Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Applied Soft Computing 128 (2022) 109422

Contents lists available at ScienceDirect

Applied Soft Computing

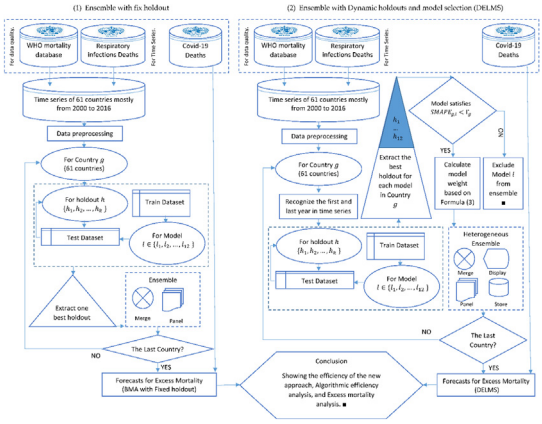
journal homepage: www.elsevier.com/locate/asoc

An ensemble learning strategy for panel time series forecasting of excess mortality during the COVID-19 pandemic☆

Afshin Ashofteh [a,*], Jorge M. Bravo [b,c], Mercedes Ayuso [d]

[a] NOVA Information Management School (NOVA IMS), Nova University Lisbon, Campus de Campolide, 1070-312 Lisboa, Portugal
[b] NOVA Information Management School (NOVA IMS), ISCTE-IUL BRU, CEFAGE-UE, Portugal
[c] Université Paris-Dauphine, France
[d] Department of Econometrics, Statistics and Applied Economy, University of Barcelona, Spain

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 31 January 2022
Received in revised form 20 June 2022
Accepted 24 July 2022
Available online 1 August 2022

Keywords:
Layered learning
Ensemble learning
Multiple learning processes
Time series
Bayesian model averaging (BMA)
Forecasting
Machine learning
Respiratory disease deaths
SARS-CoV-2
Panel data

ABSTRACT

Quantifying and analyzing excess mortality in crises such as the ongoing COVID-19 pandemic is crucial for policymakers. Traditional measures fail to take into account differences in the level, long-term secular trends, and seasonal patterns in all-cause mortality across countries and regions. This paper develops and empirically investigates the forecasting performance of a novel, flexible and dynamic ensemble learning with a model selection strategy (DELMS) for the seasonal time series forecasting of monthly respiratory disease death data across a pool of 61 heterogeneous countries. The strategy is based on a Bayesian model averaging (BMA) of heterogeneous time series methods involving both the selection of the subset of best forecasters (model confidence set), the identification of the best holdout period for each contributed model, and the determination of optimal weights using out-of-sample predictive accuracy. A model selection strategy is also developed to remove the outlier models and to combine the models with reasonable accuracy in the ensemble. The empirical outcomes of this large set of experiments show that the accuracy of the BMA approach is significantly improved with DELMS when selecting a flexible and dynamic holdout period and removing the outlier models. Additionally, the forecasts of respiratory disease deaths for each country are highly accurate and exhibit a high correlation (94%) with COVID-19 deaths in 2020.

☆ This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

* Corresponding author.
E-mail address: aashofteh@novaims.unl.pt (A. Ashofteh).

https://doi.org/10.1016/j.asoc.2022.109422

1. Introduction

In the time series forecasting literature, two techniques compete: forecast model selection [1] and forecast model combination [2]. The traditional approach to forecasting seasonal and non-seasonal time series is to select a single best model from a pool of candidate models based on certain criteria or employing a given technique [3], potentially neglecting model risk. The ensemble prediction method is widely considered a promising strategy and it has been used with considerable success in research and industry thanks to the availability of a wide variety of individual models. Since Bates and Granger's [4] seminal study, a growing number of linear and non-linear univariate and multivariate times series methods [5,6] and statistical machine learning techniques [7–9] have been proposed to increase short- and long-term predictive accuracy in relation to a wide range of problems, including stochastic population – mortality, fertility, and net migration – forecasting [10], epidemiological and excess mortality forecasting [11], meteorology [5], and finance [12,13], among others. Indeed, several comprehensive theoretical and empirical studies have confirmed the superior predictive performance of ensemble methods exploiting a variety of approaches, including stacking and blending to improve predictions, bagging to decrease variance, boosting to decrease bias [14,15] and the Bayesian model averaging (BMA) [16–18]. When adopting this empirical strategy, choices have to be made as to which models to include in the combined pool and as regards the contribution (weight) of each model to the final prediction. Here, a significant body of literature has examined optimal model combination weights [19,20], by focusing on the selection of optimal combination schemes and weights [21,22], assigning equal weights to the set of superior models [23], or selecting a subset of best models from 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 [24,25]. Similarly, to cope with concept drift, memory, change detection, learning, and loss estimation, adaptive algorithms have been proposed [26].

However, experiments are usually conducted with a holdout set so as to pick pools of models manually that perform best for a given series type [27]. This is often motivated by a lack of computational power, as well as by limitations of time for checking and allocating one specific holdout from a holdout set to each individual model, and preferring, therefore, to consider a fixed holdout for all models and, by so doing, facilitating their combination in an ensemble model. In fact, different holdouts for different models results in different lengths for each model, which means the combination of these models of different lengths becomes a challenge. However, ignoring the different holdouts for each model reduces adaptability and undermines their generalization.

This paper proposes a dynamic ensemble learning strategy (DELMS) in different layers for panel time series that not only overcomes the limitations of single-model based methods, but also addresses those of new ensemble models with fixed holdout sets and fixed thresholds for model selection. The strategy combines twelve models, including the models suggested by the M4 Competition [28] as benchmarks and standards for comparative purposes. We also consider model candidates to ensure sufficient diversification of statistical models, specifically SARIMA, and DNN models, including multi-layer perceptron (MLP) to yield more robustly accurate forecasts. We then consider different holdouts and different time series lengths in one layer, and other layers in order to select the best models for each series. In this way DELMS is able to generate effective and robust forecasts, separate the pattern from the noise, and overcome overfitting problems. The strategy is based on a Bayesian model averaging (BMA) to combine the heterogeneous models with the lowest error measure to generate an ensemble. It applies 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 contributed model, and the determination of optimal weights using out-of-sample predictive accuracy. A model selection strategy is also developed to remove the outlier models and to combine the models with reasonable accuracy in the ensemble. In short, the ensemble learning procedure proposed (DELMS) involves: (i) setting the different holdouts to be checked for each contributed model; (ii) choosing the best holdout for each model based on out-of-sample forecasting accuracy; (iii) selecting the subset of best forecasters (model confidence set), using a variable trimming scheme in which a multiple of the forecasting accuracy metric range obtained across all candidate models is used as the threshold for model exclusion; (iv) determining the posterior probabilities (weights) of each model, using the normalized exponential (Softmax) function; and, finally, (v) obtaining ensemble forecasts based on the law of total probability, considering the model confidence set and the corresponding model weights. Unlike previous approaches that have focused on either selecting optimal combination schemes and weights or equally weighting a subset of best forecasters, our novelty ensemble procedure involves identifying the best holdout period for each model, selecting the best forecasting models and determining the optimal weights based on the out-of-sample forecasting performance for each dataset.

To demonstrate empirically the robustness of our approach, we use monthly respiratory disease death data for 61 heterogeneous countries to estimate excess mortality during the COVID-19 pandemic. Excess mortality is the number of deaths attributable to all causes above and beyond mortality predictions under normal (baseline) circumstances for a given period in a population. Clearly, quantifying and analyzing excess mortality attributable to the coronavirus 2 (SARS-CoV-2) pandemic is of great relevance for policymakers, public health officials and epidemiologists [29,30] and, in this sense, any improvement in such forecasts are to be welcomed. Excess mortality is typically measured by national or supranational statistical agencies using the absolute, relative (P-score) or standardized (Z-score) number of "excess" deaths, where the benchmark is often computed quite naïvely, by using, for instance, the simple average of the previous year's deaths. The EuroMOMO project (https://www.euromomo.eu) is a notable example of this, with baseline mortality modeled using a generalized linear model corrected for over dispersion assuming that the number of deaths follows a Poisson distribution. However, this approach does not account for differences in the level, long-term secular trends, and seasonal patterns in all-cause mortality across countries and regions. Additionally, empirical studies show that it is hard to find a single, widely accepted forecasting method (if, indeed, one exists) that performs consistently well across all datasets and time horizons [31]. Besides, data quality is another concern, being responsible for biased and inconsistent parameter estimates and leading to flawed conclusions [32]. This is a matter of untold concern for forecasters of ensemble learning predictive models when seeking to predict, for example, numbers of deaths or when an economy should be re-opened [33,34], among others. Moreover, it makes excess mortality a highly appropriate case for comparing the experiences of different countries or regions, where either the degree of misdiagnosis/underreporting or the problems of data quality may differ [35]. Our hypothesis is that the approach proposed herein leads to a decrease in the individual error of ensemble members compared to that provided by normal model selection with equal holdouts for selected models and without overly decreasing the diversity between them. We examine the run times, accuracy, level of contribution, and error metric of the ensemble techniques proposed and compare

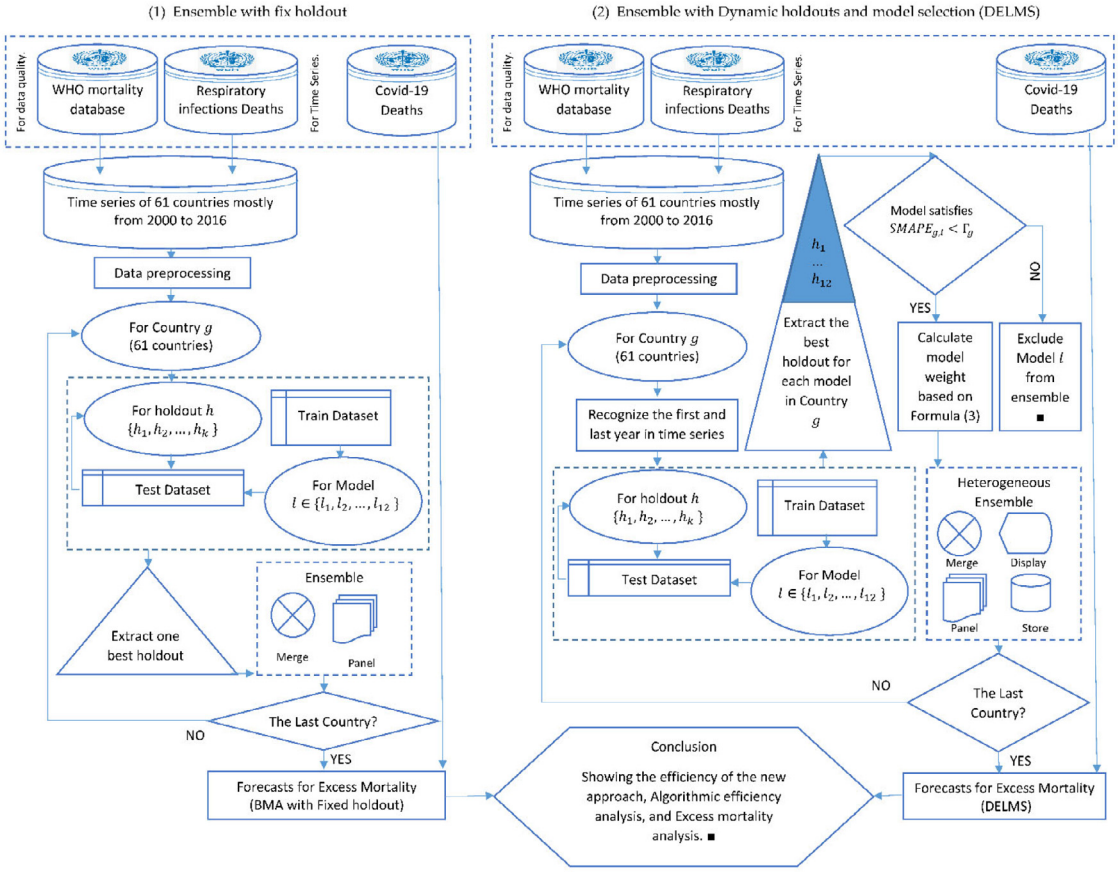


Fig. 1. Graphical overview of the dynamic ensemble learning strategy.

