Iraqi Journal of Statistical Sciences

www.stats.mosuljournals.com

Comparing SVR and Random Forest Forecasting based on Autoregressive Time Series with Application

Naam Salem Fadhil and Zinah Mudher ALbazzaz

Department of Statistics and Informatics, College of Computer Science and Mathematics, University of Mosul, Mosul, Iraq

Article information

Article history:
Received, September ,1 ,2023
Accepted ,November ,5,2023
Available online ,December, 1,2023

Keywords:
ARIMA model
SVR
Random Forest
Forecasting
Relative Humidity

Correspondence:
Naam Salem Fadhil
naamsalem@uomosul.edu.iq

Abstract

The accuracy of forecasting the time series of relative humidity in its maximum and minimum cases is important for controlling environmental impacts, damages and risks. In this study, the support vector regression (SVR) method and the random forest (RF) method will be used, depending on the principle of auto regressive (AR) and the autocorrelation (AC), which is the main characteristic of time series in general. The Lags of original time series will be depended as the explanatory (input) variables while the original series will be as target variable. This structure is fitted with the AC principle because the current observation will be depending on time lags in each time step of time series variable. Comparisons of the forecasting results will be performed by using SVR , RF methods and compared to the classical method of analysing time series which is the integrated autoregressive and moving average (ARIMA) model. The SVR and RF methods were employed due to their importance in improving the forecast results, as they are the ideal solution to the problem of non-linearity of the data, as well as the problem of heterogeneity in the climate data, especially as a result of the fact that they contain many seasonal and periodic compounds, which may lead to inaccurate forecast. The forecast of the time series of relative humidity in its minimum and maximum cases was studied in this study for one of the agricultural meteorological stations in the city of Mosul-Iraq. The results of this study reflected the superiority of both SVR method and RF method compared to the classical method represented by the ARIMA model. The results also included the superiority of the RF method in forecasting the training period compared to the SVR method, which was more balanced despite that, as it superiority the results of ARIMA in forecasting the training period and the testing period, while it was its forecast performance is slightly better than the forecast results of the RF method in the test period.

DOI: https://doi.org/10.33899/iqjoss.2023.0181220

Introduction

The diversity of the many patterns contained in the climatic time series data in general and the maximum and minimum relative humidity data in particular, which include the non-linearity of the data and the seasonal and periodic fluctuations that affect the homogeneity of the data will lead to complications in the analysis of the time series, which may negatively affect the accuracy of the forecast results or may be improper results. The accuracy of the forecast results for the relative humidity in the maximum and minimum cases time series may depend mainly on the selection of the appropriate method and its use for forecasting.

ARIMA models are among the most famous time series models often used to study and analyze univariate time series as a traditional statistical method for forecasting, although it may not be suitable for solving problems of nonlinearity of data, but

despite this it is used a lot with climate data as used by (1, 2) to forecast many pollutants and variables Meteorology, including the variable time series of relative humidity. The Box-Jenkins method will be used as an optimal method for analyzing the time series of relative humidity using ARIMA model. Because of the problem of data heterogeneity due to the effects of seasonal and cyclical patterns, the data of the study, which are in their minimum and maximum cases, will be divided into two seasons, the first being the hot season. It includes data for the months (May-June-July-August-September) and a cold season that includes data for the months (November-December-January-February-March). Months (April and October) will be neglected. As they are two relatively mild months, they may be biased towards the hot season at times, and at other times the cold season. SVR and RF methods will be proposed as two non-linear methods that deal better with non-linear data and therefore they are used to improve the forecast results of the maximum and minimum relative humidity data for the hot and cold seasons. Where many researchers have used SVR in addition to the ARIMA model with many time series data, whether climatic or otherwise, as in (3, 4). As for the random forest method, it has been used in many previous studies, in addition to the ARIMA model, to forecast many time series data, as in (5, 6).

Material and methods

In this study, the time series data of the relative humidity in the maximum and minimum cases will be studied in the city of Mosul / Iraq, which includes 372 observations for the hot season and 303 observations for the cold season as daily data. Approximately 30% of the time series ends of each season's data will be extracted as data for the test period and used as hypothetical future data used to test models built on the training data. Therefore, it will include the training data for the hot season (262 observations), while it will be (110 observations) at the end of the hot season as test data. As for the training period for the cold season, it will contain (212 observations) and (91 observations) of the cold time series as test data.

Framework of study

The framework for this study will include the following:

a. Evaluation of the total period of the time series of relative humidity for the maximum and minimum cases, into two hot and cold seasons.
b. Evaluation of each group into two periods, the first for training and the end for the test.
c. Modeling the training data for each season and for the minimum and maximum cases using the ARIMA model and estimating its parameters and characteristics.
d. Testing the ARIMA model that was created in the previous point c, using the data of the testing period.
e. Using training data and modeling it using the SVR method for each season and for the minimum and maximum cases, depending on the principle of auto-regression.
f. Testing the SVR model that was created in the previous point, through the test data for each season and for the minimum and maximum cases.
g. Using training data and modeling it using the RF method for each set of data based on autoregressive.
h. Testing the RF model that will be created in the previous point g through the test data for each set of data.
i. Comparison of forecast results for the training and testing periods and for all seasons for the minimum and maximum cases using ARIMA, SVR and RF methods through forecast error criteria. This framework can be clarified such as in figure 1 below.

