Advances in Computer Science and Information Technology (ACSIT)
p-ISSN: 2393-9907; e-ISSN: 2393-9915; Volume 4, Issue 1; January-March, 2017, pp. 25-32
© Krishi Sanskriti Publications
http://www.krishisanskriti.org/Publication.html

Prediction of Seasonal Precipitation using Ordinary Kriging and Regression Kriging Techniques: Case Study: Jalisco, Mexico

Soham Mukherjee

Geoinformatics Department, Indian Institute of Remote Sensing, ISRO
E-mail: saion523@gmail.com

Abstract—*This paper presents how spatial interpolation techniques can be used to predict precipitation amount of a certain region and the variation of precipitation of that area in different seasons throughout the year: ordinary kriging and regression kriging with different variogram fitting models are used and assessed in the study. The techniques have been illustrated using monthly precipitation data of sample points along with the elevation of that point inside a region (106oW; -101oW; 18oN; 23oN) encompassing an area of approximately 20,000 km2 that covers the state of Jalisco, northwest Mexico. Cross-validation has been used to evaluate the prediction performances of the kriging method.*

Regression kriging is performed and also compared in this study with ordinary kriging. Elevation of the sample point is the dataset is used as the covariate in the prediction. Regression kriging also shows whether precipitation has any significant relation with elevation in terms of prediction. Moreover, this work also shows the usability of kriging technique in terms of detecting a trend in the data.

The results obtained for prediction errors have been low and as desired. The RMSE (Root Mean Square Error) for ordinary kriging consisting of 101 samples has been 0.01116; while the mean error is 0.0003 and for regression kriging, RMSE is 0.02193 and the mean error is 0.0021. Kriging produces a good prediction of precipitation at unsampled locations. Methods of data analysis, like skewness removal of data, are also discussed in the paper. The analysis of the data, variogram model fitting and generation of prediction map through kriging are performed in R software.

1. INTRODUCTION

In meteorology, precipitation can be defined as the yield of condensed water vapour in the atmosphere that falls to ground due to gravity (Wilson, 1983). Precipitation can be of various forms. Major ones among them are drizzle, rain, sleet, snow and hail. Total amount of precipitation in an area plays a vital role in vegetation growth in that area. Agricultural agencies can use the precipitation information to predict the vegetation/agricultural growth of a certain area. Thus it is necessary to predict the precipitation amount in different seasons in a year of a certain place. It gives the quantitative overview of precipitation throughout the year which can be helpful in predicting suitable season for specific agricultural activities. Spatial interpolation techniques can be used to predict the precipitation through the year in un-sampled locations. These techniques are applied when ground data has not been collected all over the region. The statistical interpolation techniques merged with spatial properties create a robust method for predicting at un-sampled locations. In this study, ordinary kriging and regression kriging techniques are used to interpolate precipitation in un-sampled locations. For better kriging results different experiments are carried out to find the optimal variogram model for the chosen data-set.

2. OBJECTIVE

The following are the objectives of the work:

a) To determine the best variogram model that has the best fit for data.

b) Use of ordinary kriging (OK) to interpolate amount of precipitation (mm) in un-sampled locations.

c) Use of regression kriging (RK) by incorporating a suitable covariate in the model (elevation of observation).

d) Comparison of accuracy obtained from OK and RK to select the optimal one.

3. DATA

Precipitation data of a square (106^{o}W; -101^{o}W; 18^{o}N; 23^{o}N) encompassing an area of approximately 20,000 km2 that covers the state of Jalisco, northwest Mexico, was taken into consideration for the interpolation approach. The data consists of monthly precipitation data of sample points along with the elevation of that point. The following figure on the left side (Figure1) shows the location of Jalisco with respect Mexico and observed precipitation locations are shown in following figure (Figure 2).

The data set has total 101 observations. The original data is skewed. So, a logarithmic transformation is performed on the data to remove skewness (shown in Figures 3a, 3b, 4a and 4b for the month of January as example).

