Fall 10-16-2021

A Novel Layered Learning Approach for Forecasting Respiratory Disease Excess Mortality during the COVID-19 pandemic

Afshin Ashofteh
Universidade Nova de Lisboa NOVA IMS, aashofteh@novaims.unl.pt

Jorge M. Bravo
Universidade Nova de Lisboa NOVA IMS, jbravo@novaims.unl.pt

Mercedes Ayuso
University of Barcelona, mayuso@ub.edu

A Novel Layered Learning Approach for Forecasting Respiratory Disease Excess Mortality during the COVID-19 pandemic

Afshin Ashofteh, NOVA Information Management School (NOVA IMS), Universidade Nova de Lisboa, Portugal, aashofteh@novaims.unl.pt

Jorge M. Bravo, NOVA Information Management School (NOVA IMS), Universidade Nova de Lisboa & MagIC & CEFAGE-UE, Portugal & Université Paris-Dauphine, France, jbravo@novaims.unl.pt

Mercedes Ayuso, Department of Econometrics, Statistics and Applied Economy, University of Barcelona, Spain, mayuso@ub.edu

Abstract

Forecasting model selection and model combination are the two contending approaches in the time series forecasting literature. Ensemble learning is useful for addressing a given predictive task by different predictive models when direct mapping from inputs to outputs is inaccurate. We adopt a layered learning approach to an ensemble learning strategy to solve the predictive tasks with improved predictive performance and take advantage of multiple learning processes into an ensemble model. In this proposed strategy, we build each model with a specific holdout and make the ensemble model of time series with a dynamic selection approach. For the experimental section, we studied more than twelve thousand observations in a portfolio of 61-time series of reported respiratory disease deaths to show the amount of improvement in predictive performance of excess mortality. Then we compare the forecasting outcome of our model with the corresponding total deaths of COVID-19 for selected countries.

Keywords: *Time Series method; Machine Learning; Ensemble Bayesian Model Averaging (EBMA); Forecasting; Excess Mortality.*

1. INTRODUCTION

The customary approach to seasonal and non-seasonal time series forecasting is to adopt a single believed to be best model for each series chosen from the set of candidate models using some criteria or procedure (e.g., information criteria, forecasting accuracy measure, cross validation, bootstrapping, construction of confidence intervals, hypothesis testing for nested models), often neglecting model and parameter risk for statistical inference purposes. To this end, a growing number of linear and non-linear univariate and multivariate times series methods and statistical machine learning techniques (Ashofteh, 2018; Ashofteh & Bravo, 2019, 2021a) are proposed to increase the short- and long-term predictive accuracy on a wide range of problems, including stochastic population – mortality, fertility, net migration - forecasting (Bravo & Coelho, 2019; Hyndman et al., 2013), epidemiological and excess mortality forecasting (Scortichini et al., 2020) and longevity-linked securities pricing (Bravo & Nunes, 2021).

Empirical studies in multiple areas show that it is hard to find (if exists) a single widely accepted forecasting method that performs consistently well across all data sets and time horizons (Aiolfi &

Timmermann, 2006; Chatfield, 2016). The use of different selection methods, different fitting periods, alternative accuracy measures, structural breaks in the data generating process and misspecification problems can lead to different model choices and time series forecasts (Ashofteh & Bravo, 2021c). To tackle the model risk problem, i.e., the uncertainty regarding the identification of the true data generating process and the best fitting or forecasting method, to improve the forecasting accuracy, to deal with the limitations of some methods and to generate comparable cross-country and/or subnational forecasts, an alternative approach is to use an ensemble of heterogeneous time series models.

Since the original work of Bates and Granger (1969), several comprehensive theoretical and empirical studies have confirmed the superior predictive performance of ensemble methods using different approaches (Breiman, 1996; Makridakis & Winkler, 1983; Ueda & Nakano, 1996), including stacking and blending to improve-predictions, bagging to decrease variance or boosting to decrease bias (Akyuz et al., 2017) and Bayesian Model Ensemble (Bravo et al., 2021; Ayuso et al., 2021a,b; Bravo & Ayuso, 2021; Bravo, 2020, 2021; Raftery et al., 2005). When adopting this empirical strategy, choices must be made with regards to which models to include in the combination pool and with regards to each model contribution (weight) in the final prediction. A significant body of literature has examined optimal model combination weights (see, e.g., Aiolfi et al., 2010), focusing either on assigning equal weights to the set of superior models (Samuels & Sekkel, 2017), selecting a subset of best models among the set of candidates (model confidence set) using a dynamic trimming scheme and considering the model's out-of-sample forecasting performance in the validation period (Bravo & Ayuso, 2020), or using meta-learning (Brazdil et al., 2009) and regret minimization (Cesa-Bianchi & Lugosi, 2006) approaches to choose the best models for contributing to the ensemble model. Theoretically, any potential model carrying useful information may be considered in the pool of models. Building better model combinations to solve real-world time series problems has become a critical and active research area in recent years (Khairalla et al., 2018).

In this paper we develop and empirically investigate the forecasting performance of a novel flexible and dynamic ensemble learning strategy for seasonal time series forecasting. The strategy is based on a Bayesian Model Ensemble (BME) of heterogeneous models involving both the selection of the subset of best forecasters (model confidence set) to be included in the forecast combination, the identification of the best holdout period for each individual contributed model, and the determination of optimal weights using the out-of-sample predictive accuracy. A model selection strategy is also developed to remove the outlier models and combine the models with reasonable accuracy in the ensemble. The novel approach is empirically investigated using monthly respiratory diseases deaths data for 61 heterogeneous countries. The pool of candidate models includes traditional linear and non-linear univariate time series methods and novel statistical machine learning techniques. We examine and compare run times, accuracy, level of contribution and error metric of the proposed technique in comparison with individual forecasting models.

