Regional and Business Studies (2022) Vol 14 No 2, 5-20
Hungarian University of Agriculture and Life Sciences Kaposvár Campus
Kaposvár
doi: 10.33568/rbs.4434

AN OVERVIEW OF PERFORMANCE MEASUREMENT FOR DEMAND FORECASTING BASED ON ARTIFICIAL NEURAL NETWORKS

Steffen ROBUS[1], Zsolt KŐMÜVES[2], Virág WALTER[2]
[1] Hungarian University of Agriculture and Life Sciences, Doctoral School of Management and Organizational Science, 7400 Kaposvár, Guba Sándor u. 40., Hungary
[2] Hungarian University of Agriculture and Life Sciences, Institute of Agricultural and Food Economics, 7400 Kaposvár, Guba Sándor u. 40., Hungary

ABSTRACT

Demand forecasting is an essential task to match supply and demand. From a supplier's view, demand forecasting is important to optimize supply chains and thus maximize profits. The ever-increasing availability of data that can be used as input factors for predictive models allows more and more sophistication for diverse forecasting tasks in the context of demand forecasting. On the one hand, increasingly complex models have been used for demand forecasting over the last years, from simple exponential smoothing methods and ARIMA models up to complex, hybrid (deep) artificial neural networks. On the other hand, little attention is paid to the methods that evaluate the forecasting performance of these models, which are essential for the selection from among potential forecasting models. In this article, we aim to answer the question of what are the most favourable measurements in recent literature on applied neural network demand forecasting for supply chain management. To this end, we analyzed 193 relevant publications in which demand forecasting was applied using artificial neural networks. We found that in artificial neural network demand forecasting used to evaluate forecasting performance, Mean Absolute Percentage Error, Root Mean Squared Error, Mean Squared Error and Mean Absolute Error are by far the most popular methods. Furthermore, we found that when forecasting performance measurements are combined, the most common combination is the combination of Mean Absolute Error, the Root Mean Squared Error and the Mean Absolute Error.
Keywords: artificial neural networks, decision support systems, performance measurements, supply chain management

INTRODUCTION

From a higher level, demand forecasting is an essential task to match supply and demand. From a supplier's view, demand forecasting is key to optimising supply chains and thus maximising profits. The planning of the expected demand is the first step in putting together a business model and consequently, the basis of all planning activities (*Haberleitner et al.*, 2010). The increase in the global competition meant that, on the one hand, storage costs were cut while the availability of products improved. This requires a high level of forecasting of the expected demand (*Lewis*, 1997;

Carbonneau et al., 2008). In the retail sector, demand forecasting is used to estimate which products the seller will provide and in what quantity (e.g. *Fildes et al.*, 2022). If the forecast is too low, the customers who cannot be served will change vendors and possibly the image will be damaged. If the forecast exceeds demand, the products will remain in stock and spoil, or their storage will incur storage costs. In both cases, there is financial damage. After all, in the retail sector, a time gap between supply and demand can be compensated for by proper warehousing. This is not the case in other sectors. In the case of electricity, an existing demand must be covered by an offer from the energy supplier. It goes without saying that the potential damage from large-scale power blackouts is significant (*Suganthi & Samuel*, 2012; *Ghalehkhondabi et al.*, 2017). In the tourism sector, a good demand forecast is also essential. Based on the expectations in the sector of tourism, logistics capacities in the form of flights, hotel capacities, but also capacities for staff and food are provided in advance. If the forecast is above the demand, inefficiencies arise because these capacities are not called up and cause financial damage on the one hand, and on the other are not available elsewhere. If the forecast is too low and the capacities provided based on it is also too low, customers will be dissatisfied because the quantity or quality of service is insufficient. This causes substantial damage to the respective holiday region (*Burger et al.*, 2001; *Andrawis et al.*, 2011). These three examples show that an accurate forecast of the demand is very important. The ever-increasing availability of data that can be used as input factors for prediction allows more sophisticated models to be constructed and operated in practice. This means that increasingly complex models have been used for demand forecasting over the past ten years, from simple exponential smoothing methods, and ARIMA models to complex hybrid (deep) artificial neural networks (*Schmidhuber*, 2015). Forecasters must always ask themselves which forecasting model is the best for their respective area of application. This raises the question of how the accuracy of a forecasting model is measurement and how different forecasting models can be compared. Forecasting performance measurements or forecasting accuracy measurements quantify the accuracy of forecasts. Dealing with these methods and selecting the appropriate forecasting performance measurement for the respective application is very important, as it has a direct influence on the choice of the forecasting model and consequently, on the result that is to be achieved with forecasting (*Makridakis*, 1993). Given the importance of forecasting performance measurements, this article aims to answer the question of which are the most favored forecasting performance measurements are in the recent literature on artificial neural network demand forecasting. For this purpose, we conducted an analysis of relevant research in which demand forecasting was carried out using artificial neural network methods. In this article, we want to answer the following research questions.

Research question 1: What are the most frequent forecasting performance measurements used for applied artificial neural network demand forecasting?

Research question 2: What are the most frequent combinations of forecasting performance measurements?

We performed a Google Scholar search and analyzed 193 papers in the context of applied artificial neural network demand forecasting in detail for their used methods,

their area of application and their forecasting performance criteria. With this dedicated study, we can provide a complete overview of the forecasting performance measurements used in demand forecasting over the last ten years. As far as we know, there is no systematic treatment of the forecasting performance measurements used in the field of demand forecasting using methods of artificial intelligence. We want to help the users of forecasting methods to critically deal with the properties of the forecasting performance measurements and use them consciously and specifically in the context of a given forecasting framework. We explain the advantages and disadvantages of the most frequently used forecasting performance measurements and which alternative methods are the more suitable. Our article is organized as follows. We provide an overview of the most important research on demand forecasting using artificial neural networks and forecasting performance measuring. Then we describe the methodology we used to conduct our systematic literature review. We then discuss the results in detail. Finally, we summarize the most important findings and give an outlook on possible further research.

