Forecasting Monthly Sales Retail Time Series: A Case Study

Giuseppe Nunnari
Dipartimento di Ingegneria Elettrica Elettronica e Informatica
Universitá degli Studi di Catania
Viale A. Doria, 6, 95125 Catania (Italy)
giuseppe.nunnari@dieei.unict.it

Valeria Nunnari
Zambon Company
Via Lillo del Duca, 10
20091 Bresso (Italy)
valeria.nunnari@zambongroup.com

***Abstract*—This paper presents a case study concerning the forecasting of monthly retail time series recorded by the US Census Bureau from 1992 to 2016. The modeling problem is tackled in two steps. First, original time series are de-trended by using a moving windows averaging approach. Subsequently, the residual time series are modeled by Non-linear Auto-Regressive (NAR) models, by using both Neuro-Fuzzy and Feed-Forward Neural Networks approaches. The goodness of the forecasting models, is objectively assessed by calculating the bias, the mae and the rmse errors. Finally, the model skill index is calculated considering the traditional persistent model as reference. Results show that there is a convenience in using the proposed approaches, compared to the reference one.**

I. Introduction

Sales forecasting is of paramount importance for companies and for this reason, many of them devote significant human and financial resources to reliably perform this task. An effective sales forecasting model helps the supply chain management process and thus increasing profits. On the contrary, inaccurate sales forecasting may cause product backlog, inventory shortages, and unsatisfied customer demands. Therefore, it is important to develop effective sales forecasting models in order to generate accurate and robust forecasting results. Recent work on this field was carried out by [1] who proposed applying computational intelligence methods for predicting the sales of newly published books in a real editorial business management environment. A clustering-based sales forecasting scheme by using extreme learning machine and ensemble linkage methods has been proposed by [2]. A hybrid seasonal autoregressive integrated moving average and quantile regression approach for daily food sales forecasting was proposed by [3]. Forecasting German car sales using Google data and multivariate models was proposed by [4]. A simulation study concerning demand forecasting and inventory control of automotive spare parts was described by [5]. An interesting study concerning the electricity price forecasting, with a comprehensive review of the state-of-the-art in this field was given by [6]. The value of competitive information in forecasting Fast-Moving Consumer Goods (FMCG) retail product sales and the variable selection problem was proposed by [7]. A Support Vector Regression approach for newspaper and magazine sales forecasting was proposed by [8]. The problem of multi-step sales forecasting in automotive industry based on structural relationship identification was described by [9]. Sales forecasting for printed circuit board based on integration of genetic fuzzy systems and data clustering was described by [10]. A case study concerning forecasting of U.S. car sales and car registrations in Japan was described by [11]. An adaptive network-based fuzzy inference system to forecast automobile sales was proposed by[12]. A hybrid intelligent model for medium-term sales forecasting in fashion retail supply chains using extreme learning machine and harmony search algorithm was proposed by [13].

In this paper we describe a case study concerning the implementation of forecasting models for a set of monthly retail time series recorded by the Census Bureau of US from 1992 to 2016, which can be freely downloaded from **http://www.census.gov/retail/**. One of the peculiarity of this data set is that it refers to the whole US market since 1992. Furthermore, since at the present each time series consists of about 300 hundred monthly samples it is possible the use of automatic learning algorithms for the implementation of Non-linear Auto-Regressive (NAR) forecasting models.

TABLE I
Monthly Retail and Food Services Sales by Kind of Business recorded by the Census Bureau of US considered in this paper.

id	NAICS Code	Kind of Business
1	44111	New Cars Dealers
2	44112	Used Cars Dealers
3	443	Electronics and appliance stores
4	44312	Computer and software stores
5	444	Building mat. and garden equip. and supplies
6	445	Food and beverage stores
7	4453	Beer, wine, and liquor stores
8	446	Health and personal care stores
9	44611	Pharmacies and drug stores
10	447	Gasoline stations
11	4481	Clothing stores
12	451	Sporting goods, hobby, book, and music stores

II. Analysis of retail sales time series

Data recorded by the Census Bureau of US refer to about one hundreds of different set of goods, but in this study we selected 12 time series, as listed in Table (I). The time series include both essential and non-essential goods. Some of considered monthly retail time series are shown in Figure 1. A