The ensemble learning procedure involves: (i) setting the different holdouts to be checked for each contributed model; (ii) choosing the best holdout for each model based on the out-of-sample forecasting accuracy; (iii) selecting the subset of best forecasters (model confidence set) using a variable trimming scheme in which the midrange of the set of forecasting accuracy metric values obtained for all candidate models is used as the threshold for model exclusion; (iv) the determination of each model posterior probabilities (model weights) using the normalized exponential (softmax) function; (v) finally, ensemble forecasts are obtained based on the law of total probability considering the model confidence set and the corresponding model weights. Contrary to previous approaches focusing either on the selection of optimal combination schemes and weights or equally weighting a subset of best forecasters, our ensemble procedure involves, for each dataset, both the identification of the best holdout period for each model, the selection of the best forecasting models and the determination of optimal weights based on the out-of-sample forecasting performance.

Our empirical results show proposed approach leads to a decrease in the individual error of ensemble members in comparison with normal model selection with equal holdouts for selected models, and without overly decreasing the diversity among them. Hopefully, this article brings more clarity on which time series techniques contribute better to ensembles, and present a suitable ensemble time series with improved predictive accuracy.

The remaining sections of the paper are organized as follows. In section 2, we provide the materials, methods and related works considered in this research. Section 3 describes our proposed method. The results of an extensive set of experiments on respiratory disease deaths of 61 countries are given and discussed in Section 4. Finally, discussion and the main conclusion is presented in section 5.

2. MATERIALS AND METHODS

The proposed method is based on a meta-learning approach to adopt the ensemble to the best combination of forecasting models. The candidate models are extracted from different layers with the best holdout for each contributed model and each panel member. We use multiple learning processes to improve the predictive performance of the ensemble. It is built by an ensemble learning approach from the addressed candidates with the last layer. In this section, we discuss these techniques in brief and highlight their contributions as well.

2.1. Layered learning and the proposed ensemble learning strategy

The layered learning approach in time series consists of breaking a forecasting problem down into simpler subtasks in several layers. Each layer addresses a different predictive task and the output of one layer could be used as the input of the next layer (Cerqueira et al., 2020). In this research, the first task is to obtain a direct mapping from the time series of different countries, combining the intractable time series algorithms, and predicting the ensemble model as the final output. Therefore, the task of the first layer is finding the best holdout for each individual panel member and for each time series algorithm. It facilitates the task of model selection in the second layer, which facilitates the identification of the model confidence set of best forecasters in the last layer. It is useful to maximize the forecasting accuracy in panel time series dynamically and adopt the learning process of the model to possible unexpected shocks.

Following Ashofteh and Bravo (2021) and Bravo et al. (2021), let each candidate model be denoted by $M_l, l = 1, \ldots, K$ representing a set of probability distributions in which the "true" data-generating process is assumed to be included, comprehending the likelihood function $L(y|\theta_l, M_l)$ of the observed data y in terms of model specific parameters θ_l and a set of prior probability densities for said parameters $p(\theta_l|M_l)$. Consider a quantity of interest Δ present in all models, such as the future observation of y. The marginal posterior distribution across all models is

$$p(\Delta|y) = \sum_{k=1}^{K} p(\Delta|y, M_k)p(M_k|y) \tag{1}$$

where $p(\Delta|y, M_k)$ denotes the forecast PDF based on model M_k alone, and $p(M_k|y)$ is the posterior probability of model M_k given the observed data. The posterior probability for model M_k is denoted by $p(M_k|y)$ with $\sum_{k=1}^{K} p(M_k|y) = 1$. The weight assigned to each model M_k is given by its posterior probability

$$p(M_k|y) = \frac{p(\mathrm{y}|M_k)p(M_k)}{\sum_{l=1}^{K} p(y|M_l)p(M_l)}. \tag{2}$$

The workflow of our proposed method is presented in Figure 1. To identify the model confidence set and compute model weights, for each dataset we first specify the different holdouts to be checked for each contributed model. Let $H = \{h_1, h_2, \ldots, h_m\}$ represent the set of holdout periods to be considered in the estimation procedure.

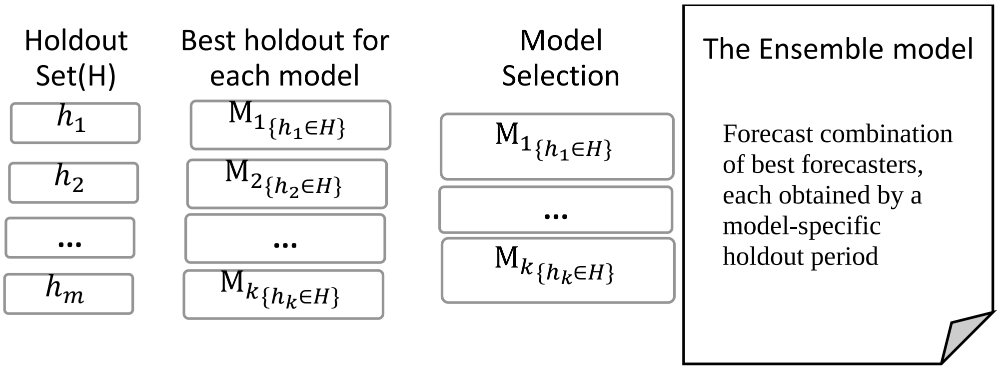

Figure 1 – Proposed strategy of ensemble learning.

We use the symmetric mean absolute percentage error (SMAPE) as forecasting accuracy measure. For choosing the best holdout for each individual model, we tested the different values of the holdout set ($H = \{3,5,7\}$ years) and compared the SMAPE's values at each iteration, keeping the model with the lowest SMAPE as the candidate for the model confidence set selection step.

The subset of best forecasters is selected using the best holdout period and a variable trimming scheme in which the midrange of the forecasting accuracy metric is used as the threshold for model exclusion, i.e., using

$$\Gamma_g = \frac{\max\{SMAPE_{g,l}\}_{l=1,\dots,K} - \min\{SMAPE_{g,l}\}_{l=1,\dots,K}}{2}, \tag{2}$$

where $SMAPE_{g,l}$ is the SMAPE value for model l in dataset (country) g. For each dataset, if the forecasting accuracy of a candidate model is greater than the midrange indicator, i.e., if $SMAPE_{g,l} > \Gamma_g$ the model is excluded from the model confidence set and the ensemble forecast computation receives a zero weight in (1). In this case, the far forecasting models will be removed from the ensemble. If the models are all close to the original time series, then although some of them would be removed, however, the mean of remaining models could be considered as a good candidate for even removed ones. This could be magnificent to avoid overfitting and control the redundancy in the output of the ensemble model. The intuition is the removal of only models, which are extremely far from other candidate models. It will save the diversity of the selected models and prevent the overfitting problem.