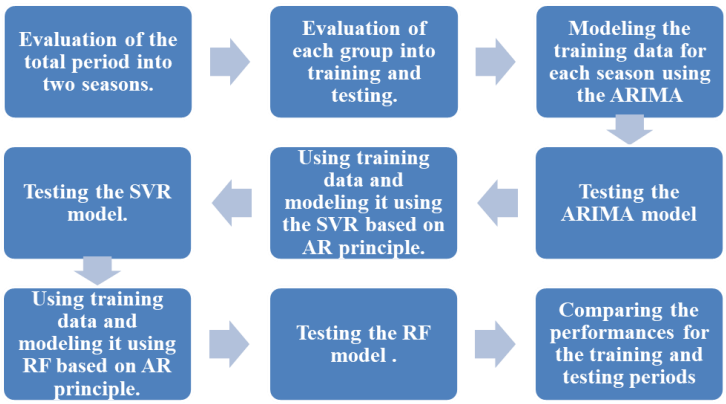

Figure 1: **T**he general framework of study

Autoregressive integrated moving average (ARIMA) model

The ARIMA model is one of the most popular traditional models for time series forecasting and it will be used in this study to forecast the hot and cold seasons for the minimum and maximum cases of relative humidity time series data. The Box-

Jenkins method will be adopted in its four phases: identification, parameter estimation, diagnosing checking, forecasting. As an optimal method for analyzing time series, use the ARIMA model, which takes the general mathematical formula, which is called the double seasonal model and symbolizes it: ARIMA(p,d,q)(P,D,Q).

$$\phi(B)\,\Phi(B)\,Gt = \ominus(B)\,\Theta(B)\,e_t \quad (1)$$

$$G(t) = (1-B)^D(1-B)^d Y_t$$

$$\phi(B) = \left(1-\phi_1 B-\phi_2 B^2-\cdots-\phi_p B^p\right),$$

$$\Phi(B) = (1-\Phi_1 B^S-\Phi_2 B^{2S}-\cdots-\Phi_P B^{PS}),$$

$$\ominus(B) = \left(1-\ominus B-\ominus_2 B^2-\cdots-\ominus_p B^q\right),$$

and Y_t is time series variable, t is current time index, B is (backshift operator) , S is the iterative seasonal period, $\phi, \ominus$ parameters of autoregressive and non-seasonal moving averages, respectively, while Φ, Θ parameters of autoregressive and seasonal moving averages, respectively, (*p,q*)(*P,Q*) represent the numbers of the parameters for non-seasonal and seasonal autoregressive and moving averages, respectively, (*d,D*) are the numbers of consecutive and seasonal differences respectively.

e_t is an i.i.d random error or residuals. The first step in the Box-Jenkins methodology, which is identification, includes achieving stability for the mean and variance, where the stability of the model is detected by drawing the time-series, as well as by drawing the autoregressive and partial autoregressive functions (ACF, PACF).

Within the ACF drawing or the PACF drawing, the instability of the data is detected if one of them is slowly declining (slow dyingout), which requires appropriate measures to be taken to achieve stability by taking power transformations to achieve the stability of the variance or conducting regular and seasonal differences to achieve stability mean after achieving stability, it is possible to identify the ranks of the ARIMA model, which are (p, q, P, Q).

Estimating the parameters of the previously recognized ARIMA model is the second step of the Box-Jenkins methodology procedures. The third step of the Box-Jenkins methodology is conducting diagnostic tests for the model after identifying and estimating the parameters, and it includes the significance of the estimated parameters. It also includes ACF insignificance for series residuals. As the residual series, all its autocorrelation must be non-significant. This is followed by the last step of the Box-Jenkins methodology, which is forecasting, and the minimum mean square error (MMSE) method will be used to calculate the accuracy of the forecast by first step ahead (7, 8).

Support vector regression (SVR)

It is one of the techniques used in forecast as a special type of support vector machine (SVM) when the output of the algorithm is a quantitative variable and it represents the forecast that was inferred from the inputs while the output of the SVM method is a categorical variable, it represents a variable for classification depending on the input variables. The SVR method is a method of forecast for linear and non-linear data, and it is considered one of the methods of supervised learning. Therefore, it is considered one of the machine learning algorithms, due to the possibility of using it in regression, forecast, and classification, because it is often called (Support vector machine regression) and an acronym called SVR, and it was referred to for the first time by (9, 10), and it was suggested later also by (11). A SVR method is a model that is trained based on a training data set that contains explanatory variables in addition to one dependent variable, which is considered as a target variable whose observations are responses to corresponding values of two or more explanatory variables. There are two types of SVR methods, the first of which is linear and the second of which is non-linear, and they are as shown below (9), as SVR method is based in particular, as in SVM method, in searching for the best way to separate the data into the hyper plane.

Therefore, it has a high performance, especially in decision-making issues when using the optimal hyper plane, which depends on points called support vectors, and not all training data (12). Also, the greater the distance between the two margins on the edges of the hyper plan at which the (Support vector) points are located, the greater the probability of classification and forecast of new data with higher accuracy. There are two types of SVR algorithms, namely (Linear SVR) and (non-Linear SVR). In the linear type, groups are separated using a straight line representing the hyper plane. As for the non-linear SVR method, it was proposed to provide higher performance and more accurate results than the application on non-linear data with complex compound patterns, as it depends on converting the second space that contains the data into higher-dimensional spaces, and thus the data can be distinguished into its groups and the forecast of new data in a more professional way, because Such data shall not be separable linearly. The main task of regression within the SVR method is to construct a nonlinear function (*f:Rn>R*). When *y=f(x)*, the estimation function and loss function are respectively defined as follows:

$$y = f(x) = [\omega.\phi(x)] + b, \quad (2)$$

As a regression model where ω is as matrix of weights (regression parameters) and b as the biased (hypothetical error).