2378-1971/17 $31.00 © 2017 IEEE
DOI 10.1109/CBI.2017.57

IEEE computer society

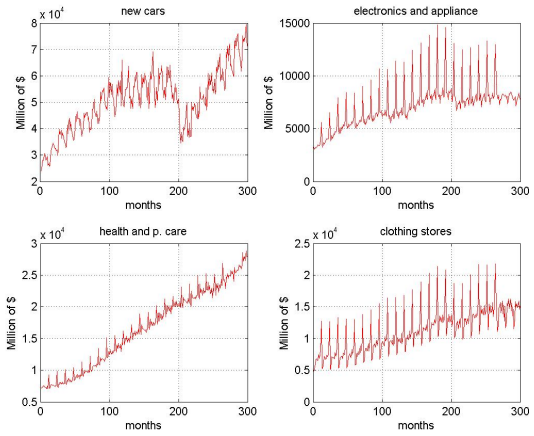

Fig. 1. Monthly sales of new cars, electronic and appliance, health and personal care and clothing stores in US, from 1992 to 2016.

pre-processing step was performed in order to handle outliers and missing data. Despite the wide variety of retail time series (e.g., sales of new/used cars, computers and software, food and beverage, health and personal care etc.) they all share some features. Indeed, they are affected by a trend (usually increasing) and exhibits a marked annual frequency component. All trends of the considered time series, obtained by averaging on a moving windows of 12 months, are shown in Figure 2. It can also be observed that the natural trend of particular time series, is occasionally interrupted by particular economic episodes. For example a fairly evident change of slope can be observed in retail time series such as new cars, gasoline and buildings materials, caused by the US subprime banking mortgage crisis, which contributed to the U.S. recession between December 2007 and June 2009. The presence of an annual component

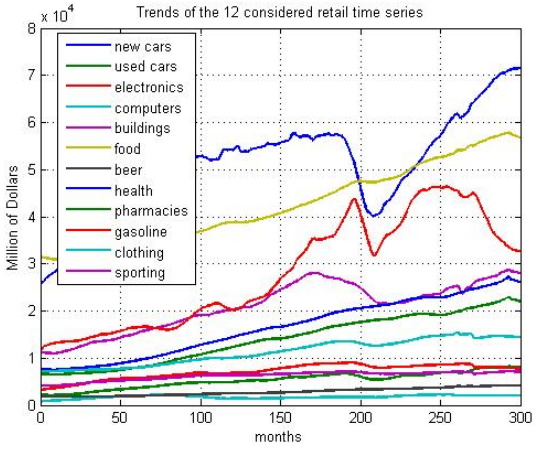

Fig. 2. Trends of the considered time series.

in the considered time series can be appreciated both visually, looking at the oscillations around the trend (see Figure 1), or, more precisely, by performing the spectral analysis. For instance, the left and right sub-figure of the upper Figure 3, show the spectrum density of the original and detrended sales retail time series of new cars, respectively. Indeed, the highest peak in the upper-right Figure 3 corresponds to the frequency of $0.0823 \cdot 10^{-}1\ month^{-1} \approx 1\ year^{-1}$. The others lower peaks are multiple of this fundamental component, which is a well known effect of the Discrete Fourier Transform (DFT) algorithm used for computing the spectrum.

Samples of the original retail time series, are well autocor-

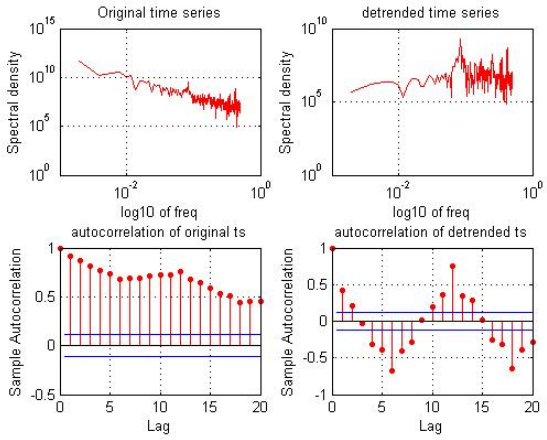

Fig. 3. Power spectral density of the original and detrended new cars sales time series (up), Autocorrelation of the original and detrended new cars sales time series (down).

