Statistical Analysis and Forecasting Model for Monthly Mean Temperature in Colombo.

EDT Somarathne

Faculty of Agriculture, Rajarata University of Sri Lanka, Puliyankulama, Anuradhapura, Sri Lanka
deshani_s@yahoo.com, deshani_s@agri.rjt.ac.lk

Abstract- *Colombo is an industrial city in western province of Sri Lanka which has a high population density. Also, it is one of the cities which is warming faster due to industrial processes and lack of forestry. According to the Koppen climate classification, Colombo has a tropical rainforest climate and the city consists of a geography based on a mix of land and water. The climate of Colombo is fairly temperate all throughout the year and consists of intra-year seasonal and cyclic temperature patterns. The scope of this research is to develop a time based regression model to forecast monthly mean temperature in Colombo with respect to the most related weather component. In` this study, monthly mean temperature anomalies(in Celsius) of Colombo city for the period of 1986 to 2008 were analyzed and furthermore monthly mean temperatures for the year 2009 were forecasted using the estimated model. Data were collected from the weather station in Colombo attached to department of meteorology. An intra year (within year) seasonal pattern was observed in January, February, May-July, November and December. The weather components average wind (Km/h) and average humidity (%) indicated higher correlations to the monthly mean temperature than average rainfall (mm) and cloud (oktas).Stepwise regression was performed to find the most accurate model by using stationary set of series. The most accurate model consisted of average wind and 11 numbers of dummy variables generated due to the seasonality. The significance of each predictor and indicator for the final model were checked. The Gaussian correlated residuals were modeled using linear auto regressive model AR (2) derived from Box Jenkins methodology. The constant and the coefficients of final estimated model for forecasting monthly mean temperature were 27.1, 0.0893, 0.388, 1.04, 1.33, 1.47, 1.08, 0.838, 0.874, 0.812, and 0.374. Accordingly, the fitted regression model was accepted as the best model with R^2 of 58.5 %.*

Keywords: Koppen Climate Classification, Multiple Regression, Linear Auto Regressive model

I. INTRODUCTION

Human beings are often related with the nature or with their surroundings. They are often facing to the various changes of the environment. When considering these environmental changes, the climate and weather changes cannot be neglected as they are highly related with the activities and lives of the mankind. Though the terms climate and weather are seemed to be having the same meaning there is a considerable difference. Weather is the meteorological day-to-day variations of the atmosphere and their effects on life and human activity. It includes temperature, pressure, humidity, clouds, wind, precipitation and fog. Also, the climate is the weather averaged over a long period of time. Simply, Weather is the combination of events in the atmosphere and climate is the overall accumulated weather in a certain location.

A. Significance of the study

When weather is taken into consideration the concept of weather forecasting cannot be ignored .Weather forecasting is the application of science and technology to predict the state of the atmosphere for a future time and a given location [1]. Forecasts based on temperature and precipitation are important to agriculture and then to their productions, and also most of forecasts are used to protect life and property. Since outdoor activities are severely curtailed by wetness or dryness, sun shine, people extremely pay their attention to temperature forecasting. Temperature forecasts are used by utility companies to estimate demand over coming days. On an everyday basis, people use weather forecasts on temperature to determine what to wear on a given day. Forecasts can be used to plan activities around the people’s events and specially health and to plan ahead and survive them. Also, not only for the humans but also some kinds of animals are severely damaged by environmental temperature. Therefore temperature forecasting is also important to beasts. So in some countries there are special plans to save animals from the unsuitable temperature.

B. Background of the Study

Weather forecasting is mainly done by the Department of Meteorology in Sri Lanka and it manages 22 weather stations around the country. For this project monthly mean temperature was considered in degrees Celsius and average wind speed in km/h, average rainfall in mm (millimeters) average humidity as a percentage (%) and Cloud in oktas in Colombo area. Accordingly the relevant data from 1986 to 2009 were obtained from the weather station in Colombo for the statistical analysis in this research. Monthly mean temperature and the other factors were computed by taken the mean of the daily readings recorded in each month.

