Software Impacts 20 (2024) 100644

Contents lists available at ScienceDirect

Software Impacts

journal homepage: www.journals.elsevier.com/software-impacts

Original software publication

alPCA: An automatic software for the selection and combination of forecasts in monthly series

Carlos García-Aroca, Mª. Asunción Martínez-Mayoral, Javier Morales-Socuéllamos, José Vicente Segura-Heras *

I.U. Operations Research Center, University Miguel Hernandez of Elche, Avda. Ferrocarril s/n, 03202 Elche, Spain

ARTICLE INFO

Keywords:
Forecasting
Time series
Forecasting combination
Principal component analysis

ABSTRACT

alPCA is a software coded in R and designed to automatically combine predictions from a collection of individual forecasting methods that integrate it. It employs three categories of weights derived from the PCA scores, and decision rules to determine the optimal combination of these methods. alPCA serves as an automated component within the artificial intelligence toolkit for monthly time series processing with the objective of obtaining the best forecast.

Code metadata

Current code version	v2.0
Permanent link to code/repository used for this code version	https://github.com/SoftwareImpacts/SIMPAC-2024-36
Permanent link to Reproducible Capsule	
Legal Code License	GNU General Public License (GPL)
Code versioning system used	none
Software code languages, tools, and services used	r
Compilation requirements, operating environments & dependencies	R (4.2.1, RStudio 2022.07.0-548)
If available Link to developer documentation/manual	https://doi.org/10.1016/j.eswa.2023.121636
Support email for questions	carlos.garcia62@goumh.umh.es

1. Introduction

alPCA is an intelligent tool that combines forecasts provided by 11 forecasting methods and models, running through all their respective configurations (52 in total). These components that integrate alPCA are available in R and are: Naïve, ETS, auto.arima, Theta, STL, Croston, Prophet, NNAR, TBATS, GRNN, MLP.

Experimental results show that combinations based on a wide diversity of forecasting methods achieve forecasts with fewer errors [1,2]. This achievement is influenced by the type of methods that are combined and, consequently, the quality of the forecasts provided by each method, as well as the estimation of the combination weights assigned to each forecast [3,4]. Combining techniques have evolved from simple methods without estimation (arithmetic mean) to sophisticated approaches that include time-varying weights, nonlinear combinations, correlations between components or cross-learning [5,6].

The development of software for forecast combination has provided good forecasting results even in complex environments. Especially, they have proven to be very effective in time series forecasting in various real-world domains, such as sales, production, demand, energy or health [7–13].

Throughout this article, we will show how alPCA carries out the process of combining or aggregating forecasts provided by experts (individual methods or models), and we will do so from the point of view of programming and applied statistics. Other packages that employ aggregation techniques such as those described here are also available in R. These packages include Opera (Online Prediction by ExpeRt Aggregation) [10], ForecastComb [14], and ForecastHybrid [15]. The ForecastHybrid package (version 5.0.19) utilizes forecasts generated from experts such as ARIMA, ETS, theta method, NNETAR, STLM, TBATS, and SNAIVE (naive seasonal). The characteristics and detailed comparison between them, for the case of the seven monthly time series we attach, are shown in [16].

DOI of original article: https://doi.org/10.1016/j.eswa.2023.121636.
* Corresponding author.
E-mail addresses: carlos.garcia62@goumh.umh.es (C. García-Aroca), asun.mayoral@umh.es (M.A. Martínez-Mayoral), j.morales@umh.es (J. Morales-Socuéllamos), jvsh@umh.es (J.V. Segura-Heras).

https://doi.org/10.1016/j.simpa.2024.100644
Received 15 March 2024; Accepted 4 April 2024

2. Software details

This software is the coding in R [17] (4.2.1, RStudio 2022.07.0-548) of the algorithm described in the following Section 3. It has been originally run on an Intel(R) Core(TM) i5-10400 CPU @ 2.90 GHz.RAM: 12 GB system working sequentially in a Windows 11 environment. It has been subsequently run-on other systems to verify its reproducibility and uploaded to the Ocean platform (https://doi.org/10.24433/CO.1623552.v2).

The software has a modular structure and incorporates 11 individual prediction methods (52 configurations), all of them available in R. At the start of the application the needed libraries are loaded (assuming the user has them all previously installed).

In addition, it includes six own functions, programmed to perform calculations and statistical analyses complementary to the methods, in order to achieve the best combination depending on the characteristics of the time series to predict: trends, seasonal and nonseasonal cycles, intermittency, pulses and steps, and outliers. These functions included in alPCA in order to achieve the best prediction are:

1. Calculo1(). Sets each configuration of a model in period T_1 (Training phase/sample), and generates the predictions for period T_2 (Validation phase/sample) for each configuration.
2. Calculo2(). Calculates the loss functions in T_1 and T_2.
3. Calculo3(). Constructs the error matrix (with RMSE, MASE, SMAPE and OWA).
4. Calculo4.m(). Calculates the PCs and Manhattan distance to the method with the smallest error.
5. Calculo5(). Calculates the weights associated to each combination according to the PC analysis.
6. Calculo6(). Obtains the predictions for the period $T_1 + T_2$.

