Response Letter

ID: NEUCOM-D-22-03187
Title: ”Exploring the association between time series features and forecasting by temporal aggregation using machine learning”
Editor in Chief: Leandro L. Minku.

1 Letter to Editor and summary of changes

Dear Editor,

Thank you for the opportunity to revise our paper for the Neurocomputing journal. We have taken a positive approach to the thoughtful comments of the authors, who we sincerely thank for helping us with this paper. This has lead to a major revision of the paper that we summarise below in our response to each referee. We quote below the detailed referees' comments that are followed by our response. In all cases, we point out, where necessary, the corresponding amendments in the new version of the paper and highlighted them in blue.

2 Response to Associate Editor

AE's comments: Please carefully revise the paper according to the suggestions if they are acceptable. BTW, R3 suggested to have a close look at the time series classifier, and it is a suggestive concern, please work on it. thanks.

Response: Thank you for giving us the chance to revise the manuscript. We have addressed all the issues and provide a detailed answers in this document. This has lead to a major revision of the paper that we summarise below in our response to each referee. We look forward to your feedback on our revised paper in due course.

Our responses are shown in black, while the corrections in the revised manuscript are highlighted in blue.

We have corrected grammatical typos, incorporated a body of literature suggested by reviewers, modify several figures, add several new models recommended by referees and finally add new results in corresponding tables and figures.

We have included these new models in the experiment: 1) Google Brain TensorFlow model (TensorFlow), 2) Facebook's Deep learning Torch model based on tensors & neural networks (DL Torch), 3) extreme gradient boosting (XGBoost), 4) recurrent neural network (RNN), 5) convolution neural network (CNN), 6) feedforward neural network (FNN) and 7) Dynamic Time Warping (DTW). That means we had to rerun the experiment and include the results of these models in the corresponding tables and figures. If the caption of a Table or a Figure is highlighted in blue, that means a change has been made inside those tables and figures, they may include new results, or just a modification suggested by referees.

3 Response Letter to Reviewer 1

- *General comment: This paper revealed that the two basic approaches, AF and AD, for time series forecasting are both not uniformly superior, and by constructing an experiment framework based on machine learning models, the author presented time series features that influence two approaches most. However, it still needs further improvement. I provide some suggestions below for the revised version*

  Response: Thank you very much for providing valuable feedback to improve the manuscript. We have addressed all comments and detailed responses are provided below.

- *Comment 1: The author should check the English expressions and punctuation in the introduction. For instance, in the fourth paragraph of introduction, the "generated" in "time series be used to generated the forecast" should be revised as "generate". A comma should be added ahead of "or should we first aggregate the time series". In the fifth paragraph of introduction, "as" should be added in "which is referred to aggregation level".*

  Response: Thank you very much for the comment. We have now checked the article and corrected multiple linguistic errors throughout the paper. Our revised manuscript has also been proofread by a native competent English speaker. Please see below corrected text in the revised paper:

  time series be used to generate the forecast
  , or should we first aggregate the time series
  which is referred to as aggregation level

- *Comment 2: In page 5, the first paragraph of experiment framework, the "method" in "then introduce the forecasting approaches and method" should be revised as "one method" to specify the ETS method in the latter. Otherwise, it may cause ambiguity.*

  Response: Thank you. To avoid ambiguity, we clarify which approaches and forecasting method is used in this paragraph. We have included the following text in the revised manuscript at the beginning of section 3:

  We first present the study framework, then describe AD and AF approaches and the ETS forecasting method, followed by the forecasting error metrics.

- *Comment 3: In page 5 and page 6, the experiment framework, the author state that using two approaches to get the forecast (AF for generating original times series forecast while AD for generating temporally aggregated time series). However, the content of steps 2-5 is not consistent with the Figure 2. The author should state the specific method "the same forecasting* *method in step 2" of generating the forecast for temporally aggregated time series in step 5 to keep the consistence between figure 2 and steps. Otherwise, it confuses readers.*

Response: Thank you for raising this issue. We have made some changes in the content of step 2-5 to make sure they are aligned with the Figure 2. The changes are highlighted in the following text and incorporated into the revised paper. Please refer to the end of page 5 and the begining of page 6:

- The original monthly time series is transformed into temporally aggregated series for a given aggregation aggregation level, m;
- The ETS forecasting method is applied to the original series to generate forecast for m periods;
- Forecasts are generated using ETS model for temporally aggregated series (forecasts from AD);
- Several machine learning (ML) models are built to accurately predict the superiority of AF/AD approaches using the data created in step 7. These models include: 1) Logistic regression (LR), 2) Linear discriminant analysis (LDA), 3) Quadratic discriminant analysis (QDA), 4) K-Nearest Neighbors (KNN), 5) Lasso, 6) Generalised Additive Model (GAM), 7) Boosting, 8) Support Vector Machine (SVM), 9) Random Forest (RF), 10) Google Brain TensorFlow model (TensorFlow), 11) Facebook's Deep learning Torch model based on tensors & neural networks (DL Torch), 12) extreme gradient boosting (XGBoost), 13) recurrent neural network (RNN), 14) convolution neural network (CNN), 15) feedforward neural network (FNN) and 16) Dynamic Time Warping (DTW). Researchers are referred to James et al. (2021) for a detail description of some of these approaches.

- *Comment 4: In page 6, the second paragraph of forecasting approaches, the author sate "These forecasts are then aggregated, because forecasting at the aggregation horizon level is of interest.". However, whether this forecasting is of interest or not has no logical relationship with the context. The author should revise it, or it will lead to logical overlap.*

  Response: Thank you. This has now been removed to avoid the confusion. We included the following text in the revised paper, please refer to the second paragraph in section 3.1, in page 7:

  This approach first involves generating base forecasts for m periods ahead. Base forecasts are then aggregated to create forecasts at the aggregation horizon level.

- *Comment 5: In page 7, forecast accuracy measures, the description of RMSSE, misclassification and F statistic should be added in more details to specify that the meaning of each parameter in the vision of this paper. Otherwise, it confuses readers.*

Response: Thank you for the valuable comment. We have added the required description. We included the following changes in section 3.3, page 8 9:

RMSSE is given by:

$$\text{RMSSE} = \sqrt{\text{mean}(q_j^2)},$$

where

$$q_j^2 = \frac{e_j^2}{\frac{1}{T-m} \sum_{t=m+1}^{T} (y_t - y_{t-m})^2},$$

where e_j is the forecast error, the difference between an observed value and its forecast. y_t is an observed value at period t, T is the length of time series, and m is the seasonal length (e.g. for monthly m=12. For non-seasonal data, $m = 1$.

The misclassification error is calculated as following:

$$Misclassification_{error} = 100\% \times \left(1 - \frac{(t_p + t_n)}{N}\right),$$

where N is the total number of cases to predict, t_p is the true positive (i.e. when the model correctly predicts the outperformance of AF approach over AD) and t_n is true negative (i.e. when the model correctly predicts the underperformance of AF approach over AD)

F statistic is defined as:

$$F_{statistics} = 2 \times 100\% \times \left(\frac{(tp/(t_p + f_p) \times t_p/(t_p + f_n))}{(t_p/(t_p + f_p) + t_p/(t_p + f_n)}\right),$$

Where f_p is false positive and f_n is false negative.

The false positive rate represents the fraction of cases in which AD was incorrectly classified as a right model to use, while the AF was the correct one. Similarly, the false negative rate represents the fraction of cases in which AF was incorrectly classified as a right model to use, while AD was the correct one.

We can also illustrate the trade-off between false positives and true positives using a curve. It is called a receiver operating characteristic curve, or ROC curve (James et al., 2021). The ROC curve is a plot of the true positive rate (t_p, sensitivity) versus the false positive rate (f_p, 1 - specificity) for a set of thresholds. We can quantify this by calculating the area under the curve, or AUC. The higher AUC, the better the model does at predictions. The maximum value of AUC is 1, which would be considered as a perfect prediction. Conversely, a model that performs no better than chance will have an AUC of 0.5 and would be considered as a poor one.

- *Comment 6: In page 8, the second paragraph of time series features, "," should be added ahead of "to develop an accurate prediction model".*

  Response: Thank you, we have now changed the following paragraph to fix the issue raised here. Please see the new paragraph in the revised paper at the beginning of section 4.1, page 9:

  Given the complexity of the relationship between the time series features and the performance of temporal aggregation approaches, we have considered all 42 features as predictor rather than using only few limited numbers of features known to users such as the strength of trend or seasonality, to develop an accurate prediction model. Using a reduced number of features either selected based on their interpretability or using dimensional reduction techniques might reduce the predictability power of the model, but this could be an interesting avenue for future research.

- *Comment 7: Please pay attention to the image format. The images, figures 3-6 and figures 15-16, need to be added with sub-annotation.*

  Response: Thank you for the suggestion. We have now added annotation to these figures in the revised version. Please refer to these figures in the revised version to see changes.