Fourth, the best forecasters model posterior probabilities (model weights) are computed using the Softmax function, i.e., we compute $p(M_k|y)$ using

$$p(M_k|y) = \frac{exp(-|\xi_k|)}{\sum_{l=1}^{K} exp(-|\xi_l|)}, \qquad k = 1, \dots, K, \tag{3}$$

with $\xi_k = S_k / max\{S_l\}_{l=1,\dots,K}$ and $S_k := SMAPE_{g,k}$. The Softmax function is a generalization of the logistic function often used in classification and forecasting exercises using traditional, machine learning and deep learning methods as a combiner or an activation function. The function assigns larger weights to models with smaller forecasting error, with the weights decaying exponentially the larger the error. Fifth, the Bayesian model ensemble forecasts are obtained based on the law of total probability (1) considering the model confidence set and the corresponding model weights (3). The pseudo-code of the proposed methodology is listed in Table 1.

```
INPUT panel time series (panel members = countries); OUTPUT ensemble model;
STATEXPLORE time series decomposition;
IMPUTE[missing] = TRUE;
First_year = 2000 (for most of time series but some of them start later)
Last_year = 2016
Target_year = 2020
Confidence_level = 0.95
Holdout_set={3, 5, 7}
Ensemble_criteria_for_computing_weights = "Symmetric Mean Absolute Percentage Error (SMAPE)"
Set.seed()
Model_list ={SNAIVE, RWF, HWA, HWM, ETS, ARIMA, TBATS, STL, NNAR, MLP, ELM, SSA, ENS)
FUNCTION model_weights (error)
    Pr = error/max(error)
    exp(-abs(pr))/sum(exp(-abs(pr)))
# First loop repeat for each country
FOR each panel in list of countries DO
    SET panel.data = SUBSET dataset(country = panel & Year > First_year & Months="Jan-Dec"
    SET Year_min = min(Year of panel.data)
    panel_data = MISSING VALUE IMPUTATION by na_seasplit
    SET (START of the run-time calculation)
# Second loop for selecting holdouts
    FOR each holdout in Houldout_set DO
        IF ( ymax-ho+1 < ymin+3 ) { break }
        ELSE
            SET train_dataset WINDOW (START = Year_min , END = Last_year – holdout)
            SET test_dataset   WINDOW (START = Last_year – holdout + 1)
            FIT models in Model_list
            CALCULATE accuracy (model , holdout)
            IF accuracy (model[holdout]) > last_accuracy (model[holdout – 1]) THEN
                SET model = model[holdout]
            ELSE
                SET model = model[holdout -1]
    CALCULATE error(ALL models), min_error(ALL models), max_error(ALL models)
    CALCULATE id_error = (min_error + max_error)/2
    FOR model in Model_list
        IF ( error_model > id_error) THEN
            PRINT ("Model is excluded!")
        ELSE
             ADD model into selected_model_list
# The model ensemble
IF selected_model_list = NULL {next country}
ELSE
     CALCULATE model_weights for ensemble
     SET First_year  based on model with min_holdouts
     SET First_month based on model with min_holdouts
     CALCULATE ensemble
     SET (END of the run-time calculation)
# The outputs
PRINT GRAPHS; SAVE OUTPUTS
```

Table 1 – Pseudo Code of the Proposed Ensemble Strategy.

2.2. *The learning algorithms*

This section summarizes the characteristics of the individual candidate learning algorithms (times series methods) used in this study. For a detailed presentation and discussion of the methods see, for instance, Hyndman and Athanasopoulos (2021). Table 2 summarizes the hyper-parameters of the algorithms used in this study.

ALGORITHM	PARAMETERS	VALUE
Seasonal Trend Decomposition using Loess	lambda t.window s.window biasadj	"auto" 6 6 TRUE
Seasonal naive	drift lambda level biasadj	F 0 clevel TRUE
The Auto-Regressive Integrated Moving Average	Auto	
The Exponential Smoothing State Space Model The ETS method with automatic and ZZA parameter setting from the forecast statistical software R package (Hyndman et al., 2020), and the TBATS method, which includes Box-Cox transformation, ARMA errors, trend and seasonal components (de Livera, Hyndman, & Snyder, 2011).	 Model Box-Cox tran. Multiplicative trend restricted for the models with infinite variance	{ETS, TBATS} ZZA TRUE Allow TRUE
Holt-Winters' multiplicative method	Seasonal level	Multiplicative clevel
Holt-Winters' additive method	Seasonal level	Additive clevel
Random Walk Forecasts	Drift Lambda Level biasadj	F "auto" clevel TRUE
Extreme Learning Machines	type hd comb reps difforder	Lasso 500 mean 200 NULL
Multilayer Perceptron for time series	Comb hd.auto.type hd.max	Mode Valid 5
Neural network model to a time series	P size decay lambda repeats MaxNWts	2 1 0.001 Auto 100 2000
Singular spectrum analysis	Kind svd.method L neig force.decompose	1d-ssa Auto 12 NULL

	mask	TRUE NULL

Table 2 – Algorithms and hyper-parameters choices.

The model fitting, forecasting and simulation procedures have been implemented using a R statistical software using libraries such as the `forecast` library (Hyndman et al., 2020).

3. Empirical Experiments

In this study, we use cause-of-death data from the World Health Organization (WHO) mortality database (World Health Organization, 2018), which includes the death time series of different countries for all genders. First, we distinguished the quality of data for each country according to the metadata of the dataset. We ranked the data quality of countries as is shown in Table 3.