THEORETICAL BACKGROUND

The literature, including *Armstrong* (2001) and *Thonemann* (2010), distinguishes between three basic approaches to predicting customer demand: qualitative forecasts, causal forecasts, and time series forecasts. In this paper, we focus on research using time series demand forecasting. Time Series Forecasting is based on historical data using time series analysis methods. A connection between the past and the future of a variable is assumed. Applied to the prediction of customer demand, this approach draws a conclusion on future demand based on historical demand. The methods for forecasting time series and their fields of application are quite numerous. A good overview can be found in *De Gooijer & Hyndman* (2006). We first provide a brief overview of forecasting time series using artificial neural networks and then briefly review the literature related to forecasting performance measurements.

Artificial Neural Networks for Demand Forecasting

Artificial Neural Networks (ANNs) have become a very popular approach for modelling and predicting time series. ANNs are based on the idea that by networking many individual calculations, the functioning of the human brain can be simulated and thus the ability to solve a variety of (non-linear) problems is created. The versatile applicability is exactly the reason why ANNs became so popular. *McCulloch & Pitts* (1943) were the first to model the artificial neuron. Combinations of these neurons then form an artificial neural network. With the perceptron, which consists of a single artificial neuron with adjustable weights and a threshold value, *Rosenblatt* (1958) published the simplest form of an artificial neural network. The building blocks of an ANN are the artificial neurons, which are arranged in different layers. From the input layer, information enters the network in the hidden layers, which process the information, and finally in the output layer, which gives the information out. Since the first neural networks, research on artificial neural networks has progressed steadily and has been applied to a wide range of problems. A good overview of the

development of ANNs can be found in *Schmidhuber* (2015) and *Goodfellow et al.* (2018). The neural networks can be roughly divided into the following categories: shallow neural networks (e.g. *Aggarwal*, 2018), multilayer perceptrons, also called deep neural networks (e.g. *Schmidhuber*, 2015), convolutional neural networks (e.g. *Gu et al.*, 2018), recurrent neural networks (e.g. *Salehinejad*, 2017), long-short term memory neural networks (e.g. *Yu et al.,* 2020), attention based neural networks (e.g. *Wang et al.*, 2016) and generative adversarial network (e.g. *Creswell et al.*, 2018). Numerous applications of artificial neural network demand forecasting can be found in the literature. *Ryu et al.* (2016) used deep learning to forecast the short term electricity demand. *Constantino et al.* (2016) forecast tourism demand by ANNs. *Wanchoo* (2019) proposed a deep learning model to forecast retail demand. Babai (2014) used an ANN to forecast intermittent demand. *Panapakidis & Dagoumas* (2017) forecast natural gas demand one day in advance using a hybrid artificial neural network. *Ke et al.* (2017) have examined the demand for on-demand ride services, and *Kilimci et al.* (2019) used a deep learning model to forecast demand in a supply chain framework.

Forecasting Performance Measuring

As there are a lot of different approaches to forecast demand, from the point of view of a decision-maker, the question now arises as to which of the different forecasting models is to be preferred. To assess this, *Granger & Pesaran* (2000a) remarked that a decision-maker needs one or more criteria to compare the performance of given forecast models. *Granger & Pesaran* (2000a) and *Granger & Pesaran* (2000b) noted that it is crucial for decision-making based on forecasts, that the forecasts are linked to a cost- or loss function, which quantifies the forecasting error in terms of the specific forecasting problem. Therefore, a fundamental problem for the decision-maker is the selection of a suitable forecasting criterion, to measurement the accuracy of forecasting or a loss function which quantifies the forecasting error. *Makridakis* (1993) wrote that from an ex-ante perspective, the decision-maker cannot judge which forecasting model is the best model, since a sample of forecasts must be made to assess the forecasting performance using a forecasting performance measurement. This means that the forecasting performance measurement that best provides information about future forecasting performance should be used. From an ex-post situation, an evaluation of different forecasting models is possible by forecasting performance measurements. It is important to realize that future forecasting performance would be influenced by the choice of a forecasting performance measurement. This happens insofar as the forecasting performance measurement is used to decide regarding the forecasting method used in the future, based on the perceived performance today. This means, that when deciding between two forecasting models, one forecasting measurement can prefer one model and another measurement can prefer another one, *Makridakis* (1993) remarked. Another dependency is the scope of the forecasting horizon. So, the outcome of a forecasting performance measurement can change with different forecasting horizons. This shows that a consistent selection of forecasting model is not easily possible, as *Clements & Hendry* (1993) showed this in the example of the mean square forecast error (MSE). To overcome this problem, *Diebold & López* (1996) formulated properties that optimal forecasts should possess. The following properties

