BAUET JOURNAL
Published by
Bangladesh Army University of Engineering & Technology
Journal Homepage: http://journal.bauet.ac.bd

Modeling and Prediction of Rainfall in the U.K. Using Nonparametric Approach

A.H.M. Rahmatullah Imon[1*], Noora N Saleh[2], and Sirajum Munira Khan[3]

[1]Department of Mathematical Sciences, Professor, Ball State University, Muncie, IN 47306, USA
[2]Department of Mathematics, Assistant Lecturer, Al-Qadisiyah University, Ai-Diwaniyah, Iraq
[3]Department of Mathematical Sciences, Research Student, Ball State University, Muncie, IN 47306, USA

Abstract: Modeling and prediction of rainfall has been an area of a great deal of attention to the climatologists for a long time. Varieties of approaches have been in use in the literature to model and predict rainfall. Linear regression model is the most commonly used method for fitting a rainfall data. But the usefulness of this method heavily relies on some standard assumptions which do not usually hold in the reality. Transformed regression models are also suggested in the literature when linear regression models are found inappropriate. However, sometimes the structure of the data could be so complex that even the best possible transformation cannot eliminate all of the potential shortcomings. Nonparametric regression is the one of the ways in this situation. In this paper, we fit the U.K. rainfall data by linear, transformed and nonparameric regression models. In addition to that, we report a cross validation study to investigate the quality of forecasts of rainfall generated by these methods. We found the nonparametric regression performs much better than other methods in both fitting and forecasting rainfall. Another important finding of our study is sun hours, air frost days; season and temperature have a very significant impact on rainfall in the U.K.

Keywords: *Regression diagnostics; Box-Cox transformation; nonparametric regression; LOWESS; Cross validation.*

Introduction: Modeling and prediction of rainfall is still a huge challenge to the climatologists. It is the most important component of a climate system. Most of the burning issues of our time like global warming, floods, draught, heat waves, soil erosion and many other climate issues are directly related with rainfall. The rainfall forecast information is an essential requirement to support water resources management especially when it is related to climate change in different regions. Agriculture is still the main source of economic activities in most of the countries of the world. In order to increase crop production and protecting crops, human life, ecosystem there is an increasing demand from the policymakers for a reliable prediction of rainfall. Therefore, it is really very important to be able to predict and finding the determinants of rainfall correctly. The modelling and prediction of rainfall in different parts of the world have been studied by many authors. The rainfall in Bangladesh is studied by [1]. [9] and [12] have studied rainfall in India. The rainfall in Japan, New Zealand, and Taiwan have been studied by [11], [14], and [16] respectively. According to [1,9,14] the climate variables that usually affect rainfall are temperature, evaporation, humidity, monsoon, geographic location, air frost days, sun hours etc. There are different techniques used for the prediction of rainfall such as regression analysis [1,9,14], clustering [12], artificial neural networks (ANN) [11], deep learning [16] etc. Among the regression techniques, the linear regression and the transformed linear regression models are very commonly used [1, 14]. The application of logistic regression in the prediction of rainfall is studied in [9]. We will see later in this paper that the standard assumptions required for the validity of linear and or transformed linear model may not hold in reality. According to [3], the nonparametric regression is the most appropriate choice. This motivates us to model and predict rainfall using a nonparametric regression called LOWESS.

The paper is organized as follows. In section 2, we briefly discuss the linear, the transformed linear and the nonparametric regression LOWESS methods. A variety of regression diagnostics techniques such as tests for normality, goodness of fit, heteroscedasticity, and multicollinearity are discussed here. This section also gives a brief description of cross validation. In section 3, we fit the rainfall data by different models. We employ a variety of diagnostic tools to compare the goodness of fits of different methods. We also determine the climate variables which have significant impacts on rainfall. Section 4 reports a cross validation study which is designed to evaluate the performance of different methods used for generating forecasts of rainfall in the U.K.