RANK	EVALUATION	DESCRIPTION
1	Excellent quality	These countries may be compared, and time series may be used for priority setting and policy evaluation.
2	Moderate quality	Data have low completeness and/or issues with cause-of-death assignment, which likely affect estimated deaths by cause and time trends. Comparisons among countries should be interpreted with caution.
3	Low quality	Data have severe quality issues. Comparisons among countries should be interpreted with caution.
4	Unacceptable	Death registration data are unavailable or unusable due to quality issues. Estimates may be used for priority setting; however, they are not likely to be informative for policy evaluation or comparisons among countries.
5	Unacceptable	Data should be ignored

Table 3 – Different levels of quality allocated for the reported respiratory disease deaths by countries. Source:(World Health Organization, 2018).

We considered only countries with quality ranked in the first three categories. Some countries reported the total death for three months in one row for some years. We divided this aggregate value to three equal values for each corresponding month. We filtered the datasets for respiratory diseases and considered the death variable as a univariate time series with monthly sampling frequency. Table 4 shows the codes that were classified as respiratory infections.

CODE	DESCRIPTION
380	For Respiratory infections (This code is the aggregate of 390 and 400)
390	For Lower respiratory infections
400	For Upper respiratory infections
410	Otitis media: Acute otitis media (AOM) is a common complication of upper respiratory tract infection whose pathogenesis involves both viruses and bacteria.

Table 4 – Metadata of code of diseases categorized as respiratory disease. Source: (World Health Organization, 2018).

For obtaining the total deaths caused by respiratory diseases, we had to aggregate either the codes 380 and 410 or equivalently the codes 390, 400 and 410. From this, we calculated the proportion of deaths caused by respiratory diseases. To estimate the number of monthly deaths caused by respiratory diseases, we multiply the annual proportion by the total forecasted deaths each month. The procedure provided us a dataset with more than twelve thousand observations in a pool of 61 panel members' time-series (countries) from 2000 to 2016. These panel time series cover the different possible situations of stationarity, non-stationarity, increasing trends, seasonality and structural breaks to evaluate the accuracy improvement of candidate and ensemble models in different scenarios comprehensively.

4. Results

4.1. Forecasting accuracy comparison

We present three approaches in Table 5. In the first approach entitled "only holdout", we only use a set of different forecasting models to make the ensemble model by different holdouts. As we can see, there are some models exhibiting better performance when compared with the ensemble model. Even in average error, the TBATS shows lower error than the ensemble.

The second approach is named "Holdout and selection". This approach uses the midrange of the SMAPE values to evaluate the distance of each model to the remaining others as shown above in the Pseudo code (Table 1). Model's with SMAPE value higher than the midrange indicator are considered poor forecasters and eliminated from the ensemble forecast. The results in Table 5 clearly highlight the improvement in the accuracy of the Bayesian model ensemble (BME) when pursuing the Holdout and selection approach, ranking first among all tested methods.

The final proposed approach is a combination of the two previous ones. It combines the best forecasting models fitted using each model optimal holdout selection. It makes the models free of the equal holdout restriction. The accuracy of the ensemble is dramatically improved, leaving the individual learning algorithms at a reasonable distance.

MODELS	THE MODEL'S ERROR (SMAPE) IN AVERAGE									TOTAL ERROR OF MODELS	RANK
	(1) ONLY HOLDOUT				(2) HOLDOUT AND SELECTION				(3)MODEL SELECTION; DYNAMIC HOLDOUTS		
	HO = 3	HO = 5	HO = 7	AVERAGE	HO = 3	HO = 5	HO = 7	AVERAGE			
BME	0.112	0.181	0.191	0.161	0.103	0.125	0.136	0.121	0.102	0.128	1
TBATS	0.120	0.150	0.172	0.147	0.114	0.143	0.177	0.145	0.119	0.137	2
ETS	0.125	0.200	0.185	0.170	0.110	0.138	0.158	0.135	0.117	0.141	3
ARIMA	0.133	0.178	0.214	0.175	0.107	0.145	0.166	0.139	0.114	0.143	4
SNAIVE	0.124	0.181	0.212	0.172	0.114	0.142	0.164	0.140	0.121	0.144	5
STL	0.117	0.180	0.201	0.166	0.118	0.155	0.169	0.147	0.121	0.145	6

NNETAR	0.141	0.194	0.210	0.182	0.106	0.150	0.181	0.146	0.106	0.145	7
HWA	0.134	0.193	0.222	0.183	0.117	0.154	0.179	0.150	0.128	0.154	8
MLP	0.130	0.220	0.240	0.197	0.123	0.140	0.169	0.144	0.123	0.155	9
HWM	0.148	0.195	0.256	0.200	0.124	0.157	0.156	0.146	0.128	0.158	10
ELM	0.139	0.227	0.242	0.203	0.114	0.150	0.203	0.156	0.122	0.16	11
SSA	0.160	0.190	0.231	0.194	0.136	0.168	0.188	0.164	0.139	0.166	12
RWF	0.153	0.289	0.362	0.268	0.111	0.141	0.184	0.145	0.123	0.179	13

Table 5 – Ranking the models and ensembles according to the accuracy measure.

4.2. Model excluded in model selection

In Table 6, we can see that all models are excluded several times in different situations, and it shows that the model selection approach is an appropriate strategy to employ the efficient models in the ensemble. The methods are ranked base on their contribution rate in the ensemble. The vertical comparison of the results gives us an insight about the contribution of the different models to the ensemble, while the horizontal comparison is useful to assess the rate of contribution across different holdout periods. The variation in the contribution rates from the best model to the worst one and from the lowest holdout period to the highest one suggest a potential positive effect on the final forecasting accuracy of the ensemble model by selecting the best holdout for each individual model along with selecting the best forecasters to the model confidence set finally used to forecast.