- *Comment 8: In page 14, the second paragraph of section 5, "using AD using non-overlapping temporal aggregation" in "These results might be surprising, because most of the findings in the literature recommends using AD using non-overlapping temporal aggregation over aggregate forecast as discussed in section @reflit." is confusing. The author should carefully check the English expression of this sentence. What's more, what is the "@reflit" in the sentence. Is there a compilation problem with the manuscript? The author should revise it, or it confuses readers.*

  Response: Thank you, we have now changed the paragraph to fix the issue raised here. "@refl" was intended to be a cross reference but there was a typo there, this is now fixed. Please see below the new paragraph in the revised paper, also you can check the second paragraph in section 4.4 in page 16:

  These results might be surprising, because common recommendation based on practice and some research findings in the literature (see Section 2) is to use AD over AF.

- *Comment 9: In page 14, figure 8, please check the position of the image which seems to exceed the width of the text. The word "A" in "AD" in the y-axis is ambiguous. What's more, the author should explain the first letter "BU and DA" in figure 8, which seems to conflict with the content "AD and AF". Otherwise it confuses readers.*

  Response: Thank you very much for spotting that issue. Bottom-Up (BU) approach has been also be used in the literature to refer to Aggregate forecast (AF). We have now changed BU and AF and corrected spelling of AD in the caption of Figure 8. Please refer to Figure 8 to see the corrected caption.

Performance of AF and AD models evaluated using MCB test. RMSSE values are used for computing the ranks and a 95 percentile confidence level.

- *Comment 10: In page 16, figure 16, what do x-axis and y-axis represent. Do x-axis and y-axis both represent the range of values extracted for each feature? The author should state it, or it confuses readers.*

  Response: Thank you. We have now clarified the x and y axis in Figure 16. X-axis represents the range of features extracted from data, and y-axis represents the chance (i.e probability) of AF outperforming the AD as the feature changes. We have now included the following text in the revised paper. Please see the last line in page 26:

  The x-axis represents the range of values extracted for each feature and the y-axis is the probability of classifying AF versus AD. Therefore, the dependent value is the probability of correctly predicting the aggregating forecasts i.e. AF approach.

- *Comment 11: Some current related works are suggested to discuss, e.g.,*

  - A fuzzy interval time-series energy and financial forecasting model using network-based multiple time-frequency spaces and the induced-ordered weighted averaging aggregation operation;
  - An efficient forecasting method for time series based on visibility graph and multi-subgraph similarity;
  - Time series forecasting based on fuzzy cognitive visibility graph and weighted multi-subgraph similarity.

  Response: Thank you for the comment. Following your suggestion, we have now included the following paragraph discussing these related works. Please see the new paragraph below added in page 5:

  A stream of research investigates the use of features in time series forecasting based on multiple time frequency spaces in visibility graphs. Liu et al. (2020) proposed a multiple time-frequency spaces fuzzy interval forecasting model using datasets from energy and finance. The original series is decomposed into different components, which are then used to reconstruct a group of time series at different temporal scales. Next, a prediction interval forecast is generated for the different reconstructed time series that are then aggregated using the induced-ordered weighted averaging aggregation operation to generate the final forecast. Hu and Xiao (2022a) suggested a novel time series forecasting model based on a new metric measuring nodes similarity in visibility graph. In proposed model, time series is first converted into a visibility graph.next, the similarities between nodes are determined and finally forecasts are generated using the normalized similarity distribution. Hu and Xiao (2022b) investigated the features of time series to generate accuracy forecasts from the perspective of fuzzy interaction between nodes. They used a fuzzy cognitive visibility graph to convert the time series into a pair of directed weighted graphs. Then, a the weighted multi-subgraph similarity is developed to calculate the similarity between nodes. They then proposed a novel forecasting method for time series forecasts based on fuzzy similarity distribution that can efficiently capture the spatio-temporal dependency in the time series data. The empirical results confirm the benefits of leveraging fuzzy interaction for time series forecasting based on the visibility graphs.

- *Comment 12: This work is meaningful in exploring the association between time series features and basic forecasting method performance. I recommend accepting this paper after addressing the above revisions.*

  Response: Thank you very much for your positive feedback and thoughtful comments. We sincerely appreciate your time for helping us with this paper.

4 Response Letter to Referee 2

General comment: The authors have designed an empirical experiment framework to explore the performance of AF and AD using the monthly time series of M4 competition dataset. The machine learning algorithms have been adopted to investigate the connection between time series features and the performance of temporal aggregation. The experimental results have indicated the superiority of RF to other machine learning models. The partial dependence plot has been extracted to describe the contribution of time series features through a probability and the experimental results have shown that how the value of a feature may favorite AF over AD approach. Overall, the manuscript is well-organized. However, the following issues should be improved.