Article history:
Received 13 May, 2020
Received in revised form 28 June, 2020
Accepted 30 June, 2020
Available online 15 August, 2020

Corresponding author details: A.H.M. Rahmatullah
E-mail address: rimon@bsu.edu
Tel: +1 765 285 8650

Data and Methodology: We have taken climate data recorded by 26 weather stations in the U.K. These stations are Aberporth, Armagh, Bradford, Braemar, Camborne, Chivenor, Cwmystwyth, Dunstaffnage, Durham, Eastbourne, Eskdalemuir, Hurn, Lerwick, Leuchars, Lowestoft, Manston, Nairn, Oxford, Paisley, Ringway, Ross-on-Wye, Shawbury, Sheffield, Southampton, Sutton Bonington, and Tiree. The available climate data are monthly rainfall, maximum temperature, minimum temperature, average temperature, air frost days, and sun hours up to 2019. All weather stations have data for more than hundred years back, the oldest record we have is data from 1886.

We write the multiple regression model as

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} + \varepsilon_i, \quad i = 1, 2, \ldots, n \qquad \text{eq. 1}$$

where Y is the dependent variable, the X's are the independent variables, β is a (k+1) vector of parameters measuring the effect of the corresponding explanatory variables, and ε is the error term. We can express the multiple regression model （1）in matrix notation as

$$Y = X\ \beta\ +\ \varepsilon \qquad \text{eq. 2}$$

The vector of parameters β is estimated by the ordinary least squares (OLS) by minimizing the sum of squares errors of the errors yielding

$$\hat{\beta}\ =\ (X'X)^{-1} X'Y \qquad \text{eq. 3}$$

We also obtain

$$V(\hat{\beta}) = \sigma^2 (X'X)^{-1} \qquad \text{eq. 4}$$

The inferential procedure of a regression model heavily depends on several assumptions that include that the relationship between Y and X is linear but no exact linear relationship exists between two or more X's and the error term is normally distributed with zero mean and a constant variance σ^2.

Most of the standard results of linear regression model are based on the normality assumption and the whole inferential procedure may be subjected to error if this assumption is violated. An excellent review of different aspects of normality tests is available in [6]. Violation of the normality assumption may lead to the use of suboptimal estimators, invalid inferential statements and inaccurate predictions. So, for the validity of conclusions we must test the normality assumption. The simplest graphical display for checking normality in regression analysis is the normal probability plot. This method is based on the fact that if the ordered residuals are plotted against their cumulative probabilities on normal probability paper, the resulting points should lie approximately on a straight line. Graphical methods are very simple but often are very subjective. So we need some analytical tests for normality as well. In recent years the Jarque-Bera test [10] has become very popular. This is an omnibus test for normality, combining the coefficient of skewness and kurtosis in one test statistic. The standard theory tells us that if a data is normal its skewness S is 0 and its kurtosis K is 3. So a large deviation either of S from 0 or K from 3 should give a significant result, regardless of which one deviates from normal values. Keeping this in mind, [10] defined a formal test for normality as

$$\text{JB} = [n / 6]\ [S^2 + (K-3)^2 / 4] \qquad \text{eq. 5}$$

It is reported by many authors [7] that the JB statistic often suffers from possessing poor power in regression problems. To overcome this problem, [7] suggests a slight adjustment to the JB statistic to make it more suitable for the regression problems. His proposed statistic based on rescaled moments (RM) of ordinary least squares residuals is defined as

$$\text{RM} = [n c^3 / 6]\ [S^2 + c(K-3)^2 / 4] \qquad \text{eq. 6}$$

where $c = n/(n-p)$, where $p = k + 1$ is the number of parameters of the linear regression model. Both the JB and the RM statistic follow a chi square distribution with 2 degrees of freedom. If the values of these statistics are greater than the critical value of the chi square, we reject the null hypothesis of normality.

We can adopt several measures to check or test the goodness of fit of models. An excellent review of different types of measures for goodness of fit is available in [8]. The easiest and the simplest of all is the *R*-square which is a proportion of the amount of variation in the explanatory variable explained by the regression model in comparison with the total variation. Although it is very popular, it does not take into account of either the number of observations or the number of variables considered in the model. To overcome this shortcoming another measure, called adjusted *R*-square is preferred. However, neither the *R*-square nor the adjusted *R*-square follows a probability distribution. For this reason, they can be used to check the goodness of fit, the higher the value (close to 1), the better the fit, but they cannot provide formal tests. To overcome this problem another measure can be used which is the *F* value for the regression given in the ANOVA table. The *F* statistic of the regression ANOVA follows an *F* distribution with $p - 1$ and $n - p$ degrees of freedom. Akaike [2] introduced the Akaike information criterion (AIC), an information theoretic approach for model/variable selection, via Kullback-Leibler divergence. Since then the AIC is being considered as one of the most popular and commonly used measure of goodness of fit. It is defined as

$$\text{AIC} = n \ln (SSE/n) + 2p \qquad \text{eq. 7}$$

Another model/variable selection criterion via Kullback-Leibler divergence is the Bayesian information criterion (BIC) defined as

$$\text{BIC} = n \ln (SSE/n) + p \ln(n) \qquad \text{eq. 8}$$

A model that corresponds to the lowest AIC or BIC value is considered to be the best.