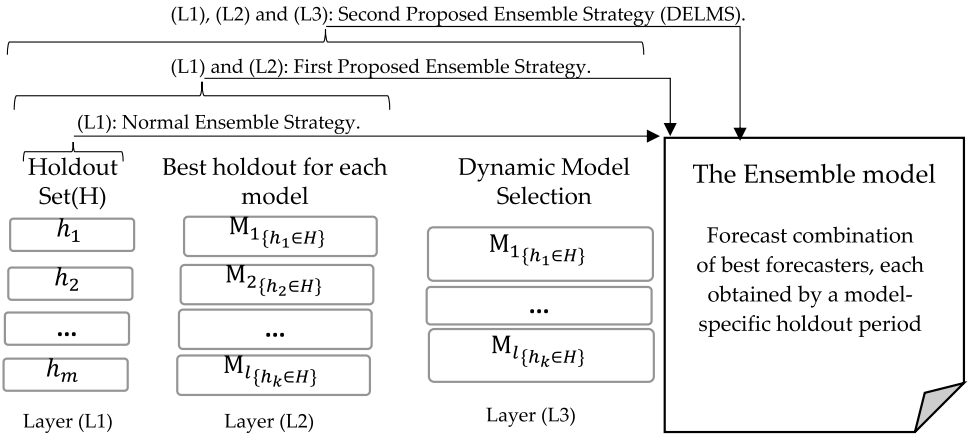


Fig. 2. Proposed ensemble learning strategy.

them with those of the well-known ensemble model without dynamic holdouts and model selection, and the individual forecasting models. This article presents a suitable ensemble time series with improved predictive accuracy and it is our belief that it helps demonstrate which time series techniques contribute more to ensembles.

The remaining sections of the paper are organized as follows. In Section 2, we describe the materials, methods, and related works considered in undertaking this study. Section 3 outlines an extensive set of experiments on respiratory disease deaths in 61 countries and the results. The main discussion and conclusions are reported and discussed in Section 4. Finally, future research proposals are presented in Section 5.

2. Materials and methods

Here, we propose a meta-learning approach for adapting 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. Figs. 1 and 2 provide graphical overviews of the materials and methods employed to develop our proposed strategy. We use multiple learning processes to improve the predictive performance of the ensemble, which is built using a learning approach for the candidates addressed in the last layer. In this section, we discuss these techniques in brief and highlight their contributions.

Table 1
Pseudocode of the proposed ensemble strategy.

```
INPUT panel time series (panel members = countries)
OUTPUT ensemble model
1.  StatExplore time series decomposition
2.  IMPUTE[missing] = TRUE
3.  First_year = 2000 (for most of time series but some of them start later)
4.  Last_year = 2016
5.  Target_year = 2020
6.  Confidence_level = 0.95
7.  Holdout_set = {3, 5, 7} and SET Teta = 0.5
8.  Ensemble_criteria_for_computing_weights = "Symmetric Mean Absolute Percentage Error (SMAPE)"
9.  Set.seed()
10. Model_list = { TBATS, ETS, SARIMA, STL, NNAR, SNAIVE, HWA, HWM, MLP, ELM, SSA, RWF}
11. FUNCTION model_weights (error)
12.     Pr = error/max(error)
13.     exp(-abs(pr))/sum(exp(-abs(pr)))
14. # First loop for selecting country
15. For each panel in list of countries do
16. {
17.     SET panel.data = SUBSET dataset(country = panel & Year > First_year & Months="Jan-Dec"
18.     SET Year_min = min(Year of panel.data)
19.     panel_data = MISSING VALUE IMPUTATION by na_seasplit
20.     SET (START of the run-time calculation)
21. # Second loop for selecting holdouts
22.     For each holdout in Holdout_set do
23.     {
24.       IF ( ymax-ho+1 < ymin+3 ) { break }
25.       ELSE
26.         SET train_dataset WINDOW (START = Year_min , END = Last_year – holdout)
27.         SET test_dataset WINDOW (START = Last_year – holdout + 1)
28.         FIT models in Model_list
29.         CALCULATE accuracy (model , holdout)
30.         IF accuracy (model[holdout]) > last_accuracy (model[holdout – 1]) THEN
31.           SET model = model[holdout]
32.         ELSE
33.           SET model = model[holdout -1]
34.     }
35.     CALCULATE error(ALL models), min_error(ALL models), max_error(ALL models)
36.     CALCULATE id_error = Teta × (min_error + max_error)
37.     FOR model in Model_list
38.     {
39.       IF (error_model > id_error) THEN
40.         PRINT ("Model is excluded!")
41.       ELSE
42.         ADD model into selected_model_list
43.     }
44. # The proposed model ensemble (DELMS)
45. IF selected_model_list = NULL {next country}
46. ELSE
47.       {
48.       CALCULATE model_weights for ensemble
49.       SET First_year based on the model with min_holdouts
50.       SET First_month based on the model with min_holdouts
51.       CALCULATE ENS as Ensemble Model
52.       SET (END of the run-time calculation)
53.       }
54. # The outputs
55. PRINT GRAPHS
56. SAVE OUTPUTS
57. }
```

2.1. Layered learning and the ensemble learning strategy

The layered learning approach as applied to time series data consists of breaking a forecasting problem down into simpler subtasks that occupy different layers. Each layer addresses a different predictive task and the output of one layer can be used as input for the next layer [36]. In this study, the first task is to obtain a direct mapping of the time series for different countries, combining the intractable time series algorithms and predicting the ensemble model as the final output. This means the first layer task is to find the best holdout for each panel member and for each time series algorithm. This facilitates the second layer task of model selection, which in turn facilitates the identification of the model confidence set of best forecasters in the last layer [37]. It is useful to maximize forecasting accuracy in panel time series – a target that is achieved dynamically – and to adapt the model's learning process to possible unexpected shocks.

Along with the layered learning approach, our ensemble method runs multiple learning algorithms to employ adaptive heuristics that combine forecasters. As a result, we obtain a better predictive performance than might be obtained from any of the constituent learning algorithms. Our strategy comprises several selected models (see Table 1 — line 10), with the best performance being based on minimum error measures. Each model considers different holdouts to solve the problem at hand and selects the best holdout in each case. This leads to a more robust overall performance of the ensemble as it increases the diversity of the holdouts; however, the time series length differs according

to the different holdouts. As such, it could constitute a problem for the ensemble layer, were we to merge the models of different lengths. Thus, we need to force all the models selected to be of equal length, so that eventually the length of the ensemble is equal to the minimum length time series in our time series set. Although this windowing strategy provides each forecaster's best prediction and, therefore, the ensemble's best performance, it is clear that to obtain the best results, the length of all the time series should be sufficiently large and almost the same.

Following Ashofteh and Bravo (2021), let each candidate model be denoted by M_l, $l = 1, \ldots, L$ 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 these 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

$$p(\Delta|y) = \sum_{l=1}^{L} p(\Delta|y, M_l)\, p(M_l|y) \tag{1}$$

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

$$p(M_l|y) = \frac{p(y|M_l)\, p(M_l)}{\sum_{l=1}^{L} p(y|M_l)\, p(M_l)}. \tag{2}$$

The workflow of our proposed method is presented in Fig. 2. To identify the model confidence set and compute model weights, we first set the different holdouts to be checked for each contributed model in each dataset. Let $H = \{h_1, h_2, \ldots, h_k\}$ represent the set of holdout periods to be considered in the estimation procedure (see Fig. 2. Layer L1). The second step involves selecting the best holdout for each candidate model based on the out-of-sample forecasting accuracy measure (see Fig. 2. Layer L2). We use the symmetric mean absolute percentage error (SMAPE) as our measure of forecasting accuracy (see Table 1 — line 8).[1]

To select the best holdout for each model, we tested the different holdout values from three to ten years — considering the holdout set ($H = \{3, 5, 7\}$ years[2]) (see Table 1 — line 7) as representative of the short-, medium-, and long-term — and compared the SMAPE values at each iteration, retaining the model with the lowest SMAPE as the candidate for the model confidence set selection step. This provides the strategy with an opportunity to cover different parts of the data space and to handle different dynamic regimes in different candidate time series. Additionally, it ensures the final ensemble model is able to manage the limitations of each in the others.

Third (see Fig. 2. Layer L3), the subset of best forecasters is selected using the best holdout period (see Table 1 — lines 21–34) and a variable trimming scheme in which a multiple θ (pre-set at 0.5) of the distance between the maximum and minimum values of the forecasting error metric is used as the threshold for model exclusion, i.e., using

$$\Gamma_g = \theta \times \left(\max\left\{SMAPE_{g,l}\right\}_{l=1,\ldots,L} + \min\left\{SMAPE_{g,l}\right\}_{l=1,\ldots,L}\right). \tag{3}$$

where $SMAPE_{g,l}$ is the SMAPE value for model l in the panel member (country) g (see Table 1 — lines 35–36). For each panel member, if the error of a candidate model is greater than the Γ_g indicator, (i.e., $SMAPE_{g,l} > \Gamma_g$) the model is excluded from the model confidence set and from the ensemble forecast computation (Table 1 — lines 37–43), i.e., it is assigned zero weight in (1).