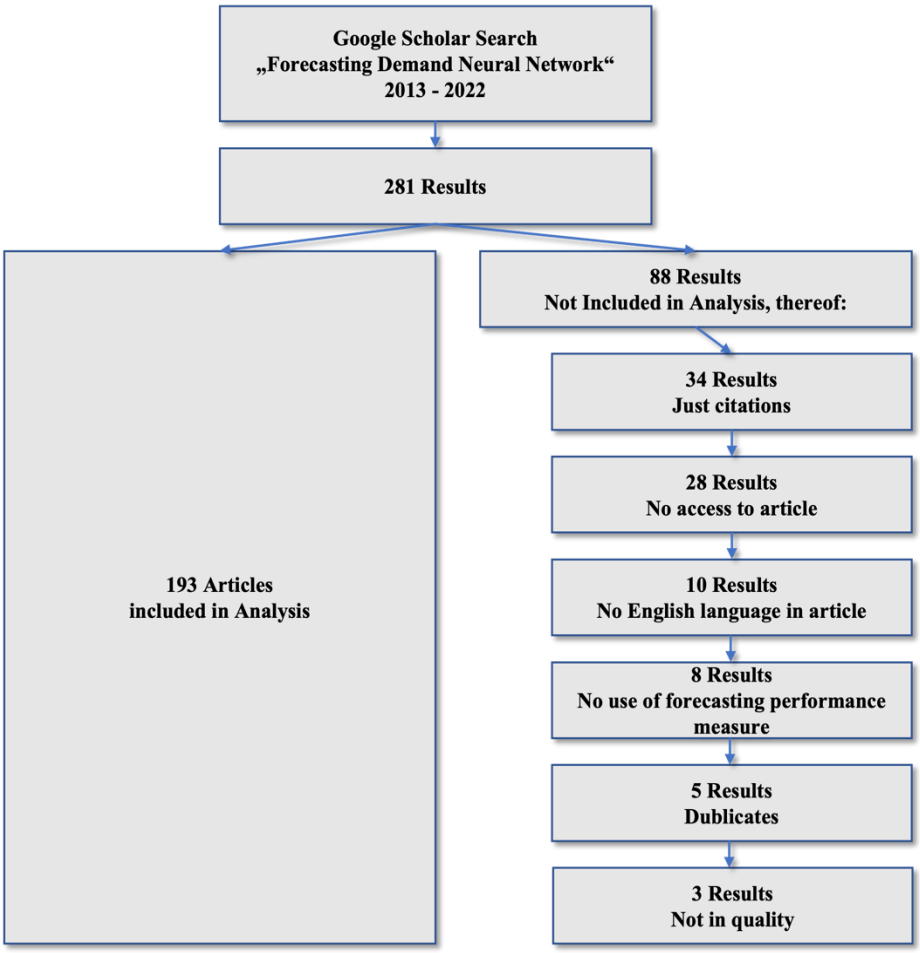
should therefore apply to the forecast errors of an optimal k-step ahead point forecast of a linear forecast model: the forecast errors have a zero mean, the 1-step ahead forecast errors following a white-noise process, the k-step ahead forecast errors following a MA(k-1) moving average process and the k-step ahead forecast error variance is not decreasing in k. So, these properties can be tested statistically. Moreover, for the ex-post comparison of two forecasts, *Diebold & Mariano* (2002) showed other approaches which test statistical significance whether one forecast has the same forecasting accuracy as the other one. However, these methods are not found in many application-related articles. Instead, and we will go into this in detail later in the Results, methods section, the Mean Average Percentage Error (MAPE), the Mean Square Error (MSE) or the Root Mean Square Error (RMSE) are used. *Hyndman & Koehler* (2006) note that the MAPE is often recommended for example by *Hanke and Reitsch* (1995); *Bowerman et al.* (2003) and *Makridakis et al.* (1982). There is a wide variety of practical applications. *Petrucci et al.* (2022); *Porteiro et al.* (2020) and *Jnr et al.* (2021) used the MAPE to benchmark models for the forecasting of electricity demand. The RMSE was used by e.g. *Zhang et al.* (2020); *Anisa et al.* (2021) and *Liang* (2022) to compare forecasts of tourism demand. *Slimani et al.* (2015); *Chawla et al.* (2019); *Herrera-Granda et al.* (2019) and *Maragkos* (2020) made their retail forecasting model selection by the MSE.

METHODOLOGY

Our research was conducted as a systematic literature review, which entails a thorough, transparent, and replicable process for literature search and analysis. This choice of method is suitable as the research questions require a quantitative overview of existing usage of methods and areas of application for demand forecasting. We made our search in Google Scholar (http://scholar.google.com). We decided to use We chose Google Scholar for the literature research because it provides a simple search, finds many sources, lists documents soon after they are published and has a good relevance ranking. The search for 'Forecasting' and 'Demand' and 'Neural Network' which are in the title of the paper was conducted from the 28th of January 2023 to the 14th of February 2023. To examine the current literature, we restricted the search to research between 2013 and 2022. With these search strings, the total number of hits was 281 publications. All hits were collated in an Excel spreadsheet as a record of the search. We then carried out a filter with regard to the type of publication, language, quality and accessibility of the publications we found. Furthermore, we only used publications that shed light on a practical forecasting problem in the supply chain context. This procedure is presented in *Figure 1*. In total, from 281 hits, we excluded 88 hits because they were just citations (34 hits), we had no access to the publication (28 hits), the publication was not written in the English language (10 hits), there was no explanation about the usage of a forecasting performance measurement (8 hits), the publication was listed twice (5 hits) or the publication was bad quality (3 hits). To be able to make our evaluations, we recorded the following details of the publication: reference, date, number of citations, the field of application of demand forecasting, forecasting method, back propagation algorithm and forecasting accuracy measurement. For the evaluation of the used

forecasting performance measurements, we analyzed the publications in our database and collected the forecasting performance measurements that were used in the research in a database. We recorded all the procedures mentioned in the publication, and we also allowed multiple entries. Then we calculated the absolute and relative frequencies across all recorded performance measurements. An overview of the various forecasting performance measurements can be found in articles by *Chen & Yang* (2004) and *Koutsandreas et al.* (2022).

Figure 1: Methodology of literature collection



RESULTS

Using a systematic literature review, we investigated the forecasting performance measurements for demand forecasting based on neural network in supply chain frameworks. We found that by far the most frequently used measurements are

MAPE, RMSE, MSE and MAE. Although these metrics have well-known weaknesses, they are widely used in relevant research. Different measurements are often used in combination with research. We have analyzed which combinations of forecasting performance measurements are used and found that the combination of MAPE, RMSE and MAE are used most frequently. In the following, we discuss these two results in detail.

Most favourite forecasting performance measurements for neural network demand forecasting

We analyzed a total of 193 publications and collected 343 database entries of forecasting performance measurements. We collected all measurements which were used. Overall, we collected 29 different forecasting performance measurements (*Table 1*). However, as we will discuss it later, only a small number of these measurements have been widely used.