Response: Thank you for your useful comments that have lead to some nice improvements to our paper. Below we response to each comment individually.

- *Comment 1: The main contributions of this work can be polished further.*

  Response: Following your comment, we have now polish further the main contributions of this work. We have included the following text in the revised paper. Please see the third paragraph in page 3:

  Despite recent developments in this area, there is still a lack indications on which temporal aggregation approach should be used to forecast a time series, given its features. To our knowledge, this is the first study that explores the association between time series features and model performance in the context of forecasting by temporal aggregation (TA). The need for such research has also been emphasised by (Babai et al., 2022) in a review article. This study contributes to the area of time series forecasting and intends to shed lights on how the performance of temporal aggregation approaches (i.e. both AF and AD) is associated with time series features. To that end, we use 48,000 time series from the monthly M4-competition dataset. First, we examine how the features of time series changes going from a high granularity level (e.g. monthly) to a low granularity (e.g. annual). We then build machine learning models to describe the association between the original time series features and the forecasting performance of temporal aggregation approaches. Next, we use models' outputs to discover which features are critical in predicting accurately the performance of temporal aggregation approaches, followed by an interpretation of features associated with the forecasting performance. This will help us to provide recommendations to forecasters and decision-makers on which approach to use.

The research objectives are as following:

1. We measure 42 features of the time series at the original level (e.g. monthly) and at various levels of temporal aggregation (e.g. quarterly, annual) using the monthly M4 competition.
2. We reveal how time series features change as we aggregate data from high frequency (e.g. monthly) to low frequency (e.g. annual).
3. We assess the forecast accuracy performance of AD and AF approaches for the forecasts generated by the the Exponential Smoothing State Space (ETS) model.
4. We build machine learning models using time series features as predictors to accurately predict which approach (AD or AF) performs better.
5. We examine the association between time series features and the forecasting performance of these approaches.

- *Comment 2:* Is the conclusion effective for other time series datasets? We suggest the authors can provide some analysis.

  Response: Thank you for raising this interesting point on how the results might be generalised. We have an extensive dataset which includes monthly time series from different sectors such as finance, retail, etc. We have also used an automatic ETS forecasting method, which is a reasonable choice for monthly time series because it can capture level, trend and seasonality, which are the most common patterns for monthly time series. We believe that the proposed framework can be generalised to any other time series data. we provide the R code through a GitHub repository to help with reproducibility.

  However, we believe that a different conclusion may be reached when using different time series granularity such sub-daily, daily or weekly time series. This will bring further complications such as the presence of long and multiple seasonal cycles, so time series features would be different, and they way they may change with increasing aggregation level can differ too. Also, the choice of forecasting method becomes important as a method like ETS can not handle long and multiple seasonalities and the changes in their values over time, which is necessary for such time series. We have now included the following paragraph in the conclusion section, page 28, to address this comment:

  The proposed framework can be generalised to be used with other time series data. We believe that the conclusion should remain true for lower frequency time series such as monthly and quarterly. However, a different conclusion might be reached when using higher frequency time series data such sub-daily, daily, or weekly. This will bring further complications such as the presence of multiple seasonal cycles and long seasonalities. Therefore, time series features and the way they may change with increasing the aggregation level may differ. Also, the choice of forecasting method becomes important as a method like ETS cannot handle those complications. This would be an important avenue for further research.

- *Comment 3: How are the machine learning models generated for time series datasets automatically? Some auto-ML methods can be extended to this issue.*

  Response: Thank you very much for the valuable comment. Following your recommendation, we have added some aditional ML models. Please also see our detailed response regarding model setup.

  - We have provided the GitHub repository that includes the R codes for each ML model setup (i.e. generating, designing, evaluating and choosing the optimal setup). Due to the space restrictions in the journal, we have only provided details for Random Forests model since it performed best in this case (please refer to sub-section 5.2).
  - Furthermore and following your recommendation, we have now included some additional ML models in the experiment, especially, some state of the art industry models developed by big tech companies like Google and Facebook. Therefore, we have now included six new models in the paper including: 1) Google Brain TensorFlow model (TensorFlow), 2) Facebook's Deep learning Torch model based on tensors & neural networks (DL Torch), 3) extreme gradient boosting (XGBoost), 4) recurrent neural network (RNN), 5) convolution neural network (CNN) and 6) feedforward neural network (FNN). Now, there are in total 16 models considered in the paper.
  - Again, for details of all models, you can refer to the GitHub repository.
  - We have added a new Appendix section to the revised version of the paper to share some information about models' setup.
  - In our paper, the ML models have not been generated automatically. Rather for each considered model we have conducted a comprehensive study and searched for the individual optimal setup. In that regard, our aims is to demonstrate the best possible performance of each ML for the given problem and show how much each model can help in explaining the choice between AD and AF forecasting approaches when forecasting many time series.