related (see the left-bottom Figure 3), essentially due to the presence of the trends, while detrended time series are much less autocorrelated at short term, since the autocorrelation falls from 1 to about 0.2 after 2 lag. The scarce autocorrelation at short of detrended time series is confirmed by computing the mutual information. Indeed, since the autocorrelation is a linear statistic and does not account for nonlinear correlations, it may be helpful to compute also the so-called mutual information, defined as in (1)

$$I = -\sum_{i,j} p_{ij}(k) ln \frac{p_{ij}(k)}{p_i p_j} \tag{1}$$

where for some partition of the time series range, p_i is the probability to find a time series values in the i_{th} interval and p_{ij} is the joint probability that an observation falls in the i_{th} interval and the observation time k later falls into the j_{th} interval. As an example, the mutual information computed for the detrended sales of new cars time series in US is shown in Figure 4. This figure shows that the mutual information falls from level 1 to level 0.2 just after the first lag, thereby meaning that forecasting of the level of sales for the next month using only past samples is rather difficult by using autoregressive models.

A. Removing the trend

It is trivial to observe that while the trend of a time series gives information concerning the long-term, the fluctuations

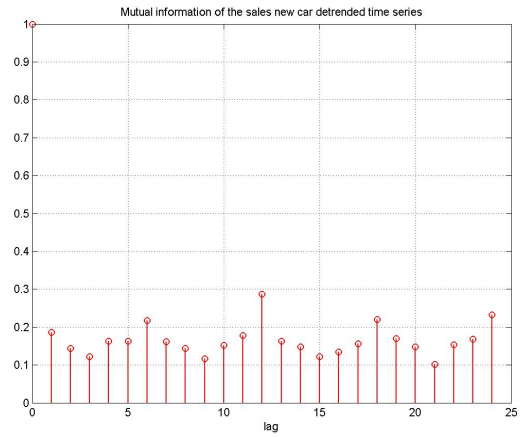

Fig. 4. Sample Autocorrelation of detrended sales of new cars time series in US.

around the trend gives information at short-term. Therefore in order to implement short-term forecasting models (a few months ahead) it is a common practice to model de-trended time series. Removing the trend of a given time series can be

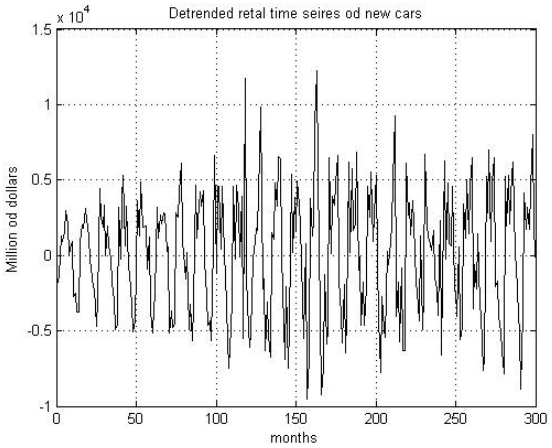

Fig. 5. Detrended sales of new cars time series in US.

performed by using various approaches. A rough approach is that of computing the best straight-line or a continuous, piece-wise linear trend, after recognizing appropriate breakpoints. A more efficient approach is that of performing the convolution between the original time series and an appropriate symmetric moving function. The application of this latter approach to de-trend the rime series of new cars, gives the result shown in Figure 5.

B. The Hurst exponent of detrended time series

A useful feature of fractal time series consists in evaluating the Hurst exponent, by using the so-called R/S analysis, originally developed by the hydrologist H.E. Hurst. It can be useful to remind here that the Hurst exponent represents a measure of long-term memory of time series. It relates to the autocorrelations of the time series, and the rate at which this decreases as the lag between pairs of values increases. A value H in the range $0.5 < H < 1$ indicates a time series with long-term positive autocorrelation, meaning that a high value in the series will probably be followed by another high value and that the values a long time into the future will also tend to be high. A value in the range $0 < H < 0.5$ indicates a time series with long-term switching between high and low values in adjacent pairs, meaning that a single high value will probably be followed by a low value and that the value after will tend to be high, with this tendency to switch between high and low values lasting a long time into the future. Finally, a value of $H = 0.5$ indicates a completely uncorrelated series. The Hurst exponent computed for the set of 12 detrended retail time series is shown in Figure (6). It is possible to see that