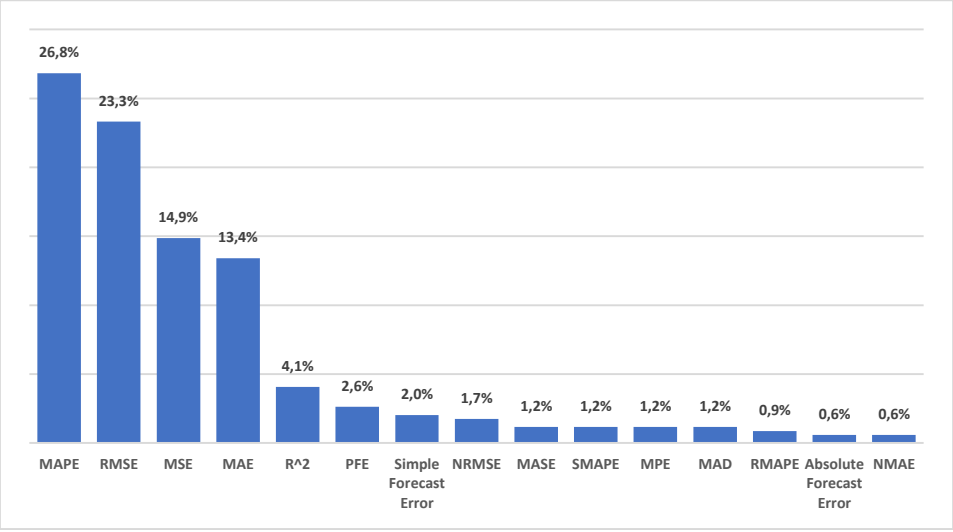
Table 1: Overview of all collected different forecasting performance measurements

Absolute Forecast Error (AFE)	Average Relative Mean Absolute Error	Coefficient of Determination (R^2)
Correlation Coefficient	Diebold-Mariano Test (DM-Test)	Mean Absolute Deviation (MAD)
Mean Absolute Error (MAE)	Mean Absolute Percentage Error (MAPE)	Mean Absolute Relative Error
Mean Absolute Relative Normalized Error (MARNE)	Mean Absolute Scaled Error (MASE)	Mean Negative Error (MNE)
Mean Percentage Error (MPE)	Mean Positive Error (MPE)	Mean Squared Error (MSE)
Nash-Sutcliff-Index	Normalized Mean Absolute Error (NMAE)	Normalized Mean Squared Error (NMSE)
Normalized Root Mean Squared Error (NRMSE)	Pearson Product-Moment Correlation Coefficient (PPMCC)	Percentage Forecast Error (PFE)
Relative Mean Absolute Percentage Error (RMAPE)	Relative Root Mean Squared Error (RRMSE)	Root Mean Squared Error (RMSE)
Root Mean Squared Scaled Error (RMSSE)	Simple Forecast Error	Sum of Squared Errors (SSE)
Symmetric Mean Absolute Percentage Error (SMAPE)	Wilcoxon-Signed-Ranks Test	

On average, every research paper used 1.78 different forecasting performance measurements. This is particularly useful when measurements with different properties are combined and thus allowing the evaluation of a forecasting model under various aspects. *Figure 2* presents the most popular forecasting performance measurements in the collected sample, which were found more than once. It can be seen that four measurements are particularly common. With a total count of 92 (26.8% of overall 343 recorded forecasting performance measurements), we can observe that the Mean Average Percentage Error (MAPE) is the most preferred

measurement. This is probably the case because the MAPE is scale independent and it is easy to interpret and compute, which makes it very popular among practitioners (e.g. *Byrne*, 2012). A significant disadvantage is that only data without zero and extreme values are necessary for the MAPE. If the true value is extremely small or large, MAPE value takes on an extreme value (*Kim & Kim*, 2016). In practice, many datasets contain zeros in the realized values, e.g., in retail, when no transaction takes place. To obtain a usable MAPE value, these observations would have to be removed as outliers from the sample used to calculate the MAPE. A specific suggestion was made by *Makridakis* (1993) who proposed to exclude values with actual values less than one or with an average percentage error values greater than the MAPE plus three standard deviations. However, since these are normal observations, the sample is distorted, and comparison is made more difficult. Another method *Makridakis* (1993) suggests is the replacement if the MAPE by the Symmetric Mean Absolute Percentage Error (SMAPE). In our analysis, however, the SMAPE was only used 4 times (1.2%). *Hyndman & Koehler* (2006) also recommended dispensing with the MAPE and using the Mean Absolute Scaled Error (MASE) instead.

Figure 2: Favourite performance measurements of demand forecasting



Sample size = 343 collected forecast performance measurements

However, the MASE was only used 4 times (1.2%) in the research papers we analyzed. In general, *Swanson et al.* (1999) have noted that measurements based on percentage errors, like the MAPE, are often highly skewed, and therefore transformations such as logarithms can make them more stable. *Clements et al.* (2004) discusses this in more detail. We did not find a measurement based on log transformation in our analysis. The second most common measurement is the Root Mean Square Error (RMSE) with a total count of 80 (23.3%). As mentioned above, the RMSE is defined as the root of the Mean Square Error (MSE), which is also a very

common measurement. Due to the square root function, the RMSE is in the same unit as the forecast and true values and is, therefore, easier to interpret than the MSE. Its popularity in our research is also consistent with former research, such as *Carbone and Armstrong* (1982) who found that RMSE is preferred by practitioners. On the other hand, *Armstrong and Collopy* (1992) found that the RMSE is not reliable in terms of the repeated application of a method that produces the same results. By using the Spearman rank-order correlation, they showed that the RMSE is not consistent in producing accurate rankings of out-of-sample forecasts of different time series extrapolation methods and performed worse than for example the MAPE. *Willmott & Matsuura* (2005) found that the RMSE approaches the mean average error for a small number of observations and increases as the number of observations increases. They, therefore, recommend that the measurement is not adequate for average model performance and do not suggest the use of it. In opposition to this, *Chai & Draxler* (2014) mentioned that RMSE is appropriate for Gaussian distributed errors resulting from the forecast models. Another measurement that is also used very frequently is the Mean Squared Error (MSE). We found that it was used 51 times in total and a share of 14.9%. The MSE is for a long time the dominant performance metric in the field of signal processing (*Wang & Bovik*, 2009). A major reason for this is probably the simple calculation, but also its properties of a valid distance metric, its physical interpretability and its excellent properties in optimization contexts. Besides, the MSE is widely used and it is therefore an established practice to compare forecasting model performance (e.g. *Wang & Bovik*, 2009). Because of the quadratic term, large errors are weighted more than small ones. The MSE is more difficult to interpret in different contexts because it is no longer in the original units of measurementment of the observed values due to the quadratic expression. The Mean Absolute Error (MAE) is the fourth most common forecasting performance measurement because the error value units match the predicted target value units. In the research articles we reviewed, the MAE was used a total of 46 times, with a share of 13.4%. The changes of the MAE are linear and therefore intuitive, unlike RMSE or MAPE. Since the error values are measurementd in the original units, the MAE is not suitable for evaluating forecasts from different units. Moreover, the errors are not weighted differently, but are treated equally. For example, MSE and RMSE penalize larger errors more. (*Schneider & Xhafa*, 2022). When measuring an average model accuracy *Chai & Draxler* (2014); *Willmott et al*., (2009) and *Willmott & Matsuura* (2005) showed that MAE outperforms RMSE in most situations, especially at Laplace distributed forecast errors, but worse in Gaussian noisy scenarios (e.g. *Qi et al*., 2020). As we observed, these four performance measurements account for more than 78% of the total measurements collected. According to the research articles we analyzed, this is mainly due to the following. First, these measurements have been in use for a long time and have therefore been widely used in the relevant research. This is likely to improve the comparability of research on forecasting. Second, they're easy to interpret as they use the same scaling as the analyzed time series, except for the mean square error. However, we have also seen that there is widespread criticism for the usage of scale-dependent measurements (e.g.: MAE, RMSE), measurements based on percentage errors (e.g.: MAPE) and measurements based on relative errors (MRAE). A good overview of this criticism is provided by *Hyndman and Koehler* (2006).

