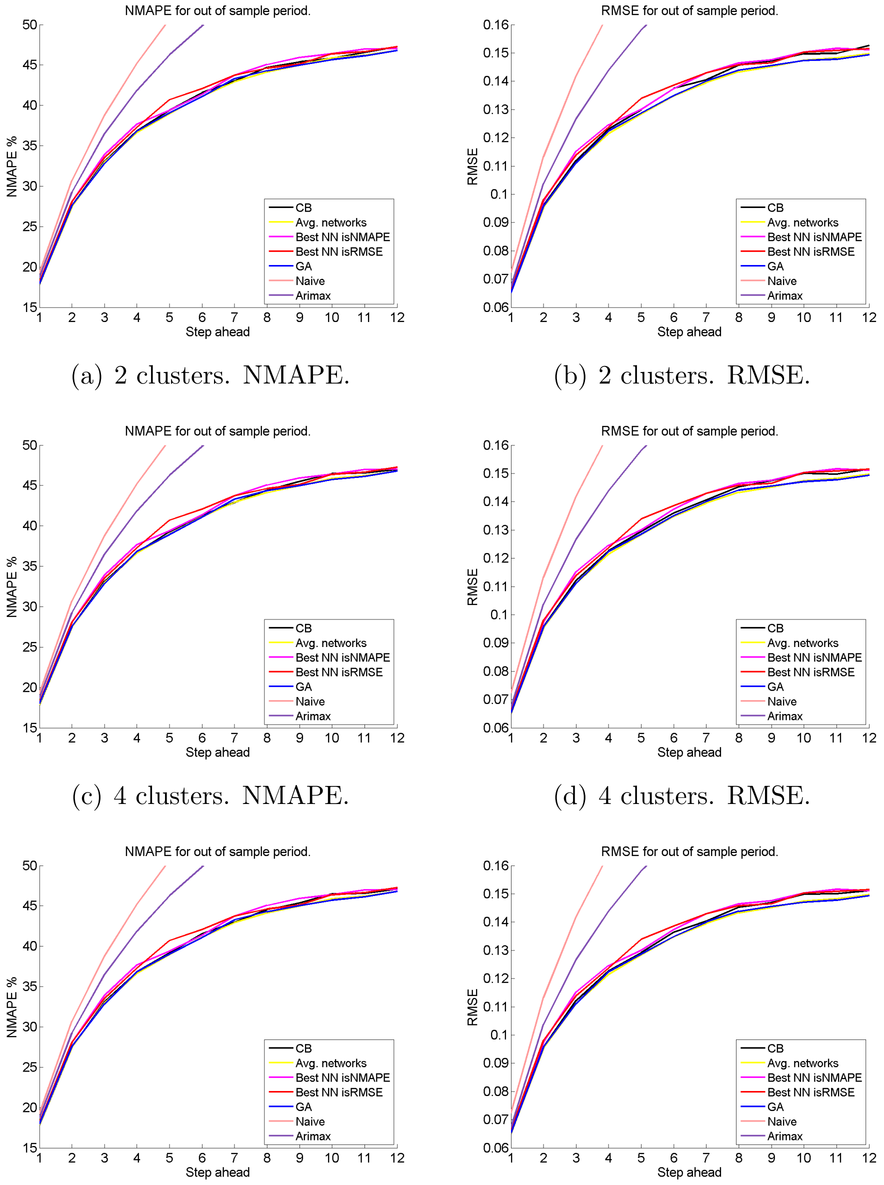
(a) Comparison with Avg. and additive Dbl. seasonal.

(b) Comparison with GA.

(c) All CB.

Figure C.6: Out-of-sample MSE for Synthetic-2S series.

# C.4 Wind Power Time Series



(a) 2 clusters. NMAPE.

(b) 2 clusters. RMSE.