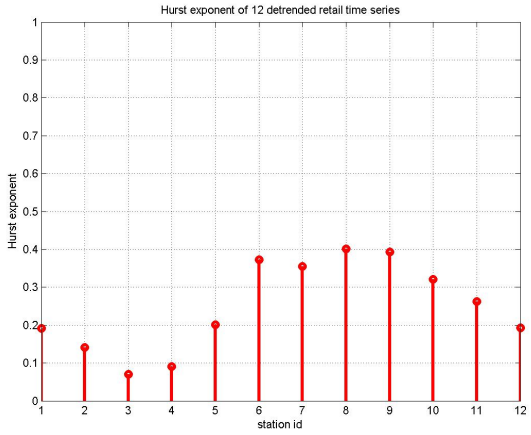

Fig. 6. Hurst exponent of the 12 detrended sales time series. The id number is that in Table I.

all de-trended time series exhibit values of the Hurst exponent in the range $0 < H < 0.5$ thus objectively assessing their switching behavior around the trend.

III. Modeling detrended retail time series

Let us indicate as $y(t)$ a generic time series. According with the Takens theorem [14], a one-step ahead forecasting model can be based on the existence of a diffeomorphism, i.e. a one-to-one differential mapping, between a finite window of the time series $([y(t), y(t-1), ..., y(t-d+1)]$ and the state of the dynamic system underlying the series. This implies that there exist a MISO (Multi-Input Single-Output) mapping $f : R^d \rightarrow R$ such that

$$y(t+1) = f(y(t), y(t-1), ..., y(t-d+1)) \qquad (2)$$

where d (dimension) is the number of considered past values. This formulation returns a state space description where, in the d dimensional state space, the time series evolution is a trajectory and each point represents a temporal pattern of length d. Such a kind of forecasting models are referred to as NAR (Nonlinear Auto-Regressive).

A. Multi-step ahead forecasting models

The problem of multi-step ahead forecasting is that of estimating $y(t+h)$, using information measured until time t. Here, h is a positive integer, referred to as the forecasting horizon. This problem can be solved in two ways: one-step ahead forecasting and iterated forecasting. In the former case, expression (2) is rewritten as (3)

$$y(t+h) = f(y(t), ..., y(t-d+1)) \tag{3}$$

where it is assumed that the samples of the time series $y(t)$ are known until time t and the problem is equivalent to a function estimation.
In the latter case the problem of estimating $y(t+h)$ is solved by iteratively using expression (2). At each step the predicted output is feedback as an argument of the f function. Hence, the f arguments consist of predicted values as opposed to actual observations of the original time series. A forecasting iterated for h times returns a *h*-step-ahead forecasting. Despite the popularity that this approach gained in the forecasting community, its design is still affected by a number of important unresolved issues, the most important being the accumulation of forecasting errors [17]. For this reason in this paper we consider the multi-step ahead forecasting problem directly approximating the map (3).

B. Mapping approximation

Regardless of which scheme will be used to perform a multi-step ahead forecasting, in order to identify a model is required to approximating the unknown map f. To this purpose, Neural Networks based approaches are among the most popular and efficient tools. In this paper, two of these kinds of approaches will be considered, namely the Neuro-Fuzzy (NF) and the Feedforward Neural Networks (NN) approaches, respectively. One of the main advantages of these approaches is that they allow to approximate nonlinear maps by various kinds of basis function, such as, sigmoidal, Gaussian, wavelet and so on. Furthermore, a wide of learning algorithms are available to automatically identify the model parameters.

C. The Neuro Fuzzy approach

One of the most interesting aspects of the NeuroFuzzy approach is that once the neural network has been trained by using automatic learning algorithms, the obtained model can be interpreted in terms of a base of *if ... then* rules. The resulting models can be represented both in linguistic form, or as multidimensional surfaces, whose coordinates are the arguments of the f function. In particular, if the rules are expressed in the so-called Takagi-Sugeno form, i.e with the consequent part expressed as a linear combination of the input mapping.

D. The FeedForward Neural Network Approach

FeedForward networks consist of a number of simple artificial neurons, organized in at least three layers. The first layer has a connection from the network input. Each subsequent layer has a connection from the previous layer. The final layer produces the network's output. Such a kind of networks can be used for many kinds of input/output mapping. A celebrated theorem [18] guarantees that a feedforward network with at least one hidden layer and enough neurons can fit any finite input-output mapping problem, provided that f is continuous. Several training algorithms can be used to learn the network, such as for instance the Backpropagation or the Levenberg-Marquardt optimization algorithm.