To select the number of method configurations used in the combination of predictions we used 3 percentiles. The 5th percentile if we want few, the 95th percentile if we want to consider almost all of them, and the 50th percentile if we want an intermediate number of them. As a result of the algorithm we show the predictions for the next 12 months in the three cases. The scores of each configuration in the selected principal components are used as weights. $w = |x_j|/(\sum_{j=1}^{m} |x_j|)$

alPCA is presented in the Ocean platform with an example of CO2 levels discussed in detail in Section 5 of this paper. The datasets are incorporated from the working directory as .rds files. The associated function file in this case is f_alPCA.monthly.r.

3. Software algorithms

procedure alPCA algorithm(S,Y, N, h, j, i)
  Let Y be a monthly time series with length N.
  Divide Y into two subsets: $T_1 = y_1, y_2, ..., y_{n-h}$, $T_2 = y_{n-h+1}, ..., y_N$ with $h = 12$
  Let M_j be one of the J considered forecasting configurations.
  for all $j \in J$ **do**
    Obtains the fitting values in T_1 and predictions at T_2
    Calculate sMape and Mase in T_1 , and sMAPE, MASE, RMSE and OWA in T_2
  end for
  Builds $E_{J,6}$ with the six loss functions and removes duplicate rows
  Obtains a PCA score
  Order the rows of $E_{J,6}$ according to their Manhattan distance to the best method*.
  Obtains quantiles I_5, I_{50} and I_{95} for empirical distribution to the distance to the best method. $I = I_5, I_{50}, I_{95}$
  Let $K_I \subset J$ be the final subset of selected configurations
  for all $i \in I$ **do**
    for all $z \in K_i$ **do**
      Determine the weight $w = |x_j|/(\sum_{j=1}^{m} |x_j|)$ to combine the z configurations
      Apply the M_z configuration to fit Y and generate the h-step-ahead forecast $\hat{y}_t^z$ for $t = N_1, ..., N_h$
    end for
  end for
  Compute the combined forecasts: $\hat{y}_{i,t} = \sum_{z \in K_i} w_t \hat{y}_t^z$
end procedure

4. Related work

Data analysis has been carried out through alPCA on the article by García-Aroca C. et al. (2024) which has facilitated the treatment of the data used. It has been applied to forecast sales of four brands of automobiles in the USA, production and consumption of electricity in the UK and to forecast CO2 levels in a given region in order to assess their economic impact. It was also used to compare the goodness of forecasts with those provided by ForecastHybrid package (version 5.0.19), ForecastComb package (version 1.3.1) and Opera (version 1.2.1).

5. Illustrative examples

The average monthly data of CO2 mole fraction (co2-mm-mlo), which represents CO2 emissions, have been taken as an illustrative example. The series is obtained from daily averages and covers from 01/01/2000 to 31/12/2019 (20 years and 240 observations). The data are from the U.S. Government's Earth System Research Laboratory, Global Monitoring Division, and are available from the Trends in Atmospheric Carbon Dioxide website https://datahub.io/core/co2-ppm#data. When the algorithm is finished, it generates a 12 × 3 dimension prediction matrix under the name result_alPCA, the first column contains the predictions obtained with the 5th percentile, the second column those obtained with the 50th percentile and the last column those obtained with the 95th percentile.

h	percentile.5	percentile.50	percentile.95
1	383.08	382.17	381.91
2	377.46	381.66	381.67
3	377.79	382.11	382.25
4	380.68	382.53	382.89
5	376.17	381.93	381.89
6	374.31	381.30	381.25
7	373.28	381.17	381.22
8	377.15	382.24	382.30
9	376.49	381.74	382.41
10	377.05	381.78	382.19
11	378.58	382.28	382.85
12	377.83	381.97	382.65

6. Impact overview

The software that implements the alPCA algorithm is the result of the research developed by our team and that emerged from the demanded search for knowledge on the spare parts consumption market at aerospace companies. In this industry, as in many others, it is critical to develop mechanisms to analyze information on consumer intentions, in order to efficiently forecast the supply of materials and production in the short, medium and long term. This objective is a constant in the automotive and energy industries, among others, and has a decisive influence on budgets, production optimization and warehousing.

In our publication, [16], we show that our proposal generates more accurate forecasts than other forecast combination packages implemented in R: ForecastComb, Opera and ForecastHybrid. In addition, we obtained a lower prediction error than the individual methods and models that integrate it. This was tested with monthly time series of the following types:

- Those constructed from actual data on parts manufactured by the aerospace industry.
- Broad subsets of the 48,000 monthly time series available in the M4 competition.

- Total new car sales in the USA, plus FORD, GM, HONDA and TOYOTA sales.
- CO2 emission levels
- Consumption of electricity.

In addition, a sensitivity analysis was performed on the results obtained to evaluate the robustness of our algorithm or its sensitivity to different specifications.