One important assumption of a classical regression model is that the error term has a constant variance for all observations. In this situation, we call the error term homoscedastic. However, there are many occasions when the assumption of homoscedastic error variance is unreasonable. For example, the monthly or seasonal variation of rainfall is not likely to be the same. If the error variance changes, we call the error heteroscedastic. When heteroscedasticity is present, the OLS estimation technique places implicit weighting to the fit and consequently the variances of the estimators tend to be higher than those should be. A good number of tests are now available in the literature for testing heteroscedasticity of errors. The residuals-fits (RF) plot gives a very simple graphical display- a funnel shape indicates the presence of heteroscedasticity. Among the analytical tests, the Goldfeld and Quandt test, the Breusch and Pagan test, and the White test have become very popular to the statisticians [8]. We would employ the White test [15] in our study because it is very easy to understand and has a wide range of applications. In this test, the square of the regression residuals are fitted on the explanatory variables and the value of R^2 is computed from the auxiliary fitted regression model. The White test is based on the fact that under homoscedasticity,

$$n R^2 \sim \chi^2 (p) \qquad \text{eq. 9}$$

Here *p* is the number of explanatory variables in the auxiliary regression model including the constant.

Multicollinearity arises when the assumption that no exact linear relationship exists between two or more explanatory variables is violated. It has several consequences such as wrong sign problem of the regression coefficients and unduly large variances and covariances for the estimators of the regression parameters. The variance inflation factor (VIF) is the most popular method for the detection of multicollinearity. Since $V(\hat{\beta}) = \sigma^2 (X'X)^{-1}$, the diagonal elements of $(X'X)^{-1}$ is the source of possible inflation of variance. One or more large VIF's indicate multicollinearity. A rule of thumb tells that there is not enough evidence of multicollinearity if VIF < 5. If VIF is in between 5 and 10, there is an evidence of moderate multicollinearity. VIF > 10 indicates the presence of severe multicollinearity [8].

We often observe that a linear regression model may not adequately fit the data. Sometimes a simple transformation on the variables can do a very good job in this situation. Nevertheless, the question still remains, how we choose the appropriate transformation? Experiences in data analysis may often help us to select an appropriate transformation, but it is very subjective. Box and Cox [4] suggested a useful class of transformations, known as the power transformation

$$Y^{(\lambda)} = \begin{cases} \left(Y^{\lambda} - 1\right)/\lambda & \lambda \neq 0 \\ \ln Y & \lambda = 0 \end{cases} \qquad \text{eq. 10}$$

where λ is a parameter to be determined. The equation (10) gives a family of transformations. For example λ = 1 gives us the OLS fit. λ = –1 suggests a reciprocal transformation. λ = ½ means use $\sqrt{Y}$ as response. But the problem is Y^{λ} approaches unity as λ approaches zero. It is meaningless to have all of the response values equal to a constant. One approach to solving this problem is to use $\left(Y^{\lambda} - 1\right)/\lambda$ as the response variable. As λ tends to zero $\left(Y^{\lambda} - 1\right)/\lambda$ goes to a limit of ln *Y*. We can play with different values of λ, preferably between -2 and 2 and can obtain the optimum fit.

In both linear and transformed linear regression we start with a tentative model and then apply diagnostics to see if the model should be modified. Sometimes it may be extremely useful to employ a strictly adaptive approach where the regression equation is determined from the data. According to [3] a nonparametric regression is a collection of techniques for fitting a curve when there is a little a priori knowledge about its shape. Here we relax all the basic assumptions and fit the data purely by empirical way. We often use a weighting function, called kernel, for obtaining a smooth estimator. In the literature, the size of the neighbourhood is referred to as the bandwidth. With a kernel smoother the weights for the X_i depend on their distance from the point of interest x_0. Specifically, the weight assigned to X_i for obtaining the predicted value at x_0 is

$$w_{0i} = \frac{c_0}{\lambda} K\left(\frac{|x_0 - x_i|}{\lambda}\right)$$

where *K*(*t*) is an even function decreasing in *t*, λ is the bandwidth, and c_0 is a constant that make the weights sum to 1. A kernel smoother is frequently written in the general form

$$\hat{y}_i = \frac{\sum_1^n K\left[\left(x_i - x_j\right)/\lambda\right] y_j}{\sum_1^n K\left[\left(x_i - x_j\right)/\lambda\right]} = \sum_{j=1}^n w_{ij} y_j \qquad \text{eq. 11}$$

