Available online at www.isas.org.in/jisas
JOURNAL OF THE INDIAN SOCIETY OF
AGRICULTURAL STATISTICS 75(3) 2021 181–188

Forecasting Wheat Yield using Wavelet-Based Multiresolution Analysis

Ranjit Kumar Paul and Dipankar Mitra

ICAR- Indian Agricultural Statistics Research Institute, New Delhi

Received 15 September 2020; Revised 24 September 2021; Accepted 29 October 2021

SUMMARY

Wavelet-based multiresolution analysis can decompose a time series into a set of components. It can improve the accuracy of forecasts. The wavelet-based multiresolution analysis augmented method (Zhang, 2017) is applied to expand wheat yield in Punjab, Haryana, and Bihar, India during the period1966 to 2017 (52 years)into a group of hierarchical series in a meaningful manner. Essentially, a regression model based on the Ordinary Least Squares (OLS) technique is used to reconcile the forecasts at different level of decomposition. Therefore, predictions at higher-level are computing by taking sum of lower-level predictions. The forecasting has been done for different rolling windows and different forecast horizons. The improvement in forecasting performance of the multi-step forecasts obtained using Multiresolution analysis has been shown in terms of minimum values of Mean absolute error (MAE) and Root mean square error (RMSE). Moreover, a comparative study for predictive performance is also carried out between wavelet-based Multiresolution augmented method and corresponding conventional approach i.e. autoregressive integrated moving average (ARIMA) model and wavelet based artificial neural network (Wavelet-ANN) hybrid model. It revealed that the wavelet based Multiresolution augmented method outperforms the other approaches for the data under consideration.

Keywords: Hierarchical time series, Multiresolution analysis, Reconciling forecast, Wavelet decomposition.

1. INTRODUCTION

Time-series forecasting has been emerged as a basis for manual and automatic planning in many real life application (De Gooijer and Hyndman, 2006). Time series can often be disaggregated in a hierarchical structure using attributes such as geographical location, product type etc. A hierarchical time series comprises of multiple times series in which the high-level observations are combined according to low-level data. Forecasting by conventional approaches use either a top-down or bottom-up method or a combination of the two. In top-down approach, the top-level data is forecasted first, and then these forecasts are disaggregated based on historical proportions; On the other hand, in bottom-up approach, the bottom-level data are forecasted first, and then additional data are included to obtain the top-level forecasts (Claeskens *et al.*, 2016; Del Negro *et al.*, 2016).If the hierarchical structure of the data is ignored and forecasting is done for all series at all levels independently, it may lead to undesirable consequence. Hyndman *et al.* (2011) developed an algorithm to compute the final forecasts by adjusting the independent forecasts. This concept for hierarchical time series forecasting can be extended to any univariate time series data. This forecasting algorithm has been applied by Pal and Paul (2016) for forecasting sorghum production in India; Mitra *et al.* (2017) used hierarchical time series approach to forecast oilseeds and pulses production in India.

To understand the structures of data, especially data including high-frequency components like financial time series, carrying out time-domain analysis is not adequate (Masset, 2008). In some sense one needs to look both frequency and time domain to catch real characteristic of data. Spectral analysis and wavelet analysis are examples of the frequency-domain analysis. The modern wavelet theory has taken shape during the late twentieth century (Boggess and Narcowich 2009). Ghosh *et al.* (2010) applied discrete wavelet transform (DWT) and multiresolution analysis (MRA) of India’s monsoon rainfall data to analyze the behaviour of trend in terms of different times and scales. Paul *et al.*

Corresponding author: Ranjit Kumar Paul
E-mail address: ranjitstat@gmail.com

(2011) studied DWT for estimation of trend in India's monsoon rainfall. Paul *et al.* (2013) demonstrated that forecasting based on wavelet is more accurate than that of the usual Autoregressive moving average (ARIMA) model. Kılıç and Uğur (2018) applied MRA to decompose and model the S&P500 time series.

Wavelet-based MRA decomposes a series into a set of series with hierarchical structure (Gencay *et al.*, 2001). Smooth and details sequences of the time series at difference scales are obtained in MRA. The traditional univariate time series prediction models can be applied to the raw data and their decomposed components independently, followed by combining and reconciling these forecasts according to the hierarchical structure. This decomposition method can improve the accuracy of forecasts of original time series data. Zhang *et al.* (2017) proposed a wavelet-based Multiresolution analysis augmented method to forecast univariate time series data.

In the present study, above approach has been applied in wheat yield data. Wheat is one of the most important cereal crops. India ranks second after China in terms of wheat production in the world. Being one of the staple foods in India, accurate forecasting of wheat yield is very important. There are few works available in literature concerning prediction of wheat yield based on time series data (Paul *et al.*, 2013; Paul *et al.* 2014; Paul, 2015). But none of these study has used the approach of MRA. The present study investigates the performance ofwavelet-based MRA augmented method using yearly wheat yield data in Punjab, Haryana and Bihar in India.