Time series arise in many applications such as in sociological statistics; in economic statistics; in meteorological statistics; in sales statistics; in statistics on defective products produced by a machine; in road traffic statistics; in measurements by psychologists on attention or body movement; in physiological measures of heart rate, respiration, and other bodily functions; in measures of river flow taken for flood-control planning; and in many other areas. A time series may trend upward or downward, as many economic series do, or may fluctuate around a steady mean, as human body temperature does. A series may contain a single cycle, or may contain several superimposed cycles. There are three major goals to be addressed in time series analysis. First, it is needed to forecast future values of a time series, using either previous values of that series or values from other series as well. Second, it is needed to assess the impact of a single event. Third, it is needed to study causal patterns, which mean the effects of variables rather than events on a series. In this research only first and third goals were addressed using the monthly mean values of each weather component. The methods proposed in this research can be used in other time series disciplines as well.

The study conducted to analyze the variations in monthly mean minimum/maximum temperatures in Antarctic region reveals the use of multiple regression models with non-Gaussian correlated errors [2] (Hughes, Raoy and Raoz, 2007).Also, the study conducted to develop time based auto regressive models uses the mathematic deviations of complex least squares in addition to real least squares [3](Gu and Jiang, 2005). The difference of this model from the conventional methods is, it consists of both real number and imaginary number. Accordingly, a better forecast model for monthly temperature were developed using complex least squares than conventional methods.

II. PRELIMINARY DATA ANALYSIS

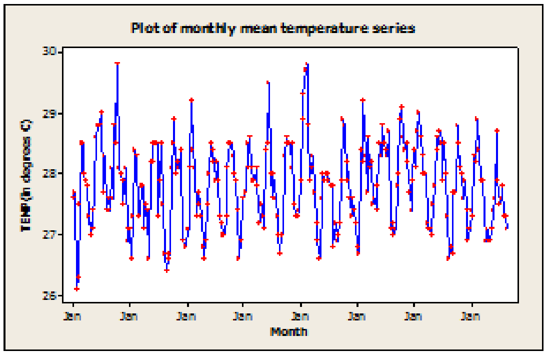

Figure 1: Plot of monthly mean temperature

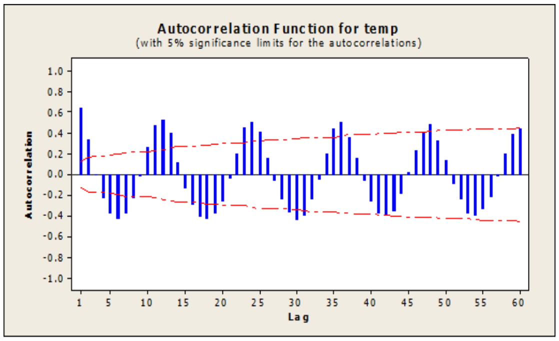

Figure 2: Seasonal patterns in monthly mean temperature series

Auto correlations of Monthly mean temperatures for months Jan, Feb, May, Jun, Jul, Nov and Dec behave in a similar pattern in each year. (Figure 2.2). This concludes that there is a within-year (intra year) seasonal pattern in monthly mean temperature data.

Table 1: Correlations of monthly mean temperature with other components

Variables	Auto correlation coefficient
Temp Vs Wind	0.112
Temp Vs Humidity	0.072
Temp Vs Rainfall	0.061
Temp Vs Cloud	0.045

Average wind and average humidity illustrate the highest correlations (.112 and .072) to monthly mean temperature. Therefore, those two components were considered for the regression procedure. (Table 2.1)

III. METHODOLOGY

A. Multiple regression models with time series data

Although multiple regression is obviously applied for non-correlated data, for some time now it is used for stationary time series data.

In this methodology,

Response Variable (Dependent variable): Monthly mean temperature in Celsius.

Predictor/s (Independent variable/s): Average wind (km/h), Average humidity (%).

According to the preliminary data analysis, it can be concluded that there is a within year seasonal pattern in monthly mean temperature as well as the other two series also consists of seasonality. So, dummy variables are used to remove the seasonality.