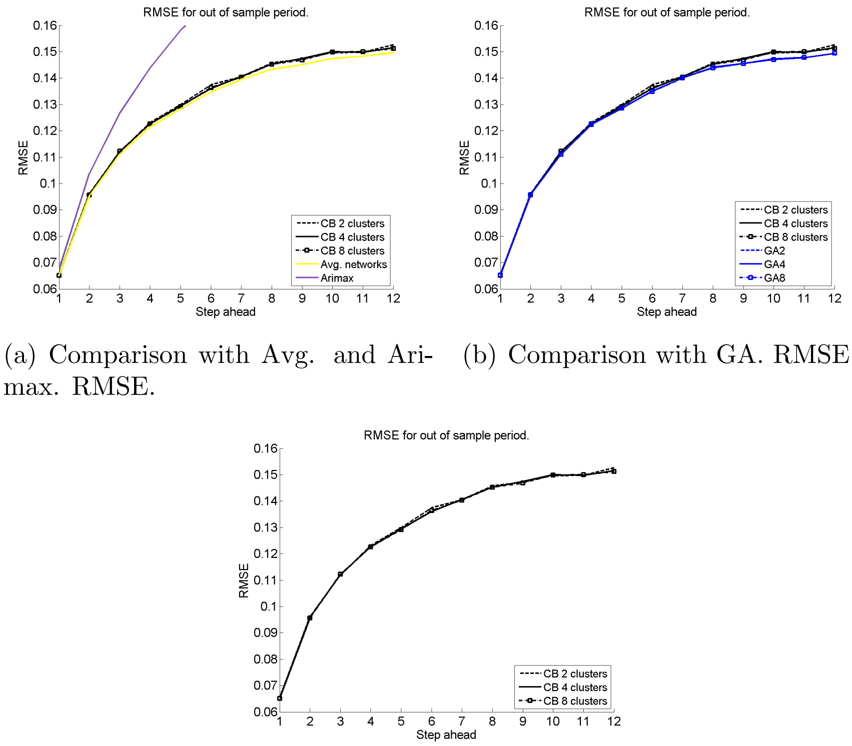
(c) 4 clusters. NMAPE.

(d) 4 clusters. RMSE.

(e) 8 clusters. NMAPE.

(f) 8 clusters. RMSE.

Figure C.7: Out-of-sample NMAPE and RMSE for Kaggle wind power production series.



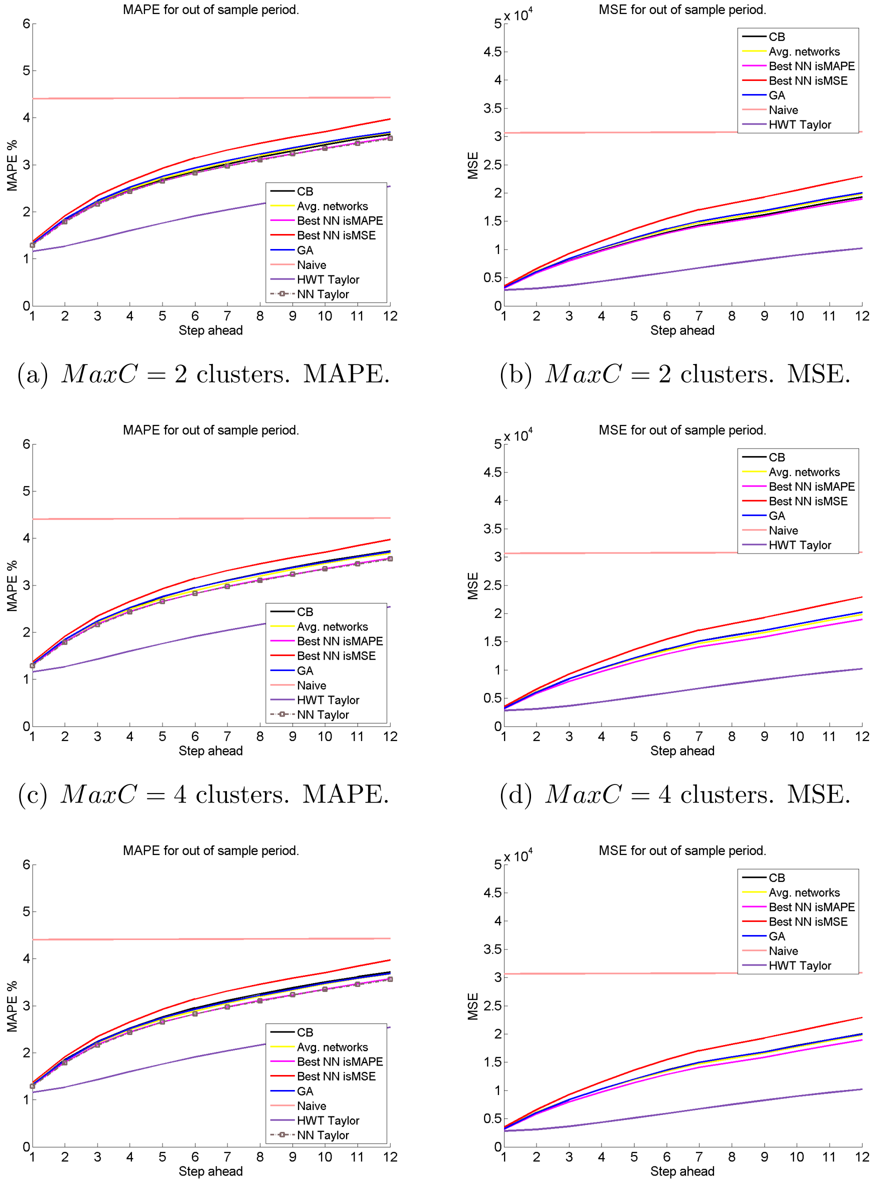
(a) Comparison with Avg. and Arimax. RMSE.