MODELS	THE MODEL'S EXCLUSION FREQUENCY								
	(2)IN MODEL SELECTION LAYER FOR EACH HOLDOUT								RANK
	HO=3		HO=5		HO=7		AVE		
	FREQ.	PROP.	FREQ.	PROP.	FREQ.	PROP.	FREQ.	PROP.	
ETS	10	4.61%	8	3.56%	8	4.30%	9	4.31%	1
TBATS	12	5.53%	13	5.78%	9	4.84%	11	5.26%	2
STL	13	5.99%	11	4.89%	11	5.91%	12	5.74%	3
ARIMA	13	5.99%	13	5.78%	14	7.53%	13	6.22%	4
SNAIVE	18	8.29%	13	5.78%	14	7.53%	15	7.18%	5
HWA	13	5.99%	19	8.44%	17	9.14%	16	7.66%	6
HWM	19	8.76%	17	7.56%	17	9.14%	18	8.61%	7
NNETAR	23	10.60%	21	9.33%	13	6.99%	19	9.09%	8
MLP	22	10.14%	24	10.67%	14	7.53%	20	9.57%	9
ELM	17	7.83%	28	12.44%	18	9.68%	21	10.05%	10
SSA	27	12.44%	21	9.33%	18	9.68%	22	10.53%	11
RWF	30	13.82%	37	16.44%	33	17.74%	33	15.79%	12

Table 6 – Contribution rate of the models in the ensemble.

Table 7 presents the contribution ranks, exclusion frequency and proportion of the selected models with the best holdout for each. As it is clear, the contribution of the models in the ensemble is changed in comparison with Table 6. It gives a proper explanation for the improvement in the accuracy of the ensemble by using the proposed method.

	TBATS	STL	ETS	HWA	ARIMA	SNAIVE	HWM	ELM	MLP	SSA	NNETAR	RWF
Frequency	13	15	17	17	18	19	19	21	23	25	29	37
Proportion	5%	6%	7%	7%	7%	8%	8%	8%	9%	10%	11%	14%
Rank	1	2	3	4	5	6	7	8	9	10	11	12

Table 7 – The model's exclusion frequency for the Ensemble with dynamic holdouts.

4.3. Algorithmic efficiency analysis

We analyse the algorithmic efficiency of each method, i.e., the amount of computational resources used by the algorithm, by measuring the time spent in fitting the ensemble model with each approach and using it to predict the maximum likely run-time of a new given time series (Table 8).

MODELS	**RUN-TIME ANALYSIS OF OBTAINING ENSEMBLE MODEL (IN MINUTE)**											
	ONLY HOLDOUT				**HOLDOUT AND SELECTION**				**MODEL SELECTION & DYNAMIC HOLDOUTS**			
	HO=3	**HO=5**	**HO=7**	**AVE**	**HO=3**	**HO=5**	**HO=7**	**AVE.**	**HO=3**	**HO=5**	**HO=7**	**AVE**
ART	2.97	2.86	2.39	2.74	3.03	2.65	2.41	2.70	3.29	2.96	2.64	2.96
STD	0.72	0.72	0.52	0.65	0.70	0.60	0.54	0.61	0.84	0.71	0.70	0.75
LCL	2.79	2.68	2.26	2.58	2.85	2.50	2.27	2.54	3.08	2.78	2.46	2.77
UCL	3.15	3.04	2.52	2.90	3.21	2.80	2.55	2.85	3.50	3.14	2.82	3.15
ART: Average run-time, STD: Standard deviation, LCL: Lower confidence limit, UCL: Upper confidence limit.												

Table 8 – The methodology effect on the run-time and computational efficiency.

If we look at the average of run-time and their mean confidence intervals for the proposed method and the other two approaches, we could see that they are not significantly different. It shows that our proposed method is efficient in terms of computation time.

4.4. Excess mortality analysis

We used the proposed ensemble learning for panel time-series with selecting strategy and dynamic holdouts to forecast the number of deaths caused by different kinds of respiratory diseases for a subset of 61 countries in 2020. Additionally, the COVID-19 deaths were extracted for the same year from the COVID-19 Weekly Epidemiological Update of the World Health Organization (WHO) with data as received from national authorities, as of 3 January 2021, which has a proper coverage on the whole period of 2020 (World Health Organization, 2021).

As it is shown in Table 9, we considered the European countries, Canada, the United States of America, and the United Kingdom from the list to calculate the correlation between actual COVID-19 deaths and our forecasts of respiratory deaths. The selection criteria was related to the official statistics maturity, the models of corruption in official statistics (Georgiou, 2021), and quality level of deaths data according to the WHO ranking discussed in section 4.1. The correlation was 94% (P-value =0.000). It could be because of a high quality of the Official Statistics in these countries as Ashofteh and Bravo (2020) show the significant quality variation of reported data about COVID-19 worldwide and the role of data science and new technologies in improving their quality (Ashofteh & Bravo, 2021b).

ROW	COUNTRY	(1) ALPHA-3	COUNTRY NO	POPULATION	(2) RD TD	(3) COVID TD	STD_RD TD	STD_COVID TD
1	Austria	AUT	40	8955.108	234	6214	-0.392	-0.227
2	Belgium	BEL	56	11539.326	1571	19693	-0.172	0.052
3	Bulgaria	BGR	100	7000.117	412	7644	-0.363	-0.198
4	Canada	CAN	124	37411.038	1766	15679	-0.14	-0.031
5	Denmark	DNK	208	5771.877	595	1345	-0.333	-0.328
6	Finland	FIN	246	5532.159	53	561	-0.422	-0.344
7	France	FRA	250	65129.731	4733	64543	0.347	0.98
8	Germany	DEU	276	83517.046	5815	34272	0.524	0.354
9	Greece	GRC	300	10473.452	2000	4921	-0.102	-0.254
10	Hungary	HUN	348	9684.68	344	9884	-0.374	-0.151
11	Iceland	ISL	352	339.037	17	29	-0.428	-0.355
12	Ireland	IRL	372	4882.498	316	2252	-0.379	-0.309
13	Italy	ITA	380	60550.092	4792	74985	0.356	1.196
14	Netherlands	NLD	528	17097.123	1206	11565	-0.232	-0.117
15	Norway	NOR	578	5378.859	528	436	-0.344	-0.347
16	Poland	POL	616	37887.771	5347	29119	0.448	0.247
17	Portugal	PRT	620	10226.178	2097	7045	-0.086	-0.21
18	Romania	ROU	642	19364.558	1484	15919	-0.187	-0.026
19	Serbia	SRB	688	8772.228	419	3288	-0.362	-0.288
20	Slovakia	SVK	703	5457.012	476	2317	-0.352	-0.308
21	Slovenia	SVN	705	2078.654	145	2889	-0.407	-0.296
22	Spain	ESP	724	46736.782	3042	50442	0.069	0.688
23	Sweden	SWE	752	10036.391	665	8727	-0.321	-0.175
24	Switzerland	CHE	756	8591.361	428	7049	-0.36	-0.21
25	The UK	GBR	826	67530.161	6943	74570	0.71	1.188
26	Ukraine	UKR	804	43993.643	1089	18854	-0.252	0.034
27	US of America	USA	840	329064.917	16554	345253	2.288	6.791