In our study, we use a special nonparametric regression technique, known as a local regression. This method was introduced by [5]. It is more popularly known as the locally weighted smoothing scatterplot (LOWESS) regression. It uses the data from a neighbourhood around the specific location. Typically the neighbourhood is defined as the span, which is the fraction of the total data points used to form neighbourhood. A span of 0.5 (which is a very popular choice) indicates that the closest half of the data points are used as the neighbourhood. The LOWESS procedure then uses the points in the neighbourhood to generate a weighted least squares (WLS) estimate of the specific response for the correction of possible heteroscedasticity. The weighted least squares procedure uses a low-order polynomial, usually a simple linear regression or a quadratic regression model.

Cross-validation is a technique for assessing how the results of a statistical analysis will generalize to an independent data set. It is mainly used in settings where the goal is prediction, and one wants to estimate how accurately a predictive model will perform in practice. One round of cross-validation involves partitioning a sample of data into complementary subsets, performing the analysis on one subset (called the training set), and validating the analysis on the other subset (called the validation set or testing set). A popular cross validation technique is data splitting [13], where we set aside some of the original data (usually 10% to 20%), develop a prediction equation using the selected data, and apply this equation to the samples set aside. These actual and predicted output (for the samples set aside) help us to compute the mean squared error if the response variable is quantitative or misclassification probability if the response is class variable. For a regular regression problem we may determine the accuracy of the prediction model by computing several popular accuracy measures [8] such as
Mean Absolute Error (MAE):

$$MAE = \frac{1}{m}\sum_{t=1}^{m}|y_t - \hat{y}_t| \qquad \text{eq. 12}$$

where y_i equals the actual value, $\hat{y}_i$ equals the fitted value, and *m* equals the number of observations in the test set.

Root Mean Squares Error (RMSE):

$$RMSE = \sqrt{\frac{1}{m}\sum_{t=1}^{m}(y_t - \hat{y}_t)^2} \qquad \text{eq. 13}$$

Mean Absolute Percentage Error (MAPE):

$$MAPE = \frac{\sum_{i=1}^{m}|(y_i - \hat{y}_i)/y_i|}{m} \times 100 \qquad \text{eq. 14}$$

For all three measures, the smaller the value, the better the forecast of the model.

Fitting of Rainfall in the U.K.: In this section, we employ different regression methods to fit the rainfall data in the U.K. We begin with a linear regression model by the method of least squares. We have data from 26 weather stations of the U.K. We have analyzed data for each weather station separately and if we report all of them it will be very long. Most of the weather stations show very similar results so reporting of all of them seems to be a repetition of the same thing. Summary results for all 26 weather stations are reported in Table 6 at the later portion of this paper. For brevity, we report extensive results from one out of 26 weather stations and that is Durham. Durham climate data contains monthly records from 1949 to 2019. At first we planned to fit rainfall on maximum and minimum temperature, sun hours, air frost days, and seasons. We have converted months into four seasons and label them 1 through 4. February, March, April constitute spring, May, June, July constitute summer, August, September, October constitute autumn, and November, December and January constitute winter. A multiple regression result shows that maximum and minimum temperature creates a severe multicollinearity problem because their corresponding VIF values are more than 70. For this reason we have dropped the maximum and minimum temperature from the model and selected the average temperature as an explanatory variable.

Table 1 reports the summary statistics of the linear regression model for Durham rainfall data. We observe that the coefficients of all of the explanatory variables considered in this fit are highly significant. Temperature and season have a positive impact on the rainfall. We also observe that sun hours and air frost days have negative impact on rainfall. The VIF values are all below 5 indicating that the multicollinearity problem does not arise here.

Table 1. Summary Statistics of the Linear Model for Durham Rainfall Data.