  We have now included the following texts to reflect these changes. Please see page 6 in the revised paper:

  We should note that we have conducted a comprehensive study and searched for optimal setups of each model rather than using automatic ML models. Additional information about the setup of models can be found in the Appendix or in the GitHub repository

5 Response Letter to Referee 3

- *General comment: The manuscript considers the task of time series forecast in cases when the granularity of the time series is different from the granularity at which the forecast is expected, e.g., when monthly sales data is available, but the forecast is expected to be made for the next year. A standard model, exponential smoothing state space, is used for the actual forecast, while the manuscript focuses on how to deal with the aforementioned difference in granularity. In particular, two strategies are studied: (i) aggregation of the time series (denoted as "AD" in the manuscript), and (ii) aggregation of forecasts (denoted as "AF"). According to the results presented in the manuscript, none of these two approaches is universally superior to the other one. Therefore, the Authors consider various machine learning techniques to automatically decide, for each time series, which of the two methods should be used. This is somewhat reminiscent to "Individualised Error Prediction" in time-series classification. The aforementioned decision task, i.e., whether to use AD or AF for forecast, is a time series classification task, which is solved by feature extraction followed by a standard classifier in the current version of the manuscript. The authors claim that random forest works best out of the examined classifiers.*

  Response: Thank you for the nice summary of the paper. We appreciate your useful comments that helped us to improve the quality of our work. Below we response to your comments individually.

- *Comment 1: My major concern is that the Authors did not try time-series classifiers for the aforementioned time-series classification task, see e.g. Buza (2018) and the references therein for a quick overview. Of course, it may happen that random forest with appropriate features is better than "usual" time-series classifiers (DTW-based methods, convolutional neural networks, etc.), nevertheless, it would be interesting to see whether this is really the case, especially because according to Fig. 11, there is still room for improvement (i.e., the forecast would be better in case of a better classifier, currently used classifiers are more-less half-way between "AD" and the "ideal classifier" which would perfectly predict whether to use AD or AF). As for the standard classifiers with domain-specific features I have to note that the Authors did not consider neural networks, which should also be discussed in my opinion.*

  Response: Thank you very much indeed for your valuable comment.

  - Following your recommendation we have added the Dynamic Time Warping (DTW) and the Convolution Neural Network (CNN) to the experiment and highlighted the related changes in the revised paper. Also, following a suggestion from the Reviewer 2, we included some other neural network models including: Google Brain TensorFlow model (TensorFlow), Facebook's Deep learning Torch model based on tensors and neural networks (DL Torch), recurrent neural network (RNN) and feedforward neural network

(FNN).

  - We have compared the accuracy of both DTW and CNN models against other models in our analysis. While both models demonstrated good results, however they failed to beat the existing (such as such as Boosting, Support Vector Machines (SVM) and random forest (RF)) and some newly added models (such as TensorFlow, DL Torch, XGBoost).
  - We have also included a paragraph to highlight the use of time-series classification in the context of temporal aggregation problem as a potential future research given its relevance to the topic of time-series classification Buza (2018) .

We have included results of new models in Table 2 and Figure 10 and Figure 11. Please refer to these figures in the revised version of the paper. Please also refer to the beginning of page 6 (item number 8), where we added the new list of the models:

These models include: 1) Logistic regression (LR), 2) Linear discriminant analysis (LDA), 3) Quadratic discriminant analysis (QDA), 4) K-Nearest Neighbors (KNN), 5) Lasso, 6) Generalised Additive Model (GAM), 7) Boosting, 8) Support Vector Machine (SVM), 9) Random Forest (RF), 10) Google Brain TensorFlow model (TensorFlow), 11) Facebook's Deep learning Torch model based on tensors & neural networks (DL Torch), 12) extreme gradient boosting (XGBoost), 13) recurrent neural network (RNN), 14) convolution neural network (CNN), 15) feedforward neural network (FNN) and 16) Dynamic Time Warping (DTW).

Please also refer to the last bullet point of the future works in the conclusion section, page 29, where we added:

Using time-series classification techniques (Buza, 2018) to classify time series data based on various factors including time series futures and investigate their link to forecast accuracy of temporal aggregation approaches might be an interesting avenue for further research.

