Muş Alparslan Üniversitesi Mühendislik-Mimarlık Fakültesi Dergisi
1(1),55-63,2020

Perform Time-series Predictions in the R Development Environment by Combining Statistical-based Models with a Decomposition-based Approach

Zeydin Pala[1*], Ahmet Faruk Pala[2]

[1]Muş Alparslan University Faculty of Architecture and Engineering,
Department of Computer Engineering, Muş, Turkey.
[2]İnönü University Faculty of Engineering,
Department of Computer Engineering, Malatya, Turkey.

*Corresponding author: z.pala@alparslan.edu.tr

Abstract

The analysis of a time-series (TS) measured or obtained by observing any area is an important step in characterizing a desired system or a phenomenon and predicting its future behavior. More precisely, predicting the value of an unknown variable is the objective of a predictive model used for TS. While doing this, it analyzes the relationships between past data well and reveals future predictions.

In this study, the prediction method contrasts the decomposition-based approach with non-decomposition-based approaches. In the comparison process, prediction metrics for assessment, such as RMSE, MAE, MPE, and MAPE were used for method achievements and the results obtained were discussed.

The experimental outcomes showed that the proposed decomposition-based approach performs better than non-decomposition-based approach in TS prediction processes.

Keywords:Decomposition-based prediction, non-decomposition-based prediction, time-series, R

1. INTRODUCTION

Modeling and forecasting of time-series (TS) has fundamental significance for different functional domains. Therefore a lot of active research work has been going on in this field for many years. Many relevant models have been suggested in the literature to enhance accuracy and effectiveness of modeling and forecasting TS.

Because of the significant relation between the past and the future, TS analysis is commonly used in many areas. Among these areas, it is possible to list many fields such as finance, sensor data analysis, speech recognition, economics, business, statistics, weather forecasting [1], engineering, environmental sciences or physical sciences [2].

There are some similarities as well as some differences between TS estimation approach and the machine learning (ML approach.

Since TS is based primarily on statistics, the need for competent TS analysis of both statistical and ML techniques increases as continuous phenomenon monitoring and data collection become more common. By combining these two fields, prediction models built TS are becoming more promising [3].

It differs from ML approaches in terms of TS, structures and the way they are analyzed. For example, to create the inputs in ML, a feature vector must first be created. However, TS prediction does not require feature generation, at least in the traditional sense. While more data in both areas is more important for training, it is possible to work with smaller-scale data sets, especially in TS.

The input data for an ML may be independent, but for the TS, this is not valid. The next entry is associated with each entry in the TS. In TS, each input is also the output of the previous observation [4] is also the product of each input. For example, in most ML models, a model is learned, tested, retrained if required until satisfactory results are obtained, and then validated with the new data set.
It is understood that when the linear regression model is established, the errors are independent from each other. In TS, however the case is different and it is accepted that each other is affected by the errors. To put it more simply, depending on the moment, the error words refer to the previous one. Therefore, relative to TS techniques, ML methods can be disadvantageous in terms of precision.

Pala et al. [5]demonstrated that decomposition-based models perform better prediction than normal models.

The aim of this research is to use a decomposition-based approach to model TS data as trend, seasonal and residual, which is a very simple but robust method for modeling and predicting TS. We hope that such new techniques will contribute to the literature in the field of prediction.

The remaining part of this research paper is structured as follows: Section 2 outlines TS and modelling. In this section, information about the structure of the TS and the working model of the decomposition method are given. Section 3 includes materials and methods. Section 4 Summarizes the experimental results with three datasets namely, mdeaths, fdeaths, and air passengers. Section 5 deals with the conclusion.

2. TS AND MODELING

TS data are time stamped values obtained from any data source such as sensors, microphones, stock markets. TS data have many important properties that need to be modeled in order to be analyzed effectively.

Comparing to other forecast algorithms, like logistic regession or linear regression, they deal with two variables, but with TS we deal a single variable which is depended on time.

If we represent the data of a TS by x, each data can be represented by xt. T, shown as an index reflects time here. In the case of t = 1, it means the first observation value, and in the case of t = T, it means the last observation value. It is possible to express the whole time set as the observation duration t = 1, 2, 3, ..., T [6]. Observations are determined over the same period of time. There can be an interval that is daily, weekly, monthly, seasonal or annual.