2. METHODOLOGY

2.1 Hierarchical time series forecasting

Let us assume that we have multiple level hierarchy, the completely aggregated series are denoted by level 0, the first level of disaggregation is denoted by level 1,and level *K*contains the most disaggregated time series. For easy identification of the individual series and the level of disaggregation, a sequence of letters is used: A denotes series A at level 1; AA denotes series A at level 2 within series A at level 1 and so on.

Assuming that time series observations are recorded at times $t = 1, 2, \ldots, n$, and we are interested in forecasting each series at each level at times $t = n+1, n+2, \ldots, n+h$. Here, we represent the observation on series X as $y_{X,t}$. Thus, $y_{AB,t}$ is the value of series AB at time t. y_t denotes aggregate of all series at time t. Therefore,

$$y_t = \sum_i y_{ij,t}\ ,\ y_{i,t} = \sum_j y_{ij,t}\ ,\ y_{ij,t} = \sum_k y_{ijk,t}\ ,$$

$$y_{ijk,t} = \sum_l y_{ijkl,t}\ ,$$

and so on. So, the observations at higher levels can be obtained by taking sum of the series at lower levels.

Let m_i denotes the total number of series at level *i* ($i = 0, 1, 2, \ldots, K$). So $m_i > m_{i-1}$ and the total number of series in the hierarchy is $m = m_0 + m_1 + \ldots + m_K$.

In terms of matrix notation, let $\mathbf{y}_{i,t}$ denotes the observations at level *i* and time t and $\mathbf{y}_t = \left[y_t, \mathbf{y}_{1,t}, \ldots, \mathbf{y}_{k,t} \right]'$. Here,

$$\mathbf{y}_t = \mathbf{S}\mathbf{y}_{k,t} \tag{1}$$

where $\mathbf{S}$ is a "summing" matrix of order $m \times m_K$ used to aggregate the lowest level series.

The main interest lies on working with forecasts rather than the actual series, in hierarchical forecasting,. Suppose that $\hat{y}_{X,n}(h)$ denotes h-step-ahead forecasts for each individual series y_X . These forecasts are based on $t = 1, 2, \ldots, n$ and hence these are the forecasts for time *n*+*h*. Therefore, $\hat{y}_{AA,n}(h)$ denotes the h-step-ahead base forecast of series y_{AA} using the sample $y_{AA,1}, y_{AA,2}, \ldots, y_{AA,n}$. For level *i*, h-step-ahead base forecasts are denored by $\hat{\mathbf{y}}_{i,n}(h)$ and the h-step-ahead base forecasts for the whole hierarchy are $\hat{\mathbf{y}}_n(h)$, which contains all of the base forecasts stacked in the same order as $\mathbf{y}_t$.

2.2 Wavelet transform and Multiresolution analysis

Wavelet analysis expands a function in terms of wavelets basis, by means of translations and dilations of *mother wavelet*. Multiresolution analysis (MRA) is used to distinct different frequency components. It is regarded as a mathematical microscope of the underlying signal (Burke, 1994). The dilation (D) and a translation (T) operators are used to define MRA. MRA decomposes any square-integrable function in detailed and smooth parts by using wavelet and scaling functions, respectively.

2.3 Multiresolution analysis augmented method

This methodology is based on Hyndman *et al.* (2011) and Zhang *et al.* (2017). Essentially, a regression model based on Ordinary Least Squares (OLS) technique is used to join the forecasts at different wavelet-decomposed series. The algorithm has four steps. First, MRA is applied to decompose the raw data into wavelet details and smooth, $\{S_1, D_1, S_2, D_2\}$ as shown in Fig 1.

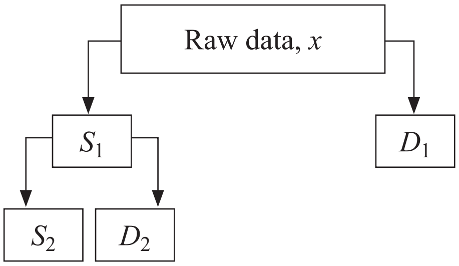

Fig. 1. Wavelet decomposition tree

Next, for a given training sample with size T , a conventional univariate model can be used to conduct first-round forecasts at all hierarchical levels independently with a horizon h. For example, $\hat{Y}_{T+h} = \left[\hat{X}, \hat{S}_1, \hat{D}_1, \hat{S}_2, \hat{D}_2\right]'$ can be predicted by Autoregressive (AR) model of order 1, where $\hat{Y}_{T+h}$ has a dimension 5-by-1 and time index $T+h$.