7. Future development of the software

We are working on incorporating some new method with its corresponding configurations. We want to simplify the selection of configurations involved in the combination to choose them by some statistical test in a unique way. We will extend the algorithm for any seasonality (yearly, quarterly, weekly, ...). Finally, we will design and implement an R package.

CRediT authorship contribution statement

Carlos García-Aroca: Investigation, Methodology, Software, Validation, Writing – original draft, Data curation. **Mª. Asunción Martínez-Mayoral:** Conceptualization, Formal analysis, Investigation, Supervision, Writing – review & editing. **Javier Morales-Socuéllamos:** Formal analysis, Investigation, Software, Validation, Writing – review & editing. **José Vicente Segura-Heras:** Conceptualization, Methodology, Software, Validation, Writing – review & editing, Data curation, Funding acquisition.

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.

Data availability

Data are public and accessible.

Acknowledgments

José V. Segura-Heras acknowledges the financial support from the Generalitat Valenciana, Spain under project PROMETEO/2021/063.

References

[1] M.E. Thomson, A.C. Pollock, D. Önkal, M.S. Gönül, Combining forecasts: Performance and coherence, Int. J. Forecast. 35 (2) (2019) 474–484, http://dx.doi.org/10.1016/j.ijforecast.2018.10.006.

[2] K.C. Lichtendahl, R.L. Winkler, Why do some combinations perform better than others? Int. J. Forecast. 36 (1) (2020) 142–149, http://dx.doi.org/10.1016/j.ijforecast.2019.03.027.

[3] A. Timmermann, Forecast combinations, in: G. Elliott, C.W.J. Granger, A. Timmermann (Eds.), in: Chapter 10 in Handbook of Economic Forecasting, vol. 1, Elsevier, 2006, pp. 135–196, http://dx.doi.org/10.1016/S1574-0706(05)01004-9.

[4] S. Cang, H. Yu, A combination selection algorithm on forecasting, European J. Oper. Res. 234 (1) (2014) 127–139, http://dx.doi.org/10.1016/j.ejor.2013.08.045.

[5] X. Li, Y. Kang, F. Li, Forecasting with time series imaging, Expert Syst. Appl. 160 (113680) (2020) http://dx.doi.org/10.1016/j.eswa.2020.113680.

[6] P. Montero-Manso, G. Athanasopoulos, R.J. Hyndman, T.S. Talagala, FFORMA: Feature-based forecast model averaging, Int. J. Forecast. 36 (1) (2020) 86–92, http://dx.doi.org/10.1016/j.ijforecast.2019.02.011.

[7] J. Stock, M. Watson, Forecasting with many predictors, in: Chapter 10 in Handbook of Economic Forecasting, vol. 1, Elsevier, 2006, pp. 515–554, http://dx.doi.org/10.1016/S1574-0706(05)01010-4.

[8] M. Gurnani, Y. Korke, P. Shah, S. Udmale, V. Sambhe, S. Bhirud, Forecasting of sales by using fusion of machine learning techniques, in: International Conference on Data Management, Analytics and Innovation, ICDMAI, 2017.

[9] C. Manescu, I. Van Robays, Forecasting the brent oil price addressing time variation in forecast performance, in: Working Paper Series NO 1735 / 2014, European Central Bank, 2014.

[10] P. Gaillard, Y. Goude, OPERA: Online prediction by expert aggregation, 2016, Retrieved from https://cran.r-project.org/web/packages/opera.

[11] B. Auder, M. Bobbia, J.-M. Poggi, B. Portier, Sequential aggregation of heterogeneous experts for pm10 forecasting, Atmospheric Pollut. Res. 7 (6) (2016) 1101–1109, http://dx.doi.org/10.1016/j.apr.2016.06.013.

[12] G. Perone, Comparison of ARIMA, ETS, NNAR, TBATS and hybrid models to forecast the second wave of COVID-19 hospitalizations in Italy, Eur. J. Health Econ. 23 (6) (2022) 917–940, http://dx.doi.org/10.1007/s10198-021-01347-4.

[13] S.M. Fortsch, J.H. Choi, E.A. Khapalova, Competition can help predict sales, J. Forecast. (2021) http://dx.doi.org/10.1002/for.2818.

[14] C.E. Weiss, E. Raviv, G. Roetzer, Forecast combinations in R using the ForecastComb package, R J. 10 (2) (2018).

[15] D. Shaub, P. Ellis, Forecasthybrid: Convenient functions for ensemble time series forecasts, 2020, Retrieved from https://github.com/ellisp/forecastHybrid.

[16] C. García-Aroca, M.A. Martínez-Mayoral, J. Morales-Socuéllamos, J.V. Segura Heras, An algorithm for automatic selection and combination of forecast models, Expert Syst. Appl. 237 (121636) (2024) http://dx.doi.org/10.1016/j.eswa.2023.121636.

[17] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2020, Retrieved from https://www.R-project.org/.