- *Comment 2: While I appreciate that "the authors are willing to share the code for reproducibility", it would be much better if they would upload their code to a publicly available repository (e.g. github).*

  Response: Thank you for your valuable suggestion. The repository is now publicly available to access the R code to reproduce results and also Rmarkdown file to reproduce the entire paper. Please refer to the reproducibility section. We have added the following text in the revised paper:

  Please refer to the newly added section on Reproducibility, in page 29, where we added:

  R code and RMarkdown file to produce all results in this paper are available at GitHub repository

- *Comment 3: Last, but not least, I have to mention that the manuscript requires careful proofreading w.r.t. typos and grammar. A few examples (the list is not complete):*

Response: Thank you very much for the comment. We have now corrected the following identified typos. Additionally, we checked the article and corrected multiple linguistic errors throughout the paper.

- ”...be used to generated...” (Section 1). **Response:** This is changed to be used to generate
- ”...an NOTA...” (Section 2). **Response:** this is changed to NOTA
- ”...one atep...” (Section 3.1). **Response:** this is changed to one step
- ”...we illustrates...” (Section 4). **Response:** This is changed to we illustrate
- ”Coeifficnet...” (Tab. 1). **Response:** this is changed to Coefficient...” (Table 1). We have also corrected few more typos in Table 1
- ”...stationary...” (Section 4.2). **Response:** this is changed to stationary
- ”...extract the futures...” (Section 4.3). **Response:** this is changed to extract the features
- ”...if the monthly series if staitonairy” (Section 4.3, two typos: ”...if the monthly series IS STATIONARY”). **Response:** this is changed to ”...if the monthly series is stationary”)
- ”...remaoins...” (Section 4.3). **Response:** this is changed to remains
- - ”...@reflit...” (Section 5). **Response:** this is now fixed
- ”...a time periods...” (Section 8). **Response:** this is changed to ”...a time period...”
- ”...that how the value...” -¿ ”...how the value...” (Section 8). **Response:** this is changed to ”how the value”
- ”...no-linearity...” -¿ ”nonlinearity...”(?) . **Response:** this is changed to ”non-linearity”
- ”...may favorite AD...” -¿ ”may favor AD” . **Response:** this is changed to ”may favor AD”

6 Response Letter to Referee 4

- *General comment: The paper deals with scenarios where a forecast of the total value over several time periods ahead is required, comparing two Temporal Aggregation (TA) approaches to produce the required forecast: i) aggregate forecast(AF) or ii) aggregate data using non-overlapping temporal aggregation (AD). The paper is well written and has a good literature review. The paper reasonably criticizes the common sense intuition that forecasts created at the higher temporal aggregation levels are more accurate.*

  Response: Thank you for the positive feedback and your valuable comments, which helped us to improve the paper.

- *Comment 1: A favorable feature of the paper is the willingness of the authors to share the code to ensure the reproducibility of the experiments. In the final version, the link to the repository can be provided in the text, to spare those interested in the code the inconvenience of having to write to the authors requesting access to the code and eventually delay in receiving a response. The results of experiments with metrics other than RMSSE can be made available there as well.*

  Response: Thank you for your valuable suggestion. The repository is now publicly available to access the R code to reproduce results and also Rmarkdown file to reproduce the entire paper. Please refer to the reproducibility section. Please also see our response to your comment 3. We have added the following text in the revised paper:

  Please refer to the newly added section on Reproducibility, in page 29, where we added:

  R code and RMarkdown file to produce all results in this paper are available at GitHub repository

  Response: Regarding making the results for other accuracy metrics available, we have already considered various accuracy metrics in the experiment, however we present results of RMSSE in the paper because it is the recommended measure in the latest M-competition (Makridakis et al., 2022), also the main conclusion remains true when using other metrics. We have not included all error metrics in the paper due to page restriction. However, for the purpose of reproducibility and for those readers who are interested in experimenting with other metrics, we have made the GitHub repository including R code and RMarkdown file, publicly available. We have included a parameter in the YAML section of the Rmarkdown file in the GitHub. This allows to change the performance metric for those who might be interested in seeing a particular metric and the results presented in Tables Figures would be automatically updated in the paper accordingly.