No of dummy variables = No of lags -1

Therefore in this case, there are 11 dummy variables due to seasonal pattern within year: $S_1, S_2, ...S_{11}$

Let Y_t be the response variable, monthly mean temperature,

W_{t-1} & H_t be predictors, average wind & average humidity,

$S_1, S_2, ...S_{11}$ be dummy variables.

$$Y_t = \alpha_0 + \alpha_1 W_{t-1} + \alpha_2 H_t + \beta_1 S_1 + \beta_2 S_2 + ... + \beta_{11} S_{11} + \varepsilon_t$$

Where Y_t =the variable to be forecast (Monthly mean temperature in Celsius)

t =the time index

W_{t-1} = Average wind (km/h)

H_t =Average humidity (%)

S_1=a dummy variable that is 1 for the second month of the year; 0 otherwise

S_2=a dummy variable that is 1 for the third month of the year; 0 otherwise

S_{11}=a dummy variable that is for the twelfth month of the year; 0 otherwise

ε_t =errors assumed to be independent and normally distributed with mean zero and constant variance.

$\alpha_0, \alpha_1, \alpha_2, \beta_0, \beta_1, \beta_2, ...\beta_{11}$ are the coefficients to be estimated.

The aforesaid coefficients are estimated by least square method. The least square estimates are denoted by $a_0, a_1, a_2, b_0, b_1, ...b_{11}$.

Therefore the estimated regression function is given by,

$$\hat{y} = a_0 + a_1 W_{t-1} + a_2 H_t + b_0 + b_1 S_1 + ...b_{11} S_{11}$$

Then the significance of each coefficient is tested.

Stepwise Regression

The Stepwise regression method adds one independent variable to the model one step at a time.

Steps:

1. All possible regressions are considered. Average wind which has the largest correlation with monthly mean temperature is inserted to the regression model.
2. The next variable average humidity which is the one which makes the largest significant contribution to the regression sum of squares is added to the model. The significance of the contribution is determined by F test.
3. Once an additional variable has been included in the equation, the individual contributions to the Regression sum of square of the other variables in the equation are checked for the significance using ***F*** statistic.
4. Steps 2 and 3 are repeated until all possible additions are no significant and all possible deletions are significant. At this point, the selection stops.

B. Test of significance of the coefficients

*1) **F-test***

H_0:
$$\alpha_o = \alpha_1 = \alpha_2 = \beta_0 = \beta_1 = \beta_2 = ...\beta_{11} = 0$$
vs. H_1: at least $\alpha_i \neq 0$ i=0, 1, 2 & $\beta_j \neq 0 : j = 0,1,...11$

tested by the F-Ratio, $F = \dfrac{MSR}{MSE}$ with *df = k, n-k-1*. At $\alpha\%$ significance level, the rejection region is, $F_{cal} > F_{n-k-1}^{k}(\alpha)$

*2) **Using p value.***

Here we compare the P value of each coefficient obtained from the analysis with α .
If P value > α then the H_0 is rejected.

*3) **Individual coefficient significance-T test***

T test is used to test the significance of individual coefficient.

H_0 : β_1=0 Vs H_1: $\beta_1 \neq 0$

$$\hat{\beta}_1 \approx N\left(\sigma^2 \Big/ \Sigma(x_i - \bar{x})^2 \right)$$

$$\frac{\hat{\beta}_1 - \beta_1}{\sigma \Big/ \sqrt{\Sigma(x_i - \bar{x})}} \sim N\,(0,\,1)$$

the estimator $\hat{\sigma}$ is used.

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma} \Big/ \sqrt{\Sigma(x_i - \bar{x})}} \sim t_{n-2,\alpha/2}$$ Where $\hat{\sigma}$ =Mean Square Error & n-2 =degrees of freedom

Test statistic T,

$$T = \frac{\hat{\beta}_1}{\hat{\sigma} \Big/ \sqrt{\Sigma(x_i - \bar{x})}}$$

At significance level α , the rejection region is,
T > $t_{n-2,\alpha/2}$
So, in this case

The coefficients $\alpha_0, \alpha_1, \alpha_2, \beta_0, \beta_1, \beta_2, ...\beta_{11}$ are tested individually for the significance using t test.