$$min\ Q = \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{n}(\xi_i^* + \xi_i) \quad (3)$$

which represents the optimal hyperplane of half the distance between the margins at nearest support vectors. Therefore:

$$yi = [wi\ k\phi(xi)] - b \le \varepsilon + \xi_i^*,$$

$$[w.k\phi(xi)] + b - y = \varepsilon + \xi_i,$$

$\xi i^* + \xi i \geq 0; i = 1, \dots, N$

When C represents the penalty coefficient that is used to determine the risk and confidence range (residuals) ε ξ_i^* represent (relaxation vectors), which are used to improve (convergence speed) but ξ represents loss function, which that are applied to estimate forecast accuracy. The formula for the loss function can be defined as follows:

$L_\varepsilon(y) = |f(x) - y| - \varepsilon$

$If\ |f(x) - y| > \varepsilon$

It is possible for $L_\varepsilon(y)$ to be equal to zero in other cases by constructing a lagerangia function when α_i and α_i^* are non-negative multiples of each observation of the variable *x*. This achieves the optimization problem by reducing the parameter values as shown below:

$$L(\alpha) = \frac{1}{2}\sum_{i,j=1}^{N}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)k(x_i - x_j^*) + \varepsilon\sum_{i=1}^{N}(\alpha_i - \alpha_i^*) - \sum_{i=1}^{N} y_i(\alpha_i - \alpha_i^*) \quad (4)$$

As a final likelihood function. On condition to be

$\sum_{i=1}^{N} \alpha_i = \sum_{i=1}^{N} \alpha_i^*$ also $0 \leq \alpha, \alpha_i^* \leq C$

The larger the value of *C*, the more we control the forecast error, the higher the accuracy, and the model is more suitable for the data. $k(x_i, x_j)$ is kernel function and one of three types of kernel will be used

1. Linear $k(x_i, x_j) = \acute{x}_i\, x_j$
2. Gaussian $G(x_i, x_j) = \exp(-\|x_i - x_j\|^2)$
3. Polynomial $G(x_i, x_j) = (1 + x_i, x_j)^q$

Where q is one of the elements of the set {2,3,…}. And because of the existence of a kernel function, the SVR method is assumed as one of the nonparametric methods. The regression function which is:

$$y(j) = \sum_{i=1}^{N-n}(\alpha_i - \alpha_i^*)k(x_i - x_j) + b \quad (5)$$

When $j=m+1,\dots,N$ represents the function that will be used to forecast ℓ^{th} steps forward, as shown below

$$x_{n+\ell} = \sum_{i=1}^{N-n}(\alpha_i - \alpha_i^*)k(\overline{x_i} - \overline{x}_{N-m+\ell}) + b \quad (6)$$

When

$\overline{x}_{N+m+\ell} = \{x_{N-m+\ell}, \dots, x_{N+\ell-1}\}$ and $\ell = 1,2, \dots, L$

Random Forest (RF)

The random forest is one of the areas similar to decision trees, and it is one of its types in the case of several decision trees gathered together to achieve one unanimous decision, and it is one of the machine learning algorithms that are supervised, that is, with the presence of a dependent variable, which is called the target variable, and depending on this principle, the RF several outputs that should match and be as close as possible to the target variable will get forecasting errors based on the principle of classification and regression tree (CART) techniques. One of the advantages of (CART), of which RF is a special case, is that it is mathematically accurate and works at a high speed, in the case of relatively large data. The RF method has been used in many previous studies as a modern technique that can be applied in different fields depending on the principle of classification and regression. The decision is used to build an accurate mathematical model that gives accurate forecasting from the data with a principle similar to the axioms of regression trees (13). Regression trees differ from classification trees in that they are a type of supervised decision tree from which we get forecasting rather than classifications.

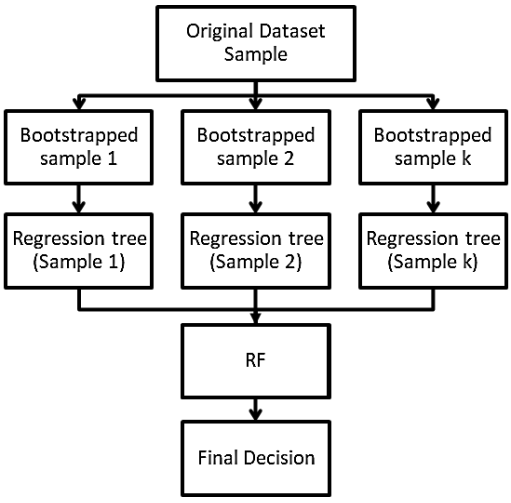

Figure 2: **T**he general framework of the RF algorithm.

Where we notice through the Figure 2. above that the method of random forest work is more generalized than the work of the CART method, because RF depends on the principle of evaluating the study sample into several sub-samples, boot strap sample sets, in order to provide all behavioral patterns of the study sample and for all different time periods and then get independent regression trees for all those different Boot Strap (BS).

Each final leaf node in the regression trees is considered as an independent decision from the rest of the decisions in the other leaf nodes. For this reason, the regression trees will be given high accuracy in fitting between outputs and forecasting compared to the target variable (the original variable), for the reasons mentioned above, the forecasting method will be developed in the time series, as well as the development of forecasting accuracy, mean obtaining optimal forecasting with less forecasting errors when using the RF method compared to traditional methods.