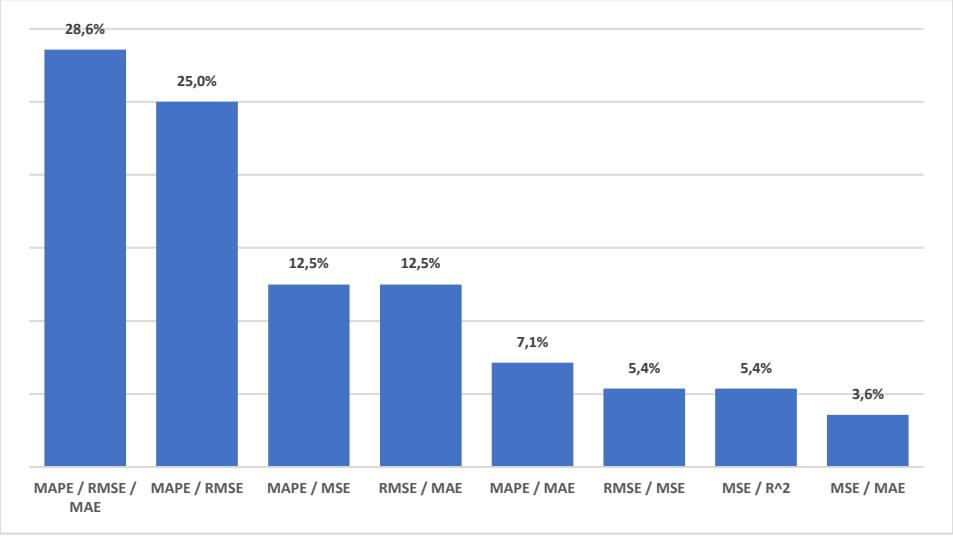
In addition, according to *Hodson* (2022) the use of RMSE and MAE can be appropriate when the measurements are chosen as a function of forecast errors – RMSE is appropriate for Gaussian errors and MAE for Laplace errors. Although less frequently, the coefficient of determination (R^2) with a total count of 14 (4.1%), the Percentage Forecast Error (PFE) with a total count of 9 (2.6%), the Simple Forecast Error with a total count of 7 (2.0%), the Normalized Root Mean Squared Error (NRMSE) with a total count of 6 (1.7%), the Mean Absolute Scaled Error (MASE) with a total count of 4 (1.2%), the Symmetric Mean Absolute Percentage Error (SMAPE) with a total count of 4 (1.2%), the Mean Percentage Error (MPE) with a total count of 4 (1.2%), the Mean Absolute Deviation (MAD) with a total count of 4 (1.2%), the Relative Mean Absolute Percentage Error (RMAPE) with a total count of 3 (0.9%), the Absolute Forecast Error with a total count of 2 (0.6%) were also used. Of the 29 discussed measurements, 15 account for a total of 95.6% of all observations.

Favorite combinations of forecasting performance measurements

In our research, we found that often more than one forecasting performance measurement was used. To examine which combinations of forecasting performance measurements were most frequently used, we analyzed 102 publications (52.8% of the complete sample of 193 publications) that used more than one single forecast measurement. Of these, a total of 54 publications (28.0%) used two forecasting measurements and 48 (24.9%) publications used three or more forecasting measurements. Each combination of forecast performance measurements (e.g. MAPE & MSE) comes from a specific publication of our database. For the analysis of clusters, we recorded all combinations (i.e. MAPE & MSE) that occurred more than once. Overall, we found 56 (54.9% of the total 102) combinations of forecasting performance measurements that occurred more than once. 46 (45.1% of overall 102) combinations were unique and were not shown separately. The results are shown in *Figure 3*. In 16 cases (28.6% of the total 56 recorded combinations), the most popular combination of forecasting performance measurements is the combination of the three measurements of MAPE & RMSE & MAE. measurement. The second most common combination with a frequency of 14 (25.0% of the total 56 recorded combinations) was the combination of MAPE & RMSE. This is followed by the combination of MAPE & MSE, with a frequency of 7 (12.5% of the total 56 recorded combinations), and the combination of RMSE & MAE with a frequency of 7 (12.5% of the total 56 recorded combinations), next, the combination of MAPE & MAE with a frequency of 4 (7.1% of the total 56 recorded combinations), then, the combination of RMSE & MSE with a frequency of 3 (5.4% of the total 56 recorded combinations), finally, the combination of MSE and R^2 with a frequency of 3 (5.4% of the total 56 recorded combinations) and the combination of MAE & MSE with one Frequency of 2 (3.6% of the total 56 recorded combinations). This result shows that the most popular combinations include the most popular measurements of MAPE, RMSE, and MAE. In general, when choosing the best forecasting model, the use of multiple forecasting performance measurements should be challenged from a decision-theoretical point of view. If a single measurement is used, the forecasting model that generates the smaller forecasting loss should be selected. When two