Variables	Coefficients	Standard Errors	*t*-value	Significance	VIF
Constant	88.04	6.42	13.71	0.000	
Temperature	2.336	0.688	3.40	0.001	4.55
Sun Hours	-0.3654	0.0374	-9.78	0.000	3.86
Air Frost Days	-2.342	0.612	-3.83	0.000	1.84
Season	5.56	1.68	3.30	0.001	1.67

Table 2. Regression Diagnostics of the Linear Model for Durham Rainfall Data.

MSE	R^2	Adj R^2	ANOVA *F*	AIC	BIC	JB	RM	White
1432	0.2842	0.2799	66.01 (0.000)	4873.71	4876.73	67.0698 (0.000)	69.5940 (0.000)	77.318 (0.000)

The diagnostics presented in Table 2 show that the F value of the regression ANOVA is highly significant (p-value is 0.000) which confirms that the overall regression is significant. From the results presented here we obtain the value of R^2 is only 0.2842 which is not satisfactory. The value of the adjusted R^2 is 0.2799. The AIC value for this model is 4873.71 and the BIC is 4876.73.

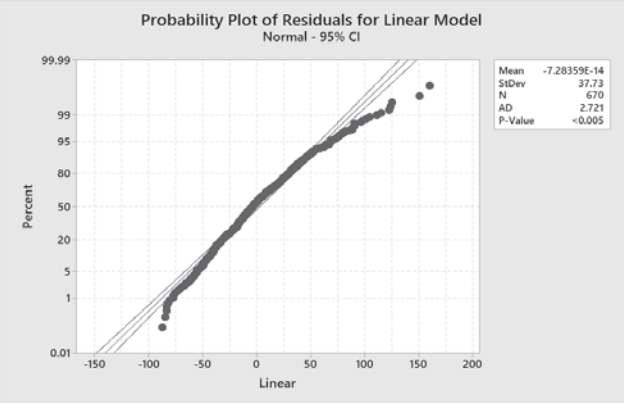


Fig. 1: Normal Probability Plot of Residuals for the Linear Regression Model

However, the normal probability plot as shown in Figure 1 indicates non-normal behavior of errors. For the residuals the sample skewness S = 0.62 and the sample kurtosis K = 3.93 yielding the value of the Jarque-Bera statistic JB = 67.0698 and the rescaled moment statistic RM = 69.594. We know that both the JB and the RM follow Chi-square distribution with 2 degrees of freedom. At the 5% level, the cut-off value for a Chi-square (2) is 5.99. Hence, there exists a very strong evidence against the normality of the errors.

In addition to normality test, we check for possible heteroscedasticity. The residual-fits plot as shown in Figure exhibits a funnel shape around 0 which is a clear indication of heteroscedasticity. For this data the value of the White test statistic is W = 77.318. This test follows a Chi-square distribution with 5 degrees of freedom. At the 5% level, the calculated value of Chi-square (5) is 11.07, so we must reject the null hypothesis of homoscedasticity and acknowledge the fact that the errors are heteroscedastic.

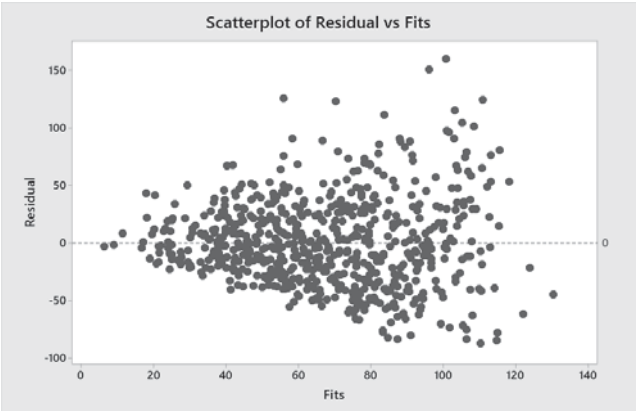


Fig. 2: Residual-Fits Plot for Durham Rainfall Data

Since the values of R-square and adjusted R-square are very low and there are strong evidences of non-normality and heteroscedasticity of errors, the linear regression model is not appropriate for this data. The immediate remedy is to do transformations on variables before the fit. We employ the optimal Box-Cox transformation on the variables and the summary statistics of this fit are presented in Table 3. We observe from this table that the summary statistics of the transformed linear regression model are very similar to those of a linear regression model.