Due to presence of an explicit hierarchical structure in the wavelet-decomposed components, the variable on each scale can be expressed as a linear combination of the lowest level (base-level) variables, $\beta = [S_2, D_2, D_1]'$, which have on descendants. For example, $X = S_2 + D_2 + D_1$, and $S_1 = S_2 + D_2$. A "summing" matrix Z, with entries of $[0,1]$ can capture the linear relationship in a given hierarchy.

$$Y = \begin{bmatrix} X \\ S_1 \\ D_1 \\ S_2 \\ D_2 \end{bmatrix} = \begin{bmatrix} S_2 + D_2 + D_1 \\ S_2 + D_2 \\ D_1 \\ S_2 \\ D_2 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} S_2 \\ D_2 \\ D_1 \end{bmatrix} = Z\beta \qquad (2)$$

In most univariate cases, the first-round independent forecasts, $\hat{Y}_{T+h} = \left[\hat{X}, \hat{S}_1, \hat{D}_1, \hat{S}_2, \hat{D}_2\right]'$, do not have the property of same hierarchical structure as the original sample, Y, i.e., $\hat{Y}_{T+h} \neq Z\hat{\beta}_{T+h}$. Therefore, those forecasts are not fundamentally consistent with each other.

As the predictions at different hierarchical levels are independent, the prediction result vector, $\hat{Y}_{T+h}$, is regressed on the corresponding summing matrix, Z. The objective of this step is to identify the set of base-level forecasts $\tilde{\beta}_{T+h} = \left[\tilde{S}_2, \tilde{D}_2, \tilde{D}_1\right]'$, that minimizes the squared deviation from the first round of independent forecasts.

$$\tilde{\beta}_{T+h} = \underset{b}{\operatorname{argmin}} \left(\hat{Y}_{T+h} - Zb\right)' \left(\hat{Y}_{T+h} - Zb\right) \qquad (3)$$

Thus, $\tilde{\beta}_{T+h}$, can be estimated using OLS. To this end, the refined optimal forecasts at all hierarchical levels can be computed as $\tilde{Y}_{T+h} = Z\tilde{\beta}_{T+h}$. Among these forecasts, one is particularly interested in the accuracy of the refined forecasts at the top level, $\tilde{X}_{T+h} = \tilde{S}_{2,T+h} + \tilde{D}_{2,T+h} + \tilde{D}_{1,T+h}$ which corresponds to the original time series, X.

2.4 Forecast evaluation

The performance of different forecasting models is compared using MAE and RMSE criteria. Mathematically MAE and RMSE are defined in Eq. (4) and Eq. (5) respectively.

$$\text{MAE} = \frac{1}{h}\sum_{t=1}^{h} \left|y_t - \hat{y}_t\right|, t = 1, 2, \ldots, h \qquad (4)$$

$$\text{RMSE} = \sqrt{\frac{1}{h}\sum_{t=1}^{h}\left(y_t - \widehat{y_t}\right)^2} \qquad (5)$$

where y_t is the actual observation for the time t and $\hat{y}_t$ is the forecast value of the series for the same time; h denotes the forecast horizon.

3. ILLUSTRATIONS

3.1 Data

To demonstrate the forecasting performance of the multi-resolution augmented method, yearly wheat yield data of Punjab, Haryana and Bihar state in India have been collected from Directorate of Economics and Statistics, Ministry of Agriculture and Framers Welfare, Government of India during the period 1966 to 2017 (52 years). The dataset is divided into two parts: training set used for model building and estimation; testing set used for model validation and evaluation. Out of total 52 observations, 44 observations (around 85%) have been kept for model building and remaining

8 observations (around 15%) for validation of the model (Khandelwal *et al.* 2015).

3.2 Application of Multiresolution analysis

A rolling window approach is applied to generate a sequence of similar out-of-sample predictions. The time plots of the dataset for three states are given in Fig 2. The plot exhibits an overall trend pattern in time series data. Wavelet-based multiresolution analyses is applied to decompose the raw yield data into wavelet details (D1, D2) and smooths parts (S1) using a level-2 Maximal Overlap Discrete Wavelet Transform (MODWT) with a Haar filter and it is exhibited in Fig 3 to Fig 5 for Punjab, Haryana and Bihar respectively. The smooth part i.e. S1 is actually the global trend for the series under consideration. The wavelet coefficients are related to differences (of various order) of (weighted) average values of portions of original series concentrated in time. Coefficients at the below(top) provide "high frequency" ("low frequency") information. Wavelet coefficients vary over time and indicates changes in the data at different time-epochs. The vertical clustering of large coefficients represents possible abnormal jumps.