measurements are combined, they should prefer the same forecasting model in order to choose a clear forecasting model. In this case, one of the forecasting performance measurements is redundant. If different forecasting model is preferred, it is difficult to choose. A first limitation of combinations with MAPE is the fact that MAPE produces extreme values with small input values (differences between forecast and true value). As explained above, this means that the choice regarding the best forecasting model cannot be made or that MAPE is no longer used as a criterion. An adjustment of small differences between forecast and true value must then also be made for the comparison models – but these can take completely different values, which means that the sample is distorted. The combination of MAPE and RMSE is complementary in that the MAPE quantifies the forecasting performance as a percentage and the RMSE quantifies it in the unit of the original time series. As we pointed out above, MAE is better for Laplace distributed errors, and RMSE is better for Gaussian distributed errors. Therefore, the combination of these two measurements makes little sense. Instead, one of the two measurements should be chosen based on the existing errors, again in the interests of better choice.

Figure 3: Favorite combinations of forecasting performance measurements



Sample size = 56.

DISCUSSION AND FURTHER RESEARCH

In our evaluation, we found that the Mean Absolute Percentage Error (MAPE), the Relative Mean Squared Error (RMSE), the Mean Squared Error (MSE) and the Mean Absolute Error (MAE) are the most frequently used performance criteria. These measurements dominate the literature for applied artificial neural network demand forecasting in supply chain contexts. This is because they are easy to compute and most of them are easy to interpret. They were also widely used in the past, therefore,

more recent research uses them as a guide. Because of their widespread use, it seems to be easy for researchers to compare their results with other research, other models and applications. However, there have been several criticisms of the measurements in the literature concerning measurement their consistency, their behaviour with outliers and their comparability. As the decision between different forecasting models, based on these forecasting measurements, depends on sample size of the dataset and the forecasting horizon, decision-making is not consistent. Although alternatives (e.g. SMAPE and MASE) with better characteristics are proposed in the literature to choose between different forecasting methods and overcome the weaknesses of the commonly used measurements, we found only a few examples of this Likewise, in our analysis, we found hardly any individual loss functions specific to an application that quantifies the economic effects of incorrect forecasts. That means that over-forecasting is just as problematic as under-forecasting. The use of symmetric performance criteria (like MAPE or RMSE) makes sense in a theoretical context and in the comparison of forecasting models. However, as explained above, this is not appropriate for the forecast of electricity demand. In principle, the choice of the forecast performance measurement should also consider the forecasting framework measurement. It follows that the choice of the forecasting model can still be optimized and so can the forecasting results, since the loss function has not been adapted to the forecasting problem. The most common combinations of forecasting performance measurements are MAPE & RMSE & MAE. This is the combination of the most widely used forecasting performance measurements. A combination of forecasting performance measurements only makes sense if the two measurements complement each other in terms of their properties. Finally, from the perspective of the decision maker, the question is which forecasting model should be preferred if the combined use of performance measurements results in different recommendations. So, the recommendation is that we should use only one "efficient" measurement. Our results detail variability in the use of forecasting performance measurements in the research of applied demand forecasting with artificial neural networks, and thus generate new insights. In our analysis of demand forecasting applications, we did not find any application for the forecast of the demand of financial services, although this is a huge industry. The application to generate optimization potential in this area seems worthwhile and would fill a research gap.

REFERENCES

Aggarwal, (2018). Machine Learning with Shallow Neural Networks. In C. C. Aggarwal (Ed.) *Neural Networks and Deep Learning* (pp. 53–104). Springer International Publishing. https://doi.org/10.1007/978-3-319-94463-0_2

Andrawis, R. R., Atiya, A. F., & El-Shishiny, H. (2011). Combination of long term and short term forecasts, with application to tourism demand forecasting. *International Journal of Forecasting*, 27(3), 870–886. https://doi.org/10.1016/j.ijforecast.2010.05.019

Anisa, M. P., Irawan, H., & Widiyanesti, S. (2021). Forecasting demand factors of tourist arrivals in Indonesia's tourism industry using recurrent neural network. *IOP Conference Series: Materials Science and Engineering*, *1077*(1), 012035. https://doi.org/10.1088/1757-899x/1077/1/012035

Armstrong, J. S. (Ed.). (2001). *Principles of forecasting: a handbook for researchers and practitioners*. International Series in Operations Research & Management Science (Vol. 30). Kluwer Academic. https://doi.org/10.1007/978-0-306-47630-3

Armstrong, J. S., & Collopy, F. (1992). Error measures for generalizing about forecasting methods: Empirical comparisons. *International journal of forecasting*, *8*(1), 69-80. https://doi.org/10.1016/0169-2070(92)90008-W

Babai, M. Z., Syntetos, A., & Teunter, R. (2014). Intermittent demand forecasting: An empirical study on accuracy and the risk of obsolescence. *International Journal of Production Economics*, *157*, 212-219. https://doi.org/10.1016/j.ijpe.2014.08.019

Bowerman, B. L., O'Connell, R. T., Murphree, E., Huchendorf, S. C., Porter, D. C., & Schur, P. (2003). *Business statistics in practice* (pp. 728-730). McGraw-Hill.

Burger, C. J. S. C., Dohnal, M., Kathrada, M., & Law, R. (2001). A practitioners guide to time-series methods for tourism demand forecasting – a case study of Durban, South Africa. *Tourism management*, *22*(4), 403-409. https://doi.org/10.1016/S0261-5177(00)00068-6

Byrne, R. F. (2012). Beyond Traditional Time-Series: Using Demand Sensing to Improve Forecasts in Volatile Times. *Journal of Business Forecasting*, *31*(2).