E. Assessing the model performances

In order to objectively evaluate the performance of forecasting models several indices can be computed. In this paper three error indices will be considered, namely the *bias*, the *mae* (mean absolute error) and the *rmse* (root square mean error), which are usually considered in literature. Such indices are defined as expressed in (4), (5) and (6), respectively.

$$bias = \frac{1}{n}\sum_{i=1}^{n}(y(i) - \hat{y}(i)) \tag{4}$$

$$mae = \frac{1}{n}\sum_{i=1}^{n}|y(i) - \hat{y}(i)| \tag{5}$$

$$rmse = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y(i) - \hat{y}(i))^2} \tag{6}$$

where n is the number of samples and the symbol $\hat{y}$ indicates the estimated sample.
It is a common practice in literature to inter-compare a forecasting model with another model assumed as reference. To this purpose, the most popular and simplest is the persistent model, formally represented by expression (7).

$$\hat{y}(t+h) = y(t) \tag{7}$$

Such a comparison, usually consists in computing the following skill index,

$$S = 1 - \frac{RMSE_{NAR}}{RMSE_{Pers}} \tag{8}$$

being $rmse_{NAR}$ and $rmse_{Pers}$ the RMSE of the NAR and persistent models, respectively. It is trivial to observe that S close to 1 are best.

IV. Numerical results

In this section performances of forecasting models of the form (3), identified by using both the NF and NN mapping approaches will be reported. For simplicity, results will be described by considering a particular time series, namely the sales of new cars. Afterwards, results obtained for the others considered retail time series will be summarized. As the autocorrelation of sales residual time series exhibits a significant periodic component with period 12 months, we set the d parameter in expression (3) to $d = 12$. To the purpose of model identification, for each time series, the data

set was divided into two adjacent sets, referred to as the training and the testing set, respectively. The former data set containing $2/3$ of the overall available samples was considered for identification of the model parameters, while the latter was exclusively considered for evaluation the performance indices. The forecasting horizon was assumed in the range $h \in [1, 12]$ months. Model identification was performed after scaling the residual time series in the range $[-1, 1]$ in order to allowing a direct comparison among performance indices obtained for different time series.

As an example, performance indices of NF, NN and persistent models for the new car scaled time series are reported in Figure (7). As expected, due to the strong annual component affecting

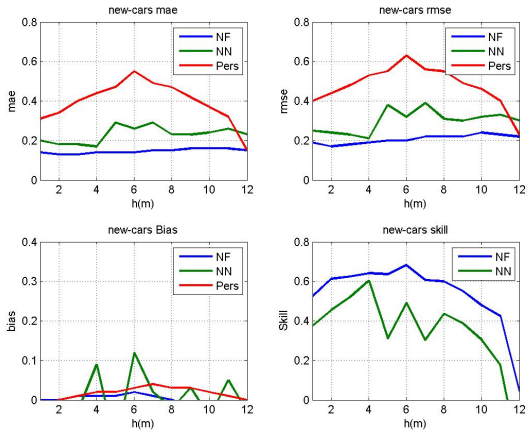

Fig. 7. mae, rmse, bias and skill indices, versus the forecasting horizon, for the scaled time series of new cars.

the de-trended time series, the mae and RMSE of the persistent model grows with the forecasting horizon until $h = 6$ and then decrease approaching h to 12. Instead, the NAR models, mainly those of NF type, exhibit a modest growth, compared to the persistent model, versus increasing horizon. The bias index is very close to zero for all considered models, thereby indicating that the estimates are not biased. Finally, the skill index indicates that for the series of new cars, the NF and NN forecasting models outperform the persistent model for a wide range of forecasting horizon. In particular, as reported in Table II and Table III, for $h = 1$ we get 0.52 and 0.40 for NF and NN models, respectively. Thus, for the series of new cars, the NF model perform slightly better than the NN ones. As concerning the others time series listed in Table I, due to the lack of room, we report the mae, rmse and skill for NF models only, in Figure 8, 9, and 10, respectively. In summary it is possible to see that $mae \leq 0.22$, $rmse \leq 0.36$ and $skill \geq 0.2$ for all considered time series and for a wide range of the forecasting horizon ($h \leq 10$). A more detailed analysis shows that among the 12 retail time series considered the best modeled ones are sporting goods, electronics and appliance, clothing, beer, computers, food and new cars, which exhibits for $h = 1$ $mae \leq 0.15$, $rmse \leq 0.22$ and $skill \geq 0.5$.