Depending on the distribution of the SMAPE values, the number of models excluded from the model confidence set will vary. From a frequentist point of view, building up a model confidence set is a way of summarizing the relative forecasting performances of the entire set of candidate models and identifying the set of statistically best forecasters. The advantage of this statistic defined in (3) is its simplicity, ease of application, and interoperability. Moreover, it falls somewhere between the time series' close and extremely distant models. In this case, the distant forecasting models are removed from the ensemble, which is ideal for avoiding overfitting and controlling the redundancy in the output of the ensemble model. Our intuition is that the models with a minimum error are closest to the actual data generating process. Yet, comparing the error measure with the mean of the errors removes only those models that are extremely distant from the other candidate models. This upholds the diversity of the selected models and avoids the overfitting problem.

Fourth, the posterior probabilities of the best forecaster model (model weights) are computed using the normalized exponential (Softmax) function

$$p(M_l|y) = \frac{\exp(-|\xi_l|)}{\sum_{l=1}^{L} \exp(-|\xi_l|)}, \quad l = 1, \ldots, L \tag{4}$$

with $\xi_l = S_l / \max\{S_l\}_{l=1,\ldots,L}$ and $S_l := SMAPE_{g,l}$. The Softmax function is a generalization of the logistic function that is often employed in classification and forecasting exercises using traditional machine learning and deep learning methods as a combiner or activation function [38]. The function assigns larger weights to models with smaller forecasting errors, with the weights decaying exponentially the larger the error (see Table 1 — lines 11–13). Fifth, the BMA forecasts are obtained based on the law of total probability (1) considering the model confidence set and the corresponding model weights (4). The sampling distribution of the ensemble forecast of the quantity of interest is a mixture of the individual model sampling distributions (see Table 1 — lines 44–53).

The pseudocode of the proposed methodology is listed in Table 1.

In the interest of reproducible science, the dataset and all methods are publicly available [39].

2.2. The learning algorithms

This section summarizes the characteristics of the individual candidate learning algorithms (times series methods) used in this study[3] (see Table 1 — line 10). We selected our models by reviewing the six top-performing hybrid or combination models in the M4 Competition, but taking into consideration our research limitations derived from the length of time series and the computational power required to build the ensemble model for a 61-member time series from 2000 to 2016 with 12 individual models and three holdouts.

[1] We avoid using the AIC and the BIC because the candidate models are in different model classes, and the likelihood is computed differently. For selected models in the same class, the BIC is useful and is used automatically by the algorithm to select, for instance, a SARIMA model among the candidate SARIMA models. Another problem associated with the error term in ensemble modeling can be avoided by using accuracy measures whose formula contains a logarithm, such as MSLE, RMSLE, and SLE. Based on the work conducted here, the program would be interrupted because some of the algorithms potentially present negative values in these measures.

[2] The results for the other holdout periods are consistent with those reported in this paper.

[3] For a detailed presentation and discussion of the methods see, for instance, Hyndman and Athanasopoulos [40].

The seasonal trend decomposition using Loess (STL) allows us to decompose a time series into its trend and seasonal components. Based on the Loess smoother, it offers a simple, versatile, and robust method for decomposing a time series and estimating nonlinear relationships [41]. The models need to be robust to the outliers detected in the multiple panel members' (countries) datasets. In specifying the STL, we use a robust decomposition so that sporadic abnormal observations do not affect the estimates of the trend-cycle and seasonal components. The time series are tested for autocorrelation using the Ljung–Box test, considering the null hypothesis that the model exhibits appropriate goodness-of-fit. The method does not handle the calendar variation automatically, and it only provides facilities for additive decompositions, which could be considered a limitation of this approach. We use the two parameters *t.window* and *s.window* to control the speed at which the trend-cycle and seasonal components can change. Smaller values allow for more rapid changes, which we need especially for some time series with strong turning points. As a result, the number six was chosen for *s.window* and *t.window* based on the results of the residual checks and the Ljung–Box test statistics.

The seasonal naive (SNAIVE) method sets the forecast to be equal to the last observed value from the same season of the year (i.e., the same month of the previous year) [42]. It is a useful benchmark for other forecasting methods and, here, it was found to be helpful in showing the recent time series trend and for adjusting the ensemble model for the trend component.

Similarly, the SARIMA and random walk forecasts (RWF) – as a SARIMA(0,0,0)(0,1,0)m model, in which m is the seasonal period – were used as state-of-the-art methods to memorize repeating monthly patterns. However, many SARIMA models have no exponential smoothing counterparts [43], and the robust univariate forecasting models, such as Holt–Winters' multiplicative method (HWM) and the exponential smoothing state space model (ETS), can be considered a good complement to SARIMA models in our final ensemble. All ETS models are non-stationary, while some SARIMA models are stationary [44]. ETS follows the last trend of the time series and it is appropriate for the ensemble model for empowering the trend parameter in the final predictions. ETS point forecasts are equal to the medians of the forecast distributions. For models with only additive components, the forecast distributions are normal, so the medians and means are equal. For multiplicative errors, or multiplicative seasonality, which perform similarly in most of the time series analyzed in this study, the point ETS forecasts are not equal to the means of the forecast distributions. In these cases, SARIMA is a better choice. On the other hand, ETS is a non-linear exponential smoothing model with no equivalent SARIMA counterpart. Therefore, we propose that the ETS model be selected automatically and the type of trend and seasonal component be additive with the restriction of finite variance. The bootstrapping method for resampled errors was employed rather than distributed errors and simulation was used rather than algebraic formulas for calculating prediction intervals. The other options for the ETS model are shown in Table 2. The TBATS – that is, (T)rigonometric terms for seasonality, (B)ox–Cox transformations for heterogeneity, (A)RMA errors for short-term dynamics, (T)rend, and (S)easonal – are also used to adopt the ensemble model with multiple seasonality of some time series.

In the case of the neural network time series algorithms, the extreme learning machines (ELM) were used with the lasso penalty. ELM theory assumes that the randomness in the determination of coefficients of neural network predictors (input weights) can feed the learning models with no iterative tuning for a given distribution as is the case in gradient-based learning algorithms. The model entails randomly defined hidden nodes and input weights without any optimization, so that only output weights need to be calibrated during the training of the ELM [45]. In the hyperparameter calibration of the ELM, we consider the maximum 500 hidden layers for 200 networks to be trained and summarized in the ELM's final ensemble forecast model.

The neural network autoregression (NNAR) refers to single hidden layer networks using the lagged values of the time series as inputs and automatic selection of parameters and lags according to the Akaike information criterion (AIC) [46]. In the NNAR model specification, we considered the last observed values from the same season as the inputs to capture the seasonality patterns and to use a size equal to one, because we have one attribute without a regressor, and by way of improvement, we used 100 networks to fit the different random starting weights and then averaged them out to produce the forecasts. Additionally, we considered the multilayer perceptron (MLP) as a kind of NNAR model. This is more complicated and advanced than the NNAR, having three components in the form of NNAR(p,P,k), in which p denotes the number of lagged values that are used as inputs and which is usually chosen based on an information criterion, like AIC, P denotes the number of seasonal lags, and k denotes the number of hidden nodes.

Finally, singular spectrum analysis (SSA) was used as one of the high-quality modeling approaches. The calibration of the SSA is an important, but not easy task, in a standalone modeling approach [47]. It depends upon two basic parameters: the window length and the number of eigentriples used for reconstruction. The choice of improper values for these parameters yields incomplete reconstruction, and the forecasting results might be misleading. In this study, we set window length equal to 12 and eigentriples equal to NULL. Table 2 summarizes the hyperparameters of the algorithms used in this study.

The model fitting, forecasting, and simulation procedures were implemented using R statistical software considering libraries such as the TSA, Metrics, nnfor, tsfknn, Rssa, rpatrec, and forecast (see, e.g., [48]).

3. Empirical experiments

3.1. Data selection and cleansing

We use cause-of-death data from the World Health Organization's (WHO) mortality database [50] to empirically demonstrate the forecasting capacity of the methodology proposed. The database collects cause-of-death statistics from country civil registration systems and estimates from the United Nations Population Division for countries that do not regularly report population data. We use an Excel file[4] of this database to evaluate the data quality of each country and a CSV file that includes the death time series of each country by gender. The first of these files identifies the quality of data for each country, using five color categories — green, dark yellow, light yellow, dark red and light red. Countries classified as green have multiple years of national death registration data with high completeness and quality of cause-of-death assignment. Estimates for these countries may be compared and time series may be used for priority setting and policy evaluation. However, this dataset only includes data for 2000, 2010, 2015, and 2016 and it is not complete for the time series. As a result, we used this dataset only to identify the countries reporting high-quality data to the WHO and ranked them according to their data quality. In line with the metadata of the dataset, the criteria used to rank the countries by data quality are shown in Table 3, coinciding, that is, with the WHO descriptors.

[4] https://www.who.int/healthinfo/global_burden_disease/GHE2016_Deaths_2016-country.xls?ua=1

Table 2
Algorithms and hyperparameter choices.

ID	Algorithm	Parameters	Value
ETS	Exponential smoothing state space model	Model	{ETS, TBATS}[a]
		Box–Cox tran.	ZZA
		Multiplicative trend restricted for the models with infinite variance	TRUE
			TRUE
SARIMA	Seasonal auto-regressive integrated moving average model	Auto	"auto"
STL	Seasonal trend decomposition using Loess	lambda	"auto"
		t.window	6
		s.window	6
		biasadj	TRUE
NNAR	Neural network model to a time series	P	2
		size	1
		decay	0.001
		lambda	Auto
		repeats	100
		MaxNWts	2000
SNAIVE	Seasonal naïve	drift	F
		lambda	0
		level	clevel
		biasadj	TRUE
HWM	Holt-Winters' multiplicative method	Seasonal	Multiplicative
		Level	Clevel
HWA	Holt-Winters' additive method	Seasonal	Additive
		Level	Clevel
MLP	Multilayer perceptron for time series	Comb	Mode
		hd.auto.type	Valid
		hd.max	5
ELM	Extreme learning machines	type	Lasso
		hd	500
		comb	mean
		reps	200
		difforder	NULL
SSA	Singular spectrum analysis	Kind	1d-ssa
		svd.method	Auto
		L	12
		neig force.decompose	NULL
		mask	TRUE
			NULL
RWF	Random walk forecasts	Drift	F
		Lambda	"auto"
		Level	clevel
		biasadj	TRUE

[a]The ETS method with automatic and ZZA parameter setting from the forecast statistical software R package [48], and the TBATS method, which includes Box–Cox transformation, ARMA errors, trend and seasonal components [49].

Table 3
Respiratory disease death data: Criteria used to rank countries by data quality.

Rank	Evaluation	Description by World Health Organization
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 between countries should be interpreted with caution.
3	Low quality	Data have severe quality issues. Comparisons between 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 between countries.
5	Ignorable	Data should be ignored.