There is a possibility that the trees in RF method are interrelated with each other sometimes, because they are counted to the same type of data, and also for the reason of adopting the principle of bagging, which is based on the BS process, where the bagging method improves the RF and makes them more robust when compiling the regression trees with each other, but after processing the correlation existing between those trees and achieving independence among them. Breiman (14) submitted a proposal that includes the growth of each tree independently and randomly, and this will thus ensure a significant increase in the accuracy of forecasting by using RF. The framework for building an RF algorithm can include three steps as follows:

1. Extracting a number of (BS) samples with a number equal to M. These samples are originally interrelated and not independent. M also represents the size of the forest and the number of its trees.
2. Each set of data extracted from point (1) will grow its own regression tree, which is symbolized by TM, by following several sequential steps until reaching the minimum leaf node of the regression trees, which is symbolized by nmin , and these steps are as follows:
   a. Choose m which represents the randomly selected number of forecasts in each part of the total number of variables p.
   b. Choosing the best forecasts from the forecasts selected in (a), denoted by the symbol m, with the selected part to which it belongs, in order to reduce the MSE value of the selected forecasts in (a).
   c. Separation of the node into two sub-nodes depending on the criterion used or the best forecasts that were chosen in (b).
3. Extracting the outputs from all regression trees by finding the set $\{\mathrm{Tb}\}_1^{\mathrm{B}}$ Finally, at a certain point X, the forecast is possible according to the following equation: (6)

$$f_{RF} = \frac{1}{B}\sum_{b=1}^{B} T_b(x) \quad (7)$$

The measurements of forecasting Error

In this study, several measures will be used to measure the accuracy of forecast through forecasting errors, and these error measures achieve the highest accuracy when their value is as low as possible, which is: Mean absolute percentage error (MAPE), Root mean squares error (RMSE), and Mean absolute error (MAE). Assuming that e_i represents the forecast error, n the number of observations, and Y_i represents the original series used as a target variable, the mathematical formulas for the three measures in a row are as follows: (15)

$$MAPE = \frac{1}{n}\sum\left|\frac{e_i}{y_i}\right| \times 100, \qquad RMSE = \sqrt{\frac{1}{2}\sum_{i=1}^{n}(e_i)^2}, \text{ and } \; MAE = \frac{1}{n}\sum_{i=1}^{n}|e_i|.$$

Studied Datasets

The ARIMA model will be studied using the Box-Jenkins method, the SVR method, and the RF random forest method in this study to analyze the time series data of the minimum and maximum relative humidity after separating it into two hot and cold seasons. The study data are data from the agricultural meteorological center. The agricultural meteorological center /Nineveh governorate /Mosul station at the specified location, longitude E = 43.16 and latitude N = 36.33. The study data included 675 observations from (5/15/2018) to (7/19/2020) and because of what it contains The data is due to the heterogeneity between the different seasons and the diversity of the seasonal seasons and their fluctuations, and to achieve greater consistency, the data has been separated into two seasons, the first being the cold months (November-December-January-February-March). As for the second season, it is hot and for the months (June-July-August-September). Also, part of the end of each time series was deducted from the data as test data for approximately 30% of the total data from each season. Thus, the cold season will contain 303 observations divided into 212 observations as training data and 91 observations as test data. Hot season it will contain 372 observations of which 262 are training data and 110 observations are test data.

Results

a. ARIMA model for the hot season maximum relative humidity is ARIMA (1,1,1)

$$(1-\emptyset_1 B)(1-B)\,Y_t = (1-\ominus_1 B)\alpha_t \quad (8)$$

Table 1 below shows the parameter values and their significance.

Table 1: The parameter values and their significance for ARIMA(1,1,1) model.

	Parameter	t	p-value
$\emptyset_1$	0.5435	3.91	0.000
$\ominus_1$	0.7633	7.20	0.000

From Table 1. above, it is clear that the ARIMA (1,1,1) parameters are significant for the hot season of the maximum relative humidity. Thus, the ARIMA (1,1,1) model is the final and its fitted for this data after passing the diagnostic checking.

b. ARIMA model for the hot season minimum relative humidity is ARIMA (0,2,1)

$(1-B)^2\, Y_t = (1-\ominus B)\alpha_t$ (9)

Table 2 below shows the model parameter values and their significance

Table 2: The parameter value and their significance for ARIMA(0,2,1) model.

	Parameter	t	p-value
$\ominus_1$	0.9912	175257.16	0.000

From Table 2. above, it is clear that the ARIMA (0,2,1) parameters are significant for the hot season of the minimum relative humidity. Thus, the ARIMA (0,2,1) model is the final and its fitted for this data after passing the diagnostic checking.

c. ARIMA model for the cold season maximum relative humidity is ARIMA (0,1,1).

$(1-B\,)\, Y_t = (1-\ominus_1 B)\alpha_t$ (10)

Table 3 below shows the parameter values and their significance for ARIMA(0,1,1) model

Table 3: The parameter value and their significance for ARIMA(0,1,1) model.

	Parameter	t	p-value
$\ominus_1$	0.4098	6.49	0.000

From Table 3. above, it is clear that the ARIMA (0,1,1) parameters are significant for the cold season of the maximum relative humidity. Thus, the ARIMA (0,1,1) model is the final and its fitted for this data after passing the diagnostic checking.

d. ARIMA model for the cold season minimum relative humidity is ARIMA (0,2,1)

ARIMA (0,2,1) = $(1-B\,)^2\, Y_t = (1-\ominus_1 B)\alpha_t$ (11)

Table 4 below shows the parameter values and their significance for ARIMA(0,2,1) model

Table 4: The parameter value and their significance for ARIMA(0,1,1) model.

	Parameter	t	p-value
$\ominus_1$	0.9901	41414.29	0.000

From Table 4. above, it is clear that the ARIMA (0,2,1) parameters are significant for the cold season of the minimum relative humidity. Thus, the ARIMA (0,2,1) model is the final and its fitted for this data after passing the diagnostic checking.

The error criteria MAE, MAPE, and RMSE were used to measure the quality of the four models above in forecasting the maximum and minimum relative humidity series for the hot and cold seasons and for the training and testing periods by measuring the forecasting errors, and the results were as in the Table 5 below.

Table 5: The error measurement of ARIMA(0,1,1) model.

