Cover letter for the revised version of IJF-D-22-00224

Étienne David et al.

November 2, 2022

We would like to thank the reviewers and the associate editor for attentively reading our work, the constructive feedback, and their appreciation of the paper. We are also grateful for the fast handling of the manuscript. We have carefully considered all the comments of the reviewers. All the minor errors and typos listed by the reviewer have been corrected as suggested and you can find detailed responses to the comments below.

Major changes

The first major change concerns the dataset description in Section 3. In section 3.1, an overview of the image recognition framework is provided with references to the main architectures used. In section 3.2, a more detailed explanation is given on how the images are translated to fashion time series and how they are normalized to remove social media bias. Finally, a clearer description of the external signal is provided in Section 3.3.

The second major change is the update of Section 4.4 concerning the M4 weekly dataset. We have added to the pool of benchmarks the top 3 models attaining the highest accuracy during the M4 competition on the weekly dataset. As these models are ensembling methods, an ensembling approach mixing 4 HERMES variations is proposed. This proposed ensembling achieves the same level of accuracy as the previous top 3 methods and illustrates that HERMES models can be easily introduced in an ensemble framework to provide more robust and accurate final predictions.

In addition to the two main changes, we want to highlight several additional updates. i) Figure 3 is updated to display a clearer big picture of the HERMES framework. ii) In Appendix A.3, a complete description of the proposed ensembling combining HERMES variations is added for completeness. iii) Finally, in Appendix B.2, we describe the entire grid search run on the M4 weekly dataset to set the hyperparameters of the HERMES architecture.

Reviewer 1

- *"In page 3, authors claim that they use cutting-edge image recognition techniques to create the "first fashion dataset". However, ... Also, the target variable to forecast,*

"share of category" is also not clearly explained or provided the exact equations used to derive these values. "

Author response: We agree perfectly with the referee and this comment has led to an important update of our paper to better describe the dataset. We have now presented more precisely the creation of the dataset: from the image recognition framework (Section 3.1) to the creation of the fashion time series and the "influencer" external signals (Section 3.2 and 3.3).

- *"Authors highlight that they avoid using ensemble-based benchmarks on M4 dataset for a fair comparison. However, HERMES also mimics a similar functionality to that of ensemble models (e.g., Boosting methodology found in the ML literature) as it uses a combination of methods."*

  Author response: We thank the referee for this sensible suggestion that motivates an improvement of Section 4.4. In addition to the existing comparison between the HERMES model and a pool of benchmarks, a second comparison has been done between the 3 ensembling methods reaching the highest accuracy on the M4 weekly dataset and an ensembling combining 4 HERMES variations. The HERMES hybrid model can easily be included in an ensembling framework and the proposed ensembling method reaches the same accuracy level as the best approaches of the M4 competition.

- *"Authors must avoid the ambiguity in the paper. For example, in page 12, authors say "existing Python or R libraries are used to estimate the different parameters". Why do you use a "or" in this sentence? Authors must be more specific about what they implement (exact packages etc)"*

  Author response: We have now corrected parts with ambiguity as suggested.

- *"Authors use a subset of the fashion time series to demonstrate the use of HERMES on smaller dataset. Again, have not explained the details of the selection of subset of data. Are they randomly selected or handpicked?"*

  Author response: There was indeed an oversight on this part. The two subsets of data are selected randomly and the text has been updated accordingly.

- *"Authors have used a grid-based methodology to determine the optimal parameters for the neural network-based methods but have not clearly mentioned the initial parameter ranges used in the experiments. In page 27, authors say that "have been trained with a range of values". To reproduce the results, it is important you provide these details (since you don't provide the source code, authors should at least make an effort to give these information)"*

Author response: To complete Appendix B.1 where we present a grid search run on the loss function, we have added in Appendix B.2 a complete example of grid search run for all the hyperparameters of the HERMES architecture on the M4 weekly dataset.

- *"Authors must use consistent benchmarks across both datasets (ES-RNN benchmark missed in the Fashion dataset). Also haven't provided a statistical test to evaluate the significance of the differences of the results (the results are statistically significant or not)"*

  Author response: We agree with the reviewer, the absence of the ES-RNN benchmark on the Fashion dataset is annoying. However, this choice has been motivated for two reasons: Firstly, the ES-RNN code base is developed in a non-usual code framework (Dynet) that could make the migration to Python difficult. A Python package exists but it is especially designed to work on the M4 dataset and we decided not to update the code to the Fashion dataset use case. Secondly, to compensate this absence: i) we partially replicate it with a method called *hermes-ets* on the fashion dataset and ii) we run numerous experiments on the M4 weekly dataset to provide a complete comparison between the true ES-RNN and the HERMES approach.

- *"In page 15, it is also not clear about the exact classification algorithms used in the study. Authors don't not refer nor provide justification for the use of proposed the threshold (5%) to label the time series into decreasing/flat/increasing trends. Of course, it helps to create a balance dataset, but that is not a valid justification to use this threshold. Need to import more evidence/justification from the fashion domain"*

  Author response: We totally understand the expectations of the referee with this remark. However, the chosen threshold was selected to match the fashion industry practices. We decided to provide additional numerical simulations, for instance in the M4 setting, to illustrate the performance of the model instead of adding results with other thresholds which would not be used in practice.

