STATISTICAL MODEL TO PREDICT CONTOUR ELEVATION USING SHUTTLE RADAR TOPOGRAPHIC MISSION DIGITAL ELEVATION DATA

Reyalt Gnanapragasam,
Department of Mathematics and Computer Science, The Open University of Sri Lanka
Email: reyalt@ou.ac.lk
Kanthi Perera,
Department of Engineering Mathematics, University of Peradeniya
Email: kanthip@pdn.ac.lk
Uditha Ratnayake,
Department of Civil Engineering, University of Peradeniya
Email: udithar@pdn.ac.lk

Abstract

This study is based on two geographical datasets, namely, Shuttle Radar Topographic Mission (SRTM) elevation and contour elevation. The SRTM data is available for all locations in Sri Lanka. It follows the shape of the actual ground but not the actual elevation of the surface. This is due to errors introduced during processing. The contour data is obtained from the actual ground survey data of contour maps and using ArcGIS software. The survey data is more reliable, but more expensive. Since it does not contain ground level variations, Sri Lanka does not have contour elevations at all locations. Measuring contour elevation for all locations is a costly procedure. Therefore, finding a method to evaluate the approximated value of contour elevation with a less costly method is essential. Thus the objective of this study is to find a statistical model to predict the contour elevation based on SRTM data. Both types of data are available only for four locations: Paddhiruppu, Kegalle, Badulla, and Katharagama, in Sri Lanka. According to the geography of Sri Lanka, three clusters are distinguishable by elevation. These are the Central Highlands, the Plains, and the Coastal belt. Since the data used in this study are for four different locations and these locations fall into three different clusters, three regression models are fitted for each cluster and the models are validated. Multiple, linear regression analysis is used to fit the models. The t- test is used to test the significance of parameters while the F- test is used to test the significance of the overall model. Residual analysis is carried out to test the normality, homoscedasticity and auto correlation of the residuals. The goodness of the fitted model is evaluated by the coefficient of determination $\left(R^2\right)$. Approximately 99% of the variation is explained by the fitted models and 82% by the validated model. Thus, if the SRTM data value is known, by choosing the appropriate model based on its cluster, the approximated contour elevation could be predicted.

Key words: SRTM, Contour, Elevation, Regression, Cluster

1. Introduction

Shuttle Radar Topography Mission (SRTM) and contour elevation data are used to evaluate their difference in elevations. Both types of elevation data are available only for four locations, which are Paddhiruppu, Kegalle, Badulla, and Katharagama, in Sri Lanka. SRTM data are released by National Aeronautics and Space Administration (2004). SRTM is the mission to map the world. Objective of this mission is to obtain RADAR data of most of the Earth's land surface to produce high-resolution topographic maps using satellites to measure the elevation. The SRTM data has more elevation data, and it is available for all locations in Sri Lanka. It follows the shape of the actual ground level, but not the actual elevation of the surface.

The contour data is obtained from 1: 50,000 maps of Survey Department created based on the actual ground survey data and therefore it is more reliable, but more expensive. A disadvantage is that it does not contain information on ground level variations at all locations. Measuring elevation for all the locations in Sri Lanka is a costly procedure. Therefore, finding a method to evaluate the approximated value of contour elevation for all locations in Sri Lanka with a less costly method is essential. This study considers available contour elevation data corresponding to the locations; Paddhiruppu, Kegalle, Badulla, and Katharagama and used together with SRTM elevation to produce a correct contour elevation data for all locations. Thus the objective of this study is to develop regression models for three clusters, which are distinguishable by elevation. Elevation of any location in Sri Lanka can be measured by using these models with less cost than getting the contour elevation data.

Dadson (1999) explained that the elevation of the contour is based on actual measurements by surveying, but the values between any two contours are generally obtained by interpolation. That is, the elevations between those two contours are estimated. Nowadays and also in this study this interpolation is performed using the ArcGIS software.

2. Materials and Methods

The SRTM and contour elevation data used in this study are for the four different locations and these locations fall into three different clusters. These data are secondary data, which are arranged in matrix form as rows and columns. The value in a cell (90m cell size) represents the elevation of the 90m×90m squared area.