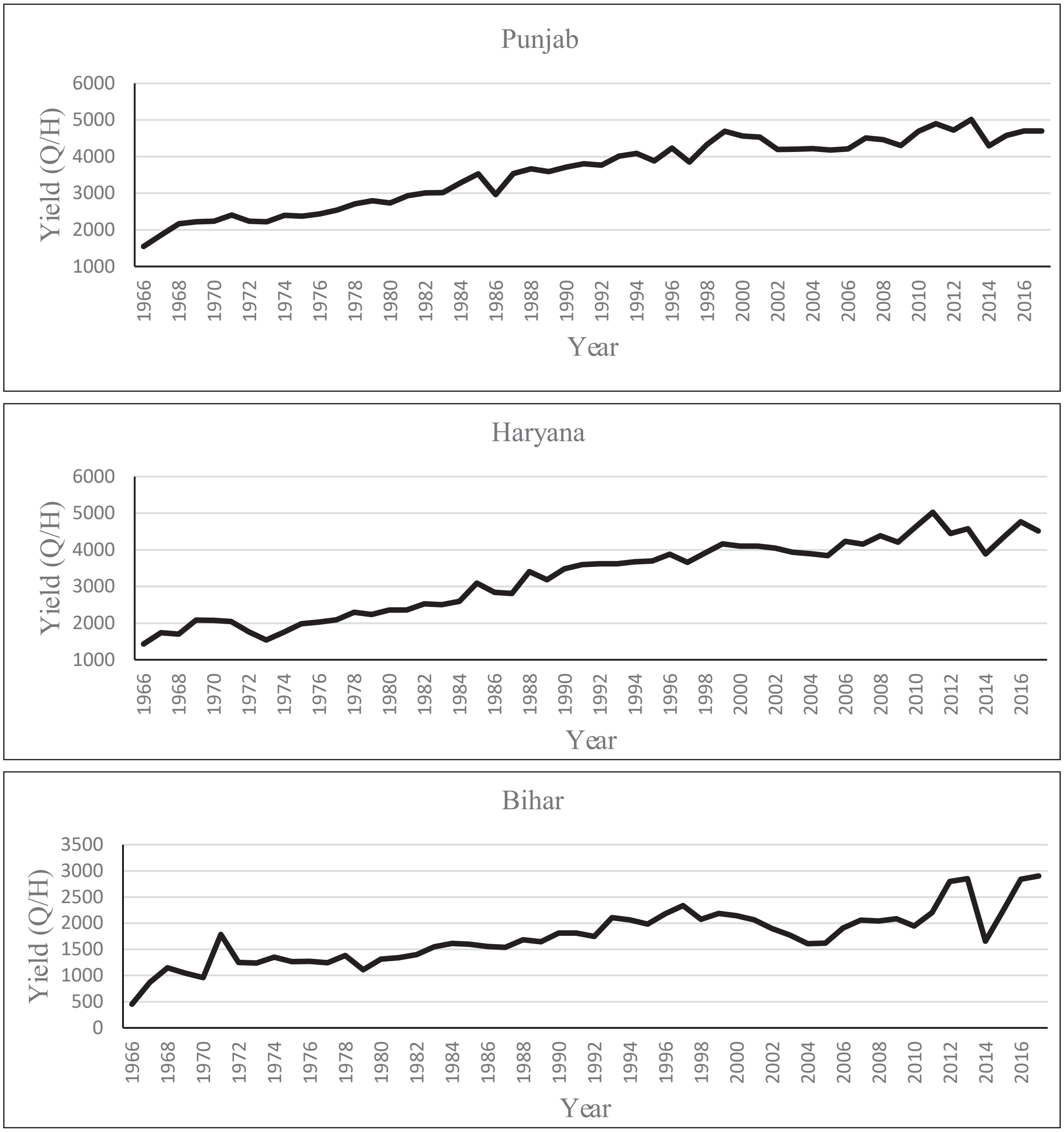

Fig. 2. Time plot of wheat yield data

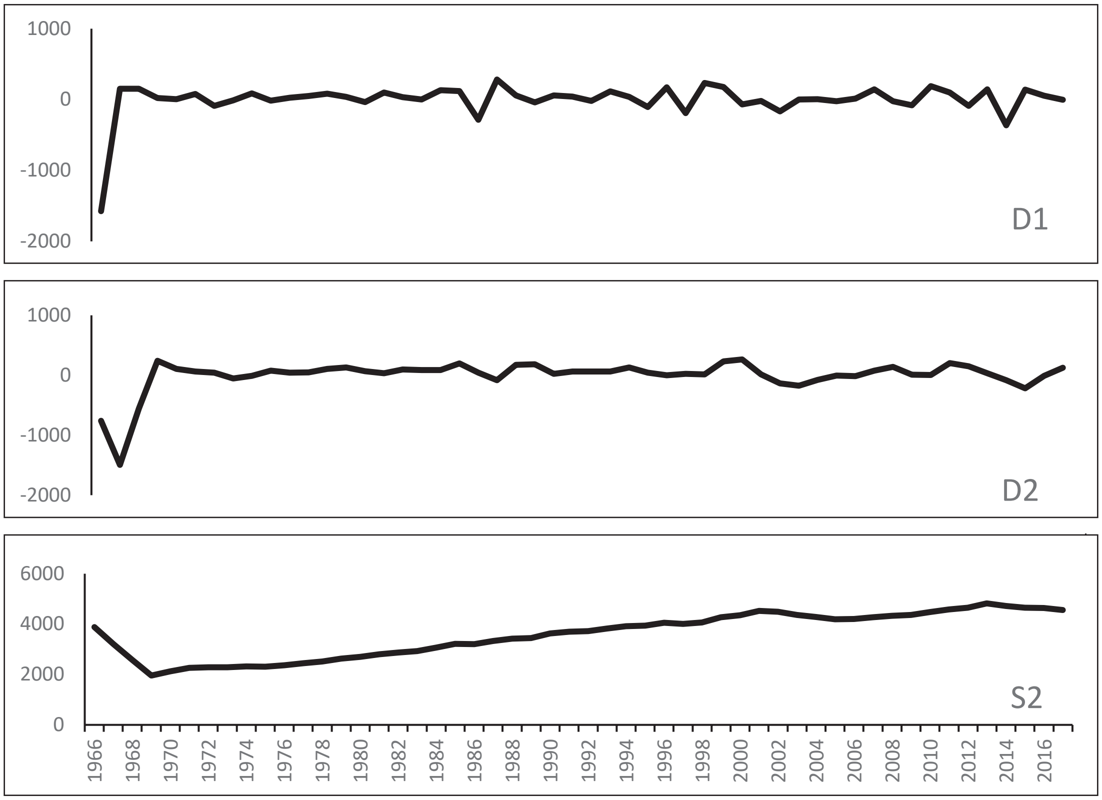

Fig. 3. MRA by level-2 MODWT with Haar filter at Punjab

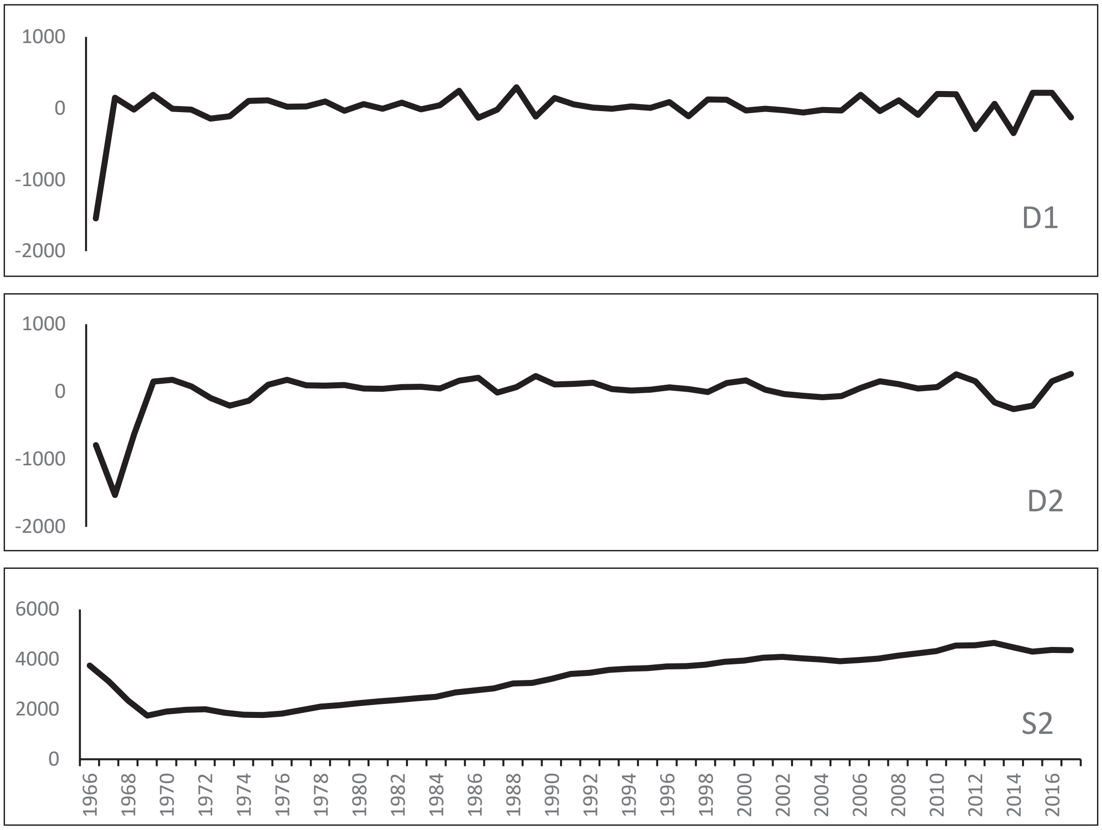

Fig. 4. MRA by level-2 MODWT with Haar filter at Haryana

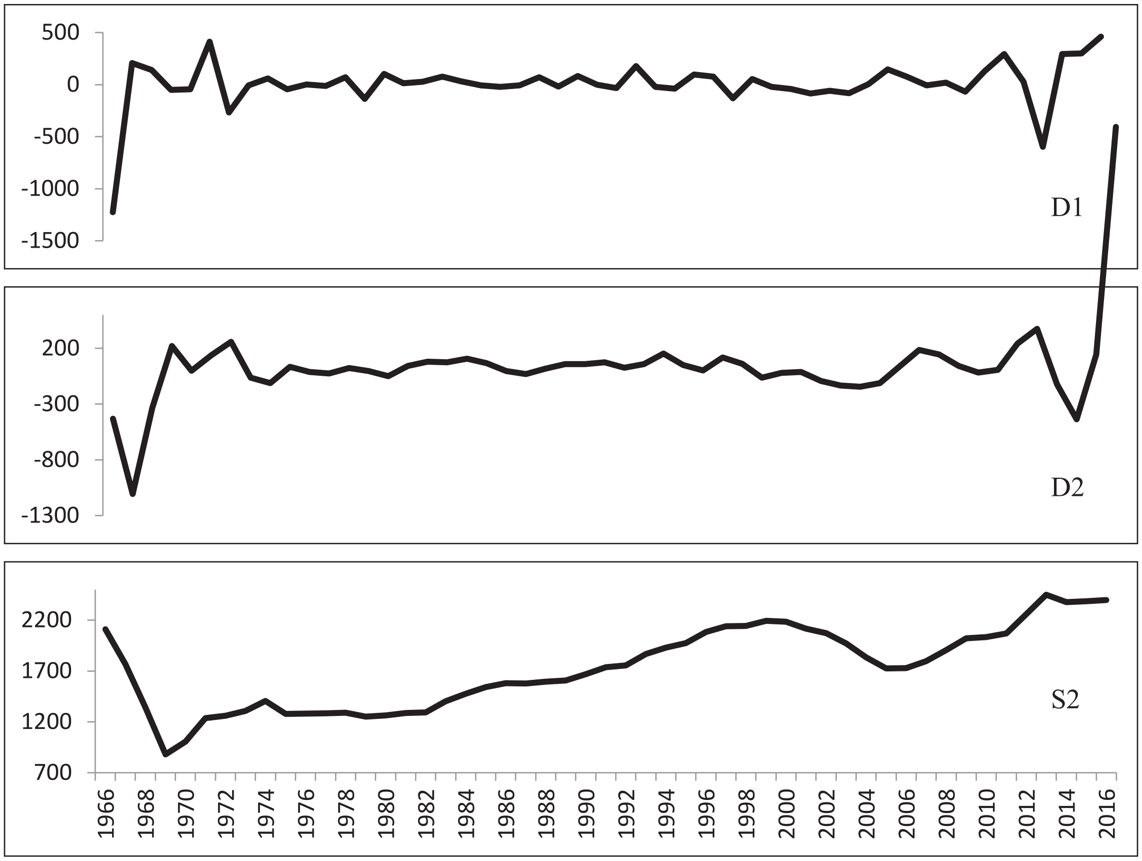

Fig. 5. MRA by level-2 MODWT with Haar filter at Bihar

2.3 Results and Discussions

The accuracy of the forecasts was compared between the methods with and without MRA using the same model. The benchmark univariate time series models include an optimal Autoregressive integrated moving average (ARIMA) model that uses the automatic algorithm of Hyndman and Khandakar (2008). ARIMA model has been fitted to the original data as well as the wavelet decomposed data using R software package. Moreover, Wavelet based artificial neural network (Wavelet-ANN) hybrid model (Paul *et al.* 2020) has been fitted to the original data for empirical comparison. Rolling window forecasts are obtained with 8, 10, and 12 periodsas the window length (Hassler,2019).The prediction performance of the multi-step ahead forecasts obtained from the conventional model and Multiresolution analysis augmented method is carried out based on MAE and RMSE as described in equation 4 and 5. To save space, the results of prediction accuracy of two competing models for the Bihar State have been shown Tables 1 and 2. The results obtained for other two states have also indicates same pattern in terms of accuracy of prediction. The last columns in Tables 1 and 2 represent the average MAE and RMSE values among all h-step forecasts. Both tables indicate the outperformance of ARIMA-MRA model over the usual ARIMA model as well as hybrid Wavelet-ANN technique in terms of lower MAE and RMSE values. It is to be noted that, ARIMA model has performed better than the other model as far as short term forecast is concerned i.e. up to two steps ahead. Otherwise, for all other horizons, ARIMA model performed poorly.