We considered only those countries with a data quality corresponding to the first three categories and eliminated various islands due to a lack of data (e.g., Åland Islands). We also cleaned the dataset by removing the total column and various rows with unknown month data and/or zero deaths. Some countries reported total deaths for three months in a row during certain years. In such instances, we assumed a uniform distribution of deaths across the quarter and allocated the corresponding value to each 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 WHO codes classified as respiratory infections. To compute the number of deaths attributable to respiratory diseases, we aggregated codes 380 and 410 or, equivalently, codes 390, 400, and 410. We also corrected the names of some of the countries (Appendix A). In this way we were able to calculate the proportion of deaths attributable to respiratory diseases. To estimate the number of monthly deaths caused by

Table 4
Metadata of the code of the disease categorized as a respiratory disease.
Source: World Health Organization, 2018.

Code	Description by World Health Organization
380	Respiratory infections (This code is the aggregate of 390 and 400)
390	Lower respiratory infections
400	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.

respiratory diseases, we multiplied the annual proportion by the total number of forecasted deaths each month. We used the fraction of annual deaths from respiratory diseases over the total number of deaths as a proportion of deaths in each month.

This procedure provided us with a dataset with more than twelve thousand observations in a pool of a 61-member panel time series (countries) from 2000 to 2016 [39] (see Table 1 — lines 3–4). These panel time series cover possible situations of stationarity, non-stationarity, increasing trends, seasonality, and structural breaks so as to undertake a comprehensive evaluation of the improvement in accuracy of candidate and ensemble models in different scenarios.

Given the varying data quality of countries/territories/areas as regards case detection, definitions, testing strategies, reporting practices, and lag times, missing values are expected in the time series dataset. To deal with this problem, we tested the *Kalman*, *seasplit* and *seadec* algorithms to impute the missing values. Of the three, the *seasplit* algorithm performed best as regards saving both the trend and the seasonality for our dataset (see Table 1 — line 19). We only imputed missing values within the time series, but not at the beginning of the time series with a start date after 2000. As a result, rather than changing the first year of the time series to our base year 2000, we used the latest year available (see Table 1 — line 17–18). To avoid the error caused by combining time series of different lengths in an ensemble model, we adapted the R code to handle different start years. The same problem arises as a result of the procedure adopted to select the best holdout for each model, which may ultimately lead to model combinations considering forecasts based on different holdouts, i.e., different time series lengths.

Finally, for comparing the superiority of the proposed DELMS model, 7137 time series models were explored. They obtained from 12 time series models plus an ensemble, 3 scenarios, and 3 holdouts for 61 countries.

3.2. Results

3.2.1. Forecasting accuracy comparison

The predictive accuracy metrics obtained for the three alternative holdout periods under investigation, using three alternative backtesting procedures, are reported in Table 5. In the case of the first approach – the "Fixed holdout" – we used a fixed holdout period equal to 3, 5, and 7 years to derive the composite (ensemble) model.

The results in the first columns show that some models exhibit better performance than that of the ensemble models with fixed holdouts. For instance, the average error of the TBATS model across two holdout periods is smaller than that of the BMA (see Table 5, Column (1)).

The second approach – the "Fixed holdout with model selection" – uses a multiple of the SMAPE values across all methods to evaluate the distance of each model to the others as detailed above in the pseudocode (Table 1). The models with SMAPE values higher than that of the introduced indicator are considered poor forecasters and eliminated from the ensemble forecast.

Table 5 presents the results aggregated across all countries, with individual country results available as supplementary material in a Mendeley dataset [39]. The results in Table 5 show that the accuracy of the BMA approach improves in robustness when pursuing the selection approach for each holdout, with the composite model now ranking first among all the methods tested. With a fixed holdout of 3 for all models, which is the classical approach, the BMA has a SMAPE value of 0.112. For the same holdout, but with model selection, this SMAPE value improves (0.103). In the third approach – "model selection plus dynamic holdouts" – that is, a combination of approaches one and two – the percentage error improves again (0.102). This approach combines the best forecasting models fitted using each model's optimal holdout selection. As a result, the accuracy of the ensemble is improved, leaving the individual learning algorithms at a reasonable distance.

Fig. 3 summarizes the above empirical results. It is apparent that the ensemble model with the new layered learning approach (DELMS) exhibits greater predictive accuracy than either of the two single forecasting methods used and either of the ensemble strategies with fixed holdouts and with fixed holdouts and model selection. It shows that the approach proposed improves the predictive performance at each step of the learning process illustrated in Fig. 2.

Finally, the Wilcoxon signed-rank test was performed to determine the significance of the superiority of the proposed model (DELMS). This test was used to determine the significance of the forecasting errors in the forecasts of the central trend made by two forecasting models with the same number of data [51]. Let e_i be the forecasting errors in the ith forecast value (i.e. countries) generated by two forecasting models (DELMS and BMA(holdout=3) (Appendix C).

$$sum\ of\ ranks = \begin{cases} r^+ & ; \quad if\ e_i > 0 \\ Eliminate & ; \quad if\ e_i = 0 \\ r^- & ; \quad if\ e_i < 0 \end{cases} \tag{5}$$

where r^+ and r^- represent the sum of ranks. For $e_i = 0$, we eliminate the comparison. The statistic W is defined as in Eq. (6):

$$W = min\{r^+, r^-\} \tag{6}$$

Table 6 presents the results of this statistical non-parametric test, by the one-tail-test at a significance level of $\alpha = 0.05$.

The proposed DELMS model significantly outperforms the BMA(ho=3), which is the ensemble model with the best performance among all ensembles with fixed holdouts (Table 5). The proposed DELMS model is significantly superior to other BMAs with respect to forecasting (P-value = 0.015).

3.2.2. Models excluded from the selection procedure

Table 7 reports the distribution of the models excluded from the selection procedure and ranks them according to their contribution to the composite model. A vertical comparison of the results offers insights into the contribution of each model to the ensemble, while a horizontal comparison enables us to assess the rate of contribution across different holdout periods.

The results show, first, that all models are excluded several times from the BMA model space as a result of the procedure to select the model confidence set, highlighting that the set of best-performing forecasters differs across the countries, i.e., their predictive accuracy is population- and period-specific. This is unsurprising and can be explained by the differential patterns observed in the respiratory disease data. The variability in the models' out-of-sample forecasting accuracy also reveals their ability to capture diverse features of mortality data. Second, the results suggest that combining models is a way to leverage their

Table 5
Ranking of the models and ensembles according to the accuracy measure.
Source: Authors' own.

Models	The model's average error (SMAPE)									Model total error	Rank
	(1) Fixed holdout (traditional ensemble)				(2) Fixed holdout with model 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
SARIMA	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
NNAR	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

(1) Fixed holdout column for the first row (BMA) shows the SMAPE for the Bayesian model averaging approach with fixed holdout. Rest of rows show individual models. (2) and (3) represent the methods proposed herein.

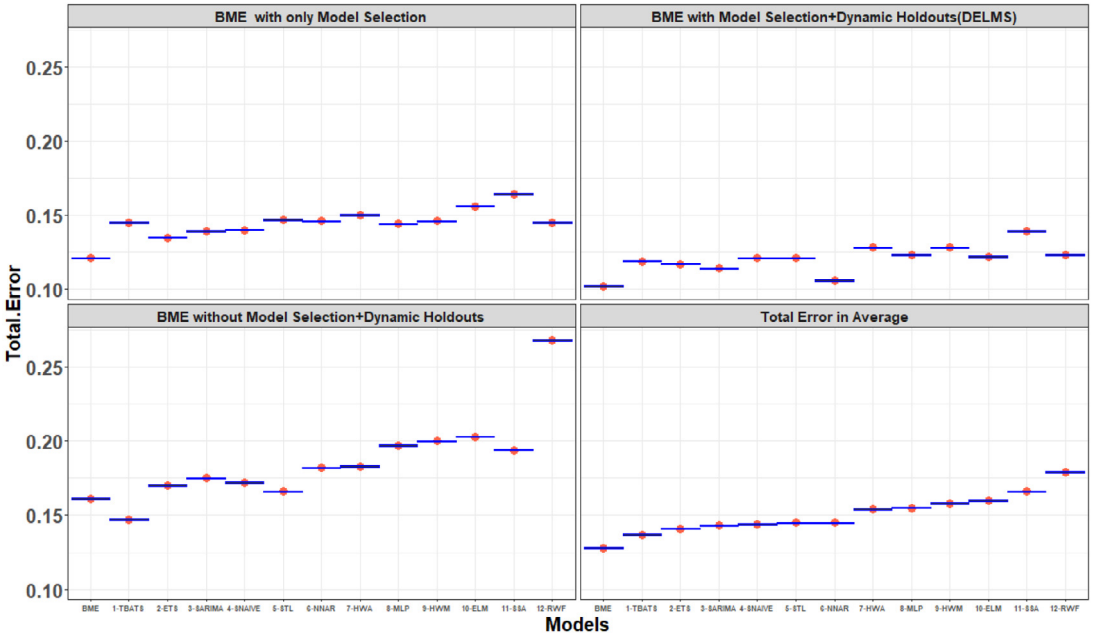


Fig. 3. Comparing the accuracy of the models.

Table 6
Results of Wilcoxon signed rank test.

Compared models	Wilcoxon signed-rank statistic	*P*-value ($\alpha = 0.05$)
DELMS versus BMA(ho = 3)	$w = 512$	0.01548**

** implies the *p*-value is lower than α.

strengths and minimize their weaknesses. The results capturing the contribution of single forecasters to the composite model show that the best contributor – the ETS model – has an exclusion rate substantially smaller than that of the worst forecaster, the RWF model. Moreover, the results suggest that increasing the holdout has a slightly positive effect on some models (the case of ETS, SNAIVE, NNAR, MLP, and RWF), a negative effect on others (the case of SARIMA, HWA, ELM, and SSA), and a neutral effect on others (the case of TBATS, STL and HWM). This variation in the contribution rates from the best to the worst model and from the lowest to the highest holdout period suggests a potentially positive effect on the final forecasting accuracy of the ensemble

Table 7
Rate of contribution of each model in the DELMS.
Source: Authors' own.

Models	Exclusion frequency of the models from the 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
SARIMA	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
NNAR	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 8
Exclusion frequency of the models for the ensemble with dynamic holdouts.
Source: Authors' own.

	TBATS	STL	ETS	HWA	SARIMA	SNAIVE	HWM	ELM	MLP	SSA	NNAR	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

model by selecting both the best holdout for each model and the best forecasters in the model confidence set finally used to make the forecast.

Table 8 presents the contribution ranks, the exclusion frequency, and the proportion of the selected models with the best holdout for the DELMS. The results show that the contribution of single learners to the ensemble changes when compared with that obtained with model selection only (Table 7), highlighting again the importance of combining model selection with holdout period calibration.