Carbone, R., & Armstrong, J. S. (1982). Note. Evaluation of extrapolative forecasting methods: results of a survey of academicians and practitioners. *Journal of Forecasting*, *1*(2), 215-217. https://doi.org/10.1002/for.3980010207

Carbonneau, R., Laframboise, K., & Vahidov, R. (2008). Application of machine learning techniques for supply chain demand forecasting. *European journal of operational research*, *184*(3), 1140-1154. https://doi.org/10.1016/j.ejor.2006.12.004

Chawla, A., Singh, A., Lamba, A., Gangwani, N., & Soni, U. (2019). Demand Forecasting Using Artificial Neural Networks – A Case Study of American Retail Corporation. In Malik, H., Srivastava, S., Sood, Y., Ahmad, A. (Eds.) *Applications of Artificial Intelligence Techniques in Engineering*. (pp. 79-89). Springer, Singapore. https://doi.org/10.1007/978-981-13-1822-1_8

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE). *Geoscientific model development discussions*, *7*(1), 1525-1534. https://doi.org/10.5194/gmdd-7-1525-2014

Chen, Z., & Yang, Y. (2004). *Assessing forecast accuracy measures*. Iowa State University.

Clements, M. P., Franses, P. H., & Swanson, N. R. (2004). Forecasting economic and financial time-series with non-linear models. *International journal of forecasting*, *20*(2), 169-183. https://doi.org/10.1016/j.ijforecast.2003.10.004

Clements, M. P., & Hendry, D. F. (1993). On the limitations of comparing mean square forecast errors. *Journal of Forecasting*, *12*(8), 617-637. https://doi.org/10.1002/for.3980120802

Constantino, H. A., Fernandes, P. O., & Teixeira, J. P. (2016). Tourism demand modelling and forecasting with artificial neural network models: The Mozambique case study. *Tékhne*, *14*(2), 113-124. https://doi.org/10.1016/j.tekhne.2016.04.006

Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., & Bharath, A. A. (2018). Generative adversarial networks: An overview. *IEEE signal processing magazine*, *35*(1), 53-65. https://doi.org/10.1109/MSP.2017.2765202

De Gooijer, J. G., & Hyndman, R. J. (2006). 25 years of time series forecasting. *International journal of forecasting*, *22*(3), 443-473. https://doi.org/10.1016/j.ijforecast.2006.01.001

Diebold, F. X., & Lopez, J. A. (1996). 8 Forecast evaluation and combination. *Handbook of statistics*, *14*, 241-268. https://doi.org/10.1016/S0169-7161(96)14010-4

Diebold, F. X., & Mariano, R. S. (2002). Comparing predictive accuracy. *Journal of Business & economic statistics*, *20*(1), 134-144. https://doi.org/10.1198/073500102753410444

Fildes, R., Ma, S., & Kolassa, S. (2022). Retail forecasting: Research and practice. International *Journal of Forecasting*, *38*(4), 1283-1318. https://doi.org/10.1016/j.ijforecast.2019.06.004

Ghalehkhondabi, I., Ardjmand, E., Weckman, G. R., & Young, W. A. (2017). An overview of energy demand forecasting methods published in 2005–2015. *Energy Systems*, *8*, 411-447. https://doi.org/10.1007/s12667-016-0203-y

Goodfellow, I. & Bengio, Y & Courville, A. (2018). *Deep Learning – Das umfassende Handbuch*. 1. Auflage. Verlags GmbH & Co. KG.

Granger, C. W. J., & Pesaran, M. H. (2000a). Economic and statistical measures of forecast accuracy. *Journal of Forecasting*, *19*(7), 537-560. https://doi.org/10.1002/1099-131X(200012)19:7%3C537::AID-FOR769%3E3.0.CO;2-G

Granger, C. W. J., & Pesaran, M. H. (2000b). A decision theoretic approach to forecast evaluation. *Statistics and finance: An interface* (pp. 261-278). https://doi.org/10.1142/9781848160156_0015

Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., ... & Chen, T. (2018). Recent advances in convolutional neural networks. *Pattern recognition*, *77*, 354-377. https://doi.org/10.1016/j.patcog.2017.10.013

Haberleitner, H., Meyr, H., & Taudes, A. (2010). Implementation of a demand planning system using advance order information. *International Journal of Production Economics*, *128*(2), 518-526. https://doi.org/10.1016/j.ijpe.2010.07.003

Hanke,John E., Arthur G. (1995). *Business forecasting*. (5th ed.). Prentice Hall.

Herrera-Granda, I. D., Chicaiza-Ipiales, J. A., Herrera-Granda, E. P., Lorente-Leyva, L. L., Caraguay-Procel, J. A., García-Santillán, I. D., & Peluffo-Ordóñez, D. H. (2019). In: Rojas, I., Joya, G., Catala, A. (Eds.) *Advances in Computational Intelligence*. (pp. 362-373). Springer Cham. https://doi.org/10.1007/978-3-030-20518-8_31

Hodson, T. O. (2022). Root-mean-square error (RMSE) or mean absolute error (MAE): When to use them or not. *Geoscientific Model Development*, *15*(14), 5481-5487. https://doi.org/10.5194/gmd-15-5481-2022

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International journal of forecasting*, *22*(4), 679-688. https://doi.org/10.1016/j.ijforecast.2006.03.001

Jnr, E. O. N., Ziggah, Y. Y., & Relvas, S. (2021). Hybrid ensemble intelligent model based on wavelet transform, swarm intelligence and artificial neural network for electricity demand forecasting. *Sustainable Cities and Society*, *66*, 102679. https://doi.org/10.1016/j.scs.2020.102679

Ke, J., Zheng, H., Yang, H., & Chen, X. M. (2017). Short-term forecasting of passenger demand under on-demand ride services: A spatio-temporal deep learning approach. *Transportation research part C: Emerging technologies*, *85*, 591-608. https://doi.org/10.1016/j.trc.2017.10.016

Kilimci, Z. H., Akyuz, A. O., Uysal, M., Akyokus, S., Uysal, M. O., Atak Bulbul, B., & Ekmis, M. A. (2019). An improved demand forecasting model using deep learning approach and proposed decision integration strategy for supply chain. *Complexity*, *2019*. https://doi.org/10.1155/2019/9067367

Kim, S., & Kim, H. (2016). A new metric of absolute percentage error for intermittent demand forecasts. *International Journal of Forecasting*, *32*(3), 669-679. https://doi.org/10.1016/j.ijforecast.2015.12.003

Koutsandreas, D., Spiliotis, E., Petropoulos, F., & Assimakopoulos, V. (2022). On the selection of forecasting accuracy measures. *Journal of the Operational Research Society*, *73*(5), 937-954. https://doi.org/10.1080/01605682.2021.1892464

Liang, Y. H. (2022). Forecasting International Tourism Demand Using the Recurrent Neural Network Model with Genetic Algorithms and ARIMAX Model in Tourism Supply Chains. *International Journal of Machine Learning and Computing*, *12*(5). https://doi.org/10.1155/2022/3376296

Lewis, C. D. (1997). *Demand forecasting and inventory control: A computer aided learning approach*. Routledge.