Table 3. Summary Statistics of the Transformed Linear Model for Durham Rainfall Data.

Variables	Coefficients	Standard Errors	*t*-value	Significance	VIF
Constant	6.287	0.226	27.80	0.000	
Temperature	0.0876	0.0242	3.62	0.000	4.55
Sun Hours	-0.01364	0.00132	-10.36	0.000	3.86
Air Frost Days	-0.0730	0.0216	-3.38	0.001	1.84
Season	0.1376	0.0593	2.32	0.021	1.67

Table 4. Regression Diagnostics of the Transformed Linear Model for Durham Rainfall Data.

MSE	R^2	Adj R^2	ANOVA *F*	AIC	BIC	JB	RM	White
1427	0.2852	0.2809	66.34 (0.000)	4871.41	4874.73	7.6855 (0.000)	7.8601 (0.000)	10.653 (0.000)

We observe slight improvements in the *F* value of the regression ANOVA, R^2, adjusted R^2, AIC, and BIC for the transformed model in comparison with the linear model in Table 4, but these improvements are negligible. However, there occurs a significant change in heteroscedasticity. The White test statistic for the transformed model is 10.653 with a *p*-value 0.059. We can say that this transformation corrects the problem of heteroscedasticity.

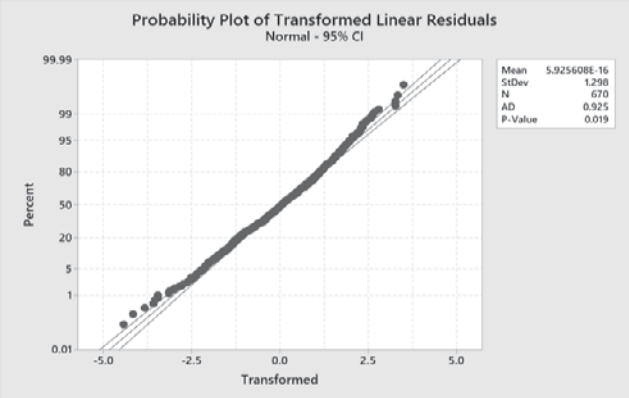


Fig. 3: Normal Probability Plot of Residuals for the Transformed Linear Model

The normal probability plot of transformed model is presented in Figure 3. Obviously, this plot looks more normal than the original model as shown in Figure 1. The Jarque-Bera and the RM test statistics for the transformed model are 7.6855 (*p*-value 0.0214) and 7.8601 (*p*-value 0.0916) respectively. These results tell that the optimal Box-Cox transformation improve the non-normality to a certain extent but it cannot eliminate the problem entirely.

Although the optimal Box-Cox transformation corrects the heteroscedasticity problem and improves the non-normality the latter problem still remains. Moreover, this makes a negligible improvement in the fit. For this reason we need to employ another model to fit this data. At this point we select a LOWESS model in R. Since LOWESS is a nonparametric model it does not assume normality of errors and consequently it does not matter whether the errors of the model follow a normal distribution or not. It uses the weighted least squares in determining the weights and thus correct the problem of heteroscedasticity automatically.

Table 5. Measures of Adequacy of Different Fits for Rainfall Data.

Models	MSE	R^2	Adj R^2	ANOVA F	AIC	BIC
Linear	1432	0.2842	0.2799	66.01	4873.71	4876.73
Optimal Box-Cox	1427	0.2852	0.2809	66.34	4871.41	4874.73
LOWESS	**852**	**0.5732**	**0.5706**	**896.38**	**4525.86**	**4528.88**

Table 5 offers a comparison to evaluate how LOWESS fits the rainfall data. We consider this with linear and transformed models. Results presented clearly show the merit of using the nonparametric LOWESS to fit the data. It outperforms the other two competitors in every respect. It produces the maximum R-square, adjusted R-square and ANOVA F, and the smallest MSE, AIC, and BIC. This method is followed by the transformed linear regression model and the linear regression method. The performance of the latter two are very similar.

Now we run all the three models to all 26 weather stations separately and the significant predictors appeared in Table 6.

Table 6. List of Significant Predictors for All 26 Weather Stations.