TABLE II
PERFORMANCE INDICES FOR NAR FN MODELS, $h = 1$

t. series	bias NF	mae NF	rmse NF	skill NF
new cars	0.00	0.14	0.19	0.52
used cars	0.02	0.19	0.28	0.36
electronics	0.00	0.05	0.09	0.64
computers	0.00	0.07	0.12	0.60
buildings	0.00	0.14	0.22	0.37
food	0.03	0.15	0.22	0.56
beer	0.01	0.09	0.16	0.60
health	0.07	0.19	0.36	0.21
pharmacies	0.06	0.18	0.32	0.24
gasoline	0.01	0.11	0.14	0.42
clothing	0.00	0.07	0.13	0.64
sporting	0.00	0.05	0.11	0.65

TABLE III
PERFORMANCE INDICES FOR NAR NN MODELS, $h = 1$

t. series	bias NN	mae NN	rmse NN	skill NN
new cars	-0.03	0.18	0.24	0.40
used cars	0.04	0.37	0.49	0.11
electronics	0.01	0.05	0.09	0.64
computers	-0.05	0.12	0.15	0.50
buildings	-0.03	0.13	0.19	0.46
food	-0.02	0.20	0.28	0.44
beer	0.01	0.11	0.17	0.57
health	0.05	0.19	0.29	0.37
pharmacies	0.01	0.16	0.24	0.43
gasoline	-0.03	0.15	0.18	0.25
clothing	0.01	0.07	0.12	0.67
sporting	0.00	0.06	0.10	0.68

V. CONCLUSIONS

In this paper we reported results of a case study concerning the modeling and forecasting of retail sales time series by using two neural network based approaches. Performance, assessed in terms of mae, rmse, bias and skill, show that, despite the scarce auto-correlation of the residual time series at short term, there is a benefit in using these kinds of models, with respect to the simple persistent model, for a large range

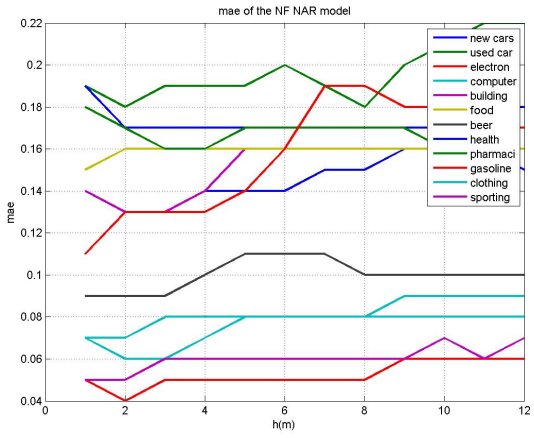

Fig. 8. mae indices, versus the forecasting horizon, for all considered scaled time series and NF kinds models.

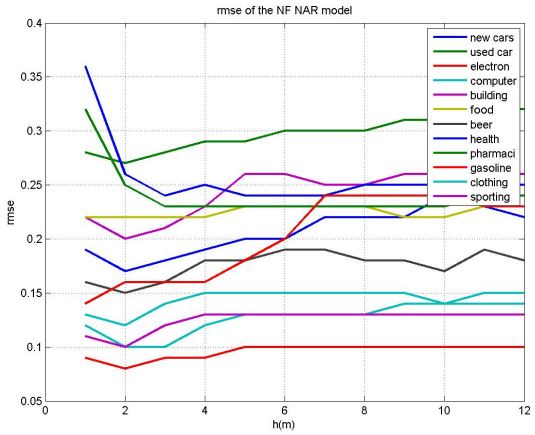

Fig. 9. rmse indices, versus the forecasting horizon, for all considered scaled time series and NF kinds models.

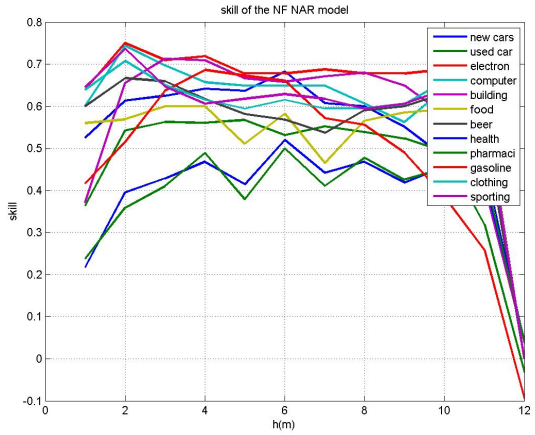

Fig. 10. skill indices, versus the forecasting horizon, for all considered scaled time series and NF kinds models.

of forecasting horizons. In particular, we found that setting the order of the NAR models to $d = 12$, is a good choice for this kinds of forecasting models. Furthermore, numerical results obtained so far, show that between the two neural networks approaches, the NF slightly outperform the NN models for the considered task.