For a time sequence, the future prediction horizon can be interpreted as T + 1, T + 2, T+ 3, ..., T + h. Here it is possible to express h = 1, 2, 3, ..., H as a forecast horizon. With the help of the decomposition model, it is possible to divide a TS into three components, as given in eq. 1:

$$\hat{X}_t = T_t + S_t + R_t \quad (1)$$

Here, $\hat{X}_t$ is the modeled or predicted value at time t, T_t is the trend component at time t, S_t is the seasonal component at time t, and R_t is the remainder component at time t.

In this study, two different monthly datasets named mdeaths and fdeaths, each consisting of 72 records, were used. Each of the datasets includes the numbers of men and women who died of emphysema and asthma disease in the UK between 1974-1979, respectively.

The decomposition method of both datasetes divided into four components as data, trend, seasonal, and remainder is shown in Figure 1, and Figure 2 respectively.

As seen in Figure 1, we can say that the average monthly deaths of the estimated trend component in 1974 started around 1550 and peaked in 1975. We can state that from 1975 to the middle of 1976, deaths dropped to around 1450, and until the middle of 1977 this trend remained steady. We can say that there was a small rise since 1978 and then the decreases continued until 1980 and fell below the 1400s.

Unlike men, steady decreases in monthly female mortality were observed until mid-1976, as shown by the trend component in Figure 2. The number of female deaths, which was around 520 per month since mid-1976, rose to 570 by 1979. Since then, the number of female deaths has declined sharply to around 510 per month.

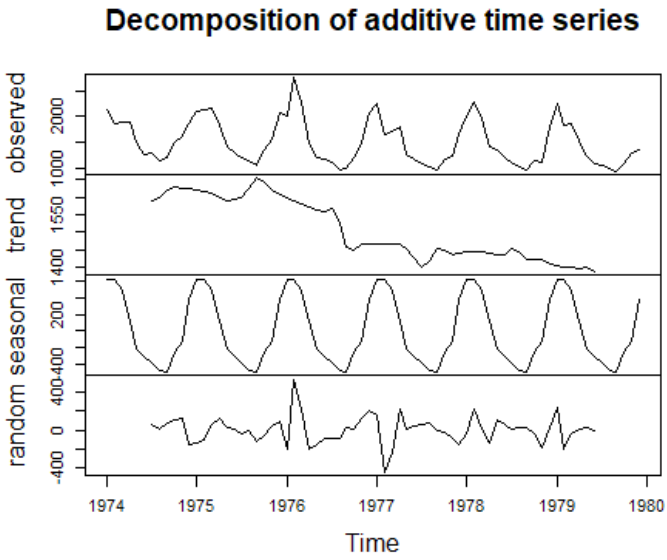

Figure 1: Monthly mdeats dataset shown as four components with the help of decomposition method

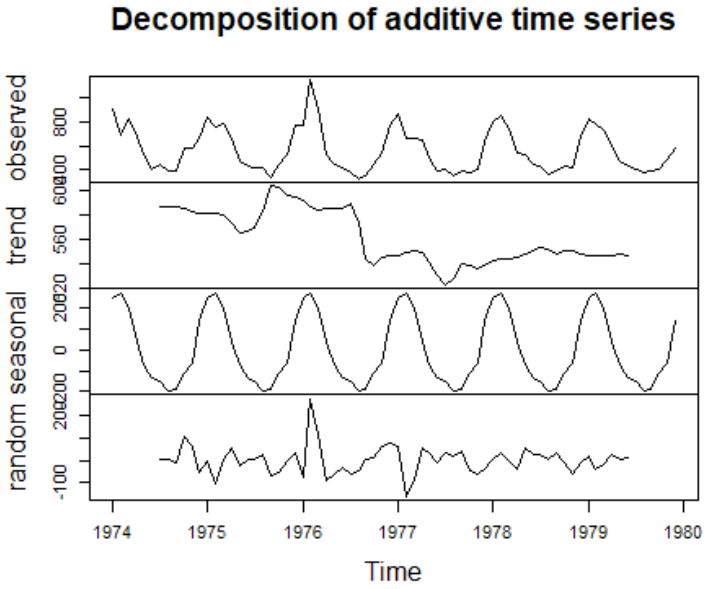

Figure 2: Monthly fdeaths dataset shown as four components with the help of decomposition method

Once seasonal and trend components are removed from the data component, the remaining component is the remainder. Gray bars to the right of each panel show the relative scales for that panel. Each gray bar represents the same length, but because the drawings are of different scales, the sizes of the bars vary.

