A Simple Weather Forecasting Model Using Mathematical Regression

Paras¹ and Sanjay Mathur²
1. Assistant Professor, 2. Associate Professor,
Department of Electronics & Communication Engineering, College of Technology, G.B. Pant University of Agriculture & Technology, Pantnagar, (India) 263 145

ABSTRACT

A simple model for weather forecasting has been described. The model is simple due to the fact that it uses simple mathematical equation using Multiple Linear Regression (MLR) equations that can be easily understood by a medium educated farmer. Weather data at a particular station is recorded which is a time-series data. The weather parameters like maximum temperature, minimum temperature and relative humidity have been predicted using the calculated features depending upon the correlation values in the weather data series over different periods from the weather parameter time-series itself. Relative humidity is also predicted using time series of maximum and minimum temperature and rainfall. The category of rainfall has been estimated using features of maximum and minimum temperature and relative humidity. The development phase of the model is to obtain MLR equations using input set and output parameter. The coefficients of these regression equations have been used to estimate the future weather conditions. The personal computer and simple data processing software like MS Excel can be used to make and validate the model by the user itself. The results obtained show that MLR model can estimate the weather conditions satisfactorily.

Key words: Weather forecasting; Time series; Features; Multiple Linear Regression; Correlation;

Weather Forecasting is an important and necessary area of investigation in human life. Weather for future is one of the most important attributes to forecast because agriculture sectors as well as many industries are largely dependent on the weather conditions. Weather conditions are required to be predicted not only for future planning in agriculture and industries but also in many other fields like defence, mountaineering, shipping and aerospace navigation etc. It is often used to warn about natural disasters are caused by abrupt change in climatic conditions.

At macro level, weather forecasting is usually done using the data gathered by remote sensing satellites. Weather parameters like maximum temperature, minimum temperature, extent of rainfall, cloud conditions, wind streams and their directions, are projected using images and data taken by these meteorological satellites to access future trends. The variables defining weather conditions like temperature (maximum or minimum), relative humidity, rainfall etc., vary continuously with time, forming time series of each parameter and can be used to develop a forecasting model either statistically or using some other means that uses this time series data (Chatfield, 1994; Montgomery and Lynwood, 1996). The application of time series data is merely not limited to weather forecast only but also in stock market forecast (Mathur, 1998; Mathur, Pant, Shukla, 2000) and in agriculture areas (Roadknight et al., 1997).

In statistical analysis, regression models are often used for estimating the future events or values. Trend extraction and curve fitting methods are also used to estimate the future behavior of the time series and to fit the future data according to the trend. Regression analysis includes parametric methods such as linear and logistic regression. Non-parametric methodologies such as projection pursuit, additive models, multivariate adaptive regression etc. have also been applied to estimation and prediction problems (Holmström et al. 1997).

In this paper Multiple Linear Regression (MLR) is used to develop a model for forecasting weather
parameters. The proposed model is capable of forecasting the weather conditions for a particular station using the data collected locally. The data is processed to obtain some statistical indicators to extract the hidden information present in the time series. These statistical indicators, also called as features viz. moving average (MA), exponential moving average (EMA), rate of change (ROC), oscillator (OSC), moments (µ2, µ3 and µ4) and coefficients of skewness and kurtosis are calculated over certain periods. On the basis of correlation, features are chosen as inputs to the models. Regression equations are obtained for the parameter to be forecast, which is termed as target. The whole data set is divided in two parts, the first is used to obtain MLR equations and remaining is used for testing the model. The power of MS Excel has been used to process the data and present the results simply in understandable form.

**MULTIPLE LINEAR REGRESSION (MLR) MODEL**

Regression explains the nature of relationship, i.e., the average probable change in one variable given by a certain amount of change in the other variable.

The general equation of regression of Y on X is

\[ Y = \alpha + \beta X + e \]  

This equation is known as the mathematical model for linear regression. As the special case the form \( Y = \alpha + \beta X \) is called the deterministic model (Agarwal, 1991).

Multiple Linear Regression equation, which describes the linear relationship with set of dependent variable Y, and k sets of independent variables X1, X2, X3,............, Xk is,

\[ Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots \ldots + \beta_k X_k + e \]  

where Y is predictant and X1, X2, X3, ....Xk are the predictors. ‘e’ is the error which is distributed normally with zero mean and variance \( \sigma^2 \), i.e. \( e \sim N(0, \sigma^2) \).

To develop the proposed model, the parameters \( \alpha, \beta_1, \beta_2 \ldots \ldots \) are estimated using the training sample sets. \( X \) is called the independent variable X. It is obtained as

\[ \beta = \frac{\text{Variance explained}}{\text{Total Variance}} \]

A high \( R^2 \) shows that there exists a linear relationship between the two variables. If \( R^2 = 1 \), it indicates the perfect relationship between the two variables.