(b) Comparison with GA. RMSE.

(c) All CB. RMSE.

Figure C.8: Out-of-sample RMSE for Kaggle wind power production series.

## C.5 Rio de Janeiro Electricity Demand Time Series



(a) $MaxC = 2$ clusters. MAPE.

(b) $MaxC = 2$ clusters. MSE.

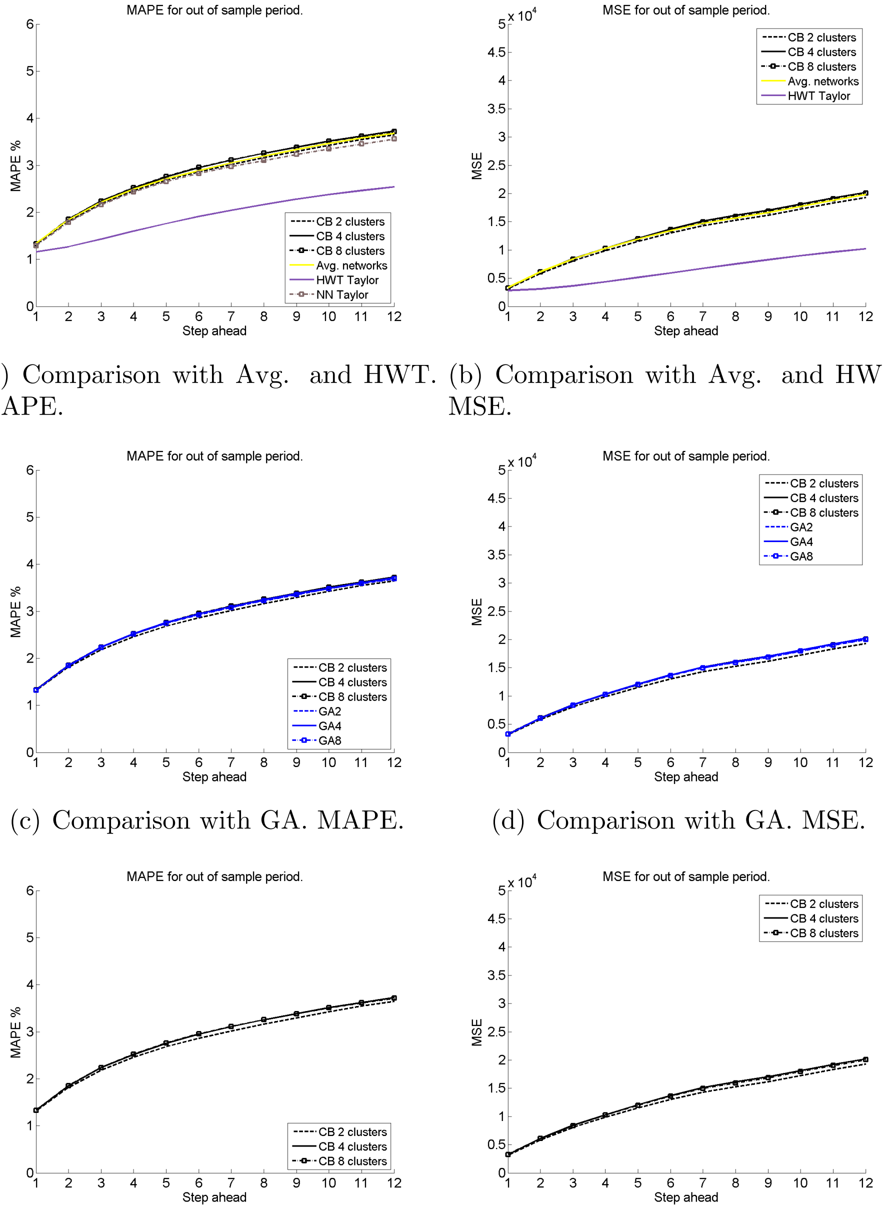
(c) $MaxC = 4$ clusters. MAPE.

(d) $MaxC = 4$ clusters. MSE.

(e) $MaxC = 8$ clusters. MAPE.

(f) $MaxC = 8$ clusters. MSE.

Figure C.9: Out-of-sample MAPE and MSE for Rio de Janeiro electricity demand series.



(a) Comparison with Avg. and HWT. MAPE.

(b) Comparison with Avg. and HWT. MSE.

(c) Comparison with GA. MAPE.

(d) Comparison with GA. MSE.

(e) All CB. MAPE.

(f) All CB. MSE.

Figure C.10: Out-of-sample MAPE and MSE for Rio de Janeiro electricity demand series.