4. CONCLUSIONS

In order to understand the underlying characteristics of the time series under consideration both time as well as frequency domain analysis is desirable. Spectral analysis and wavelet analysis are examples of the frequency-domain analysis. To analyze the series with specific scales and time intervals, wavelet followed by MRA can be implemented. The combination forecast perform well as compared to the conventional hierarchical approaches. This forecast approach can be

Table 1. Forecast performance (MAE) for wheat yield data with different forecast horizons and rolling windows

Models	Forecast horizons								
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Avg**
8-periodrolling window									
ARIMA (0,1,1)	479.52	603.82	523.14	165.66	364.07	451.98	663.67	657.31	**488.65**
ARIMA-MRA	516.41	614.16	515.93	77.85	340.09	408.23	594.04	557.75	**453.06**
Wavelet-ANN	490.97	618.42	583.6	118.22	348.2	415.01	611.85	558.32	**468.07**
10- period rolling window									
ARIMA (0,1,1)	408.41	565.90	514.13	295.86	368.04	428.01	650.78	637.15	**483.53**
ARIMA-MRA	483.70	585.94	471.28	272.47	309.90	387.77	533.65	498.59	**442.91**
Wavelet-ANN	432.55	497.91	448.64	481.88	404.32	431.80	566.00	502.74	470.73
12- period rolling window									
ARIMA (0,1,1)	512.36	487.69	439.58	389.58	427.66	352.78	342.37	286.86	**404.86**
ARIMA-MRA	554.20	502.18	421.67	320.47	399.02	319.81	284.76	254.17	**382.03**
Wavelet-ANN	544.32	531.8	429.45	352.74	407.42	339.4	294.59	263.47	395.40

Table 2. Forecast performance (RMSE) for wheat yield data with different forecast horizons and rolling windows

Models	Forecast horizons								
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Avg**
8-periodrolling window									
ARIMA (0,1,1)	487.13	614.28	572.50	367.31	427.76	543.40	668.36	657.31	**542.26**
ARIMA-MRA	529.48	664.67	563.43	86.70	407.25	486.68	597.54	557.75	**486.69**
Wavelet-ANN	522.55	697.91	569.64	281.88	414.32	531.80	596.00	592.74	525.85
10- period rolling window									
ARIMA (0,1,1)	418.59	614.96	617.39	385.40	424.35	523.28	653.89	637.15	**534.37**
ARIMA-MRA	521.47	649.30	554.71	363.36	405.27	440.02	537.49	498.59	**496.27**
Wavelet-ANN	525.89	674.08	597.02	371.76	416.70	502.23	560.30	545.15	524.14
12- period rolling window									
ARIMA (0,1,1)	495.23	653.27	563.52	578.63	453.28	542.55	481.64	423.52	**523.95**
ARIMA-MRA	564.81	682.98	502.45	480.56	410.94	518.32	402.67	358.61	**490.17**
Wavelet-ANN	496.25	677.94	536.55	555.98	442.94	535.18	469.32	418.33	493.51

extended to any univariate time series data. A wavelet-based MRA can be used to decompose the original univariate time series into a group of components (wavelet details and smooth) having hierarchical structure. Therefore, the optimal combination forecast approach can be applied to a univariate time series. The prediction accuracy obtained by different methods is compared. The conventional univariate models are considered as benchmark. For every univariate model, a pair of forecast is obtained; one based on raw data and other based on a MRA. A comparison of these forecasts are used to test the accuracy of MRA augmented method and any gain in forecast accuracy is due to wavelet decomposition method. The improved forecasting performance of the Multiresolution analysis augmented forecasting method has been demonstrated using real datasets viz. the yearly wheat yield in Punjab, Haryana and Bihar state of India. It has been found that the Multiresolution analysis based augmented method outperforms the conventional ARIMA as well as hybrid Wavelet-ANN model in terms of lower MAE and RMSE values.

ACKNOWLEDGEMENTS

The authors are thankful to the anonymous reviewers for the valuable comments which helped in improving the quality of the manuscript.