Figure 1: (Jalisco's Location Map)

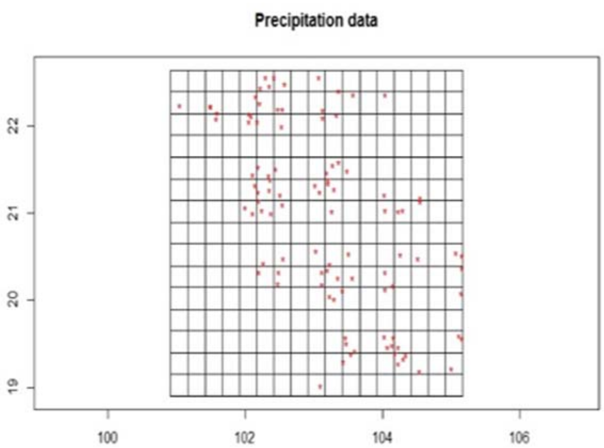

Figure 2: (Precipitation Map of Jalisco)

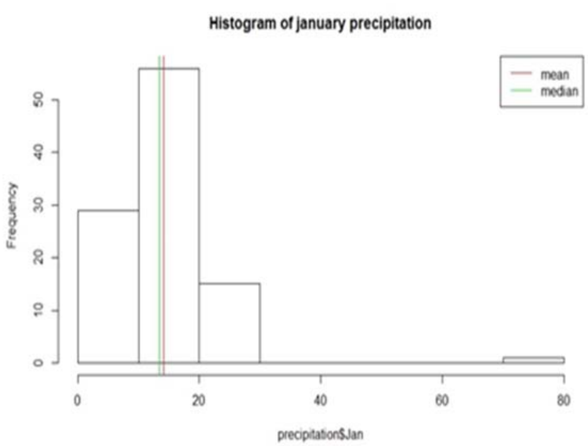

Figure 3a: Histogram of skewed Data

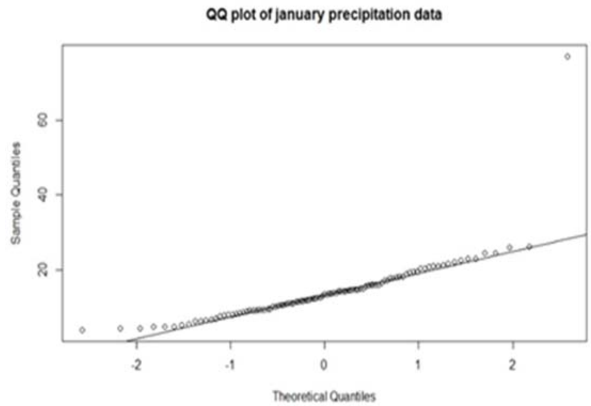

Figure 3b: Q-Q plot of skewed Data

In the transformed dataset, it can be seen that the dataset is now moreover normally distributed and closer to the Gaussian distribution that the previous one. The range of the data has also been reduced due to logarithmic transformation (shown in Figures 5a and 5b). So for further work, log-transformed data is used.

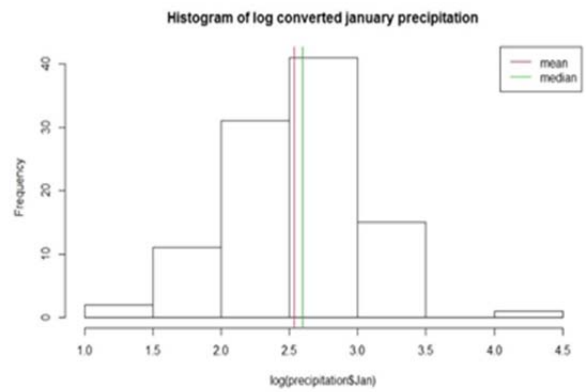

Figure 4a: Histogram of Transformed Data

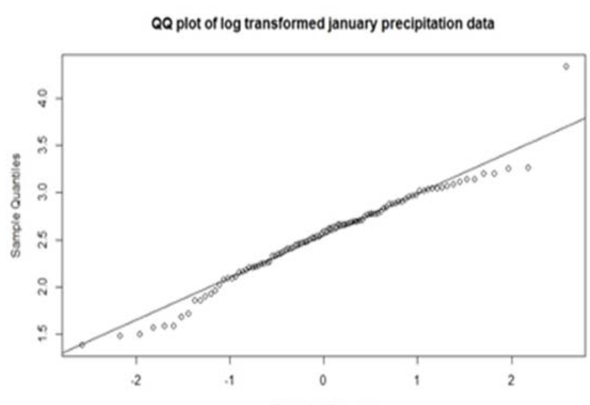

Figure 4b: Transformed (Square-Root) Data

The data set used (Figure 5) has a. minimum value of 4.0 and maximum value of 77.0, the median and mean are 14.16 and 13.4 respectively. After the logarithmic transformation (Figure 6.), the minimum value, maximum value, median and mean are 1.386, 4.344, 2.595 and 2.534 respectively.

```
> summary(precipitation$Jan)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   4.00    9.40   13.40   14.16   17.20   77.00
```

Figure 5: shows summary of original data

```
> summary(log(precipitation$Jan))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.386   2.241   2.595   2.534   2.845   4.344
```