Before proceeding to TS prediction processes, the method of decomposition is considered as an analysis stage. As an analysis process, this stage can also be considered.

The individual future values of the pattern, seasonal, and remaining components are first determined to forecast the future of a TS using a decomposition model. Then these components are again put together. With this methodology, the challenge is to find the best model for each of the components.

To make decomposition-based estimates in the R environment, we used the local regression (LOESS) technique here. STL is an abbreviation of the LOESS method, the filtering method, and is a flexible and robust method used for TS decomposition [7]. In 1990, the STL algorithm was generated by Cleveland et al. This has been suggested by [8]. With the aid of the LOESS equation, the STL algorithm utilizes polynomial regression to model trend and seasonal components. Assuming that the TS data i = 1 to n are independent for xi and dependent on yi, the LOESS regression curve, g (x), is the smoothing of y given x. It can be calculated for any x-value along the scale of the independent variable [9].

In the R versions of the STL presented in this study, a fixed number of iterations with a default of 2 is used. TS decomposition methods consist of classical, STL, X11, and X-13ARIMA-SEATS decomposition methods. Although STL has many advantages over the classical method, SEATS and X11 parsing techniques, it does not process calendar changes automatically and only provides additive decomposition facilities.

Various statistical tests, including mean percentage error (MPE), mean absolute percent error (MAPE), root mean square error (RMSE), and mean absolute error (MAE), were determined to analyze the performance of models using numerical TS values [10]. The MAPE metric represents the proportion of absolute errors occurring on average. It is independent of the measurement scale but influenced by the transformation of data. The path of error does not reveal it. Extreme variations are not panelized.

MPE is the percentage that occurred during forecasting of the average error. It has similar properties to MAPE, except that it indicates the direction of error. It is desirable that the obtained MPE should be small for a good forecast. The average absolute deviation of forecast values from original values is calculated by the MAE. The Mean Absolute Deviation is often referred to as (MAD). This illustrates the extent of the cumulative error that occurred due to forecasting. The acquired MAE should be as minimal as possible for a good forecast.

The RMSE is nothing but the measured Mean Squared Error square root of (MSE). All of the MSE assets still hold RMSE. In equations 2, 3, 4 and 5, respectively, equations for MPE, MAPE and RMSE and MAE metrics are given [11].

$$\boldsymbol{MPE} = \frac{\mathbf{1}}{\boldsymbol{T}} \sum_{t=1}^{T} \left| \frac{e_t}{y_t} \right| x100 \qquad (2)$$

$$\boldsymbol{MAPE} = \frac{\mathbf{1}}{\boldsymbol{T}} \sum_{t=1}^{T} \left| \frac{e_t}{y_t} \right| x100 \qquad (3)$$

$$\boldsymbol{RMSE} = \sqrt{\frac{\mathbf{1}}{\boldsymbol{T}} \sum_{t=1}^{T} (\boldsymbol{e_{t+h}})^2} \qquad (4)$$

$$\boldsymbol{MAE} = \frac{\mathbf{1}}{\boldsymbol{T}} \sum_{t=1}^{T} |\boldsymbol{e_{t+h}}| \qquad (5)$$

The t, T, e_t, y_t, and h given in the equations show time, number of observations, prediction error, actual value and future forecast horizon, respectively.

3. MATERIAL AND METHOD

TS analysis is generally the process of obtaining meaningful insights from TS dataset using data visualization tools, statistical applications and mathematical models.

Analyzes made here are generally performed by data collection, data preparation, descriptive analysis and predictive analysis steps.

Statistical based models, the Auto.arima, ETS, Naïve and Rwdrift models were used in this analysis, together with the STL method described in the R programming language and environment forecast [12] library.

The ARIMA model, which is popular in the field of statistics, is used as auto.arima for prediction operations in the R environment.

The ARIMA (p, d, q) model is defined by three parameters that have statistical significance in terms of model accuracy. Here p indicates the autoregressive order, d the individual order of difference, and q indicates the moving average size of the window [13]. The ARIMA (p, q) model is mathematically given in eq. 6.

$$X_t = \emptyset_1 X_{t-1} + \emptyset_p X_{t-p} + \cdots + \emptyset_1 \epsilon_{t-1} + \cdots + \epsilon_t \qquad (6)$$

Here, $\emptyset$ shows the coefficients, X is the observation values of the d degree difference, and ϵ is the error term.