		Hot		Cold	
		Max.	Min.	Max.	Min.
		ARIMA(1,1,1)	ARIMA(0,2,1)	ARIMA(0,1,1)	ARIMA(0,2,1)
Training	MAE	5.93	1.81	4.24	7.24
	MAPE	14.96	15.30	5.00	15.48
	RMSE	8.62	3.62	7.38	10.06
Testing	MAE	12.23	29.06	6.22	13.21
	MAPE	20.90	378.03	7.85	28.35
	RMSE	16.37	34.55	9.55	16.35

From Table 5 above, it is clear that the forecasting for the training period outperformed the testing forecasts for all ARIMA models, which is required to avoid the over fitting problem.

The figures of 3-6 show the compatibility between the original series and the forecasting series for the training and testing periods using the ARIMA models for the maximum and minimum relative humidity data for the hot and cold seasons.

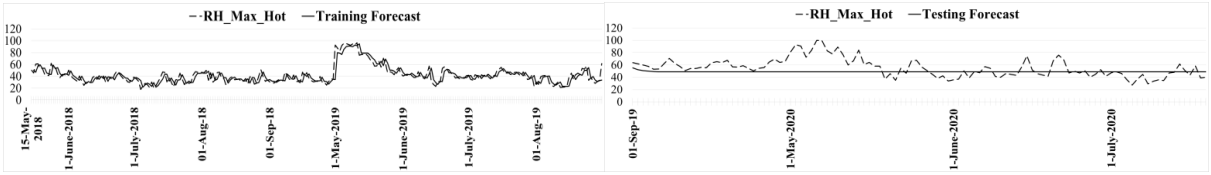

Figure 3: Original and forecasting series\ maximum RH\ hot season for training and testing respectively using ARIMA(1,1,1).

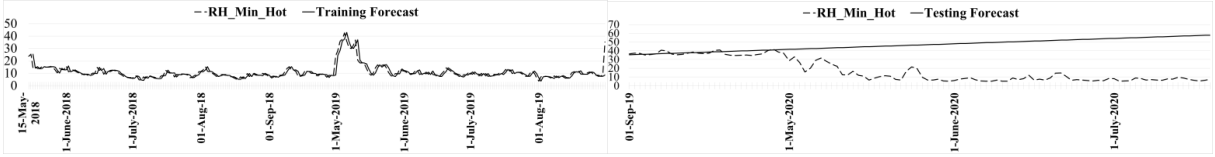

Figure 4: Original and forecasting series\ minimum RH\ hot season for training and testing respectively using ARIMA(0,2,1).

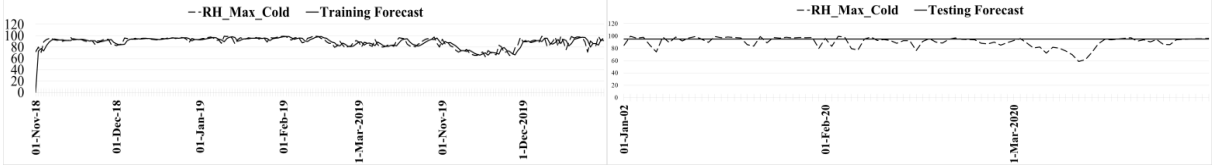

Figure 5: Original and forecasting series\ maximum RH\ cold season for training and testing respectively using ARIMA(0,1,1).

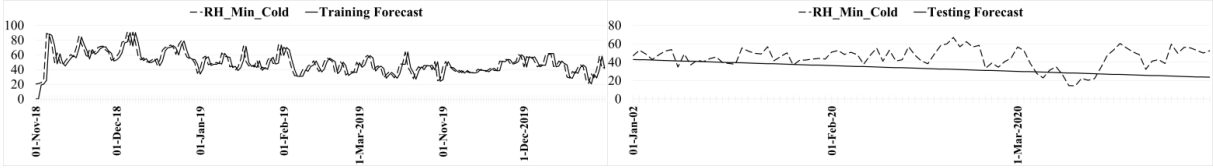

Figure 6: Original and forecasting series\ minimum RH\ cold season for training and testing respectively using ARIMA(0,2,1).

Figures (3-6) show that there is compatibility and fitting between the original series for each type of studied data using the four ARIMA models referred to earlier just for training period only.

Support vector Regression (SVR)

In this section, the results obtained through forecasting the maximum and minimum relative humidity data will be presented using the SVR method. The SVR method can be used to forecast future values for univariate and multivariate dataset, as is the case with the possibility of using SVM also to classify values. Matlab program was used to apply SVR method by following the applied algorithm to obtain the forecasting results.

1. Preparing the data as groups, the first for the input variables, which represent the explanatory variables, and the second includes the target variable, which represents the depended variable. The data for both variables above were divided into two parts, 70% of the data for the training period and 30% from the end of the time series for the testing period, and then entering those variables for the training and testing periods into workspace of Matlab to be used for modelling and forecasting.
2. Writing the special coomands for SVR method in an integrated program, such as follows.

For creating SVR model:

SVR_Model = fitrsvm(training inputs,training target)

For forecasting for the training and testing periods, respectively:

training_output = predict(SVR_Model, training inputs)

testing_output = predict(SVR_Model, testing inputs)

3. Imputing the forecasting errors for both training and testing periods such as follows.

training_residuals = training target-training_output

testing_residuals = testing target-testing_output

4. Calculation of measurements of forecasting errors, such RMSE, MAPE, MAE.
5. Logarithm will stopped if one of the two cases is reached:

a. Reaching the best indicators for forecasting errors,
b. Reaching the maximum number of iterations.

By applying the instructions of the previously mentioned SVR algorithm, Table 6. below was obtained, which includes the values of MAE, MAPE, RMSE for the training and testing periods for the hot and cold seasons, and for the minimum and maximum relative humidity data using SVR.

Table 6: The error measurements of SVR model.