Figure 6: shows summary of transformed data

4. METHODS

a) Variogram:

Variogram is used in cases where covariance function cannot be implemented and mean is not constant over the data. Variogram observes the variance of points in pairs of the observation. This variance is named as semi variance. Semi variance does not depend on the actual point locations but the distance (lag) between the points in a pair (Webster & Oliver, 2008). A low sum squared error (SSErr) denotes less uncertainty in the data. To select the best fitted model in the variogram of the data is selected based on the nugget value (Van Groenigen, 2000). A small nugget value represents low random error in the data. In ideal cases, the nugget value is presumed to be zero (Western & Blöschl, 1999).

Based on the best fit model parameters of variogram, kriging is used on the data-sets.

b) Ordinary Kriging:

Ordinary kriging of a single variable is a robust and common geostatistical method and used mostly to predict the variables at unknown location (Webster & Oliver, 2008). It is based on the assumption of unknown mean and is said to be unbiased (Webster & Oliver, 2008). Ordinary kriging deals with single variable of the observation. For this study precipitation amounts at sampled locations are used to predict precipitation amount at un-sampled locations. The prediction weights associated to points do vary from point to point; it takes spatial properties of the data into account, e.g. distance between points (Diggle & Ribeiro, 2007). Kriged prediction/estimation and associated kriging variance were obtained. RMSE (Root Mean Square Error) and ME (Mean Error) were calculated to obtain the uncertainty measure.

c) Regression Kriging:

Regression kriging is performed when multiple numbers of spatial explanatory variables are observed and to identify which of the variable effects the mean response (Diggle & Ribeiro, 2007). In regression, kriging mean is observed to be non-constant over the data. The advantage of regression kriging is its ability to include the method of regression technique and to allow separate interpretation of kriged components (Hengl, Heuvelink, & Rossiter, 2007). In the study, elevation of the observation point of precipitation is considered as a covariate and included in regression kriging to study its effect on the prediction of precipitation.

5. RESULTS AND ANALYSIS

Table1: Parameter selection for cutoff and bin width in exponential (default) model

Model	Cut-off	Width	Minimum no.ofpairs	Nugget	Psill	Range	SSErr	ME	RMSE
Exponential	default	default	39	0.000	0.208	0.134	8.07	0.0074	0.08643
Exponential	2	.10	36	0.000	0.202	0.123	6.33	0.0068	0.08276
Exponential	3	.10	36	0.000	0.205	0.127	7.34	0.0070	0.08370
Exponential	2	.15	71	0.146	0.043	0.233	5.87	0.0002	0.01498
Exponential	3	.15	71	0.143	0.051	0.244	7.15	0.0003	0.01822

a) Variogram and model selection:

Multiple combinations of cutoff and bin width are tested to choose the optimal variogram. The results are tabulated above (Table 1). The highlighted parameters show the most preferred variogram. The results differ from the default variogram in terms of variogram parameters such as no. of point pairs and variogram fitting parameters such as nugget, sill and range, and also SSErr and RMSE values. From table 1, it can be clearly seen that for a bin width of 0.10 with varying cutoff (2-3) produces a large amount of SSErr (sum of square error) and a smaller number of pair points in certain class distance. It refers to the fact that the model is poorly fitted in the variogram. Hence, these combinations were discarded first. Bin width equals to. 15 seems to fit the variogram model better. Sample variogram is also sensitive to cut off. With increasing cutoff, greater the spread of the data is covered and less correlated points come into consideration. Considering these facts, variogram with cutoff 2 and bin width. 15 seems better to be fitted by the model and there is also sufficient number of point pairs. The various combinations of bin width and cut-off are shown in Figure 7a, 7b, 7c and 7d for the data of month of January as example one.