Multiple regression analysis is used to fit the models. The t- test is used to test the significance of parameters while the F- test is used to test the significance of the overall model. Throughout the analysis a significance level $\alpha = 0.05$ is used. The goodness of the fitted models are evaluated by the coefficient of determination $\left(R^2\right)$, which if significantly closer to 1 indicates a good fit for the data. When building the regression models some assumptions on the residuals are made. It is necessary to check whether the assumptions are satisfied by the fitted models. Three standard tests are applied to detect the normality of residuals. They are Kolmogorov-Smirnov test, Shapiro-Wilk test and Anderson-Darling test. Another key assumption of the ordinary regression model is that the residuals have constant variance, or homoscedasticity. Whites General Test is used to test the homoscedasticity.

The efficiency of Ordinary Least-Squares (OLS) parameter estimates is adversely affected when the error terms are auto correlated and the standard error estimates would be biased. The Durbin-Watson (DW) statistic is used to test for the presence of first-order auto correlation in the residuals. The DW closer to 2 reveals that the residuals are uncorrelated. The SAS AUTOREG procedure is used to correct for auto correlation of the residuals. The AUTOREG procedure solved this problem by augmenting the regression model with an autoregressive model for the random error, thereby accounting for the autocorrelation of the errors. The Akaike Information Criteria (AIC) and the Mean Square Error (MSE) are used to select the best model. The best model is the one which gives the lowest AIC and MSE values.

Extensive faulting and erosion over time have produced a wide range of topographic features. According to the Geography of Sri Lanka (2010), three zones are distinguishable by elevation, which are the Central Highlands, the plains, and the coastal belt. Most of the island's surface consists of plains between 30 and 200 meters above sea level. A coastal belt about thirty meters above sea level surrounds the island. The elevation of the coastal belt is less than 30 meters, plains between 30 and 200 meters above sea level. The third cluster is the central highlands, elevation is over 200 meters. For each cluster models are developed so that the correct contour elevation could be predicted using SRTM data.

3. Results and Discussions

Hereafter throughout this paper dataset 51 will refer to Paddhiruppu dataset, dataset 53 will refer to Kegalle dataset, dataset 69 will refer to Badulla dataset and dataset 83 will refer to Katharagama dataset. Those four locations are highlighted in Figure 1 map of Sri Lanka. The map was released by the Department of Field Support (2008).

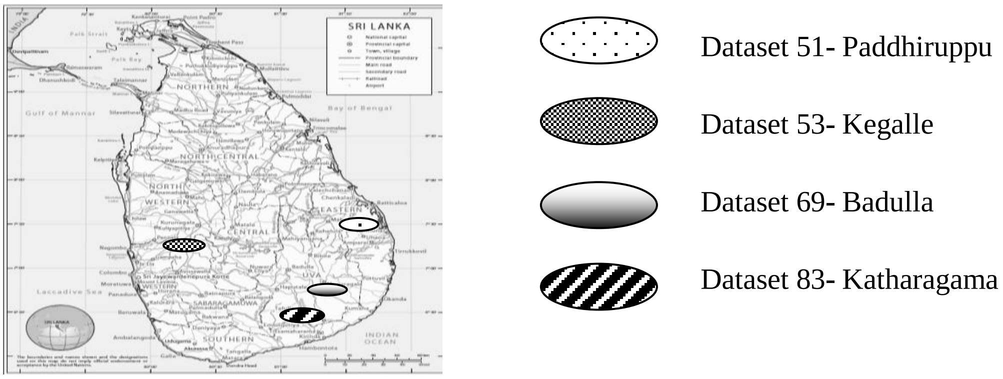

Figure 1: Map of Sri Lanka

3.1 Identifying the clusters for the datasets

Geography of Sri Lanka (2010) contains the clusters of area of Sri Lanka. Since the mean of SRTM data of the dataset 51 is 29.88m, which is less than 30m, it belongs to the Coastal Belt cluster. Also the mean of SRTM data of the datasets 53 and 83 are 106.54 and 63.30 meters respectively, which are in between 30m and 200m. Therefore those two datasets belong to the Plain cluster. But the mean of SRTM data of the dataset 69 is 1279.10m, which is greater than 200m. Therefore, this dataset belongs to the Central Highlands.