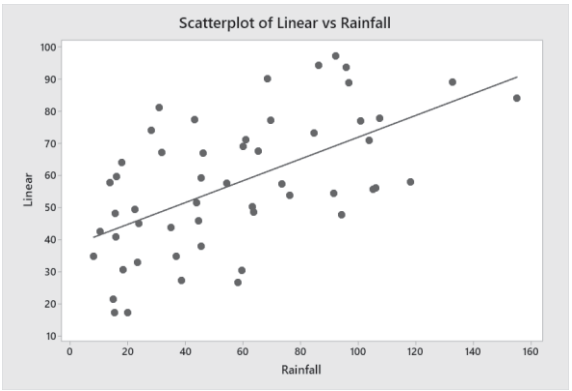
Stations	Linear / Linear Box Cox	Nonparametric
1, 2, 3	Sun***, AF***, Season***	Temp*, Sun***, AF***, Season**
4, 5	Temp*, Sun***, AF***, Season**	Temp*, Sun***, AF***, Season**
6, 7	Sun***, AF***, Season**	Sun***, AF***, Season**
8	Temp**, Sun***, AF***, Season**	Temp**, Sun***, AF***, Season**
9, 13, 14, 19. 20, 21, 22, 23, 24, 25	Temp***, Sun***, AF***, Season**	Temp***, Sun***, AF***, Season**
10	Sun***, AF***	Sun***, AF***, Season**
11	Sun***, AF***, Season**	Temp**, Sun***, AF**, Season**
12	Temp**, Sun***, AF**, Season**	Temp**, Sun***, AF**, Season**
15	AF Days***, Sun***, Temp**	Temp***, Sun***, AF***, Season**
16	Temp***, Sun***, Season**	Temp***, Sun***, AF***, Season**
17	Sun***, Season**	Sun***, AF Days***, Season**
18	Sun***, AF Days***, Season**	Temp*, Sun***, AF***, Season**
26	Temp***, Sun***, AF***, Season*	Temp***, Sun***, AF***, Season*

*** = significant at 1% level, ** = significant at 5% level, * = significant at 10% level

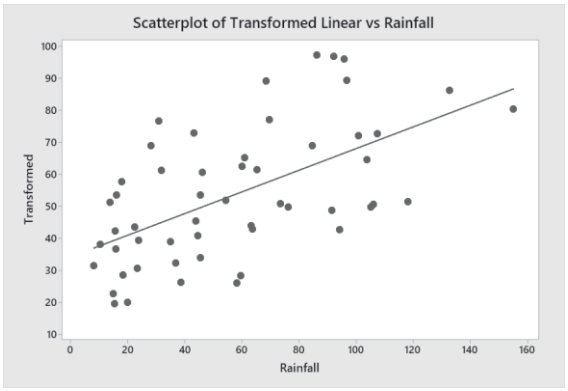
The above results show that sun hour is the most significant predictor of rainfall. It becomes significant for all three models and in all 26 weather stations. The other significant predictors are air frost days, season and temperature.

Prediction of Rainfall in the U.K.: In this section we report a cross validation study which is designed to evaluate forecasts of rainfall generated by three different models considered in our study. For all these methods we have used roughly 90% observations as a training set and the remaining 10% as a test set. Based on the data in the training set, we generate forecasts for the last 10% observations using a linear regression model, optimal Box-Cox transformation, and LOWESS.

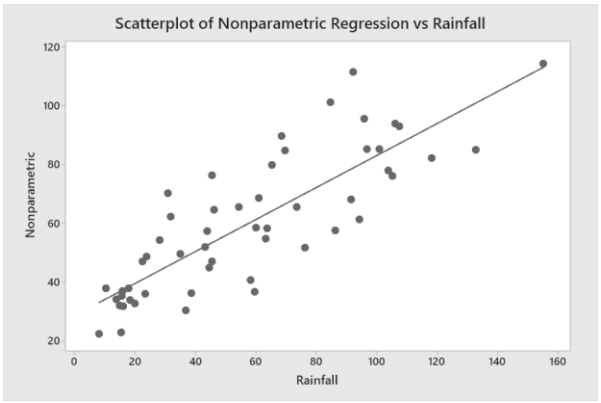
Scatterplot of the forecasted values against the original values for these three different models are presented in Figure 4. We observe a linear positive relationship between observed and forecasted rainfall for all these three methods but is clear that LOWESS produces the best fit.