Fig. 4 reports the BMA model confidence set (vertical axis) and corresponding posterior probability (horizontal axis) for selected countries.

As we used the SMAPE criterion to select the set of models and respective weights, a given weight of zero indicates excluding that individual model from the BMA forecast combination. We can observe that the model's contribution to the ensemble varies across countries and the ensemble model consistently performs well in all countries.

3.2.3. Algorithmic efficiency analysis

We analyze 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 9). The CPU used here is the Intel Core i7-7500U Processor @ 2.70 and 2.90 GHz with 16.0 GB RAM. The modeling, training, tuning, and testing are programmed in R 4.1.2. The method proposed fits the models considering three holdout periods in order to select the best holdout for each model.

Our expectation is that the method drives the run-time at least three times more than the two other approaches, which is expected given that the underlying model is a multi-step forecasting method. However, if we consider the average run-time and the mean confidence intervals for the three approaches, we see that they do not differ greatly, which indicates that our proposed method is efficient in terms of computation time.

3.2.4. Excess mortality analysis

The proposed ensemble learning for panel time series with strategy selection and dynamic holdouts (as discussed in Section 2 and here, above, in Section 3) was used to forecast the number of deaths caused by different kinds of respiratory disease for a subset of 61 countries in 2020 (see Table 1 — line 5). Additionally, COVID-19 deaths were extracted for the same year from the COVID-19 Weekly Epidemiological Update published by the World Health Organization (WHO) with data as received from national authorities, as of 3 January 2021, which provides full coverage for the period of 2020 [52].

Table B.1 (in Appendix B) presents forecasts of the total number of deaths attributable to respiratory diseases (RD TD), which is calculated as an aggregation of monthly death forecasts for each country. The last two columns show the standardized values of the total number of deaths attributable to respiratory diseases and COVID-19, respectively, used in calculating the correlation. The Pearson correlation for all 61 countries is 0.34, which is statistically significant (P-value = 0.007). As shown in Table 10, to calculate the correlation for a more limited set, we considered the European countries, including the United Kingdom, Canada, and the United States of America.

The selection criteria used were the maturity standards of their official statistics (SDDS+, SDDS, GDDS), outcomes from official statistics corruption models [53], and the quality of death data according to the WHO ranking discussed in Section 3.1.

Now, the correlation coefficient increased dramatically to 94% (P-value =0.000), which can be attributed to the higher quality of the official statistics in these countries. Ashofteh and Bravo (2020) have shown there to be significant variation in the quality of the COVID-19 datasets reported worldwide, albeit a recent study suggests that data science and new technologies can be expected to play a significant role in improving data quality from national statistical offices in the future [54].

The comparison of death forecasts attributable to respiratory diseases and COVID-19 deaths is shown in Figs. 5 and 6. For most countries, COVID-19 deaths can be said to have "replaced" the respiratory deaths that would have occurred based on extrapolations of past respiratory disease trends. Here, a study of

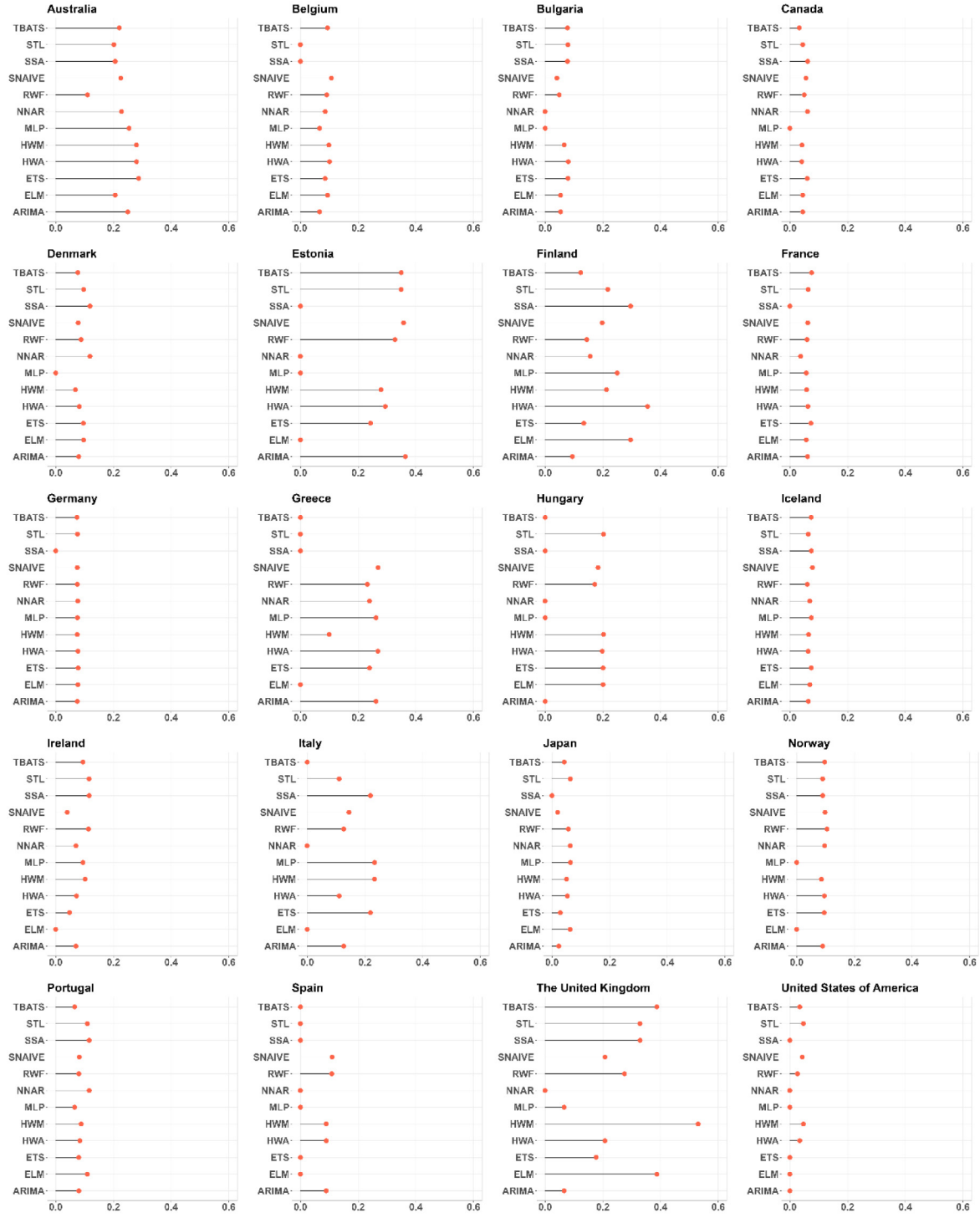


Fig. 4. BMA model confidence set and estimated weights per country.

Table 9
Effect of the methodology on run-time and computational efficiency.
Source: Authors' own.

Models	Run-time analysis to obtain ensemble model (in mins)											
	Fixed holdout				Fixed holdout + Model selection				Dynamic holdouts + Model selection (DELMS)			
	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

Notes: ART: Average run-time, STD: Standard deviation, LCL: Lower confidence limit, UCL: Upper confidence limit.

Table 10
Comparison of number of deaths forecast for respiratory diseases and actual COVID-19 deaths.
Source: Authors' own.

Row	Country	(1) Alpha-3	Country No	Population	(2) RD TD	(3) COVID TD	(4) Standardized RD TD	(5) Standardized 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	United States	USA	840	329064.917	16554	345253	2.288	6.791

Notes: (1) Abbreviated country code (three letters); (2) Respiratory Diseases Total Deaths; (3) WHO COVID-19 Total Deaths; (4) (Country Respiratory Disease Deaths − All Countries' Respiratory Disease Death Average)/All Countries Respiratory Disease Deaths STDEV; (5) (Country WHO COVID-19 Deaths − All Countries' COVID-19 Death Average)/All countries COVID-19 Deaths STDEV.

the factors affecting COVID-19 mortality shows a high correlation between respiratory deaths and COVID-19 deaths, a finding that is consistent with clinical manifestations and epidemiological studies. For example, countries with a high expectancy of respiratory diseases presented higher excess mortality, that is, at the macro (country) level. At the individual level, the higher number of deaths from respiratory diseases could be considered an indication of the population's greater susceptibility to COVID-19 symptoms and a greater risk of death. This comparative study highlights the fact that the policy effectiveness of different countries could result in an evaluation bias, without considering their past experience with respiratory diseases.

Fig. 5 shows that the countries of Europe and North America were sensitive to respiratory diseases and that this boosted the excess mortality attributable to the COVID-19 pandemic; however, Fig. 6 shows that in 2020 some countries dealt better with COVID-19 than others as regards their vulnerability to respiratory diseases. Thus, this last figure highlights that in countries in which the forecast of respiratory disease deaths significantly exceeds the confirmed COVID-19 deaths (e.g., Japan and the Philippines), the management of the pandemic crisis succeeded in reducing excess mortality. The results shown in these two figures are very much in line with a recent study indicating a much lower overall excess-mortality burden due to COVID-19 in Japan than in Europe and the USA [55]. Here, Yorifuji et al. [56] suggest that in Japan, the public health regulations aimed at preventing COVID-19 may have incidentally reduced mortality related to respiratory diseases, such as influenza, and so decreased net excess mortality.

Additionally, in addressing vulnerability to respiratory diseases, Japan and the Philippines appear to have set a good example for the rest of the world in terms of controlling the effects of respiratory death numbers on the number of COVID-19 deaths. The similarity of the situations in these two countries seems to testify to the importance of the agreements struck on their, so-called, COVID-19 Response Support. As reported on the website of the Department of Foreign Affairs in the Philippines, the Japanese Government has been unstinting in its commitment to the Philippines' recovery efforts, previously pledging over JPY100 billion assistance in emergency and standby loans and donating 1 million Japan-manufactured AstraZeneca vaccines.[5]

Fig. 6 shows a similar situation for the Republic of Korea, which geographically lies in the same vicinity as these two countries. The outcome of the comparison of death forecasts attributable to respiratory diseases and actual COVID-19 deaths for the Republic of Korea is in line with a recent study estimating mortality in Korea undertaken by Shin et al. [57], which finds that mortality in 2020 was similar to the historical trend. This similarity of outcomes reported by these neighboring countries seems to highlight the importance of international cooperation and the sharing of resources for the successful control of the effects of pandemics. Moreover, as these countries are geographically close to each other, meteorological factors might also have been influential in their respective outcomes. Clearly, more research is required.

Finally, in addition to the effect of respiratory deaths on deaths attributable to the pandemic, international cooperation, optimal scheduling and the utilization of medical resources, large-scale virus testing, protecting and managing the healthcare of the elderly, lockdowns, vaccination, and controlling the borders are examples of other factors that might result in different outcomes by country. However, accurate and timely estimations of respiratory deaths also seem to be an important factor when undertaking comparisons of multiple countries.

