Contents lists available at ScienceDirect

International Journal of Forecasting

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

GROEC: Combination method via Generalized Rolling Origin Evaluation

Jose Augusto Fiorucci [a,*], Francisco Louzada [b]

[a] Department of Statistics, University of Brasilia, Brazil
[b] Department of Applied Mathematics and Statistics, University of São Paulo, Brazil

ARTICLE INFO

Keywords:
M4 competition
Forecast combination
Theta models
ARIMA models
Exponential smoothing

ABSTRACT

Combination methods have performed well in time series forecast competitions. This study proposes a simple but general methodology for combining time series forecast methods. Weights are calculated using a cross-validation scheme that assigns greater weights to methods with more accurate in-sample predictions. The methodology was used to combine forecasts from the Theta, exponential smoothing, and ARIMA models, and placed fifth in the M4 Competition for both point and interval forecasting.

1. Introduction

Combined forecast methods have been being studied (Clemen, 1989; Newbold & Granger, 1974) and used in forecast competitions (Assimakopoulos & Nikolopoulos, 2000) for a considerable time. The results of past forecast competitions indicate that combined methods generally outperform individual components (Makridakis & Hibon, 2000). On the other hand, model selection methods have also been used, usually to choose only one model, when it presents the best fit or the most accurate predictions within the sample (Bergmeir & Benítez, 2012; Tashman, 2000).

This paper presents a new approach to the combination of time series forecast methods, in which weights are based on their in-sample accuracy performances. It can be used to combine any methods, such as statistical models and artificial-intelligence-based algorithms.

A particular setting of the proposed approach to combine four statistical models from the Theta, exponential smoothing and ARIMA families, reached the fifth place in the M4 competition (M4 Team, 2018; Makridakis, Spiliotis, & Assimakopoulos, 2018), the most extensive time series competition to date.

The remainder of this paper is organized as follows. Section 2 presents the combination process and the method utilized for the M4 competition, while Section 3 presents the test results using the M4 dataset. Section 4 presents final comments and suggestions for future research.

2. GROE combination method

The methodology used in the M4 competition for combining models set their weights as the inverse of their generalized rolling origin evaluation (GROE) results. Thus, a model that performed less effectively in the in-sample test received less weight than a better-performing one. We therefore refer to the proposed method, described in Sections 2.1 and 2.2, as the GROE combination (GROEC) method.

2.1. GROE loss function

The GROE method (Fiorucci, Pellegrini, Louzada, & Petropoulos, 2015) is a very general process for cross-validation on time series forecasting methods. It generalizes some well-known processes like fixed origin evaluation and rolling origin evaluation (Tashman, 2000).

For a time series $Y_1, Y_2, \dots, Y_n$, let $\hat{Y}_{t+j}$ denote the point forecast for Y_{t+j} using the prior information $Y_1, \dots,$

[*] Corresponding author.
E-mail address: jafiorucci@unb.br (J.A. Fiorucci).

https://doi.org/10.1016/j.ijforecast.2019.04.013
0169-2070/© 2019 Published by Elsevier B.V. on behalf of International Institute of Forecasters.

Y_t (training set) to fit a forecast method or model, where $t \in \{1, 2, \ldots, n-1\}$ is called the origin and $j \in \{1, \ldots, n-t\}$ is the required number of steps ahead to be forecast.

The GROE method is parameterized to utilizes a number p of origins, where the origins are index markets of the time series. They are referred to as $n_1, n_2, \ldots, n_p$, which are updated recursively through the equation

$$n_{i+1} = n_i + m, \tag{1}$$

for $i = 1, \ldots, p-1$, where m denotes the number of points observed between consecutive origins. Since Eq. (1) allows us to compute the origins, n_1, m and p are the only parameters required to set all origins. An additional parameter is included in order to define the number of predictions for each origin, which is denoted by H.

For example, if a time series has length $n = 30$ and the parameters are configured as $n_1 = 20$, $m = 5$, $p = 2$ and $H = 7$, the origins will be the indexes $n_1 = 20$ and $n_2 = 25$, where for the first origin, the method/model will be fitted from the first observed point until the 20th point and the prediction errors will be computed for seven points ($H = 7$), i.e., from the 21st point until 27th point. For the second origin ($n_2 = 25$), the method/model will be fitted from the first point until the 25th point and the prediction errors will be computed for five points (since only five are available), i.e., from index 26 until index 30.