IV. RESULTS AND DISCUSSION

Table 2: Correlations-Temperature (in degrees C), AVERAGE HUMIDITY (%), and Average Wind

	Temp	***Avg Humidity***
Avg Humidity	*-0.048* *0.427*	
Avg Wind	*0.291* *0.000*	*-0.240* *0.000*
Cell Contents: Pearson correlation *P-Value*		

Average wind was the most correlated variable with monthly mean temperature and it was significant for the regression model. Average humidity was not significant for the regression model. Also, multicollinearity was observed between average wind and average humidity.

A. Multiple Regression Analysis-Stepwise Regression

According to the highest of accuracy measure R-sq the model *T =f (w, h, z)* was the best model (Table 4.2), but VIF of average wind was always larger than that of differenced average humidity and it has a negative regression coefficient every time because of the effect of multicollinearity.

Model	***R-sq value***	***F value***	***Model Significance***
T=f(w,z)	*58.5%*	*31.31*	*Significant*
T=f(w,h,z)	*65.3%*	*37.83*	*Significant*
T=f(h,z)	*64.7%*	*40.22*	*Not significant*
T – monthly mean temperature *W – average wind* *H – average humidity* *Z – dummy variables at seasonal lags*			

Table 3: Significance of models in stepwise regression

Furthermore, average humidity was removed from the final regression model though it increased the R^2 of the model since it was less related to monthly mean temperature than average wind and it presented multicollinearity within the model. The model with the second highest R-sq was rejected since it was not significant. Accordingly, the model T=*f (w,z)* was accepted as the best model with R^2 of 58.5 % .

B. Significance of the Coefficients

Dummy variables Z_{10} and Z_{11} were not significant for the final model.

$$TEMP = 27.1 + 0.0893\ Wind + 0.388\ z_1 + 1.04\ z_2 + 1.33\ z_3 + 1.47\ z_4 + 1.08\ z_5 + 0.838\ z_6 + 0.874\ z_7 + 0.812\ z_8 + 0.374\ z_9$$

C. *Residual Analysis*

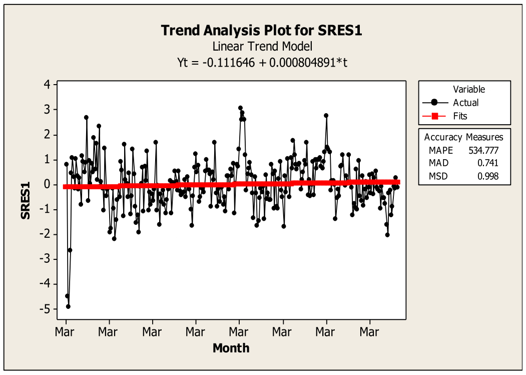

Figure 3: Plot of Trend Analysis of Residuals

Figure 5.1 illustrates that the series of the residuals is stationary.

D. *Test for correlation- Durbin Watson test*

Durbin-Watson statistic = 1.12354

H_0: No correlation Vs H_1: correlation exists
For n=200, dl=1.748 and du=1.789
Let dw=1.12354
If dl=1.748 and du=1.789
then 4-dl = 2.252 and 4-du=2.211
Accordingly, **0< dw < dl**
So, H_0 is rejected at 5% significance level and it can be concluded that there exists a positive serial correlation in residuals.
Therefore, the final conclusions on the standard residuals of the fitted regression were,

- Stationary and,
- Positively correlated.

Since the residuals are correlated they were modeled using Box Jenkins methodology.

E. *Box Jenkins Methodology*

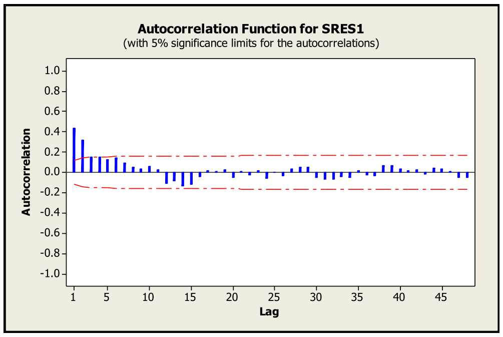

Figure 4: Correlations of the residual series.