4. Conclusions

We have tested a new ensemble learning technique (DELMS) for the panel time series forecasting of respiratory diseases and

5 https://dfa.gov.ph/dfa-news/dfa-releasesupdate/29206-philippines-and-japan-sign-agreements-on-covid-19-response-support-and-on-scholarship-grants-for-civil-servants

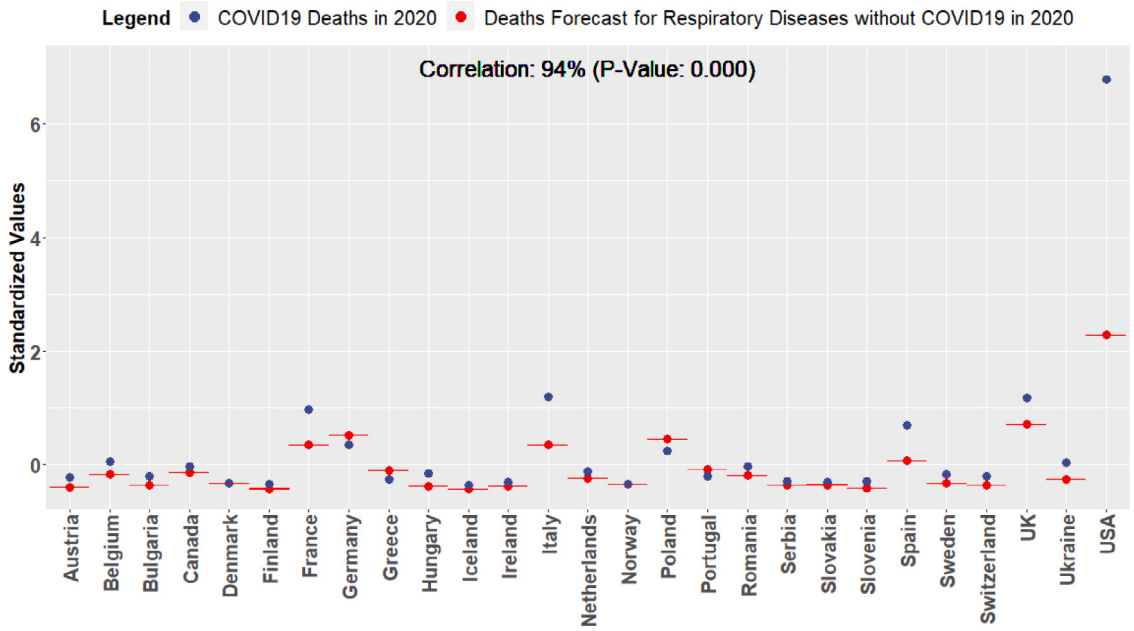


Fig. 5. Respiratory disease deaths and COVID-19 deaths for Europe and North America in 2020.

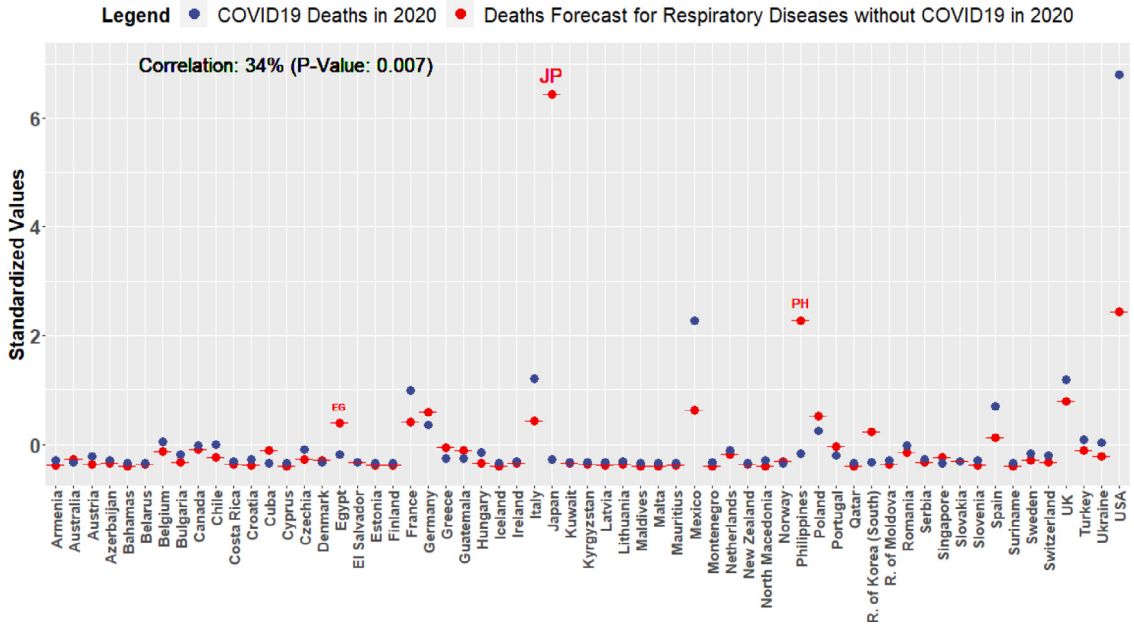


Fig. 6. Respiratory diseases deaths and COVID-19 deaths for Each Country in 2020.

we summarized the empirical results obtained when using individual models, a simple ensemble model, an ensemble with model selection, and an ensemble with model selection and dynamic holdouts (DELMS). Our goal in so doing was to obtain a benchmark for evaluating the excess mortality related to COVID-19 that might serve as a common framework for all countries.

Based on the performance outcomes of the models (Table 5) and results of Wilcoxon signed-rank test (Table 6), on average, the ensemble with model selection and dynamic holdouts (DELMS) performs significantly better than the other methods. Our results provide clear evidence of the competitiveness of this method in terms of its predictive performance when compared to the state-of-the-art approaches and even the ensemble model without the dynamic holdout and model selection layer.

Our analysis of the contribution of each of the candidate models to the ensemble (Tables 7 and 8) highlights the positive effect on overall prediction accuracy of selecting the best holdout for each model and excluding the outlier models from the ensemble. Moreover, it was evident that some of the state-of-the-art approaches outperformed the neural network time series models. A possible explanation for the underperformance of the complex neural network approaches might lie in the non-stationary elements, for example, the trend component and their pre-set hyperparameters. However, neural network time series models have been shown to perform much better when the time series data are nonlinear and stationary and present sudden changes in their layering hierarchy [58]. For this reason, they can be expected to add value to the ensemble in the case of mostly detrended

time series. Additionally, recurrent neural networks, such as LSTM and GRU, have the potential to outperform time series models and their use could be usefully explored for the ensemble in future studies with sufficient computation resources or less panel members.

The variation in the performance of each model stresses the need to improve each of them individually by selecting the best holdout and, moreover, to determine the best models to contribute to the ensemble without overfitting. The indicator proposed here in Formula (3) removes only those models that are very distant from the other models and, by so doing, we are able to avoid the significant bias in the set of candidate forecasters. The final ensemble model shows a significant improvement in overall accuracy when compared with the other ensembles and with each individual state-of-the-art approach. The superiority of the proposed DELMS model was explored by comparing 7137 time series models obtained from 61 countries, 12 time series models plus an ensemble, 3 scenarios, and 3 holdouts.

Here, we have used the new ensemble strategy to forecast the number of deaths from respiratory diseases in 2020 for a sample of 61 countries. The correlation between the standardized values of deaths from respiratory diseases and those from COVID-19 was positive and statistically significant. Based on this outcome, it is apparent that we should consider death forecasts from respiratory diseases as a covariate for evaluating the management strategies employed by different countries, be they lockdown rules or the relaxation of border control regulations. On the basis of our study, Japan and the Philippines are candidates for further investigation in this regard; indeed, they are more eligible than other countries that only record a low death toll. It may well be that the experience of these countries with high mortality attributable to respiratory diseases played a more than relevant role in their management of the pandemic.

Indeed, in the case of the COVID-19 pandemic it might be more relevant to focus on the death toll rather than on the cumulative number of patients. Given the nature of pandemics, the challenge usually lies in being able to control its spread; however, here the primary concern might be said to have been controlling the severe cases and caring for the patients facing the greatest likelihood of death. Those countries presenting a high number of cases of respiratory disease and which successfully managed the pandemic, therefore, could be better targets for further studies that compare their health policies and strategies with those implemented by countries presenting only a low rate of mortality.

In short, the study described here represents an initial attempt at developing a new approach to ensemble forecasting tasks. The main motivation for this paper was the observation that the performance of the ensemble model might potentially be enhanced by selecting the best holdout for each candidate model and by choosing the best outcomes based on the dynamics of the observed values of the main series. In experiments using the 61-member panel time series of respiratory disease deaths recorded between 2000 and 2016, the aggregation of selected forecasting models employing our approach provides a consistent advantage in terms of accuracy and leads to better predictive performance. Moreover, our study provides a correction of the total number of positive cases of COVID-19, in accordance with the expected number of deaths attributable to respiratory diseases as identified by our ensemble model.

Finally, this study has highlighted the pandemic experiences of Japan and the Philippines, identifying them as candidates for further exploration. The two countries present a high degree of vulnerability to the COVID-19 pandemic; yet, despite this, they succeeded in managing it well. Thus, regardless of higher death tolls than those recorded in other countries, their policy response should be examined to extract best practices. Finally, and of particular interest, is the fact that in most countries COVID deaths seem to have "replaced" the deaths attributable to respiratory disease that appear likely to have occurred in the absence of the pandemic, based, that is, on an extrapolation of past trends of such deaths.

5. Future research

Future studies could usefully seek the optimization of θ in Formula (3), that is, investigate the dynamic selection of optimum θ to ensure better performance. Additionally, as the usual neural networks fail to model time series adequately, especially in the case of incomplete/limited data during the onset of the epidemic [59], the study of recurrent neural networks, such as LSTM and GRU, would constitute an interesting future step if necessary computational power is available. This research should examine their impact on predictive accuracy, computation time and other resources, given the potential of these mechanisms to outperform ensemble time series models with no more than a reasonable increase in the computation power requirement. Indeed, the consideration of a non-linear meta-learning approach, as opposed to a linear approach, and of prediction intervals, as opposed to a point forecast, could constitute a fruitful next step. Moreover, the use of classification techniques to analyze heterogeneous and homogeneous countries could be considered another layer following the application of the forecasting methods. As such, a clustering analysis might usefully be implemented based on the notion of excess mortality. Finally, the countries of Japan and the Philippines standout, and their policy response should be subject to an epidemiological examination to determine what lessons might be learned.