Makridakis, S. (1993). Accuracy measures: theoretical and practical concerns. *International Journal of Forecasting*, *9*(4), 527-529. https://doi.org/10.1016/0169-2070(93)90079-3

Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, E. & Winkler, R. (1982). The accuracy of extrapolation (time series) methods: Results of a forecasting competition. *Journal of forecasting*, *1*(2), 111-153. https://doi.org/10.1002/for.3980010202

Maragkos, N. (2020). *Retail Demand Forecasting with Artificial Neural Networks* [Doctoral dissertation, Aristotle University of Thessaloniki].

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, *5*, 115-133. https://doi.org/10.1007/BF02478259

Panapakidis, I. P., & Dagoumas, A. S. (2017). Day-ahead natural gas demand forecasting based on the combination of wavelet transform and ANFIS/genetic algorithm/neural network model. *Energy*, *118*, 231-245. https://doi.org/10.1016/j.energy.2016.12.033

Petrucci, A., Barone, G., Buonomano, A., & Athienitis, A. (2022). Modelling of a multi-stage energy management control routine for energy demand forecasting, flexibility, and optimization of smart communities using a Recurrent Neural Network. *Energy Conversion and Management*, *268*, 115995. https://doi.org/10.1016/j.enconman.2022.115995

Porteiro, R., Hernández-Callejo, L., & Nesmachnow, S. (2022). Electricity demand forecasting in industrial and residential facilities using ensemble machine learning. *Revista Facultad de Ingeniería Universidad de Antioquia*, (*102*), 9-25. https://doi.org/10.17533/udea.redin.20200584

Qi, J., Du, J., Siniscalchi, S. M., Ma, X., & Lee, C. H. (2020). On mean absolute error for deep neural network based vector-to-vector regression. *IEEE Signal Processing Letters*, *27*, 1485-1489. https://doi.org/10.1109/LSP.2020.3016837

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*(6), 386-408. https://doi.org/10.1037/h0042519

Ryu, S., Noh, J., & Kim, H. (2016). Deep neural network based demand side short term load forecasting. *Energies*, *10*(1), 3. https://doi.org/10.3390/en10010003

Salehinejad, H., Sankar, S., Barfett, J., Colak, E., & Valaee, S. (2017). Recent advances in recurrent neural networks. *arXiv preprint arXiv:1801.01078*. https://doi.org/10.48550/arXiv.1801.01078

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, 61, 85–117. https://doi.org/10.1016/j.neunet.2014.09.003

Schneider, P., & Xhafa, F. (2022). *Anomaly Detection and Complex Event Processing Over IoT Data Streams: With Application to EHealth and Patient Data Monitoring.* Academic Press.

Slimani, I., El Farissi, I., & Achchab, S. (2015). Artificial neural networks for demand forecasting: Application using Moroccan supermarket data. *2015 15th International Conference on Intelligent Systems Design and Applications (ISDA)* (pp. 266-271). IEEE. https://doi.org/10.1109/ISDA.2015.7489236

Suganthi, L., & Samuel, A. A. (2012). Energy models for demand forecasting – A review. *Renewable and sustainable energy reviews*, *16*(2), 1223-1240. https://doi.org/10.1016/j.rser.2011.08.014

Swanson, N.R., Ghysels, E. & Callan, M. (1999). A Multivariate Time Series Analysisof the Data Revision Process for Industrial Production and the Composite Leading Indicator. In R.F. Engle & H. White (Eds.), *Cointegration, Causality and Forecasting: A Festschrift in Honor of Clive W.J. Granger*. (pp.45-75). Oxford University Press.

Thonemann, U. (2010). *Operations Management*, 2., aktualisierte Auflage, Pearson.

Wanchoo, K. (2019, March). Retail demand forecasting: a comparison between deep neural network and gradient boosting method for univariate time series. *2019 IEEE 5th International Conference for Convergence in Technology (I2CT)* (pp. 1-5). IEEE. https://doi.org/10.1109/I2CT45611.2019.9033651

Wang, B., Liu, K., & Zhao, J. (2016). Inner attention based recurrent neural networks for answer selection. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1288-1297). https://doi.org/10.18653/v1/p16-1122

Wang, Z., & Bovik, A. C. (2009). Mean squared error: Love it or leave it? A new look at signal fidelity measures. *IEEE signal processing magazine*, *26*(1), 98-117. https://doi.org/10.1109/MSP.2008.930649

Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate research*, *30*, 79-82. https://doi.org/10.3354/cr030079

Willmott, C. J., Matsuura, K., & Robeson, S. M. (2009). Ambiguities inherent in sums-of-squares-based error statistics. *Atmospheric Environment*, *43*(3), 749-752. https://doi.org/10.1016/j.atmosenv.2008.10.005

Yu, D., Li, Z., Zhong, Q., Ai, Y., & Chen, W. (2020). Demand management of station-based car sharing system based on deep learning forecasting. *Journal of Advanced Transportation*, *2020*, 1-15. https://doi.org/10.1155/2020/8935857

Zhang, B., Li, N., Shi, F., & Law, R. (2020). A deep learning approach for daily tourist flow forecasting with consumer search data. *Asia Pacific Journal of Tourism Research*, *25*(3), 323-339. https://doi.org/10.1080/10941665.2019.1709876

Corresponding author:

Steffen ROBUS
Hungarian University of Agriculture and Life Sciences
Doctoral School of Management and Organizational Science
7400 Kaposvár, Guba Sándor u. 40., Hungary
e-mail: steffenrobus@gmail.com