		Hot		Cold	
		Max.	Min.	Max.	Min.
Training	MAE	5.97	5.47	4.24	6.47
	MAPE	14.69	6.32	4.91	13.65
	RMSE	8.48	7.30	6.81	9.11
Testing	MAE	7.37	2.62	5.47	6.50
	MAPE	14.93	18.22	6.32	16.13
	RMSE	9.44	3.50	7.30	9.91

From the above tables 5 and 6, it is clear that the accuracy of the forecasting in the training and testing of the time series data of the maximum and minimum relative humidity exceeds by using SVR method comparing to the accuracy of the forecast results for the same time series and the same periods using the ARIMA method for both the hot and cold seasons. Testing forecasts with SVR has more accuracy comparing to same period using ARIMA model.

The figures (7-10) show the compatibility between the original series and the forecast series for the training and testing periods, and for the hot and cold seasons using SVR method with data of the maximum and minimum relative humidity.

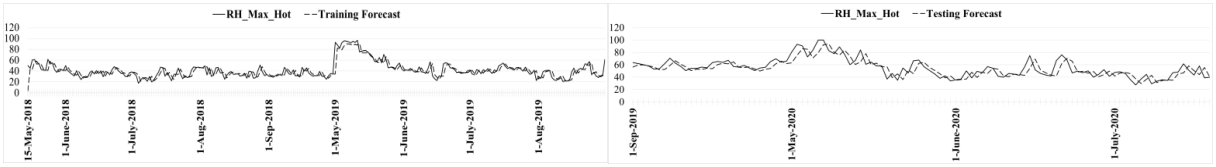

Figure 7: Original and the forecasting series\ maximum RH\ hot season for training and testing respectively using SVR.

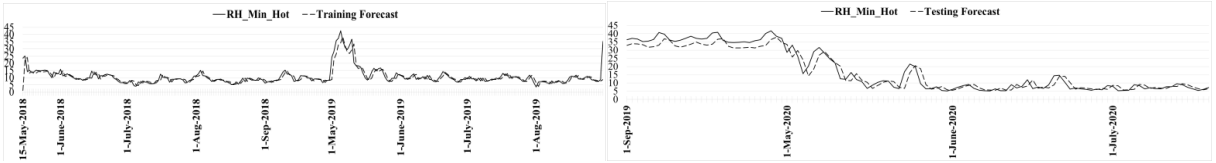

Figure 8: Original and the forecasting series\ minimum RH\ hot season for the training and testing respectively using SVR.

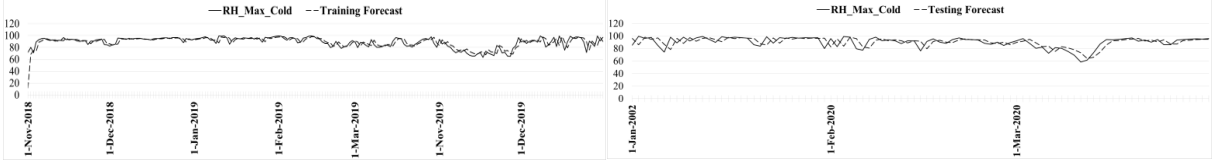

Figure 9: Original and the forecasting series\ maximum RH\ cold season for training and testing respectively using SVR.

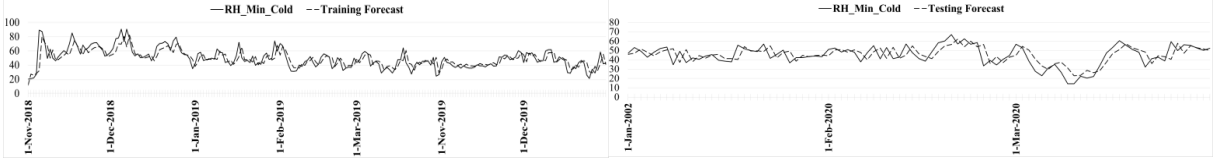

Figure 10: Original and the forecasting series\ minimum RH\ cold season for training and testing respectively using SVR.

From the figures (7-10), It is clear that there is high compatibility and fitting between the original series of each type of data using SVR method, compared to compatibility and fitting using ARIMA, as shown in the figures (3-6).

Random Forest

In this item, the random forest method is used to obtain the forecast results of the maximum and minimum relative humidity data for the hot and cold seasons for the training and testing periods. The instruction will be used in the Matlab program in order to build the regression ensemble model by entering time series data for the explanatory variables of the input

data with one dependent variable as a target variable according to the principle of autocorrelation that was referred to in SVR method. The principle of fit ensemble in the Matlab program is as it was previously passed in building an regression ensemble model in the methods and material item, noting the following:

1. The least-squares boosting (LSBoost) algorithm, which includes finding sums for the best equations and models that fit the time series data, will be relied upon in this study at each stage of the LS-Boost algorithm. New learning will be done and a new regression model will be found, and then the forecast error will be found through the difference between the real data (target variable) and the accumulated forecast from output learning steps, while ensuring that the smallest measure of MSE forecast errors is obtained through the use of models collected from several regression trees to obtain more accurate solutions and results than if using traditional methods.
2. Depending on the principle of aggregation and combination between models. in the independent trees within the random forest, the use of 10 parts of data will be adopted as a default number when using the fit ensemble directive as a maximum to obtain the best and most accurate results.
3. Finally, 100 regression trees were reconciled to obtain the best predictions.
4. The predict directive will be used to obtain forecasting for the training and test periods and for both the hot and cold seasons for the time series of maximum and minimum relative humidity.

Table 7 below shows the values of the RMSF, MAPE, and MAE of forecasting errors for the training and testing data for the hot and cold seasons, maximum and minimum relative humidity, using RF.

Table 7: The error measurements of RF method.