ACKNOWLEDGMENT

This work was supported by the Universitá degli Studi di Catania under the grant FIR 2014.

REFERENCES

[1] P. A. Castillo, A. Mora, H. Faris, J.J. Merelo, P. Garcia-Sanchez,A. J. Fernandez-Ares, P. De las Cuevas, M.I. Garcia-Arenas, Applying computational intelligence methods for predicting the sales of newly published books in a real editorial business management environment, Knowledge-Based Systems 115 (2017) 133-151.

[2] C. Lu and L. Kao, A Clustering-based sales forecasting scheme by using extreme learning machine and ensambling linkage methods with application to computer server, Engineering Applications of Artificial Intelligence 55 (2016) 231–238 .

[3] N. S. Arunraj, D. Ahrens, A hybrid seasonal autoregressive integrated moving average and quantile regression for daily food sales forecasting, Int. J. Production Economics 170 (2015) 321-335.

[4] D. Fantazzini, Z. Toktamysova, Forecasting German car sales using Google data and multivariate models, Int. J. Production Economics 170 (2015) 97-135.

[5] J.R. do Rego, M. A. de Mesquita, Demand forecasting and inventory control: A simulation study on automotive spare parts, Int. J. Production Economics 161 (2015) 1–16.

[6] R. Weron, Electricity price forecasting: A review of the state-of-the-art with a look into the future, International Journal of Forecasting 30 (2014) 1030–1081.

[7] T. Huang, R. Fildes, D. Soopramanien, The value of competitive information in forecasting FMCG retail product sales and the variable selection problem, European Journal of Operational Research 237 (2014) 738–748.

[8] X. Yua, Z. Qi,Y. Zhao, Support Vector Regression for Newspaper/Magazine Sales Forecasting, Procedia Computer Science 17 (2013) 1055–1062.

[9] A. Sa-ngasoongsong, S.T.S. Bukkapatnam, J. Kim, P. S. Iyer, R.P. Suresh, Multi-step sales forecasting in automotive industry based on structural relationship identification, Int. J. Production Economics 140 (2012) 875–887.

[10] E. Hadavandi,H. Shavandi, A. Ghanbari, An improved sales forecasting approach by the integration of genetic fuzzy systems and data clustering: Case study of printed circuit board, Expert Systems with Applications 38 (2011) 9392–9399.

[11] C. Pierdzioch, J.C. Rulke, G. Stadtmann, Forecasting U.S. car sales and car registrations in Japan: Rationality, accuracy and herding, Japan and the World Economy 23 (2011) 253–258.

[12] F. Wang,K. Chang, C. Tzeng, Using adaptive network-based fuzzy inference system to forecast automobile sales, Expert Systems with Applications 38 (2011) 10587–10593.

[13] W.K. Wongn, Z.X. Guo, A hybrid intelligent model for medium-term sales forecasting in fashion retail supply chains using extreme learning machine and harmony search algorithm, Int. J. Production Economics 128 (2010) 614–624.

[14] F. Takens, Detecting strange attractors in turbulence. In D. A. Rand and L.-S. Young, Dynamical Systems and Turbulence, Lecture Notes in Mathematics, vol. 898. Springer-Verlag., 1981, pp. 366-381.

[15] L. Ljung, System Identification - Theory for the User, 2nd Edition, Prentice-Hall, Upper Saddle River, N J, 1999 ISBN 0-13-656695-2, 607 pages.

[16] S. A. Billings, Nonlinear System Identification: NARMAX Methods in the Time, Frequency, and Spatio-Temporal Domains, Wiley, ISBN 978-1-1199-4359-4, 2013.

[17] Q. Zhang, L. Ljung, Multiple steps prediction with nonlinear arx models, In Proc. NOLCOS 2004 - IFAC Symposium on Nonlinear Control Systems, Stuttgart, Germany, September 2004.

[18] G. Cybenko, Approximations by superpositions of sigmoidal functions, Mathematics of Control, Signals, and Systems, 2 (4), 303–314, 1989.