The sum of the errors may be written as the following loss function:

$$GROE = \sum_{i=1}^{p} \sum_{j=1}^{\min(H,\, n-n_i)} g(Y_{n_i+j}\ ,\ \hat{Y}_{n_i+j}), \tag{2}$$

where $g(.,.)$ denotes an error function, either the absolute error ($g(x, y) = |x - y|$) or the squared error ($g(x, y) = (x - y)^2$).

The GROE parameters and parametric spaces are:

- $n_1 \in \{2, \ldots, n-1\}$ denotes the index of the first origin;
- $m \in \{1, \ldots, n-n_1\}$ denotes the distance between consecutive origins;
- $p \in \{1, 2, \ldots, p_{max}\}$ denotes the number of origins, where $p_{max} = 1 + \lfloor (n - n_1)/m \rfloor$ and $\lfloor x \rfloor$ denotes the biggest integer number that is lower than $x \in \mathbb{R}$;
- $H \in \mathbb{N}$ denotes the maximum number of evaluation points after each origin.

2.2. GROEC method used in the M4 competition

Four forecast models were combined by GROEC for the M4 competition: two statistical models, namely the dynamic optimized Theta model (DOTM) and the optimized Theta model (OTM) proposed by Fiorucci, Pellegrini, Louzada, and Petropoulos (2016); and two expert systems, ETS and ARIMA (also called Auto-ARIMA) described by Hyndman and Khandakar (2008). All code was implemented in the R software (R Core Team, 2017) using the functions `dotm()` and `otm()` from the *forecTheta* package (Fiorucci, Louzada, & Yiqi, 2016) and the functions `ets()` and `arima()` from the *forecast* package (Hyndman et al., 2018).

2.2.1. Error function

We obtain an error function that is as close as possible to the overall weighted average (OWA) metric, the official metric for point forecast of the M4 competition (M4 Team, 2018), by developing the following error function and combining it with the GROE cross-validation process given in Eq. (2):

$$g(Y_{n_i+j}, \hat{Y}_{n_i+j}) = 0.5\frac{sAPE(Y_{n_i+j}, \hat{Y}_{n_i+j})}{sMAPE_i^*} + 0.5\frac{ASE(Y_{n_i+j}, \hat{Y}_{n_i+j})}{MASE_i^*}, \tag{3}$$

where *sAPE* and *ASE* denote the error functions, the symmetric absolute percentage error and the absolute scaled error, which are given by

$$sAPE(Y_{n_i+j}, \hat{Y}_{n_i+j}) = \frac{100\,|\hat{Y}_{n_i+j} - Y_{n_i+j}|}{(|\hat{Y}_{n_i+j}| + |Y_{n_i+j}|)/2}$$

and

$$ASE(Y_{n_i+j}, \hat{Y}_{n_i+j}) = \frac{|\hat{Y}_{n_i+j} - Y_{n_i+j}|}{\frac{1}{n_i - f}\sum_{j=f+1}^{n_i} |Y_t - Y_{t-f}|},$$

respectively, where f denotes the time series frequency (i.e., $f = 1$ for yearly series, $f = 12$ for monthly series).

The OWA metric computes relative differences from the results produced by the method called Naive 2 (M4 Team, 2018). We reproduce this metric in the cross-validation scheme by including in the error function in Eq. (3) the symmetric mean absolute error metric and the mean absolute scaled error, both computed for the Naive 2 method, which are given by

$$sMAPE_i^* = \frac{1}{\min(H,\ n - n_i)} \sum_{j=1}^{\min(H,\, n-n_i)} sAPE(Y_{n_i+j}, \hat{Y}^*_{n_i+j})$$

and

$$MASE_i^* = \frac{1}{\min(H,\ n - n_i)} \sum_{j=1}^{\min(H,\, n-n_i)} ASE(Y_{n_i+j}, \hat{Y}^*_{n_i+j}),$$

respectively, where $\hat{Y}^*$ denotes the point forecast produced by the Naive 2 method.