		Hot		Cold	
		Max.	Min.	Max.	Min.
Training	MAE	0.05	0.04	0.01	0.06
	MAPE	0.13	0.27	0.01	0.22
	RMSE	0.08	0.21	0.02	0.37
Testing	MAE	20.76	3.72	5.77	9.78
	MAPE	19.50	29.93	6.50	27.00
	RMSE	13.41	5.11	7.59	12.68

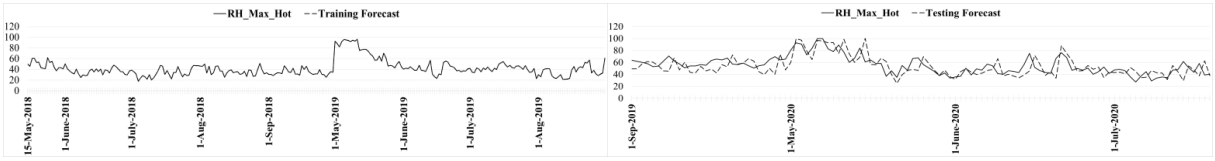

Figure 11: Original and the forecasting series\ maximum RH\ hot season for the training and testing respectively using RF.

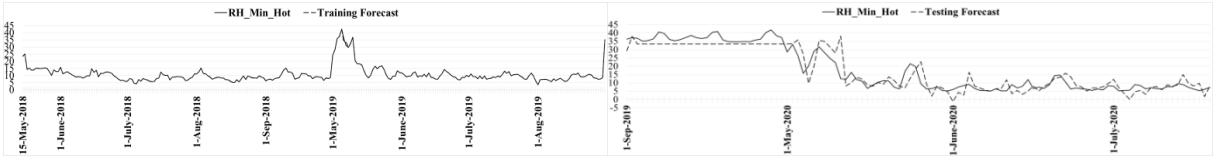

Figure 12: Original and the forecasting series\ minimum RH\ hot season for the training and testing respectively using RF.

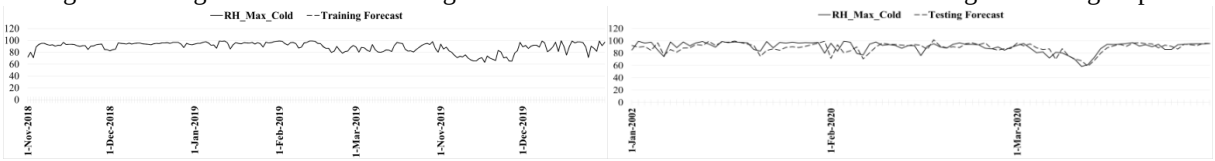

Figure 13: Original and the forecasting series\ maximum RH\ cold season for training and testing respectively using RF.

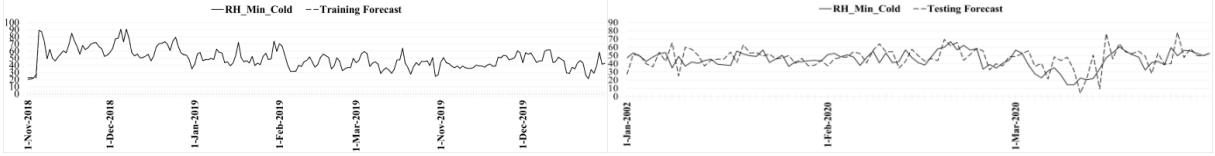

Figure 14: Original and the forecasting series\ minimum RH\ cold season for the training and testing respectively using RF.

Discussion

Through Table 7 above, it is clear that the accuracy of forecasting for the training period using RF is superior if it is compared with the accuracy of forecasting using SVR for the training period in table 6 for maximum and minimum relative humidity, and if it is compared with the accuracy of forecasting using ARIMA for the training period as in table 5, and this is evidence the reliability of the forecasting of the RF model compared with the rest of the models and the quality of output with a behavior very close to the behavior of the target variable in the training period only.

While for the testing period, the performance of SVR is superior through the accuracy of forecasting compared to each of the ARIMA and RF methods in the table referred to above. This reflects the stability of the SVR methodology, regardless of the changes in the data and its characteristics, due to its use of the optimal hyperplane and maximum gap principle, or the so-called upper and lower margins that surround the optimal hyperplane as explained previously.

In general, the performance of both SVR and RF methods, based on the principle of autocorrelation, is clearly outperformed the accuracy of forecasting for the training and testing periods over the performance of the traditional method represented by the ARIMA model. This is confirmed by figures (7-10) of the SVR method, which includes the compatibility of the original series with the forecasting series, as well as figures (11-14) listed above for RF if compared with the concordances of both the original series and the forecasting and use of the ARIMA method confirmed by figures (3-6).

Conclusion

In this research, the traditional method represented by ARIMA models and the modern methods represented by the SVR method and the RF method were used to forecast the time series data of the maximum and minimum relative humidity after dividing them into two hot and cold seasons, and into two training and testing periods. Through the results and discussions presented in the previous section, it is possible to conclude the preference of using both the SVR method and the RF method to forecast relative humidity data in particular or similar other climatic data, due to their outperforming in forecasting compared to the traditional ARIMA method, where the ability of these two methods to deal with the non-linearity of the data, given that the division of the data into two seasons, hot and cold, has addressed the problem of heterogeneity in the data due to the diversity of seasonal and periodic compounds in the data. For the forecasting of the training data, it is better to use the RF method, which will outperform both ARIMA and SVR methods, while it is better to use the SVR method to forecast the testing data, which will outperform both ARIMA and RF methods in forecasting accuracy. We conclude from the foregoing the constancy and stability of the SVR methodology, regardless of the data and its characteristics change, depending on its use of the gap principle or the so-called margins and its selection of the optimal hyperplane on the basis of achieving the largest gap between the level and the support vectors.