**EXTRACTION AND SELECTION OF FEATURES FOR THE MODELS**

The prediction model performs better when some hidden features are presented which enhance its adapting capability. The features considered in this study are statistical indicators calculated for different time frames and explained as below:

(i) **Moving Average (MA):** It is calculated progressively as an average of \( N \) number data values over the certain period. Data set is represented by \( d_t, d_{t-1}, d_{t-2}, \ldots \ldots , d_0 \) where \( d_t \) is present and \( d_0 \) is the first data value, the moving average with a sliding window of period \( N \) is

\[ \text{MA} = \frac{d_t + d_{t-1} + d_{t-2} + \ldots \ldots + d_{t-N}}{N} \]

(ii) **Exponential Moving Average (EMA):** It is defined as

\[ \text{EMA} = (1-a) \times EMA_{t-1} + a \times X_t \]

where ‘a’ is called the smoothing constant having value \( 0 < a < 1 \).

(iii) **Oscillator (OSC):** Oscillator is used to indicate the rising or trailing trend present in the time series. It is defined as difference of moving averages or exponential moving averages of two different periods.

\[ \text{OSC} = \text{MA}_{N_1} - \text{MA}_{N_2} \]

where \( N_1 \) and \( N_2 \) are different periods and \( N_1 > N_2 \).

(iv) **Rate of Change (ROC):** It indicates the rate of change of the variable at present, as compared to the value of the variable at certain period back. Thus percentage ROC at ‘a’ times back is given by,

\[ \text{ROC} = \frac{d_t - d_{t-a}}{d_t} 100\% \]

(v) **Moments:** The \( r \)th moment of \( X \), \( \mu_r \) is given by

\[ \mu_r = \frac{1}{N} \sum_{i=0}^{N} (X_i - \mu)^r \]

where \( X \) is total data set and \( \mu \) is the mean of \( N \) data variables.

Second, third and fourth moments correspondingly to \( r = 2, 3 \) and \( 4 \) are considered.

(vi) **Beta Coefficient of Skewness (\( \beta_2 \)):** It is defined as

\[ \beta_2 = \frac{\mu_3^2}{\mu_2^3} \]

(vii) **Beta Coefficient of Kurtosis (\( \beta_2 \)):** It is defined as

\[ \beta_2 = \frac{\mu_4^2}{\mu_2^4} \]

For estimation of different weather parameters certain time frames are decided to gather the information inherited by time series. Three different time frames chosen are 15 weeks (w15), 30 weeks (w30) and 45 weeks (w45) over which these features are extracted.
In certain cases, the weather parameter to be forecast can be estimated on the basis of the features of same parameter. However, in some cases the parameter to be forecast exhibits a strong dependence on other weather parameters. In such cases the model should include the features extracted from other weather parameters also. Therefore, to forecast each parameter, independent variables are decided to include all these features exhibiting strong trends established by them over the period of observation.