Source: Author's preparation. Notes: (1) Abbreviation code of the country (Three letters); (2) Respiratory diseases deaths; (3) WHO COVID-19 deaths.

Table 9 – Comparison between forecasting deaths for respiratory diseases and actual COVID-19 deaths.

The comparison of respiratory diseases and COVID-19 deaths are shown in Figures 2 and 3.

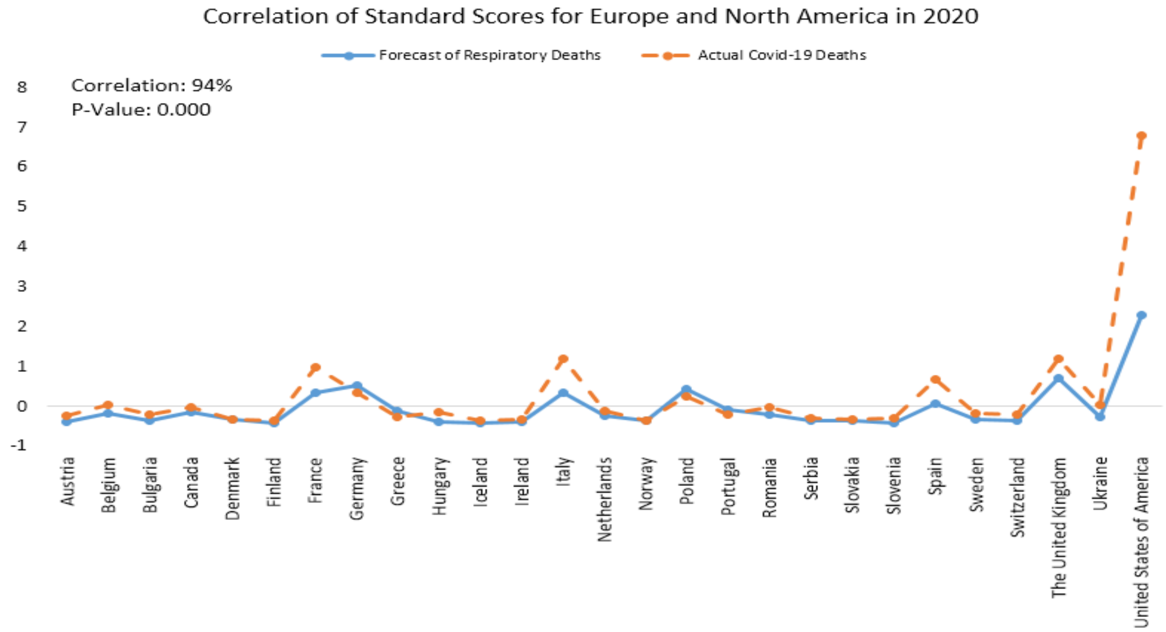

Figure 2 – Respiratory diseases deaths and COVID-19 deaths for Europe and North America in 2020.

Figure 2 shows that some countries in our sample have dealt with this COVID-19 in 2020 better than others in respect to their vulnerability to respiratory diseases. The countries with the forecast of respiratory diseases significantly higher than the COVID-19 deaths show notable performance to manage this chaotic year despite the statistically significant positive correlation between these two indicators. Although it is not the case for the European countries and North America, we can see in Figure 3 that Japan and the Philippines are two fitted examples for this case.

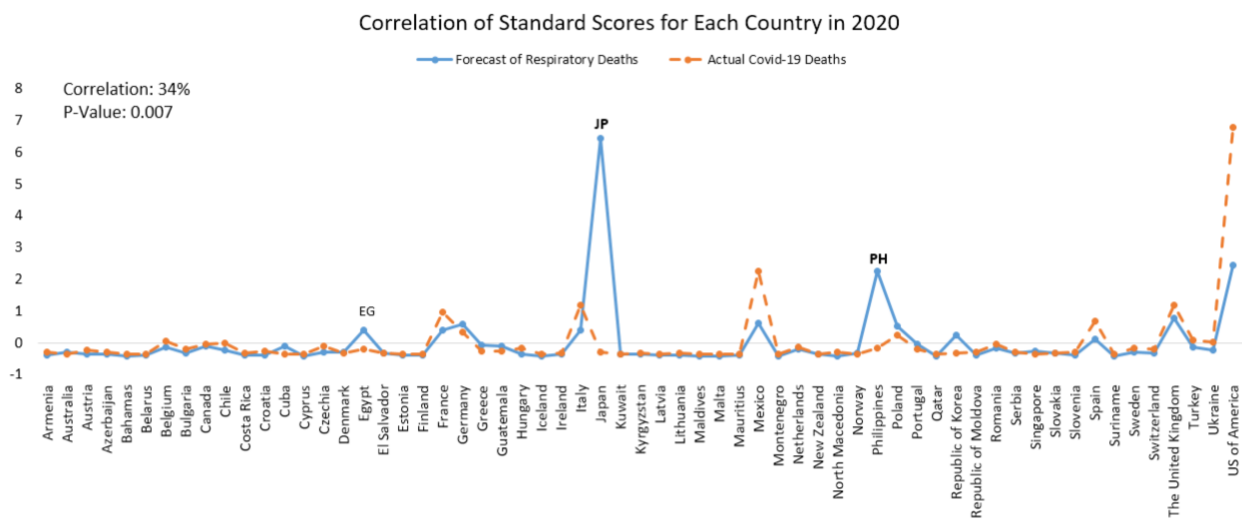

Figure 3 – Respiratory diseases deaths and COVID-19 deaths for Each Country in 2020.