3.2 Development of the Models

Since the aim of this study is to predict contour elevation using SRTM elevation, the contour elevation was selected as the response variable and SRTM elevation was selected as the regressor variable when fitting the multiple regression models. Listed below are the three regression models considered for all three clusters in Sri Lanka:

MODEL1: $contour = \beta_0 + \beta_1 * srtm$

MODEL2: $contour = \beta_0 + \beta_1 * srtm + \beta_2 * srtm^2$

MODEL3: $contour = \beta_0 + \beta_1 * srtm + \beta_2 * srtm^2 + \beta_3 * srtm^3$

Table 1: Test results using REG procedure for the dataset 51

Model	*Durbin- Watson* *(Test Statistic)*	*Significance of the model* *F-Test (p-value)*	R^2 *- value*
MODEL1	*0.2704*	*<0.0001*	*0.91*
MODEL2	*0.2813*	*<0.0001*	*0.91*
MODEL3	*0.2829*	*<0.0001*	*0.91*

REG procedure is used for the Ordinary Least Square Estimation. In the Table 1, p - values of F-Test for all the models are less than 0.05. Therefore the models are significant at 5% significance level. But the Durbin Watson test statistics for three considered models are not closer to 2 indicating that the auto correlation exists among the residuals. To remove the serial correlation among residuals AUTOREG procedure is used and the results are reported in the Table 2.

Table 2: Test results using AUTOREG procedure for the dataset 51

Model	*No. of Lags*	R^2 *- value*	DW *- value*	*Significance of the Parameters (p-value)*				*Significance of the model (p-value)*
				β_0	β_1	β_2	β_3	
MODEL1	***2***	***0.99***	***1.98***	*<0.0001*	*<0.0001*			*<.0001*
	3	*0.98*	*2.01*	*<0.0001*	*<0.0001*			*<.0001*
MODEL2	***2***	***0.98***	***1.92***	*<0.0001*	*<0.0001*	*<0.0001*		*<.0001*

	3	*0.98*	*1.96*	*<0.0001*	*<0.0001*	*<0.0001*		*<.0001*
MODEL3	***2***	***0.98***	***1.91***	*<0.0001*	*<0.0001*	*<0.0001*	*<0.0001*	*<.0001*
	3	*0.98*	*1.94*	*<0.0001*	*<0.0001*	*<0.0001*	*<0.0001*	*<.0001*

As seen in the Table 2, p- values of the F-test for three considered models are less than 0.05. Therefore it can be concluded that all three models are significant at 5% significance level.

Also the p-values of the t-test of three considered models are less than 0.05, which verified that all the parameters of three models are significant at 5% significance level. Even though the R^2 - values in all the models are very high, the models with less number of lags (No. of Lag is 2) and the DW value closer to 2 are selected for further analysis.

Table 3: Test results for the selected models of the dataset 51

Model	*Normality Test* *(p- values)*			*Homoscedasticity* *Test(p- value)*	*AIC*	*MSE*
	Kolmogorov Smirnov	*Shapiro Wilk*	*Anderson Darling*	*Whites General Test*		
MODEL1	*0.2632*	*0.1577*	*0.1416*	*0.1592*	*494076.33*	*16.85*
MODEL2	*0.0312*	*<0.0050*	*0.0274*	*0.1921*	*494788.64*	*16.98*
MODEL3	*0.2571*	*0.0356*	*0.1675*	*0.2421*	*497185.74*	*17.46*

According to the Table 3, the p-values of the Whites General Test for three selected models are greater than 0.05. Therefore it can be concluded that the underlined assumption that the constant variance of residuals is satisfied at 5% significance level. Since the p- values of all three normality tests of MODEL2 and the p- value of Shapiro- Wilk test of MODEL3 are less than 0.05, the normality assumption of the residuals is not satisfied by MODEL2 and MODEL3. Thus MODEL2 and MODEL3 are not considered. Also MODEL1 has the least AIC and MSE values. Therefore MODEL1 is selected as the best model for the Coastal Belt cluster (dataset 51) and the fitted model is:

$$contour\,51 = 7.5690 + 0.8403 * (srtm51)$$

The same procedure and techniques are used for the other two datasets 69 and 83 and the best fitted models for other two clusters are given below:

The model for the Central Highlands cluster is:

$$contour\,69 = 617.4692 + 0.00009686 * (srtm69) + 0.00361 * (srtm69)^2$$

The model for the Plains cluster is:

$$contour\,83 = 47.9823 + 0.4459 * (srtm83)$$

3.3 Model Validation

The dataset 53 is used for the validation purpose. Since datasets 83 and 53 belongs to Plain cluster, the selected model for dataset 83 is used to predict the contour elevation data of the dataset 53.