Linear



Transformed



LOWESS

Fig. 4: Scatter Plot of Original and Forecasted Rainfall for Different Models

Table 7. Accuracy Measures of Different Forecasts for Rainfall Data

Models	MAE	RMSE	MAPE	Correlation
Linear	32.25	40.2907	76.20	0.583
Optimal Box-Cox	32.76	41.4856	69.50	0.582
LOWESS	**18.11**	**20.9468**	**52.26**	**0.831**

Table 7 clearly shows the merit of using the LOWESS in forecasting the rainfall data. It outperforms the other two competitors in every respect. LOWESS produces forecasts that yield the minimum mean absolute error (MAE), the root mean (sum of) squares (of) errors (RMSE), and the mean absolute percentage error (MAPE) and also gives the maximum correlation coefficient with the true rainfalls. This method is followed by transformed linear and linear regression models although the previous one is marginally better than the latter one.

Conclusions: The main objective of our research is to find the most appropriate method to fit and predict rainfall in the U.K. and to determine the climate variables that have significant impacts on rainfall. Regression diagnostics reveal that the most popular and commonly used linear model suffer from nonmorality and heteroscedasticity problems and produces the worst fit. Even the optimal transformation of linear model did not improve things as expected; however, nonparametric LOWESS method performs extremely well not only in fitting the data, but also in generating the best quality forecasts. It outperforms the other methods in every respect. Sun hours, air frost days, season and temperature emerge as the most significant determinant of rainfall for the most part of the U.K.

Acknowledgements: The authors express their thanks and gratitude to the anonymous reviewers for giving some useful suggestions that led to considerable improvement in the methodology and presentation of the results.

References:

[1] R. Ahmed, S. Karmakar, Arrival and withdrawal dates of the summer monsoon in Bangladesh, International Journal of Climatology, 13 (1993) 727–740.
[2] H. Akaike, A new look at the statistical model identification, IEEE Transactions on Automatic Control, 19 (1974) 716–723.
[3] N. S. Altman, An introduction to kernel and nearest neighbor nonparametric regression, The American Statistician, 46 (1992) 175-185.
[4] G. E. P. Box, D. R. Cox, An analysis of transformation (with discussions), Journal of the Royal Statistical Society, Series B, 26 (1964) 211-252.
[5] W. S. Cleveland, Robust locally weighted regression and smoothing scatterplots, Journal of the American Statistical Association, 83 (1979) 596-610.
[6] K. R. Das, A. H. M. R. Imon, A brief review of tests for normality, American Journal of Theoretical and Applied Statistics, 5 (2016) 5–12.
[7] A. H. M. R. Imon, Regression residuals, moments, and their use in tests for normality, Communications in Statistics—Theory and Methods, 32 (2003), 1021–1034.
[8] A. H. M. R. Imon, Introduction to Regression, Time Series, and Forecasting, Nandita Prokash, Dhaka, 2017.
[9] A. H. M. R. Imon, M. C. Roy, S. K. B. Bhattacharjee. Prediction of rainfall using logistic regression, Pakistan Journal of Statistics and Operation Research, 8 (2012) 655-667.
[10] C. M. Jarque, A. K. Bera, Efficient tests for normality, homoscedasticity and serial independence of regression residuals, Economics Letters, 6 (1980) 255–259.
[11] Tomoaki Kashiwao, Koichi Nakayama, Shin Ando, KenjiIkeda Moonyong Lee, Alireza Bahadori. A neural network-based local rainfall prediction system using meteorological data on the Internet: A case study using data from the Japan Meteorological Agency, Applied Soft Computing, 56 (2017) 317-330.
[12] R. S. Kumar, C. Ramesh. A study on prediction of rainfall using datamining technique, International Conference on Inventive Computation Technologies (ICICT), India, 2016.
[13] D. Montgomery, E. Peck, G. Vining, Introduction to Linear Regression Analysis, 4th ed., Wiley, New York, 2006.
[14] M. J. Salinger, G. M. Griffiths, Trends in New Zealand daily temperature and rainfall, International Journal of Climatology, 21 (2001) 1437-1452.
[15] H. White, A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity, Econometrica, 48 (1980) 817–838.
[16] Meng-Hua Yen, Ding-Wei Liu, Yi-Chia Hsin, Chu-En Lin, Chii-Chang Chen. Application of the deep learning for the prediction of rainfall in Southern Taiwan, Scientific Reports, 9 (2019), Article number: 12774, https://doi.org/10.1038/s41598-019-49242-6.