Acknowledgment

The authors are very grateful to the University of Mosul/ College of Computer Sciences and Mathematics for their provided facilities, which helped to improve the quality of this work.

Conflict of interest

The author has no conflict of interest.

References

1. Chaudhuri, S. and D. Dutta. Mann–Kendall trend of pollutants, temperature and humidity over an urban station of India with forecast verification using different ARIMA models. *Environmental monitoring and assessment*. 2014; *186*: 4719-4742.
2. Eymen, A. and Ü. Köylü. Seasonal trend analysis and ARIMA modeling of relative humidity and wind speed time series around Yamula Dam. *Meteorology and Atmospheric Physics*. 2019; *131*: 601-612.
3. Al-Musaylh, M.S., et al. Two-phase particle swarm optimized-support vector regression hybrid model integrated with improved empirical mode decomposition with adaptive noise for multiple-horizon electricity demand forecasting. *Applied energy*. 2018; *217*: 422-439.
4. Xu, D., et al. Application of a hybrid ARIMA–SVR model based on the SPI for the forecast of drought—a case study in Henan Province, China. *Journal of applied meteorology and climatology*. 2020; *59(7)*: 1239-1259.
5. Kumar, M. and M. Thenmozhi. Forecasting stock index returns using ARIMA-SVM, ARIMA-ANN, and ARIMA-random forest hybrid models. *International Journal of Banking, Accounting and Finance*. 2014; *5(3)*: 284-308.
6. Noureen, S., et al. A comparative forecasting analysis of arima model vs random forest algorithm for a case study of small-scale industrial load. *International Research Journal of Engineering and Technology*. 2019; *6(09)*: 1812-1821.
7. Das, R.C. Forecasting incidences of COVID-19 using Box-Jenkins method for the period July 12-Septembert 11, 2020: A study on highly affected countries. *Chaos, Solitons & Fractals*. 2020; *140*: 110248.
8. Shukur, O.B. and M.H. Lee. Daily wind speed forecasting through hybrid KF-ANN model based on ARIMA. *Renewable Energy*. 2015; *76*: 637-647.
9. Cortes, C. and V. Vapnik. Support-vector networks. *Machine learning*. 1995; *20*: 273-297.
10. Zhang, F. and L.J. O'Donnell, *Support vector regression*, in *Machine learning*. 2020, Elsevier. p. 123-140.
11. Ahmed, H.U., et al. Support vector regression (SVR) and grey wolf optimization (GWO) to predict the compressive strength of GGBFS-based geopolymer concrete. *Neural Computing and Applications*. 2023; *35(3)*: 2909-2926.
12. Li, Z.-C., et al. Prediction of protein structure class by coupling improved genetic algorithm and support vector machine. *Amino acids*. 2008; *35*: 581-590.
13. Shumway, J., et al. Machine learning to improve the prioritization and effectiveness of pre-treatment physics chart checks. *International Journal of Radiation Oncology, Biology, Physics*. 2020; *108(3)*: S54-S55.
14. Breiman, L. Random forests. *Machine learning*. 2001; *45*: 5-32.
15. Hyndman, R.J. and A.B. Koehler. Another look at measures of forecast accuracy. *International journal of forecasting*. 2006; *22(4)*: 679-688.

مقارنة تنبؤ انحدار المتجه الداعم والغابة العشوائية اعتماداً على الانحدار الذاتي للسلاسل الزمنية مع التطبيق

نعم سالم فاضل و زينة مضر البزاز

قسم الاحصاء والمعلوماتية ، كلية علوم الحاسوب والرياضيات ، جامعة الموصل ، الموصل ، العراق

الخلاصة

إن دقة التنبؤ بالسلاسل الزمنية للرطوبة النسبية في حالاتها العظمى والدنيا أمر مهم للتحكم في التأثيرات والأضرار والمخاطر البيئية. سيتم في هذه الدراسة استخدام طريقة الانحدار المتجه الداعم (SVR) وطريقة الغابة العشوائية (RF)، وذلك بالاعتماد على مبدأ الانحدار الذاتي القائم على مبدأ الارتباط الذاتي الذي يعتبر السمة الرئيسية للسلاسل الزمنية بشكل عام. سيتم إجراء مقارنات لنتائج التنبؤ باستخدام طرق SVR و RF ومقارنتها بالطريقة التقليدية لتحليل السلاسل الزمنية وهي نموذج الانحدار الذاتي والمتوسطات المتحركة المتكامل (ARIMA). تم استخدام طريقتي SVR و RF نظرا لأهميتهما في تحسين نتائج التنبؤ حيث تعتبران الحل الأمثل لمشكلة عدم خطية البيانات وكذلك مشكلة عدم التجانس في البيانات المناخية خاصة وذلك لاحتوائها على العديد من المركبات الموسمية والدورية، مما قد يؤدي إلى تنبؤات غير دقيقة. تمت في هذه الدراسة دراسة تنبؤات السلاسل الزمنية للرطوبة النسبية في حالاتها الدنيا والعظمى لإحدى محطات الأرصاد الجوية الزراعية في مدينة الموصل–العراق. عكست نتائج هذه الدراسة تفوق كل من طريقة SVR وطريقة RF مقارنة بالطريقة التقليدية المتمثلة في نموذج ARIMA. كما تضمنت النتائج تفوق طريقة RF في التنبؤ بفترة التدريب مقارنة بطريقة SVR التي كانت أكثر توازناً رغم ذلك، كما تفوقت نتائج ARIMA في التنبؤ بفترة التدريب وفترة الاختبار، في حين كانت تنبؤاتها أفضل قليلاً من نتائج التنبؤ باستخدام طريقة RF في فترة الاختبار.

الكلمات المفتاحية: نموذج اريما،SVR ، الغابة العشوائية، التكهن، الرطوبة النسبية.