3.3.1 Geometric Interpretation

Since the dataset has 117,694 data points, it is very difficult to view the plot which contains all the data points. Even though the line graph is plotted partially (10,000 points in a graph), still the plot is not clear. Note that in the following graph series1 represents the original contour values (contour values in the dataset 53) and series2 represents the predicted values (from the model of dataset 83).

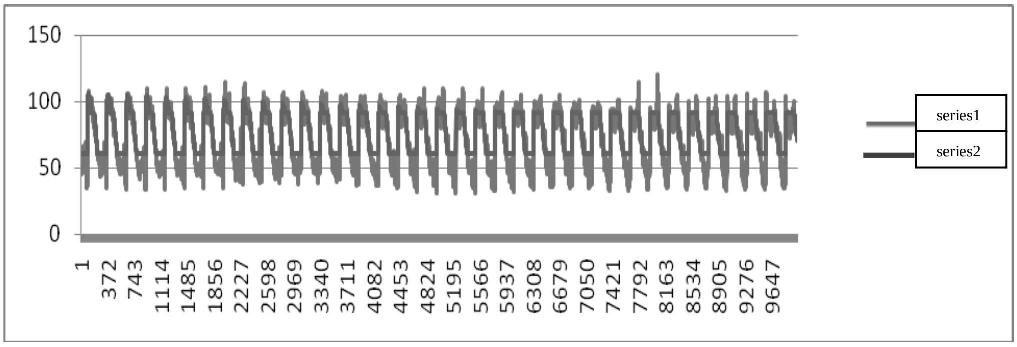

Figure 2: Partially Plotted Line Graph of Predicted and Contour data

In the Figure 2, it is difficult to judge that whether predicted value is bigger or smaller than original contour values. Therefore, when the data set is large, to get a better idea the algebraic method is more appropriate.

3.3.2 Algebraic Interpretation

The following results are obtained to calculate the R^2 value:

The total number of data points in the dataset $n = 117694$

Mean of the observed value (mean of contour data) $\overline{Y} = 104.109$

Then the sum of squares of residuals, $\sum \varepsilon^2 = 275375527.74$

The sum of squares of total, $\sum Y^2 = 2802656610.047$

Thus $R^2 = 1 - \dfrac{\sum \varepsilon^2}{\sum Y^2 - n\overline{Y}^2} = 0.8197$

It appeared that, approximately 82% of the variation can be explained by the fitted model.

4. Conclusion

Based on the Geography of Sri Lanka (2010), the dataset 51 (Paddhiruppu) belongs to the Coastal Belt cluster, the dataset 69 (Badulla) belongs to the Central Highlands cluster and the dataset 83 (Katharagama) belongs to the Plains cluster. Thus the fitted models to predict contour elevation using SRTM elevation for the three clusters are given below:

The model for the Coastal Belt cluster is

$$contour = 7.5690 + 0.8403 * srtm \qquad R^2 = 0.99$$

The model for the Central Highlands cluster is

$$contour\,69 = 617.4692 + 0.00009686 * (srtm69) + 0.00361 * (srtm69)^2 \qquad R^2 = 0.99$$

The model for the Plains cluster is

$$contour\,83 = 47.9823 + 0.4459 * (srtm83) \qquad R^2 = 0.99$$

It concludes that, if SRTM data value is known (any location in Sri Lanka) by choosing the appropriate model based on its cluster, the approximated contour elevation data value can be predicted.

References

Dadson, S. (1999). Digital Terrain Modelling, *Geog 516*, (available online http://www.geog.ubc.ca/cources/geos516/notes/dtm/html [Accessed on 29/05/2010])

Department of Field Support, Cartographic section (March 2008)*; Map No 4172 Rev. 3*, *UNITED NATIONS*, (available online http://www.un.org/Depts/Cartographic/map/profile/srilanka.pdf [Accessed on 20/05/2012])

Geography of Sri Lanka (November 2010), (available online http://en.wikipedia.org/wiki/Geography_of_Sri_Lanka [Accessed on 13/08/2011])

National Aeronautics and Space Administration (2004), *Jet Propulsion Laboratory*, *Californiya Institute of Technology*, (available online http://www2.jpl.nasa.gov/srtm/index.html [Accessed on 28/08/2010])

Shuttle Radar Topography Mission (July 2009), (available online http://en.wikipedia.org/wiki/Shuttle_Radar_Topography_Mission [Accessed on 07/08/2010])