For the auto.arima the auto-search function executes the search for the best possible model according to the constraints of the given degree of the equation. The forecasts for ETS points are equal to the medians of the distributions of the projection. The forecast distributions are common for models with only additive parts, so the medians and means are equal. For multiplicative error or multiplicative seasonality ETS models, the point forecasts will not be equal to the mean of the predicted distributions.

As shown in eq. (7), we simply set all forecasts to the value of the last observation for naïve forecasts. That is,

$$\hat{y}_{T+h|T} = y_T. \quad (7)$$

For several financial and financial time series, this technique works surprisingly well. For simplicity, the naive function [14] is simply a wrapper to rwf. Forecast package offers algorithms and modeling frameworks for automatic univariate TS prediction in the R environment. In addition to code and functions, the R packet can contain data sets that support any of the formats (data frame, TS, matrix, etc.) specified as R.

The dataset used in this study was supplied from the Datasets library of the Rstudio development environment and 4.3.0 version of the R programming language was used for TS analysis. A desktop computer with a 64-bit Windows 10 based, intel core i5 dual-core (3.10 GHz) processor and 8 GB main memory was used in the analysis.

The first dataset used here includes the numbers of male patients who died of emphysema in the UK between 1974-1979 and consists of 72 records.

The minimum value, average, and maximum value of the mdeaths monthly dataset used here are 940,1496, and 2750, respectively. While the highest number of deaths occurred in February of 1976, the lowest number of deaths occurred in September 1979.

The second dataset used here includes the numbers of female patients who died of asthma in the UK between 1974-1979 and consists of 72 records.

The minimum value, average, and maximum value of the mdeath monthly dataset used here are 330, 560.7, and 1141, respectively. While the highest number of deaths occurred in February of 1976, the lowest number of deaths occurred in August 1976.

In order to evaluate the performance of the TS estimates, it is necessary to divide them for training and testing operations at certain rates. Although there is no standard for these rates, rates such as (80%, 20%), (75%, 25%) or (70%, 30%) were generally preferred for training and testing. The first 54 records, 75% of the mdeath and fdeaths TS, which is used in this study and consists of 72 records, are reserved for training, while the remaining 18 records, that is, 25%, are reserved for testing.

4. FINDINGS AND DISCUSSION

First, the models were trained with training data and 18-month future predictions were made, and then the performance was evaluated by comparing the forecast results with test data. The 18-month prediction charts of mdeaths and fdeaths datasets, each containing 72 months of data, are shown in Figure 3, and Figure 4 with the help of four different algorithms used with STL. Except for seasonal periods, estimates in general have produced similar results.