5. DISCUSSION AND CONCLUSION

According to the performance of the models, we provided clear evidence on the competitiveness of our method in terms of predictive performance when compared to the state of the arts and even the usual ensemble models with fix holdouts for all models and without our proposed model selection layer. In comparison of candidate models to contribute to the ensemble, Tables 5 and 6 show the positive effect on prediction accuracy by selecting the best holdout for each model and removing the outlier models from the ensemble. The proposed ensemble model shows a significant improvement in the accuracy in comparison with the other ensembles and each individual state-of-arts.

We used this new ensemble strategy to forecast the number of death for respiratory diseases for 2020 of our sample, included 61 countries. The correlation between the standardized values of the respiratory diseases death and the COVID-19 deaths were positive and statistically significant. It recommends us to consider the forecasted values of the respiratory diseases as a covariate to evaluate the effective strategies of different countries, such as lockdown rules or relaxing of border control regulations. Japan and the Philippines are candidates with our study for more investigation in this regard, and they are more eligible than other countries with only a low death toll. It could be probable that the experience of these countries with high mortality of respiratory diseases played a role in managing the pandemic.

It is considerable in this pandemic to focus more on the death toll than the cumulative number of patients. According to the nature of pandemics, it is challenging to control its spread, however the main concern could be controlling the severe cases and the patients with a high likelihood of death. These countries with a high number of respiratory diseases that could manage the pandemic reasonably could be more recommendable for further studies on their policies and health strategies in comparison with the countries with only a low rate of mortality.

ACKNOWLEDGMENTS

The authors are grateful to the anonymous reviewers for their constructive comments. Their critical and constructive remarks were precious to improve the final paper. Jorge M. Bravo was supported by Portuguese national science funds through FCT under the project UIDB/04152/2020-Centro de Investigação em Gestão de Informação (MagIC). Additionally, M. Ayuso is grateful to the Secretaria d'Universitats i Recerca del departament d'Empresa i Coneixement de la Generalitat de Catalunya for funding received under grant 2020-PANDE-00074. It's a research project directly related to COVID and economy.

REFERENCES

Aiolfi, M., Capistrán, C., Timmermann, A., Aiolfi, M., Capistrán, C., & Timmermann, A. (2010). Forecast Combinations. *Working Paper*. Retrieved from https://econpapers.repec.org/RePEc:bdm:wpaper:2010-04

Aiolfi, M., & Timmermann, A. (2006). Persistence in forecasting performance and conditional combination strategies. *Journal of Econometrics*, *135*(1–2), 31–53. https://doi.org/10.1016/j.jeconom.2005.07.015

Akyuz, A. O., Uysal, M., Bulbul, B. A., & Uysal, M. O. (2017). Ensemble approach for time series analysis in demand forecasting: Ensemble learning. In *Proceedings - 2017 IEEE International Conference on INnovations in Intelligent SysTems and Applications, INISTA 2017* (pp. 7–12). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/INISTA.2017.8001123

Ashofteh, A. (2018). Mining Big Data in statistical systems of the monetary financial institutions (MFIs). In *International Conference on Advanced Research Methods and Analytics (CARMA)*. Valencia: Universitat Politecnica de Valencia. https://doi.org/10.4995/carma2018.2018.8570

Ashofteh, A., & Bravo, J. M. (2019). A non-parametric-based computationally efficient approach for credit scoring. In *Atas da Conferencia da Associacao Portuguesa de Sistemas de Informacao*. Associacao Portuguesa de Sistemas de Informacao. Retrieved from https://www.scopus.com/record/display.uri?eid=2-s2.0-85086641145&origin=inward&txGid=0e87a8c228db37a09073b1441dfffe9e

Ashofteh, A., & Bravo, J. M. (2020). A study on the quality of novel coronavirus (COVID-19) official datasets. *Statistical Journal of the IAOS*, *36*(2), 291–301. https://doi.org/10.3233/SJI-200674

Ashofteh, A., & Bravo, J. M. (2021a). A conservative approach for online credit scoring. *Expert Systems with Applications*, *176*, 114835. https://doi.org/10.1016/j.eswa.2021.114835

Ashofteh, A., & Bravo, J. M. (2021b). Data science training for official statistics: A new scientific paradigm of information and knowledge development in national statistical systems. *Statistical Journal of the IAOS*, *Preprint*(Preprint), 1–19. https://doi.org/10.3233/SJI-210841

Ashofteh, A., & Bravo, J. M. (2021c). Life Table Forecasting in COVID-19 Times: An Ensemble Learning Approach. In *16th Iberian Conference on Information Systems and Technologies (CISTI)* (pp. 1–6). IEEE. https://doi.org/10.23919/CISTI52073.2021.9476583

Ayuso, M., Bravo, J. M., & Holzmann, R. (2021). Getting life expectancy estimates right for pension policy: Period versus cohort approach. *Journal of Pension Economics and Finance*, *20*(2), 212–231, doi: 10.1017/S1474747220000050. https://doi.org/10.1017/S1474747220000050

Bates, J. M., & Granger, C. W. J. (1969). The Combination of Forecasts. *Journal of the Operational Research Society*, *20*(4), 451–468. https://doi.org/10.1057/jors.1969.103

Bravo, J. M. (2021). Pricing Participating Longevity-Linked Life Annuities: A Bayesian Model Ensemble Approach. *European Actuarial Journal*, *Revised an*.