- *Comment 2: In AF, the prediction with larger granularity can be used to understand how the accuracy is being affected as the prediction moves away from the data (probably it will fall). This sensitivity is lost in AD.*

  Response: Thank you for your observation, we hope that we have correctly understood your interpretation here. If our understanding is correct, we then totally agree with your interpretation here. By definition when we aggregate data to lower frequency domain such as annual, we loose some information for instance and AD is losing some of the sensitivity compared to AF. For example, if you have daily data and aggregate to yearly, surely some important information is lost there as we would not be able to see intra-week seasonality at yearly level. AF can pick up those information at the original level to predict accurately. Additionally, if you forecast far ahead in the future like forecasting one year ahead (i.e. 365 days), the AF might not be able to pick far ahead days accurately and it might seem rational to use AD approach. We have added the following paragraph at the begining of page 3:

When we aggregate data to lower frequency domain such as annual, we loose some information and AD is losing some of the sensitivity compared to AF. AF may better capture detailed information at the higher frequency to predict accurately. However, this might be also affected by forecast horizon, as using AF might not be useful when forecasting far ahead into the future.

- *Comment 3: Only one error metric was considered in the validation, but a prediction may have had a smaller error, without the predicted series being more similar to the ground truth, having had a smaller error just because the mean of its points is closer to the mean of ground truth points. To have a more fair comparison, simultaneously the error metric should be used as some measure of time series similarity, such as Dynamic Time warping (which is not a metric, as it does not respect the triangular inequality), Pearson correlation coefficient, or compression- based dissimilarity.*

  Response: Thank you for your comment. We definitely agree with you in principles that we need to use multiple error metrics simultaneously when evaluating the model's performance. In this study we have included multiple error metrics in the experiment as well. In the following, we have discussed briefly why we presented the result of RMSSE and how readers can access other metrics used in this experiment:

  1. First, in the study experiment we have evaluated the time series forecast accuracy and bias. To compare the forecast accuracy we have used several performance metrics including Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) and Mean Error (ME), Mean Absolute Percentage Error (MAPE), Mean Absolute Scaled Error (MASE) and Root Mean Squared Scaled Error(RMSSE). We analysed the forecasting performance of approaches using all metrics, and observed that accuracy metrics are consistent. Therefore, regardless of which accuracy metric we use, the overall conclusion remains the same.
  2. Second, we use the monthly data from M-Competition) to empirically evaluate the models' performance. There has been a long debate around which forecast performance metric should be used, and the last series of forecasting competition known as M5 (Makridakis et al., 2022) suggested RMSSE to evaluate point forecast accuracy compared to others. We used RMSSE in this study following this recommendation and also to be consistent with the M competition guidelines, given that it is the source of data used in the experiment.
  3. Third, there is a page restriction in terms of how much we can present and discuss in the paper. Therefore, for the purpose of reproducibility and for those readers who are interested in experimenting with other metrics, we have made the GitHub repository of this study publicly available.
  4. Fourth, we have included a parameter in the YAML section of the Rmarkdown file in the GitHub. This allows to change the performance metric for those who might be

interested in seeing a particular metric and the results presented in Tables Figures would be automatically updated in the paper accordingly.

We have added the following paragraph in section 3.3, page 8:

We only present the results for RMSSE due to the space limit of the journal and also because it is the recommended error metric in M5 competition (Makridakis et al., 2022). We share the R code and the Rmarkdown file through the GitHub repository, which provide the possibility of changing the error metric in the YAML section of the Rmarkdown to obtain results for other metrics. We should also note that similar conclusions are reached when using other accuracy metrics.

- *Comment 4: Recurrent neural networks and transformers were missing from the list of compared models, as they are relevant options for any current time series analysis task. The use of attentional mechanisms can influence the benchmark between PA and AD, so this should be considered at least for future work.*

  Response: Thank you very much for your valuable comment. Following your recommendation, we have included recurrent neural network (RNN) in the revised paper. Moreover and following some suggestions from other reviewers, we have extended the list of models implemented for the purpose of this experiment. The new models implemented for the revised version of the paper includes: i) Google Brain TensorFlow model (TensorFlow), ii) Facebook's Deep learning Torch model based on tensors & neural networks (DL Torch), iii) extreme gradient boosting (XGBoost), iv) recurrent neural network (RNN), v) convolution neural network (CNN), and vi) feedforward neural network (FNN). Additionally, we believe that the current research could be further extended to investigate the application of transformers neural networks and attentional mechanisms in the context of this problem.

  Please refer to Figure 10, 11 and Table 2, where we added results for new models. Also refer to the beginning of page 6 (item number 8) where we added the list of models considered.,

  These models include: 1) Logistic regression (LR), 2) Linear discriminant analysis (LDA), 3) Quadratic discriminant analysis (QDA), 4) K-Nearest Neighbors (KNN), 5) Lasso, 6) Generalised Additive Model (GAM), 7) Boosting, 8) Support Vector Machine (SVM), 9) Random Forest (RF), 10) Google Brain TensorFlow model (TensorFlow), 11) Facebook's Deep learning Torch model based on tensors & neural networks (DL Torch), 12) extreme gradient boosting (XGBoost), 13) recurrent neural network (RNN), 14) convolution neural network (CNN), 15) feedforward neural network (FNN) and 16) Dynamic Time Warping (DTW).