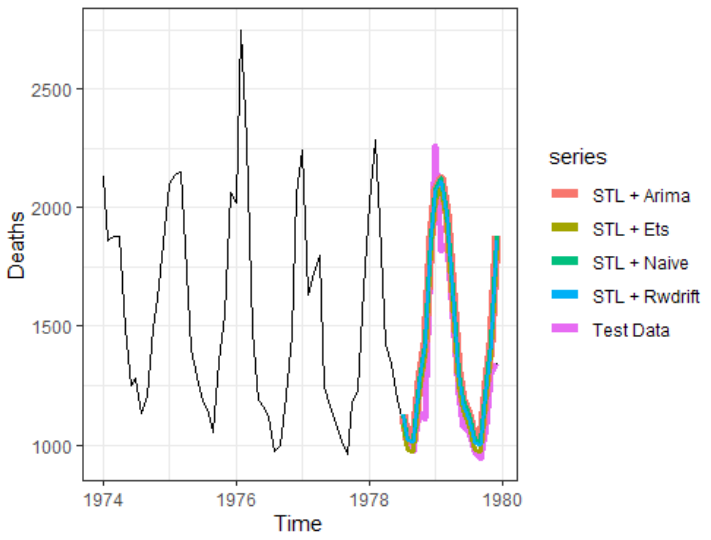

Figure 3:18-month comparative forecasts for monthly mdeaths dataset

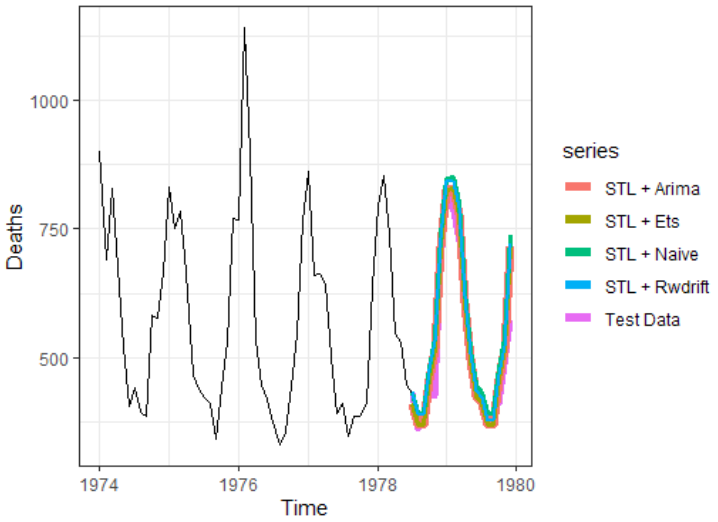

Figure 4:18-month comparative forecasts for monthly fdeaths dataset

In Figure 5, and Figure 6, 18-month estimates are given for the test data with the support of the STL + ETS model, which makes the best estimation. The graph on the left shows the estimate of the number of male deaths, while the graph on the right shows the estimate of the number of female who died. In both graphs, it is seen that the parts other than the peaks are quite well predicted.

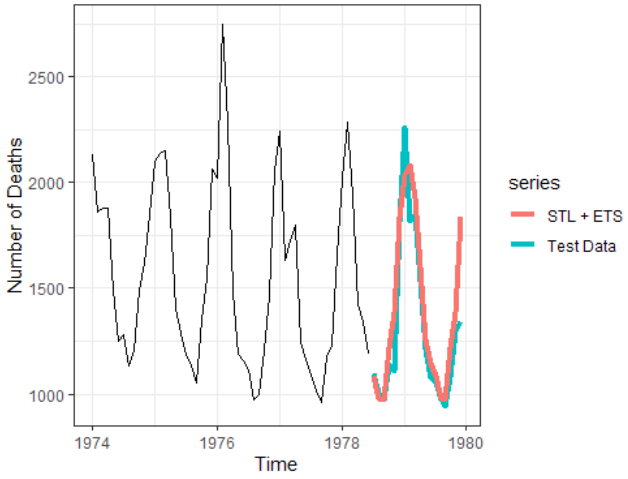

Figure 5: Prediction graphics plotted for 18-month test data of the mdeaths dataset where STL and ETS models are used together

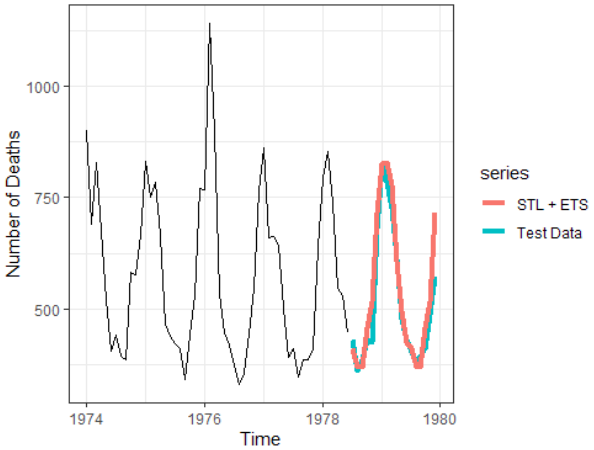

Figure 6: Prediction graphics plotted for 18-month test data of the fdeaths dataset where STL and ETS models are used together

For the mdeaths dataset, with the help of RMSE and MAPE measurement metrics, Figure 7 and Figure 8 were created to show the performance of the algorithms visually and more clearly. When the graphics are examined, it is seen that the performances of the STL_Arima and STL_Naive models with the decomposition approach are quite close.

In addition, the achievements of two different methods, RMSE, MAE, MAPE and MPE metric values are given in Table I, Table II for mdeaths, and fdeaths dataset respectively. In comparison to the naive model, when both the data in the table and the output charts are analyzed, we can assume that in the non-decomposition process, the decomposition approach makes more accurate predictions.