REFERENCES

Boggess, A., and Narcowich, F.J. (2009). *A first course in wavelets with Fourier analysis.* Hoboken, NJ:Wiley.

Burke, B. (1994).The mathematical microscope: Waves,wavelets, and beyond. In M. Bartusiak (Ed.), *Scientific discovery at the frontier* (pp. 196–235). Washington: National Academy Press.

Claeskens, G., Magnus, J.R., Vasnev, A.L., and Wang, W. (2016). The forecast combination puzzle: A simple theoretical explanation. *International Journal of Forecasting*, **32(3)**, 754-762.

De Gooijer, J.G., and Hyndman, R.J. (2006). 25 years of time series forecasting. *International Journal of forecasting*, **22**(3), 443-473.

Del Negro, M., Hasegawa, R.B., and Schorfheide, F. (2016). Dynamic prediction pools: An investigation of financial frictions and forecasting performance. *Journal of Econometrics*, **192(2)**, 391-405.

Gençay, R., Selçuk, F., and Whitcher, B. J. (2001). *An introduction to wavelets and other filtering methods in finance and economics*. Elsevier.

Ghosh, H., Paul, R.K., and Prajneshu. (2010). Wavelet frequency domain approach for statistical modeling of rainfall time-series data. *Journal of Statistical Theory and Practice*, **4(4)**, 813-825.

Hassler, U. (2019). Time Series Analysis with Long Memory in View. Wiley.

Hazewinkel, M. (2013). *Encyclopaedia of mathematics: Coproduct—Hausdorff—Young inequalities*. Berlin: Springer.

Hyndman, R.J., and Khandakar, Y. (2008). Automatic time series for forecasting: the forecast package for R (No. 6/07). Clayton VIC, Australia: Monash University, Department of Econometrics and Business Statistics.

Hyndman, R.J., Ahmed, R.A., Athanasopoulos, G., and Shang, H.L. (2011). Optimal combination forecasts for hierarchical time series. *Computational Statistics and Data Analysis*, **55(9)**, 2579-2589.

Hyndman, R.J., Lee, A.J., and Wang, E. (2016). Fast computation of reconciled forecasts for hierarchical and grouped time series. *Computational Statistics and Data Analysis*, **97**, 16-32.

Khandelwal, I., Adhikari, R. and Verma, G. (2015). Time Series Forecasting using Hybrid ARIMA and ANN Models based on DWT Decomposition. *Procedia Computer Science*, **48**, 173-179

Kılıç, D.K., and Uğur, Ö. (2018). Multiresolution analysis of S&P500 time series. *Annals of Operations Research*, **260**(1-2), 197-216.

Mallat, S.G. (1989). A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (7), 674- 693.

Masset, P. (2008). *Analysis of financial time-series using Fourier and wavelet methods. Fribourg*: University of Fribourg.

Mitra, D., Paul, R.K., and Pal, S. (2017). Hierarchical time-series models for forecasting oilseeds and pulses production in India. *Economic Affairs*, **62(1)**, 103.

Pal, S. and Paul, R.K. (2016). Modelling and Forecasting Sorghum Production in India using Hierarchical Time-Series Models. *Indian Journal of Agricultural Sciences*, **86(6)**, 803-808.

Paul, R.K., Prajneshu and Ghosh, H. (2011). Wavelet methodology for estimation of trend in Indian monsoon rainfall time-series data. *Indian Journal of Agricultural Sciences*, **81(3)**, 290-292.

Paul, R.K., Prajneshu, and Ghosh, H. (2013). Wavelet Frequency Domain Approach for Modelling and Forecasting of Indian Monsoon Rainfall Time-Series Data. *Journal of the Indian Society of Agricultural Statistics*, **67(3)**, 319-327

Paul, R.K., Prajneshu, and Ghosh, H. (2013). Statistical modelling for forecasting of wheat yield based on weather variables. *Indian Journal of Agricultural Science*, **83(2)**, 180-183

Paul, R.K., Ghosh, H. and Prajneshu (2014). Development of out-of-sample forecast formulae for ARIMAX-GARCH model and their application. *Journal of the Indian Society of Agricultural Statistics*, **68(1)**, 85-92

Paul, R.K. (2015). ARIMAX-GARCH-WAVELET Model for forecasting volatile data. *Model Assisted Statistics and Application*, **10(3)**, 243-252

Paul, R.K., Paul, A.K. and Bhar, L.M. (2020). Wavelet-based combination approach for modeling sub-divisional rainfall in India. *Theoretical and Applied Climatology*, **139(3-4)**, 949-963.

Zhang, K., Gençay, R., and Yazgan, M.E. (2017). Application of wavelet decomposition in time-series forecasting. *Economics Letters*, **158**, 41-46.