2.2.2. GROE parameters

For each time series from the M4 competition, we set the following configuration for the GROE parameters:

- H as the required number of forecast points, e.g., $H = 6$ for yearly data, $H = 18$ for monthly data, etc.;
- $p = 6$, i.e., six origins for each time series;
- $m = \lfloor H/p \rfloor$, where $\lfloor x \rfloor$ denotes the biggest integer number lower than x; and
- $n_1 = \begin{cases} n - H, & \text{if } n - H \geq 5, \\ 5, & \text{o.w.} \end{cases}$

This parameter configuration was chosen based on the results presented by Fiorucci et al. (2015), who suggest that setting H as the number of forecasts required and the first origin as $n_1 = n - H$ are the best choices. Fiorucci et al.'s results also suggest setting $p = p_{max}$ and $m =$

1; however, given the large number of time series in the competition and the complexity of the forecasting methods involved, we understand that the present configuration ($p = 6$ and $m = \lfloor H/p \rfloor$) would be relatively close to this for the competition, without making the computational cost extreme.

2.2.3. *GROEC forecast algorithm*

For the M4 Competition data, we present the steps that are utilized for computing the point and interval forecasts using the GROEC method below:

Step 1 (Individual forecasts): Compute the required number of point and interval forecasts for each individual method (DOTM, OTM, ETS and ARIMA).

Step 2 (Cross-validation): Compute the GROE loss function in Eq. (2), combined with the error function in Eq. (3), for each individual method, letting $l_1, \ldots, l_4$ be these results.

Step 3 (Weights): Set $s_i = 1/l_i$, for $i = 1, \ldots, 4$, as the individual scores. Then, individual weights are defined as $w_i = s_i/s$ for $i = 1, \ldots, 4$, where $s = \sum_{i=1}^{4} s_i$.

Step 4 (Combination): The point and interval forecasts produced in Step 1 from the four methods are then combined with the respective weights.

Step 5 (Remove negative forecasts): Negative point and interval forecasts are replaced with zero values, if no time series observations are negative.

Although another loss function could be elaborated for combining the interval forecasts obtained by the four methods, we used the same weights as were derived for the point forecasts.

For the expert systems (ETS and ARIMA), we reduced the computational costs by utilizing the model selected in Step 1 to compute the GROE loss function in Step 2. For example, if the ETS algorithm chooses an SES model (simple exponential smoothing), which is a particular case of the ETS, for a given time series, then Step 2 is performed just for the SES model. The code ran on a Dell Optiplex 7050 computer with Windows 10, an i7-7700 processor, and 16 GB of RAM.

The aggregate time to run the 100,000 time series was 61.64 h, giving a mean of 2.22 s per series.

Table 1 provides weight distribution data for all models and indicates that each made approximately the same contribution, viz., about 25% of the overall mean, with around 23% and 27%, respectively, for the first and fourth quartiles. The proximity of the weight distribution suggests that the predictions generated within the sample in the process of cross-validation are close to carrying similar weights for most of the time series. However, when we observe the values of the minimum and maximum weights, we can note that, at least for a small number of series, the GROEC process assigns extreme weights to only one model.

Table 1
GROEC distribution of weights for the M4 dataset.

	DOTM	OTM	ETS	ARIMA
Minimum	0.000	0.000	0.001	0.000
First quartile	0.236	0.229	0.229	0.222
Median	0.251	0.250	0.250	0.249
Mean	0.251	0.247	0.252	0.250
Third quartile	0.268	0.269	0.270	0.273
Maximum	0.768	0.757	1.000	0.997

3. Some posterior tests for the M4 dataset

The GROEC method described in Section 2.2 was derived from tests using the M3 dataset, which took as its reference the results presented by Fiorucci et al. (2015) and Fiorucci, Pellegrini et al. (2016). For the M3 dataset, we considered the premise of selecting the best model from among exponential smoothing, ARIMA, and Theta, using the results for the minimum GROE function. However, preliminary tests indicated that selecting among models would not obtain as good a result as combining one model chosen from each, and that including two theta models (DOTM and OTM) rather than one improves the performance. This is reflected in GROEC, under the conviction that the behavior observed on the M3 dataset would be replicated on the M4 dataset.