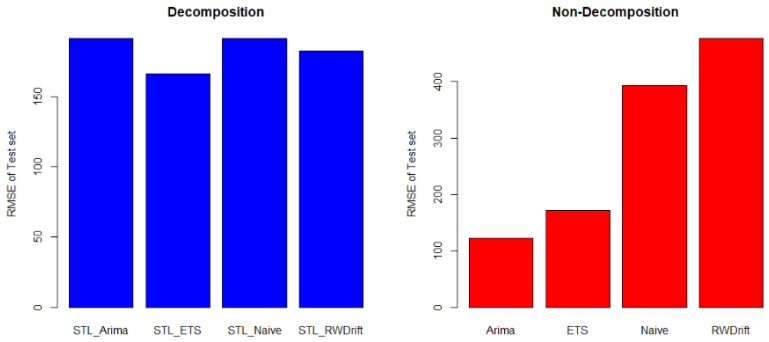

Figure 7: For the mdeaths dataset, showing the RMSE performances obtained for test data of two different approaches with bar graph support

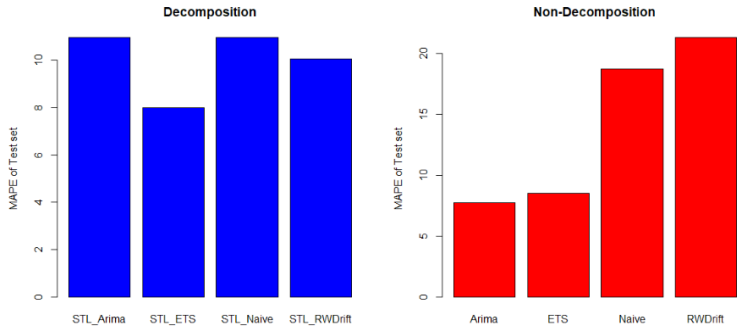

Figure 8: For the mdeaths dataset, showing the MAPE performances obtained for test data of two different approaches with bar graph support

Table I: RMSE, MAE, MAPE and MPE performance numerical values obtained using test data for two different approaches using mdeaths dataset

mdeaths			**Decomposition approach**				**Non-Decomposition approach**	
Method	**RMSE**	**MAE**	**MAPE**	**MPE**	**RMSE**	**MAE**	**MAPE**	**MPE**
Ets	166.32	109.26	7.99	-6.45	171.69	115.05	8.48	-7.12
Rwdrift	182.41	133.60	10.06	-10.05	476.13	334.66	21.30	-3.14
Arima	191.38	144.18	10.96	-10.05	121.95	95.99	7.76	3.44
Naive	191.37	144.18	10.97	-9.10	393.07	280.38	18.74	17.08

ETS, RWDRIFT, ARIMA and NAİVE, respectively, made the best predictions among the four algorithms used for prediction using mdeaths dataset. While the ETS model made a 7.99 % error with the decomposition method of the MAPE metric, the non-decomposition approach made an 8.48% error.
The efficiency of the same algorithm calculated with the MAPE metric for the two approach is 92.01% and 91.52% respectively. To put it more clearly, the decomposition-based approach made 0.49% less errors for the ETS model and thus was more successful than the non-decomposition-based approach.

Table II: RMSE, MAE, MAPE and MPE performance numerical values obtained using test data for two different approaches using fdeaths dataset

fdeaths			**Decomposition approach**				**Non-Decomposition approach**	
Method	**RMSE**	**MAE**	**MAPE**	**MPE**	**RMSE**	**MAE**	**MAPE**	**MPE**
Ets	200.75	129.55	6.83	-4.28	208.18	132.30	6.88	-1.54
Rwdrift	231.16	167.99	9.09	-7.01	683.56	494.00	22.91	23.77
Arima	247.56	186.83	10.24	-4.28	165.57	138.14	8.20	1.01
Naive	247.56	186.84	10.25	-8.88	549.79	391.66	18.65	6.77

ETS, RWDRIFT, ARIMA and NAİVE, respectively, made the best predictions among the four algorithms used for prediction using fdeaths dataset. While the ETS model made a 6.83% error with the decomposition method of the MAPE metric, the non-decomposition approach made an 6.88% error.

The efficiency of the same algorithm calculated with the MAPE metric for the two versions is 93.17% and 93.12% respectively. To put it more clearly, the Decomposition approach made 0.05% less errors for the ETS model and thus was more successful than the non-decomposition approach.

In addition, in the 18-month forecast process using fdeaths dataset, the ETS model performed better than the analysis using mdeaths dataset under the same conditions.