  We also included the following section in the conclusion section, page 28:

  Given that the exact relationship between time series features and the optimal temporal aggregation approach is not known, one direction for future research could focus on using meta-learning and other advanced techniques such as transformers neural networks and attentional mechanism to further shed lights on this problem

- *Comment 5: The main setup information for the models should be presented in an appendix.*

  Response: Thank you for the suggestion. We have added a new table in the appendix that represents the general methodology we followed while generating the optimal setup for each model. In our research, there are several models considered. Due to the space restrictions in the journal, we have only provided details about the generation process for the random forests model since it performed best on a given problem. For full details about each model designing process, you can refer to the GitHub repository.

  Please refer to see page 6 where we addded:

  Additional information about the setup of the models can be found in the Appendix or in the GitHub repository

- *Comment 6: For me it was not clear if AF is multi-output or not. If so, it could be considered to use some form of more realistic validation, such as those proposed by "Rafael de Oliveira Werneck et al., 'Data-driven deep-learning forecasting for oil production and pressure.' Journal of Petroleum Science and Engineering 210 (2022)", which avoiding mixing multiple prediction confidences over long forecasts.*

  Response: Thank you very much for the valuable comment and literature reference. In our research, AF doesn't have multi-outputs, but we believe it is possible to define the problem of temporal aggregation as a form of multi-output. We have added the following paragraph to the conclusion section of the paper as a potential future avenue for this research.

  In this paper, we define the AD and AF approaches to forecast a cumulative number of periods ahead, therefore it is considered as a single output approach. It might be interesting to investigate forecasting by temporal aggregation as a multi-output scenario (de Oliveira Werneck et al., 2022), where a sequence of two or more future data points are of interest.

- *Comment 7: The lack of unanimity between AF and AD seems to suggest that the use of meta-learning or some other technique for model selection would be indicated for this type of problem. This could be discussed in the conclusion.*

  Response: Thank you very much for the valuable suggestion. We have now discussed this in the conclusion section following your suggestion:

  Given that the exact relationship between time series features and the optimal temporal aggregation approach is not known, one of the directions of future research could focus on using meta-learning and other advanced techniques such as (transformers neural networks and attentional mechanism) for unveiling the mystery surrounding a given research question.

- *Comment 8: typos: i) Page 3 - "In particular, given the features of the time sires," ii) Page 4 - "time serieslevel)" - this parenthesis was not opened, iii) Page 9 - "staitonairy time series"*

Response: Thank you. We have now checked the article and corrected multiple linguistic errors throughout the paper, including those suggested in this comment.

References

Babai, M.Z., Boylan, J.E., Rostami-Tabar, B., 2022. Demand forecasting in supply chains: a review of aggregation and hierarchical approaches. International Journal of Production Research 60, 324–348.

Buza, K., 2018. Time series classification and its applications, in: Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics, pp. 1–4.

Hu, Y., Xiao, F., 2022a. An efficient forecasting method for time series based on visibility graph and multi-subgraph similarity. Chaos, Solitons & Fractals 160, 112243.

Hu, Y., Xiao, F., 2022b. Time series forecasting based on fuzzy cognitive visibility graph and weighted multi-subgraph similarity. IEEE Transactions on Fuzzy Systems .

James, G., Witten, D., Hastie, T., Tibshirani, R., 2021. Statistical learning, in: An introduction to statistical learning. Springer, pp. 15–57.

Liu, G., Xiao, F., Lin, C.T., Cao, Z., 2020. A fuzzy interval time-series energy and financial forecasting model using network-based multiple time-frequency spaces and the induced-ordered weighted averaging aggregation operation. IEEE Transactions on Fuzzy Systems 28, 2677–2690.

Makridakis, S., Spiliotis, E., Assimakopoulos, V., 2022. M5 accuracy competition: Results, findings, and conclusions. International Journal of Forecasting 38, 1346–1364.

de Oliveira Werneck, R., Prates, R., Moura, R., Gonçalves, M.M., Castro, M., Soriano-Vargas, A., Júnior, P.R.M., Hossain, M.M., Zampieri, M.F., Ferreira, A., et al., 2022. Data-driven deep-learning forecasting for oil production and pressure. Journal of Petroleum Science and Engineering 210, 109937.