In what follows, we verify the behavior of our method against that of a simple average, as well as when one of the four individual methods are disregarded. Table 2 presents some results, now considering the M4 competition database, for six different types of combinations, where GROEC denotes that the algorithm (as shown in Section 2.2) was used to combine the models that appear within parentheses and Simple indicates that a simple mean was used to combine the models. Table 2 also includes the results of the Naive 2 model (M4 Team, 2018), which is used as a reference for computing the out-of-sample performance OWA accuracy metric. Moreover, we also included the sMAPE and MASE metric results.

The results show that the combination of four methods via GROEC weights presents the best results according to all three metrics when all series are considered, with the gain relative to the simple weights method being 0.006 according to the OWA metric. The GROEC combinations of three models revealed that the DOTM, OTM, and ARIMA models contributed most to the performance in the competition, since discarding the ETS model led to little loss of accuracy of the method GROEC (DOTM, OTM, ARIMA) against the GROEC (DOTM, OTM, ETS, ARIMA). However, the GROEC (OTM, ETS, ARIMA) presented the best results for monthly data. The Naive 2 model was outperformed in all situations presented.

4. Final comments and future research

This study proposes a method for combining time series forecast models in which the weights are proportional to the quality of the cross-validation results of each. We have approximated the OWA metric of the M4 competition in the cross-validation process by developing a new

Table 2
Metrics over several model combinations for the M4 database.

	sMAPE				
	Yearly	Quarterly	Monthly	Others	All
GROEC (DOTM, OTM, ETS, ARIMA)	13.676	9.808	12.704	4.352	11.815
Simple (DOTM, OTM, ETS, ARIMA)	13.770	9.867	12.776	4.419	11.889
GROEC (OTM, ETS, ARIMA)	13.975	9.836	12.664	4.322	11.870
GROEC (DOTM, ETS, ARIMA)	13.947	9.824	12.763	4.325	11.908
GROEC (DOTM, OTM, ARIMA)	13.571	9.899	12.737	4.368	11.829
GROEC (DOTM, OTM, ETS)	13.617	9.888	12.925	4.530	11.936
Naive 2	16.342	11.012	14.427	4.754	13.564
	MASE				
	Yearly	Quarterly	Monthly	Others	All
GROEC (DOTM, OTM, ETS, ARIMA)	3.043	1.122	0.908	2.967	1.553
Simple (DOTM, OTM, ETS, ARIMA)	3.069	1.128	0.916	3.029	1.568
GROEC (OTM, ETS, ARIMA)	3.120	1.119	0.900	2.962	1.566
GROEC (DOTM, ETS, ARIMA)	3.114	1.119	0.905	2.964	1.567
GROEC (DOTM, OTM, ARIMA)	3.014	1.138	0.916	2.981	1.555
GROEC (DOTM, OTM, ETS)	3.020	1.138	0.932	3.028	1.566
Naive 2	3.974	1.371	1.063	3.169	1.912
	OWA				
	Yearly	Quarterly	Monthly	Others	All
GROEC (DOTM, OTM, ETS, ARIMA)	0.801	0.855	0.867	0.926	0.842
Simple (DOTM, OTM, ETS, ARIMA)	0.807	0.859	0.874	0.942	0.848
GROEC (OTM, ETS, ARIMA)	0.820	0.855	0.862	0.922	0.847
GROEC (DOTM, ETS, ARIMA)	0.818	0.854	0.868	0.922	0.849
GROEC (DOTM, OTM, ARIMA)	0.794	0.864	0.872	0.930	0.843
GROEC (DOTM, OTM, ETS)	0.797	0.864	0.886	0.954	0.850
Naive 2	1.000	1.000	1.000	1.000	1.000

error function that takes the accuracy of the Naive 2 model as a reference for measuring the in-sample error.