To assess whether a specific feature is suitable for the model or not its correlation with the target is obtained. Therefore, different features may be formed suitable for estimating different weather parameters. For example, for estimating maximum temperature ($T_{\text{max}}$) and minimum temperature ($T_{\text{min}}$) five features namely, MA, EMA, OSC, ROC and $\mu$3 are selected. For example, 0.498, -0.161, 0.679, -0.246 and 0.579 for $T_{\text{max}}$ and 0.591, -0.178, -0.725, 0.527, -0.202 for $T_{\text{min}}$ are the values of correlation coefficients for a particular period only. However, in the case of estimation of relative humidity (RH), two cases are considered. In the first case, $T_{\text{max}}$, $T_{\text{min}}$ and rainfall (RF) are considered as input variables because relative humidity exhibits a strong dependence on these parameters. In the second case, RH is estimated using the features, MA, EMA, OSC, ROC and $\mu$3 are selected.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Regression Equation obtained</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maximum temperature estimation using features of maximum temperature calculated over period of 15 weeks</td>
<td>$Y = 0.339 + 0.635654X1 - 0.15217<em>X2 + 0.180097</em>X3 - 0.02085<em>X4 + 0.095199</em>X5$</td>
</tr>
<tr>
<td>Maximum temperature estimation using features of maximum temperature calculated over period of 30 weeks</td>
<td>$Y = 0.924849 - 0.45934<em>X1 + 0.147329</em>X2 + 0.016252<em>X3 + 0.095199</em>X5$</td>
</tr>
<tr>
<td>Maximum temperature estimation using features of maximum temperature calculated over period of 45 weeks</td>
<td>$Y = 0.723898 - 0.3241<em>X1 + 0.079132</em>X2 - 0.43317<em>X3 + 0.125765</em>X4 + 0.398313<em>X5 - 0.18627</em>X6 - 1.61159<em>X7 + 0.469794</em>X8 - 0.21661*X9$</td>
</tr>
<tr>
<td>Minimum temperature estimation using features of minimum temperature calculated over period of 15 weeks</td>
<td>$Y = 1.101016 - 1.00977<em>X1 + 0.510472</em>X2 - 0.21753<em>X3 + 0.225351</em>X4 - 0.19177<em>X5 - 0.00953</em>X6$</td>
</tr>
<tr>
<td>Minimum temperature estimation using features of minimum temperature calculated over period of 30 weeks</td>
<td>$Y = 1.677708 - 1.4282<em>X1 + 0.157355</em>X2 - 0.01824<em>X3 + 0.405171</em>X4 + 0.52492<em>X5 - 0.3173</em>X6$</td>
</tr>
<tr>
<td>Minimum temperature estimation using features of minimum temperature calculated over period of 45 weeks</td>
<td>$Y = 1.267648 - 1.06202<em>X1 + 0.361068</em>X2 + 0.060438<em>X3 - 0.08303</em>X5 - 0.07609*X5$</td>
</tr>
<tr>
<td>Relative humidity estimation using extracted features of relative humidity calculated over period of 15 weeks</td>
<td>$Y = 1.03459 - 0.18114<em>X1 + 0.671148</em>X2 + 0.016252<em>X3 + 0.245718</em>X5 - 0.22175*X6$</td>
</tr>
<tr>
<td>Relative humidity estimation using extracted features of relative humidity calculated over period of 30 weeks</td>
<td>$Y = 1.677708 - 1.4282<em>X1 + 0.157355</em>X2 - 0.01824<em>X3 + 0.405171</em>X4 + 0.52492<em>X5 - 0.3173</em>X6$</td>
</tr>
<tr>
<td>Relative humidity estimation using extracted features of relative humidity calculated over period of 45 weeks</td>
<td>$Y = 1.129623 - 0.32981<em>X1 + 0.198254</em>X2 - 0.03391<em>X3 - 0.044988</em>X4 + 0.398313<em>X5 - 0.18627</em>X6 - 1.61159<em>X7 + 0.469794</em>X8 - 0.21661*X9$</td>
</tr>
<tr>
<td>Relative humidity estimation using $T_{\text{max}}$, $T_{\text{min}}$ and RF</td>
<td>$X_1, X_4, X_7$ are MA of $T_{\text{max}}$, $T_{\text{min}}$ and RH, respectively</td>
</tr>
<tr>
<td>Rainfall estimation using features of $T_{\text{max}}$, $T_{\text{min}}$ and RH</td>
<td>$X_2, X_5$ and $X_8$ are EMA of $T_{\text{max}}$, $T_{\text{min}}$ and RH, respectively</td>
</tr>
<tr>
<td>$X_3, X_6$ and $X_9$ are OSC of $T_{\text{max}}$, $T_{\text{min}}$ and RH, respectively</td>
<td></td>
</tr>
</tbody>
</table>
OSC, ROC, \(\mu_2\) and \(\mu_3\) extracted from relative humidity time series itself. The correlation coefficients of these features with RH as target are obtained as 0.239, -0.274, 0.623, 0.528, 0.232 and -0.392, respectively for 45 week period. For forecasting the rainfall, the extent of rainfall is categorized into five categories namely, very low, low, medium, high and very high. This categorization is made on the basis of observation of rainfall data. The model therefore estimates the rainfall only for a particular category.

**IMPLEMENTATION OF THE MODEL**

The weather data available for this study have been recorded at Pantnagar Station since April 1996 to March 1999, obtained as an average of all the daily-recorded values over the week. The whole data set is taken in two parts. The first part (also called training data which is approximately 80%) is taken for obtaining the regression equations. The features are taken as inputs and target variable is taken as output for these regression equations. The model therefore estimates the rainfall only for a particular category.

Table 1

| Regression Equations obtained from regression analysis |

**RESULTS AND DISCUSSION**

The data used are first used to obtain the regression equations that are shown in Table 1 and rest of the data (also called as test data) are used to check the model. The regression equations are used to obtain the output parameters estimated by the regression model for training as well as for testing data. The scatter diagrams show the linear relationship between the actual weather parameter and the output of the MLR model.

Fig 1(a) shows the training target and output from MLR model for estimating maximum temperature over period of 15 weeks. Fig. 1(b) shows the output predicted by MLR model. During training over 61% outputs are in error range less than 5% (68/110).