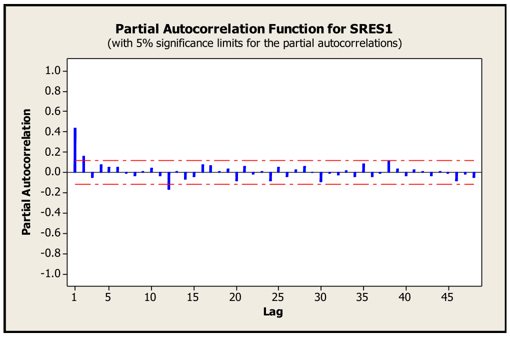

Figure 5: Partial correlations of the residual series.

i. *Model Identification*

Partial autocorrelation and autocorrelation plots are analyzed to identify the tentative models.

Autocorrelation	**Partial autocorrelation**	**Model**
Dies out	Cuts off at lag 2	ARIMA (2, 0, 0)
Cuts off at lag 2	Dies out	ARIMA (0, 0, 2)
Dies out	Dies out	ARIMA (2, 0, 2)

Table 5.1: Model Accuracy/Adequacy

Model	***Significance of coefficient***
ARIMA(2,0,0)	*Significant*
ARIMA(0,0,2)	*Significant*
ARIMA(2,0,2)	*Not Significant*

Accordingly, ARIMA (2, 0, 0) and ARIMA (0, 0, 2) models were eligible for the conclusion of final model (Table 5.1).

Model accuracy was compared using Akaike Information Criterion (AIC) and Bayesian information criterion (BIC) or Schwarz criterion.

Table 4: Checking for Model Accuracy

Criterion	*ARIMA(2,0,0)*	*ARIMA(0,0,2)*
AIC	*721.6518*	*724.8457*
SBC	*732.5021*	*735.696*

The lowest AIC and SBC indicated ARIMA (2, 0, 0) and that was the best fitted model for residuals (Table 5.2).

Therefore, the final model for the residuals is:

$Y'_t = 0.3596 Y'_{t-1} + 0.1726 Y'_{t-2} + \eta_t$

2) Significance of Coefficients

H_0 : $\phi_i = o$ Vs. H_1 : $\phi_i \neq 0$; i =1,2

Let $\phi_1 = 0.3596, \phi_2 = 0.1726$

Since P values of ϕ_1 , ϕ_2 < 0.05

H_0 is rejected at 5% significance level.

All coefficients are significant at 5 % significance level.

F. Development of final model

The final forecasting model were developed using the combination of regression model and the auto regressive model as follows,

$T_t = a_0 + a_1 W_t + a_2 Z_1 + a_3 Z_2 + a_4 Z_3\ a_5 Z_4 + a_6 Z_5 + a_7 Z_6 + a8 Z_7 + a_9 Z_8 + a_{10} Z_9 + (a_{12}\,\varepsilon_{t-1} + a_{13}\,\varepsilon_{t-2})$

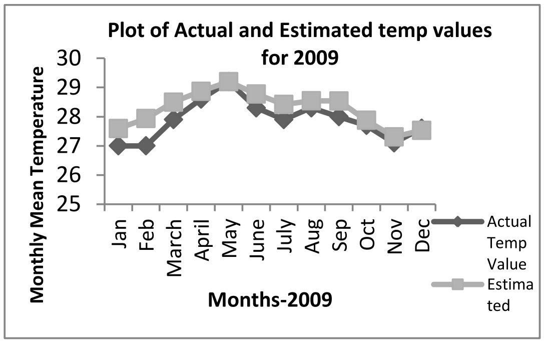

Figure 6: Plot of actual values vs. estimated values for 2009

Table 5: Coefficients of final model

a_0	*27.1*	a_7	*0.838*
a_1	*0.0893*	a_8	*0.874*
a_2	*0.388*	a_9	*0.812*
a_3	*1.04*	a_{10}	*0.374*
a_4	*1.33*	a_{12}	*0.3596*
a_5	*1.47*	a_{13}	*0.17265*
a_6	*1.08*		