The results demonstrate that the forecast points that our method obtains are superior to those in which equal weights are used. According to the OWA metric, the gain that can be attributed to GROEC weights relative to simple weights was equal to 0.006. Although this difference may seem insignificant, the difference between the second-place and sixth-place competitors in the M4 Competition was equal to only 0.01 (Makridakis et al., 2018), which shows that a difference in the third decimal place could change our rank in the competition.

We believe that the good performance of our method in the M4 Competition was due to the choice of the four models to be combined, as well as the attribution of weights that the GROEC allows. They tend to be the same for most of the time series, but the method is also able to assign insignificant weights when one of the models being used performs poorly in the in-sample tests.

Further research should examine parameter configurations according to different types of time series and combining other forecast models, which could enhance the results. A loss function that takes into account interval errors in the cross-validation process could also improve GROEC's performance, particularly for interval forecasts.

The R code for reproducing the results of the GROEC method in the M4 competition is available at github.com/M4Competition/M4-methods.

Acknowledgments

This work was partially supported by the Brazilian organizations CNPq and FAPESP.

References

Assimakopoulos, V., & Nikolopoulos, K. (2000). The theta model: a decomposition approach to forecasting. *International Journal of Forecasting*, *16*(4), 521–530.

Bergmeir, C., & Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. *Information Sciences*, *191*, 192–213.

Clemen, R. T. (1989). Combining forecasts: a review and annotated bibliography. *International Journal of Forecasting*, *5*(4), 559–583.

Fiorucci, J. A., Louzada, F., & Yiqi, B. (2016). Forectheta: forecasting time series by theta models. R package version 2.3. https://CRAN.R-project.org/package=forecTheta.

Fiorucci, J. A., Pellegrini, T. R., Louzada, F., & Petropoulos, F. (2015). The optimised theta method. arXiv preprint arXiv:1503.03529.

Fiorucci, J. A., Pellegrini, T. R., Louzada, F., Petropoulos, F., & Koehler, A. B. (2016). Models for optimising the theta method and their relationship to state space models. *International Journal of Forecasting*, *32*(4), 1151–1161.

Hyndman, R., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., Petropoulos, F., Razbash, S., Wang, E., & Yasmeen, F. (2018). Forecast: forecasting functions for time series and linear models. R package version 8.3. http://pkg.robjhyndman.com/forecast.

Hyndman, R., & Khandakar, Y. (2008). Automatic time series forecasting: the forecast package for R. *Journal of Statistical Software*, *27*, 1–22.

M4 Team (2018). M4 competitor's guide: prizes and rules. https://www.m4.unic.ac.cy/wp-content/uploads/2018/03/M4-Competitors-Guide.pdf.

Makridakis, S., & Hibon, M. (2000). The M3-competition: results, conclusions and implications. *International Journal of Forecasting*, *16*(4), 451–476.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). The M4 competition: results, findings, conclusion and way forward. *International Journal of Forecasting*, *34*(4), 802–808.

Newbold, P., & Granger, C. W. (1974). Experience with forecasting univariate time series and the combination of forecasts. *Journal of the Royal Statistical Society. Series A (General)*, *137*, 131–165.

R Core Team (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, https://www.R-project.org/.

Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: an analysis and review. *International Journal of Forecasting*, *16*(4), 437–450.

José Augusto Fiorucci is a Full Professor at Department of Statistics, University of Brasilia (UnB), Brazil. He holds Bachelor's and Master's degree in Computational Mathematics from University of São Paulo (USP) and Ph.D. degree in Statistics from Federal University of São Carlos (UFSCar). Current research are related to time series analysis and econometrics. He has supervised M.Sc students.

Francisco Louzada is a Professor of Statistics at the Institute of Mathematical Science and Computing, University of São Paulo (USP), Brazil, and director for the Center for Mathematics and Statistics Applied to Industry (CeMEAI), USP, Brazil. He received his Ph.D. degree in Statistics from the University of Oxford, UK, his M.Sc degree in Computational Mathematics from USP, and his B.Sc. degree from UFSCar, Brazil. His main interests are data science and statistics. He has published books, papers in scientific journals and book chapters. He has supervised PosDoc, Ph.D. and M.Sc. students.