CRediT authorship contribution statement

Afshin Ashofteh: Conceptualization, Methodology, Software, Validation, Formal analysis, Data curation, Writing- Original draft, Visualization, Writing – reviewing and editing. **Jorge M. Bravo:** Conceptualization, Methodology, Software, Validation, Formal analysis, Data curation, Writing- Original draft, Visualization, Writing – reviewing and editing. **Mercedes Ayuso:** Conceptualization, Methodology, Software, Validation, Formal analysis, Data curation, Writing- Original draft, Visualization, Writing – reviewing and editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors are grateful to the anonymous reviewers for their constructive comments. Afshin Ashofteh and Jorge M. Bravo were supported by Portuguese national science funds made available through the FCT under project UIDB/04152/2020-Centro de Investigação em Gestão de Informação (MagIC). Additionally, Mercedes Ayuso is grateful to the Spanish Ministry of Science and Innovation for funding received under grant PID2019-105986GB-C21 and 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 (research project directly related to COVID-19 and economy). ■

Appendix A

Corrections made to the dataset

- Name of countries:
  - CuraÃ§ao is changed to Curaçao;
  - Falkland Islands (Malvinas) is changed to Falkland Islands [Islas Malvinas];
  - North Macedonia is changed to North Macedonia [FYROM];
  - Republic of Korea is changed to Republic of Korea (South);
  - Reunion is changed to Réunion;
  - Saint Helena ex. dep. is changed to Saint Helena;
  - United Kingdom of Great Britain and Northern Ireland is changed to The United Kingdom;
  - Venezuela (Bolivarian Republic of) is changed to Venezuela;
  - Wallis and Futuna Islands is changed to Wallis and Futuna;
  - Åland Islands was not found in death table. We do not have data for the proportion of respiratory disease in this region.
- The following countries report for a period of fewer than eight years and we did not consider their time series. Albania; Bahrain; Barbados; Bosnia and Herzegovina; Brazil; Brunei; Georgia; Panama; Saint Lucia; Seychelles; Tajikistan; Trinidad and Tobago; Uruguay; Uzbekistan; Mongolia; Saint Vincent and the Grenadines; Venezuela.
- Turkey has data for fewer than eight years; however, it has data for the most recent years considered. For this reason, it could be included in the study.
- Kazakhstan and Russian Federation were removed because they did not report data for recent years up to 2016.
- Deaths for China were not reported in the UN data. As a result, China is not in our final dataset.

Appendix B

See Table B.1.

Appendix C

```
# Paired Samples Wilcoxon Test
bias_BMA_ho3 < −c(−18.17144646, 0.845229748,
−0.648753395, −10.52927, 2.640979819,
−6.047415139, −10.09648659, −0.319100827, 0.26282477,
−0.412075608, 0.873710439, −1.372537678,
−0.200553074, −2.301345913, 0.828796298, −0.236458205,
0.422081767, −0.034221056, −0.090795426,
−4.275662757, −0.064372261, −3.703538265, −8.311840917,
−0.008717382, −4.60124742, −0.00465072, 0.009560071,
15.9249704, −22.95882253, 0.386195568, −34.78938463,
−4.593744983, −13.21814867,
−38.06280128, 147.2735756, 1.551344272, −45.85641334,
−0.014675746, −3.921082105, −0.086582781,
−1.055021314, 0.243349847, −2.871760309, −5.708225223,
−1.948395992, 1.868894907, −0.354732773, 9.097734439,
5.570868831, 1.387560297, 2.349852967, 0.650763477,
−4.846550659, 5.076776565, 1.447481937)
bias_BMA_delms < −c(−16.46339149, 2.489256218,
0.447586255, −8.368886752, 3.609450275,
−3.732700619, −7.68550753, 0.193324403, 0.569831253,
−0.162919411, 1.09325965, −1.213764392,
−0.052303056, 0.530358968, 3.891338424, −0.188220835,
3.987774606, −0.022094779, −0.079898738,
−0.297984813, −0.058321039, 0.855019376, −3.811265776,
−0.005188413, 1.361724355, −0.007053147,
−0.001949911, 23.49176553, −12.66403877, 0.346276863,
−22.82631969, −4.69635284, 0, −8.707822341, 174.3253852,
1.388558066, −2.579278071, −0.281251108, −4.192272506,
−0.363675082, 0.168771577,
−0.054404215, −3.336746928, −6.398988299, −2.883322348,
0.835158324, −1.525576287, 7.692591962, 3.927723027,
−0.323862101, 0.543333355, −1.739072039, −7.993404212,
0.620743283,−3.318268994)
myData <−
    data.frame(group = rep(c("bias_BMA_ho3", "bias_BMA_
delms"), each = 55), weight = c(bias_BMA_ho3, bias_BMA_delms))
    result = wilcox.test(weight ∼group, data = myData, paired =
TRUE, alternative = "less")
    print(result)
```

Table B.1
Comparison between forecasting deaths for respiratory diseases and actual COVID19 deaths.

Row	Country	Alpha-3	Country No	Population	RD TD	COVID TD	Standardized RD TD	Standardized COVID TD
1	Armenia	ARM	51	2957.728	83	2850	−0.417	−0.297
2	Australia	AUS	36	25203.2	16554	909	2.288	−0.337
3	Austria	AUT	40	8955.108	234	6214	−0.392	−0.227
4	Azerbaijan	AZE	31	10047.719	294	2703	−0.382	−0.3
5	Bahamas	BHS	44	389.486	21	170	−0.427	−0.352
6	Belarus	BLR	112	9452.409	205	153	−0.397	−0.353
7	Belgium	BEL	56	11539.326	1571	19693	−0.172	0.052
8	Bulgaria	BGR	100	7000.117	412	7644	−0.363	−0.198
9	Canada	CAN	124	37411.038	1766	15679	−0.14	−0.031
10	Chile	CHL	152	18952.035	992	16724	−0.268	−0.01
11	Costa Rica	CRI	188	5047.561	170	2185	−0.403	−0.311
12	Croatia	HRV	191	4130.299	68	4072	−0.419	−0.272
13	Cuba	CUB	192	11333.484	1689	146	−0.153	−0.353
14	Cyprus	CYP	196	1198.574	19	129	−0.427	−0.353
15	Czechia	CZE	203	10689.213	738	11960	−0.309	−0.108
16	Denmark	DNK	208	5771.877	595	1345	−0.333	−0.328
17	Egypt	EGY	818	100388.076	4626	7741	0.329	−0.196
18	El Salvador	SLV	222	6453.55	452	1351	−0.356	−0.328
19	Estonia	EST	233	1325.649	68	244	−0.419	−0.351
20	Finland	FIN	246	5532.159	53	561	−0.422	−0.344
21	France	FRA	250	65129.731	4733	64543	0.347	0.98
22	Germany	DEU	276	83517.046	5815	34272	0.524	0.354

(continued on next page)

Table B.1 (*continued*).

Row	Country	Alpha-3	Country No	Population	RD TD	COVID TD	Standardized RD TD	Standardized COVID TD
23	Greece	GRC	300	10473.452	2000	4921	−0.102	−0.254
24	Guatemala	GTM	320	17581.476	1726	4827	−0.147	−0.256
25	Hungary	HUN	348	9684.68	344	9884	−0.374	−0.151
26	Iceland	ISL	352	339.037	17	29	−0.428	−0.355
27	Ireland	IRL	372	4882.498	316	2252	−0.379	−0.309
28	Italy	ITA	380	60550.092	4792	74985	0.356	1.196
29	Japan	JPN	392	126860.299	39818	3548	6.107	−0.282
30	Kuwait	KWT	414	4207.077	291	937	−0.383	−0.337
31	Kyrgyzstan	KGZ	417	6415.851	253	1359	−0.389	−0.328
32	Latvia	LVA	428	1906.74	103	668	−0.414	−0.342
33	Lithuania	LTU	440	2759.631	185	1644	−0.4	−0.322
34	Maldives	MDV	462	530.957	5	48	−0.43	−0.355
35	Malta	MLT	470	440.377	39	220	−0.424	−0.351
36	Mauritius	MUS	480	1269.67	82	10	−0.417	−0.356
37	Mexico	MEX	484	127575.529	5956	126507	0.547	2.263
38	Montenegro	MNE	499	627.988	12	690	−0.428	−0.342
39	Netherlands	NLD	528	17097.123	1206	11565	−0.232	−0.117
40	New Zealand	NZL	554	4783.062	232	25	−0.392	−0.355
41	North Macedonia	MKD	807	2083.458	36	2522	−0.425	−0.304
42	Norway	NOR	578	5378.859	528	436	−0.344	−0.347
43	Philippines	PHL	608	108116.622	15580	9253	2.128	−0.164
44	Poland	POL	616	37887.771	5347	29119	0.448	0.247
45	Portugal	PRT	620	10226.178	2097	7045	−0.086	−0.21
46	Qatar	QAT	634	2832.071	12	245	−0.428	−0.351
47	Republic of Korea	KOR	410	51225.321	3712	962	0.179	−0.336
48	Rep. of Moldova	MDA	498	4043.258	221	3020	−0.394	−0.293
49	Romania	ROU	642	19364.558	1484	15919	−0.187	−0.026
50	Serbia	SRB	688	8772.228	419	3288	−0.362	−0.288
51	Singapore	SGP	702	5804.343	906	29	−0.282	−0.355
52	Slovakia	SVK	703	5457.012	476	2317	−0.352	−0.308
53	Slovenia	SVN	705	2078.654	145	2889	−0.407	−0.296
54	Spain	ESP	724	46736.782	3042	50442	0.069	0.688
55	Suriname	SUR	740	581.363	39	123	−0.424	−0.353
56	Sweden	SWE	752	10036.391	665	8727	−0.321	−0.175
57	Switzerland	CHE	756	8591.361	428	7049	−0.36	−0.21
58	The UK	GBR	826	67530.161	6943	74570	0.71	1.188
59	Turkey	TUR	792	83429.607	1658	21295	−0.158	0.085
60	Ukraine	UKR	804	43993.643	1089	18854	−0.252	0.034
61	US of America	USA	840	329064.917	16554	345253	2.288	6.791

References

[1] D.C. Montgomery, C.L. Jennings, M. Kulahci, Introduction to Time Series Analysis and Forecasting, second ed., John Wiley & Sons, 2015.

[2] F. Chan, L.L. Pauwels, Some theoretical results on forecast combinations, Int. J. Forecast. 34 (1) (2018) 64–74.

[3] B. Fatimah, P. Aggarwal, P. Singh, A. Gupta, A comparative study for predictive monitoring of COVID-19 pandemic, Appl. Soft Comput. 122 (2022) 108806.

[4] J.M. Bates, C.W.J. Granger, The combination of forecasts, J. Oper. Res. Soc. 20 (4) (1969) 451–468.

[5] Z. Wang, H. Chen, J. Zhu, Z. Ding, P.M. Daily, 2.5 And PM10 forecasting using linear and nonlinear modeling framework based on robust local mean decomposition and moving window ensemble strategy, Appl. Soft Comput. 114 (2022) 108110.