Decomposition approach gives more successful results, especially in seasonal TS. For example, we can obtain more effective results by using the air passengers dataset as a seasonal TS in the R datasets library. When we allocate the first 108 records of the air passengers TS consisting of 144 records for training and the remaining 36 months for testing, the MAPE metric for decomposition and non-decomposition approaches is 5.80% and 7.94%, respectively. This result shows that the decomposition approach makes 2.14% less error than the non-decomposition approach.

5. CONCLUSION

In this study, initially two different monthly datasets named mdeaths and fdeaths, each consisting of 72 records, were used and 18-month estimates were made for each dataset. The datasets used here include the numbers of men and women who died of emphysema and asthma disease in the UK between 1974-1979, respectively. The air passenger was used as a third dataset to show that the metric performances of the results obtained by modeling and estimating the seasonal TS with the decomposition method were better.

In this research, open source R programming language and libraries were used for the analysis of TS.

The proposed decomposition-based approach has shown that the results obtained with the help of evaluation metrics such as RMSE, MAE, MAPE, and MPE are generally more successful than the results obtained with the non-decomposition-based approach. The best result among the four statistical based algorithms used in the prediction process was obtained with the ETS model using mdeaths, and fdeaths datasets. For the mdeaths dataset, the performance of the same model measured by the MAPE metric is 92.01% and 91.52% for the two approaches, respectively. More clearly, using mdeaths dataset the decomposition approach made 0.49% less errors for the ETS model and was more successful than the non-decomposition approach. For the fdeaths dataset, for the two methods, the output of the same model measured by the MAPE metric is 93.17% and 93.12%. More clearly, the decomposition method made 0.05% fewer errors for the ETS model using the fdeaths dataset and was more effective than the non-decomposition strategy.

It was also concluded that the decomposition-based approach is more successful especially on seasonal datasets such as Air passenger.

REFERENCES

[1] Artasanchez A., Joshi P. Artificial Intelligence with Python, Packt Publishing, Birmingham, UK, 2020.
[2] Pala Z., Atici R. Forecasting Sunspot Time Series Using Deep Learning Methods, Solar Physics, 294 (50), 1-14, 2019.
[3] Nielsen A. Practical Time Series Analysis Prediction with Statistics and Machine Learning, O'Reilly Media, Inc., Sebastopol, CA, USA, 2020.
[4] Pala Z. Using Decomposition-based Approaches to Time Series Forecasting in R Environment, International Conference on Data Science, Machine Learning and Statistics, Van, June 26-29, 2019.
[5] Pala Z., Ünlük İ.H., Şahin Ç. Forecasting Low Frequency Electromagnetic Fields Values Time Series Using Python, International Conference on Innovative Engineering Applications (CIEA' 2018), 20-22 Sep, Sivas, Turkey, 2017.
[6] Mills T.C. Applied Time Series Analysis A Practical Guide to Modeling and Forecasting, Academic Press, London, UK, 2019.
[7] Hyndman R.J., Athanasopoulos G. Forecasting: Principles and Practice (second ed.), Monash University, Australia, 2018.
[8] Cleveland, R. B., Cleveland, W. S., McRae, J. E., & Terpenning, I. J. STL: A seasonal-trend decomposition procedure based on loess. Journal of Official Statistics, 6 (1), 3–33, 1990.
[9] Yameogo B. L. M., Charlton D. W., Doucet D., Desrosiers C, O'Sullivan M, Tremblay C. Trends in Optical Span Loss Detected Using the Time Series Decomposition Method, in Journal of Lightwave Technology, 38 (18), 5026-5035, 2020.
[10] Namin S.S., Namin A.S. Forecasting economic and financial time series: Arima vs. LSTM, Texas Tech University, TX, USA, 2018.
[11] Adhikari R., Agrawal RK. An Introductory Study on Time Series Modeling and Forecasting, arXiv:1302.6613 [cs.LG], LAP Lambert Academic Publishing, Germany, 2013.
[12] Hyndman R.J., Khandaka Y. Automatic Time Series Forecasting: The forecast Package for R, Journal of Statistical Software, 27 (3), 1-22, 2008.
[13] Alghamdi T., Elgazzar K., Bayoumi M., Sharaf T., Shah S. Forecasting Traffic Congestion Using ARIMA Modeling, 2019 15th International Wireless Communications & Mobile Computing Conference (IWCMC), Tangier, Morocco, 1227-1232, 2019.
[14] Frank E., Trigg L., Holmes G. et al. Technical Note: Naive Bayes for Regression. Machine Learning 41, 5-25, 2000.