Fig 1(c) and 1(d) show the training and testing scatter diagrams. During testing about 50% outputs have error less than 5%(15/31).
Fig. 2(a) and 2(b) show the trend curve during training and testing, respectively, in estimation of minimum temperature over period 15 weeks. The curves follow the trend present in actual target. The model has produced 36% samples under 5% error ranges (40/110).

The testing outputs for minimum temperature estimation are shown in Fig. 2(c). Fig. 2(d) represents scatter diagram having linearly distributed data samples. Approximately 98% data samples are under the error range of 3% (9/31).

Fig. 3(a) shows the training target and output for estimating relative humidity over 45-week period. The output trend matches with that of target trend. The training scatter diagram in Fig. 3(b) has all the data points close on the linear curve, which shows that output and target are very similar. During training 85% outputs are in error range less than 3% (49/90).
Fig 3(c)
The testing outputs for relative humidity estimation over 45-week are shown in Fig. 3(c). The scatter diagram in Fig. 3(d) represents linearly distributed data samples about the line. 71% data samples are under the error range of 3% in this estimation (13/21).

Fig 4(b)
Fig. 4(a) and 4(b) show the trend curve and scatter diagram for training in estimation of relative humidity using Tmax, Tmin and RF. Here also the curves are very close to the actual targets. The model when tested produced 62% samples under 5% for training (56/90).

Fig 4(c)
The trend curve and scatter diagram for testing in estimation of relative humidity using Tmax, Tmin and RF are shown in Fig. 4(c) and 4(d). The target and output curves are close to the actual targets. The model when tested produced 94% samples under 3% for training (15/20).
Table 2: \( r^2 \) values for MLR model for each experiment

<table>
<thead>
<tr>
<th>Experiment Name</th>
<th>Period</th>
<th>( r^2 ) for training</th>
<th>( r^2 ) for testing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maximum temperature estimation using features of max temperature</td>
<td>15</td>
<td>0.845</td>
<td>0.809</td>
</tr>
<tr>
<td>--do--</td>
<td>30</td>
<td>0.787</td>
<td>0.703</td>
</tr>
<tr>
<td>--do--</td>
<td>45</td>
<td>0.849</td>
<td>0.678</td>
</tr>
<tr>
<td>Minimum temperature estimation using features of min temperature</td>
<td>15</td>
<td>0.959</td>
<td>0.916</td>
</tr>
<tr>
<td>--do--</td>
<td>30</td>
<td>0.942</td>
<td>0.851</td>
</tr>
<tr>
<td>--do--</td>
<td>45</td>
<td>0.950</td>
<td>0.421</td>
</tr>
<tr>
<td>Relative humidity vs. extracted features of RH</td>
<td>15</td>
<td>0.457</td>
<td>0.184</td>
</tr>
<tr>
<td>--do--</td>
<td>30</td>
<td>0.513</td>
<td>0.092</td>
</tr>
<tr>
<td>--do--</td>
<td>45</td>
<td>0.633</td>
<td>0.117</td>
</tr>
<tr>
<td>Max. Temp., min temp., rain vs. RH</td>
<td>---</td>
<td>0.752</td>
<td>0.490</td>
</tr>
<tr>
<td>Max. Temp., min temp., RH vs. rain</td>
<td>15</td>
<td>0.445</td>
<td>0.395</td>
</tr>
</tbody>
</table>

The Fig. 5(a) shows the rainfall estimation output and target using extracted features of Tmax, Tmin and RH. During training NN performed very good to produce outputs very near to target. 92% samples were in proper category of rainfall during training. Fig. 5(b) shows linear trend of output and target obtained during training (44/75).

The Fig. 5(c) and Fig. 5(b) show the rainfall estimation output and target using extracted features of Tmax, Tmin and RH during testing. 80% samples were in proper category of rainfall during testing. Fig. 5(d) shows linear trend of output and target obtained during testing (9/20).
CONCLUSION

The statistical forecasting models can be developed to estimate the future weather conditions. For maximum temperature and minimum temperature estimation, the optimum size of the period for which the features are obtained, is 15 week while in case of relative humidity estimation it is 45 week. Relative humidity estimation using maximum temperature, minimum temperature and rain as input parameters is better than taking features extracted from its own time series. The category of rainfall can be estimated using the features extracted from other parameters like maximum temperature, minimum temperature and relative humidity. It is possible to relate one weather parameter with other parameters. For example, category of rainfall can be estimated using features of Tmax, Tmin and RH. Also relative humidity can be predicted by using Tmax, Tmin and RF. These variables can be combined to extract the features that can be used to predict the future values of target variable. Technical and statistical indicators chosen are simple and capable of extracting the trends, which can be considered as features for developing the models. The model uses excel sheet data and simple processes to find the model parameters. Thus the method can be directly used by the medium educated farmers to forecast weather in their small geographical regions.

REFERENCES