[6] S. Makridakis, R.L. Winkler, Averages of forecasts: some empirical results, Manage. Sci. 29 (7) (1983) 987–996.

[7] H.J. Park, Y. Kim, H.Y. Kim, Stock market forecasting using a multi-task approach integrating long short-term memory and the random forest framework, Appl. Soft Comput. 114 (2022) 108106.

[8] A. Ashofteh, J.M. Bravo, A conservative approach for online credit scoring, Expert Syst. Appl. 176 (2021) 114835.

[9] F. Perla, S. Scognamiglio, T. Mathonsi, T.L. Van Zyl, A statistics and deep learning hybrid method for multivariate time series forecasting and mortality modeling, Forecast. 2022 4 (1) (2021) 1–25.

[10] R.J. Hyndman, H. Booth, F. Yasmeen, Coherent mortality forecasting: The product-ratio method with functional time series models, Demography 50 (1) (2013) 261–283.

[11] M. Scortichini, et al., Excess mortality during the COVID-19 outbreak in Italy: A two-stage interrupted time-series analysis, Int. J. Epidemiol. 49 (6) (2020) 1909–1917.

[12] J.M. Bravo, J.P.V. Nunes, Pricing longevity derivatives via Fourier transforms, Insurance Math. Econom. 96 (2021) 81–97, http://dx.doi.org/10.1016/j.insmatheco.2020.10.008.

[13] R.S. Tsay, Analysis of Financial Time Series, third ed., Wiley, 2010.

[14] L. Breiman, Bagging predictors, Mach. Learn. 24 (2) (1996) 123–140.

[15] A.O. Akyuz, M. Uysal, B.A. Bulbul, M.O. Uysal, 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, 2017, pp. 7–12.

[16] J.M. Bravo, M. Ayuso, Forecasting the retirement age: A Bayesian model ensemble approach, in: Advances in Intelligent Systems and Computing, 1365 AIST, 2021, pp. 123–135.

[17] M. Ayuso, J.M. Bravo, R. Holzmann, E. Palmer, Automatic indexation of the pension age to life expectancy: When policy design matters †, Risks 9 (5) (2021) 96.

[18] A.E. Raftery, T. Gneiting, F. Balabdaoui, M. Polakowski, Using Bayesian model averaging to calibrate forecast ensembles, Mon. Weather Rev. 133 (5) (2005) 1155–1174.

[19] M. Aiolfi, C. Capistrán, A. Timmermann, M. Aiolfi, C. Capistrán, A. Timmermann, Forecast combinations, in: Work. Pap, 2010.

[20] L.M. de Menezes, D.W. Bunn, J.W. Taylor, L.M. de Menezes, D.W. Bunn, J.W. Taylor, Review of guidelines for the use of combined forecasts, European J. Oper. Res. 120 (1) (2000) 190–204.

[21] V.R.R. Jose, R.L. Winkler, Simple robust averages of forecasts: Some empirical results, Int. J. Forecast. 24 (1) (2008) 163–169.

[22] R.R. Andrawis, A.F. Atiya, H. El-Shishiny, R.R. Andrawis, A.F. Atiya, H. El-Shishiny, Forecast combinations of computational intelligence and linear models for the NN5 time series forecasting competition, Int. J. Forecast. 27 (3) (2011) 672–688.

[23] J.D. Samuels, R.M. Sekkel, Model confidence sets and forecast combination, Int. J. Forecast. 33 (1) (2017) 48–60.

[24] C. Simões, L. Oliveira, J.M. Bravo, Immunization strategies for funding multiple inflation-linked retirement income benefits, Risks 9 (4) (2021) 60.

[25] A. Ashofteh, J.M. Bravo, Life table forecasting in COVID-19 times: An ensemble learning approach, in: 16th Iberian Conference on Information Systems and Technologies, CISTI, 2021, pp. 1–6.

[26] J. Gama, I. Zliobaite, A. Bifet, M. Pechenizkiy, A. Bouchachia, A survey on concept drift adaptation, ACM Comput. Surv. 46 (4) (2014) 1–37.
[27] M. Pawlikowski, A. Chorowska, Weighted ensemble of statistical models, Int. J. Forecast. 36 (1) (2020) 93–97.
[28] S. Makridakis, E. Spiliotis, V. Assimakopoulos, The M4 competition: 100,000 time series and 61 forecasting methods, Int. J. Forecast. 36 (1) (2020) 54–74.
[29] F. Checchi, L. Roberts, Interpreting and using Mortality Data in Humanitarian Emergencies, London, 2005, 52, Sep..
[30] WHO, Global excess deaths associated with the COVID-19 pandemic, technical advisory report, 2022, [Online] Available: https://www.who.int/data/stories/global-excess-deaths-associated-with-covid-19-january-2020-december-2021. (Accessed 05 Jun 2022).
[31] M. Aiolfi, A. Timmermann, Persistence in forecasting performance and conditional combination strategies, J. Econ. 135 (1–2) (2006) 31–53.
[32] A. Ashofteh, J.M. Bravo, A study on the quality of novel coronavirus (COVID-19) official datasets, Stat. J. IAOS 36 (2) (2020) 291–301.
[33] S. Cui, Y. Wang, D. Wang, Q. Sai, Z. Huang, T.C.E. Cheng, A two-layer nested heterogeneous ensemble learning predictive method for COVID-19 mortality, Appl. Soft Comput. 113 (2021) 107946.
[34] D.A. Leon, V.M. Shkolnikov, L. Smeeth, P. Magnus, M. Pechholdová, C.I. Jarvis, COVID-19: a need for real-time monitoring of weekly excess deaths, Lancet 395 (10234) (2020) e81.
[35] V. Kontis, et al., Magnitude, demographics and dynamics of the effect of the first wave of the COVID-19 pandemic on all-cause mortality in 21 industrialized countries, Nat. Med. 2020 2612 26 (12) (2020) 1919–1928.
[36] V. Cerqueira, L. Torgo, C. Soares, Early anomaly detection in time series: A hierarchical approach for predicting critical health episodes, oct, 2020.
[37] A. Ashofteh, Big data for credit risk analysis: Efficient machine learning models using pyspark, in: Proceedings of SIMSTAT 2019-10th International Workshop on Simulation and Statistics, 2019.
[38] A.T. Sergio, T.P.F. de Lima, T.B. Ludermir, Dynamic selection of forecast combiners, Neurocomputing 218 (2016) 37–50.
[39] A. Ashofteh, J.M. Bravo, M. Ayuso, Time series data for monthly respiratory diseases deaths in 61 countries from 2000 to 2016 - mendeley data, mendeley data, 2021.
[40] R.J. Hyndman, G. Athanasopoulos, Forecasting: Principles and Practice, third ed., OTexts, Melbourne, Australia, 2021, Hyndman, Rob J, Athanasopoulos, George: 8601404468544: Amazon.com: Books.
[41] R.B. Cleveland, W.S. Cleveland, J.E. McRae, I. Terpenning, STL: A seasonal-trend decomposition procedure based on loess (with discussion), J. Off. Stat. 6 (1990) 3–73.
[42] T.T.H. Phan, É. Poisson Caillault, A. Bigand, Comparative study on univariate forecasting methods for meteorological time series, in: European Signal Processing Conference, 2018, pp. 2380–2384, 2018-Septe.
[43] R.J. Hyndman, Y. Khandakar, Automatic time series forecasting: The forecast package for R, J. Stat. Softw. 27 (3) (2008) 1–22.
[44] M. Kim, Z. Gu, S. Yu, G. Wang, L. Wang, Methods, challenges, and practical issues of COVID-19 projection: A data science perspective, J. Data Sci. 19 (2) (2021) 219–242.
[45] G. Bin Huang, Q.Y. Zhu, C.K. Siew, Extreme learning machine: A new learning scheme of feedforward neural networks, in: IEEE Int. Conf. Neural Networks - Conf. Proc., Vol. 2, 2004, pp. 985–990.
[46] S.K. Safi, O.I. Sanusi, A hybrid of artificial fneural network, exponential smoothing, and ARIMA models for COVID-19 time series forecasting, Model. Assist. Stat. Appl. 16 (1) (2021) 25–35.
[47] H. Hassani, R. Mahmoudvand, Singular Spectrum Analysis, Palgrave Macmillan UK, London, 2018.
[48] R. Hyndman, et al., Forecasting functions for time series and linear models, 2020, p. 140.
[49] A.M. de Livera, R.J. Hyndman, R.D. Snyder, Forecasting time series with complex seasonal patterns using exponential smoothing, J. Amer. Statist. Assoc. 106 (496) (2011) 1513–1527.
[50] D. of I. E. and R. World health organization, global health estimates 2016: Deaths by cause, age, sex, by country and by region, 2000–2016., worksheet labeled ASDR2016, 2018.
[51] Z. Zhang, W.C. Hong, Electric load forecasting by complete ensemble empirical mode decomposition adaptive noise and support vector regression with quantum-based dragonfly algorithm, Nonlinear Dynam. 98 (2) (2019) 1107–1136.
[52] World health organization, coronavirus disease (COVID-19) situation reports, Jan, 2021.
[53] A.V. Georgiou, The manipulation of official statistics as corruption and ways of understanding it, Stat. J. IAOS (2021) 1–21, Preprint, no. Preprint.
[54] A. Ashofteh, J.M. Bravo, Data science training for official statistics: A new scientific paradigm of information and knowledge development in national statistical systems, Stat. J. IAOS 37 (3) (2021) 771–789.
[55] T. Kawashima, et al., Excess all-cause deaths during coronavirus disease pandemic, Japan, January-2020, Emerg. Infect. Diseases 27 (3) 789–795, 01-Mar-2021.
[56] T. Yorifuji, N. Matsumoto, S. Takao, Excess all-cause mortality during the COVID-19 outbreak in Japan, J. Epidemiol. 31 (1) (2021) 90–92.
[57] M.S. Shin, B. Sim, W.M. Jang, J.Y. Lee, Estimation of excess all-cause mortality during COVID-19 pandemic in Korea, J. Korean Med. Sci. 36 (39) (2021) 1–10.
[58] G.W.R.I. Wijesinghe, R.M.K.T. Rathnayaka, ARIMA and ANN approach for forecasting daily stock price fluctuations of industries in colombo stock exchange, Sri Lanka, in: Proceedings of ICITR 2020-5th International Conference on Information Technology Research: Towards the New Digital Enlightenment, 2020.
[59] S.J. Fong, G. Li, N. Dey, R.G. Crespo, E. Herrera-Viedma, Composite Monte Carlo decision making under high uncertainty of novel coronavirus epidemic using hybridized deep learning and fuzzy rule induction, Appl. Soft Comput. J. 93 (2020) 106282.