Bravo, J. M., & Ayuso, M. (n.d.). Forecasting the retirement age: A Bayesian Model Ensemble Approach. In *Trends and Applications in Information Systems and Technologies, WorldCIST 2021* (pp. 1–13, doi: 10.1007/978-3-030-72657-7_12). Springer Nature. https://doi.org/10.1007/978-3-030-72657-7_12

Bravo, J. M., & Ayuso, M. (2020). Mortality and life expectancy forecasts using bayesian model combinations: An application to the portuguese population. *RISTI - Revista Iberica de Sistemas e Tecnologias de Informacao*, *2020*(E40), 128–144, doi: 10.17013/risti.40.128-145. https://doi.org/10.17013/risti.40.128-145

Bravo, J. M., & Ayuso, M. (2021). Forecasting the Retirement Age: A Bayesian Model Ensemble Approach. In *Advances in Intelligent Systems and Computing* (Vol. 1365 AIST, pp. 123–135). Springer, Cham. https://doi.org/10.1007/978-3-030-72657-7_12

Bravo, J. M., & Coelho, E. (2019). Forecasting Subnational Demographic Data using Seasonal Time Series Methods. *CAPSI 2019 Proceedings*. Retrieved from https://aisel.aisnet.org/capsi2019/24

Bravo, J. M., & Nunes, J. P. V. (2021). Pricing longevity derivatives via Fourier transforms. *Insurance: Mathematics and Economics*, *96*, 81–97, doi: 10.1016/j.insmatheco.2020.10.008. https://doi.org/10.1016/j.insmatheco.2020.10.008

Brazdil, P., Carrier, C., Soares, C., & Vilalta, R. (2009, November 22). Metalearning - Applications to Data Mining. *Cognitive Technologies*. Springer Berlin Heidelberg. Retrieved from https://books.google.com/books?hl=en&lr=&id=-Gsi_cxZGpcC&oi=fnd&pg=PA1&dq=Brazdil,+P.B.+(Ed.),+2009.+Metalearning:+applications+to+data+mining,+Cognitive+technologies.+Springer,+&ots=wkZEoWxtMg&sig=o7ZkVPhvJJROwAdWP1hnGlHm7To

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140, doi: 10.1007/bf00058655. https://doi.org/10.1007/bf00058655

Cerqueira, V., Torgo, L., & Soares, C. (2020). Early Anomaly Detection in Time Series: A Hierarchical Approach for Predicting Critical Health Episodes. Retrieved from https://arxiv.org/pdf/2010.11595.pdf

Cesa-Bianchi, N., & Lugosi, G. (2006). *Prediction, Learning, and Games*. Cambridge University Press.

Chatfield, C. (Ed.). (2016). *The Analysis of Time Series: An Introduction* (6th ed.). CRC Press. Retrieved from https://books.google.pt/books/about/The_Analysis_of_Time_Series.html?hl=de&id=qKzyAbdaDFAC&redir_esc=y

de Livera, A. M., Hyndman, R. J., & Snyder, R. D. (2011). Forecasting time series with complex seasonal patterns using exponential smoothing. *Journal of the American Statistical Association*, *106*(496), 1513–1527. https://doi.org/10.1198/jasa.2011.tm09771

Georgiou, A. V. (2021). The manipulation of official statistics as corruption and ways of understanding it. *Statistical Journal of the IAOS*, *Preprint*(Preprint), 1–21. https://doi.org/10.3233/sji-200667

Hyndman, R., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., … Yasmeen, F. (2020). Forecasting Functions for Time Series and Linear Models. Retrieved from https://pkg.robjhyndman.com/forecast/

Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: principles and practice: Hyndman, Rob J, Athanasopoulos, George: 8601404468544: Amazon.com: Books* (3rd ed.). Melbourne, Australia: OTexts. Retrieved from https://otexts.com/fpp3/

Hyndman, R. J., Booth, H., & Yasmeen, F. (2013). Coherent Mortality Forecasting: The Product-Ratio Method With Functional Time Series Models. *Demography*, *50*(1), 261–283, 10.1007/s13524-012-0145–5. https://doi.org/10.1007/s13524-012-0145-5

Khairalla, M. A., Ning, X., AL-Jallad, N. T., & El-Faroug, M. O. (2018). Short-Term Forecasting for Energy Consumption through Stacking Heterogeneous Ensemble Learning Model. *Energies*, *11*(6), 1605. https://doi.org/10.3390/en11061605

Makridakis, S., & Winkler, R. L. (1983). AVERAGES OF FORECASTS: SOME EMPIRICAL RESULTS. *Management Science*, *29*(7), 987–996. https://doi.org/10.1287/mnsc.29.9.987

Raftery, A. E., Gneiting, T., Balabdaoui, F., & Polakowski, M. (2005). Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, *133*(5), 1155–1174. https://doi.org/10.1175/MWR2906.1

Samuels, J. D., & Sekkel, R. M. (2017). Model Confidence Sets and forecast combination. *International Journal of Forecasting*, *33*(1), 48–60. https://doi.org/10.1016/j.ijforecast.2016.07.004

Scortichini, M., Schneider Dos Santos, R., De' Donato, F., De Sario, M., Michelozzi, P., Davoli, M., … Gasparrini, A. (2020). Excess mortality during the COVID-19 outbreak in Italy: A two-stage interrupted time-series analysis. *International Journal of Epidemiology*, *49*(6), 1909–1917. https://doi.org/10.1093/ije/dyaa169

Ueda, N., & Nakano, R. (1996). Generalization error of ensemble estimators. In *IEEE International Conference on Neural Networks - Conference Proceedings* (Vol. 1, pp. 90–95). Washington, DC, USA: IEEE. https://doi.org/10.1109/icnn.1996.548872

World Health Organization. (2021). *Coronavirus Disease (COVID-19) Situation Reports*. *Coronavirus disease (COVID-19)/Situation reports*. Retrieved from https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports/

World Health Organization, D. of I. E. and R. (2018). Global Health Estimates 2016: Deaths by Cause, Age, Sex, by Country and by Region, 2000–2016. Retrieved August 19, 2019, from https://www.who.int/healthinfo/global_burden_disease/GHE2016_Death-Rates-country.xls?ua=1