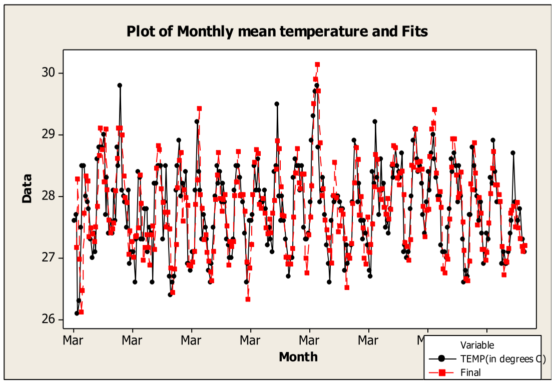

Figure 7: Plot of fitted values vs actual values of monthly mean temperature

Table 6: Forecasts for 2009 using the estimated model

Month	*Actual Temp Value*	*Estimated Temp Value*
Jan	*27*	*27.6*
Feb	*27*	*27.9*
March	*27.9*	*28.5*
April	*28.6*	*28.9*
May	*29.2*	*29.2*
June	*28*	*28.8*
July	*27.9*	*28.4*
Aug	*28.3*	*28.5*
Sep	*28*	*28.5*
Oct	*27.7*	*27.9*
Nov	*27.1*	*27.3*
Dec	*27.6*	*27.5*

V. MAJOR FINDINGS AND CONCLUSIONS

There is no any significant increase or decrease in monthly mean temperature during the period from 1986 to 2009 but seasonality exists in monthly mean temperature series. Also, a cyclic pattern is also predicted. Then the monthly mean temperature is highly related with average wind than average humidity. The accuracy R^2 of final regression model of monthly mean temperature with respect to the average wind and other seasonal indicators is 58.5%.

VI. RECOMMENDATIONS FOR FURTHER INVESTIGATION

Investigation of other influences (Other weather components such as rainfall, humidity and clouds) towards temperature changing would be helpful to develop a more accurate model. Use of daily, hourly or half hourly values rather than the use of mean values might develop a more accurate and efficient model.

ACKNOWLEDGMENT

I wish to thank department of meteorology, 383, Bauddhaloka Mawatha, Colombo 07 for assisting me by providing the monthly mean weather data from 1986 to 2009.

REFERENCES

Gillian L. Hughes, Suhasini Subba Raoy and Tata Subba Raoz, 2007, "Statistical analysis and time series models for minimum/maximum temperatures in the Antarctic Peninsula.", Volume: 463, Proceeding of the Royal Society[2]

X. Gu and J. Jiang, 2005, "A complex autoregressive model and application to monthly temperature forecasts", Volume 23, Annales Geophysicae[3]

Christoph C. Raible, Georg Bischof, Klaus Fraedrich, and Edilbert Kirk, "Statistical Single-Station short term forecasting of Temperature and probability of precipitation: Area interpolation and NWP Combination", 1999, Vol. 14 Issue 2, Weather and Forecasting

Peter J. Brockwell, Richard A. Davis, 2009, "Time Series: Theory and Methods", Springer, pg 239-258

Harry Frank, Steven C. Althoen, 1995,"Statistics- Concepts & Applications", Cambridge University Press, pg 570-586.

http://en.wikipedia.org/wiki/Weather_forecasting[1]

BIOGRAPHY OF AUTHOR

Ms. E.D.T. Somarathne is currently an instructor in computer technology attached to the faculty of agriculture, Rajarata University of Sri Lanka, Sri Lanka. She has successfully completed BSc (Physical Science) at University of Peradeniya with Computer Science and Mathematics as main subjects. She is currently expecting to continue her postgraduate studies after completion of the course work of in MSc in Applied Statistics at PGIS, University of Peradeniya .Her research interests are time series based data mining and forecasting techniques. This is her first attempt to submit and publish her own research paper to an international conference and in the meantime she has submitted an extend abstract to the international conference at Rajarata university of Sri Lanka. The author involves in statistical forecasting since 2007.