- *"The performance of the RNN on the M4 weekly time series dataset seems to be poor. Given the proposed HERMES also use this algorithm (Figure 6) as a part of the framework, it is a bit surprising to see how it improves the accuracy after the inclusion of per series TBATS model (Hermes-tbats). I can understand why RNN model is underperforming (because the time series are heterogenous with different starting and end date, and if you globally train the model, without additional/external information, it will not pick up the correct signals)."*

  Author response: We agreed with the reviewer that the M4 weekly dataset is a particularly challenging dataset for RNN approaches. For the RNN benchmark, the neural network has to rebuild the entire prediction while in the HERMES framework,

the RNN part task is only to correct an already computed prediction. Thus, the final accuracy of the 4 HERMES variations is mainly supported by their collections of per-time-series predictors. To illustrate this point, we have added Table A.7 in Appendix A which shows the accuracy of the predictors alone and included in the HERMES framework on the M4 weekly dataset. For each version, the RNN corrector always increases even slightly the final overall accuracy. Moreover, several interesting corrections of the RNN part can be noted as in Figure A.9.

Reviewer 2

- *"On p.7, in the formula above eq.1 It appears that z is centered around Ypred[T+k], with k=i-h[i/j] . I assume that i/j is an integer division. This is a strange formula, where h seems to be assumed to be equal seasonality. Also, why would you center the input by the output prediction value?"*

  Author response: We thank the referee for this remark and we have now improved this part to give more motivation about our preprocessing. We center the input by the output prediction value for two reasons: i) to include the per-time-series prediction in the RNN input ii) to remove from the RNN input the fundamental patterns already learned by the per-time-series predictor.

- *"On p.13, fig.6 What is the meaning of e.g. (None,52,50) or (None,50)?"*

  Author response: There was indeed a mistake in the figure 6, (None,52,50) and (None,50) are displayed in the graph without additional information. In order to shed more light on this point, we have decided to merge the RNN part architure and the overall HERMES architecture in a same figure (Figure 3). Information about the size of the LSTM and Dense layer are no longer represented but can be found in Section 4.1.

- *"On p.14, line 203 Why limit yourself to only one version of ETS (additive, seasonal, no trend I assume)"*

  Author response: We totally agree with the referee. We have now introduced two different versions of the ETS model on the M4 weekly dataset. The first one is an ETS model with an additive trend, seasonality and noise and the second one is the multiplicative ETS with a multiplicative trend, seasonality and noise.

- *"On p. 18, lines 255-260 I would assume that reducing the dataset size 10 times, would necessitate an update the model/training hyperparameters. No?"*

  Author response: We thank the reviewer for this sensible suggestion. In fact, for each new use case or experiment, a grid search can be run to improve the final

accuracy of the HERMES architecture. However, we have noticed that the hyperparameters, except for the choice of the learning rate, do not impact significantly the final accuracy of the model. It is why we have decided not to change the hyperparameters for the experiments on the subsamples of the fashion dataset and also on the M4 weekly dataset, where the same hyperparameters are used for all the HERMES variations.

- *"On p.2, line 302; Table 5 You mention that a model called "Uber" is added. But I do not see it in table 5. Instead there is there a strange phrase "S.Smyl Hyndman et al. (2020)""*

  Author response: We have now corrected this mistake.

- *"General remark: You fit an NN model on residuals from a tbats model. But tbats has some subversions, e.g. may or may not use Box-Cox transformation. I would suggest that passing to NN the fitted tbats model version could be useful."*

  Author response: In the hybrid model proposed by S.Smyl, the per-time-series ETS models are not fitted and the RNN part has consequently two tasks, set the parameters of the ETS models depending on the time series and correct their predictions. In the HERMES framework, the per-time-series predictor is first fitted using a Python package and then passed to the RNN corrector. This allows the proposed hybrid model to be built with any kind of per-time-series predictors, even complex one, depending on the use case.

Reviewer 3

- *"The paper gives the out of sample test results for all horizons, h=1, 2, ... 52 weeks. It is also interesting to see the performance of the methods separately for short (one week), medium (one month) and long term (one year) horizons."*

  Author response: With the applications on the Fashion dataset and the M4 weekly dataset, the HERMES approach has been tested with two different horizons: h=13 and h=52. Nevertheless, looking at the impact of the Influencer signal on the main signal (as displayed in Figure 1), it would be interesting to focus on a short term forecast. This point is not developed in this paper but is currently fueling our research work.

- *"Explain more clearly how you have constructed the values of weak signals."*

  Author response: We agree perfectly with the referee and this comment has led to a major update of our paper. We have now presented more precisely the creation of the fashion time series and the "influencer" external signal (Section 3.2 and 3.3).

- *" Page 12: Explain why you have selected the window size as 104, also the same for M4 data. Are the results sensitive to window size?"*

  Author response: So as to highlight the impact of the different hyperparameters of the HERMES architecture on the final accuracy, we have run a complete grid search on the M4 weekly data and displayed the result in Appendix B. The window size is a determinant parameter and we have noticed that giving at least two seasonal periods to the RNN corrector usually brings better results on average.

- *"Page 16, What is the Accuracy metric in table 2? Is it percentage of correct classification?"*

  Author response: We have added a sentence at the end of Section 4.2 to describe the Accuracy metric.

- *"Page 18: Did you select your 1000 and 100 time series randomly?"*

  Author response: There was indeed a oversight on this part. The two subsets of data are selected randomly and the text has been updated

- *"Write the definition of OWA."*

  Author response: We agree with the referee and we have added a section describing the OWA in Appendix A.

- *"The paper has some spelling mistakes and inconsistencies, e.g. in somes cases (Page 6) and SMAE and MASE (used in page 26). Correct the spelling and use the notations consistently."*

  Author response: We thank the reviewer and the text has been corrected accordingly.