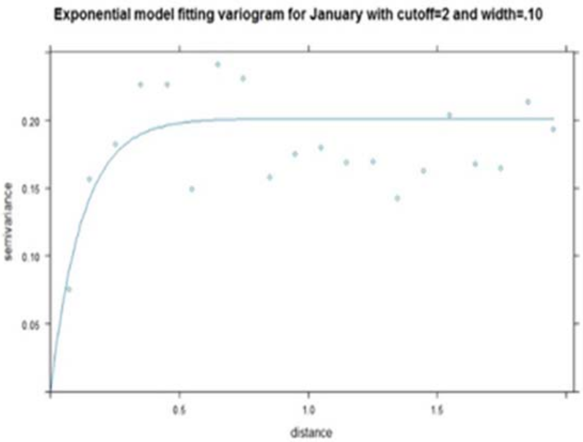

Figure 7a: variogram with cutoff=2, width=.10

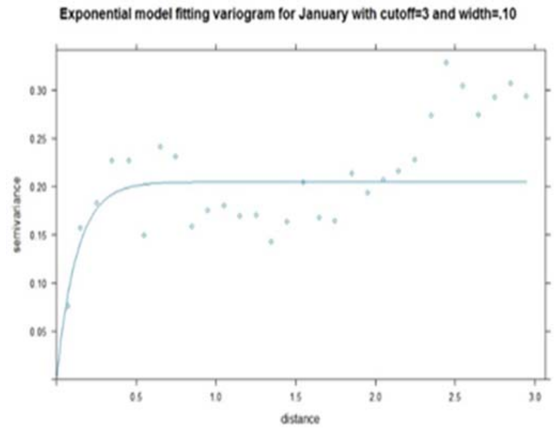

Figure 7b: variogram with cutoff=3, width=.10

Table 2: Optimal model selection to fit variogram

Model	Cutoff	Width	npairs	Nugget	Psill	Range	SSErr	ME	RMSE
Exponential	2	.15	71	0.146	0.043	0.233 Effective:0.699	5.87	0.0002	0.01498
Spherical	2	.15	71	0.143	0.042	0.316	6.24	0.0001	0.01116

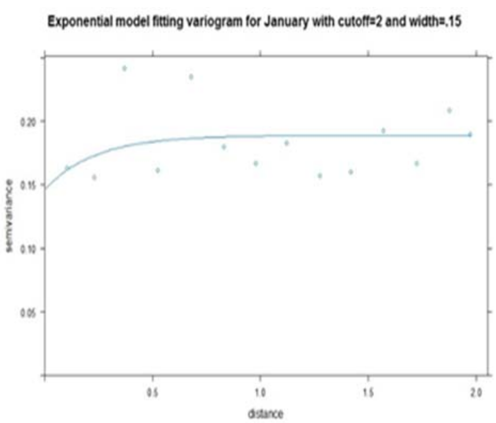

Figure 7c: variogram with cutoff=2, width=.15

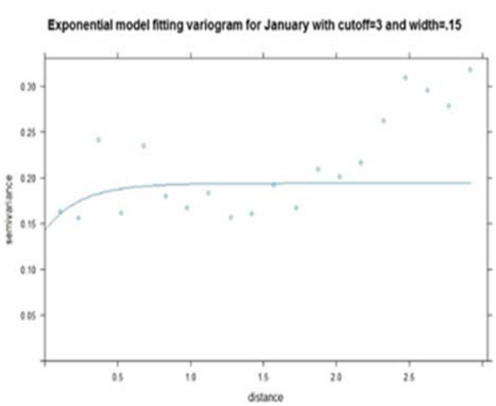

Figure 7d: variogram with cutoff=3, width=.15

With the optimal cutoff and bin width obtained from table 1, exponential and spherical model is tested to obtain a better fit. The results are tabulated in table 2 above. The highlighted parameters show the better fit. The results differ in terms of variogram fitting parameter such as nugget, sill and range, and also SSErr and RMSE values. From the table 2, keeping the cutoff = 2 and the width =. 15, The root mean square error (RMSE) is lowest for Spherical model suggesting that the accuracy of prediction is higher and intra-cluster variation is minimal. Moreover, Spherical model gives the lower nugget value which means the data are correlated and less random error is in the data. Higher nugget values lead to discontinuity and spikes representing higher variability of the non-spatial component. Since spherical model has given a better fit for spatially correlated dataset (See Figure 8a and 8b) than exponential model, it is considered to be the optimal solution for fitting the variogram and is considered for further analysis.

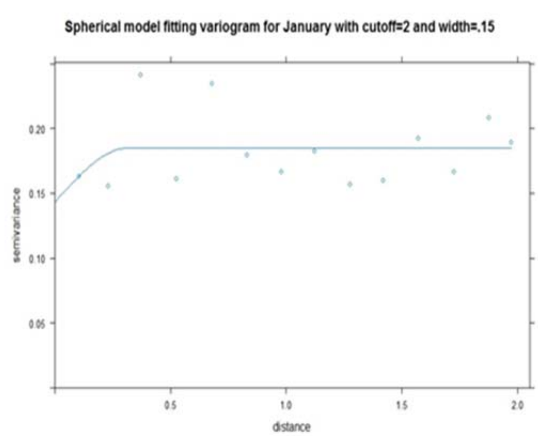

Figure 8a: Spherical model fit variogram

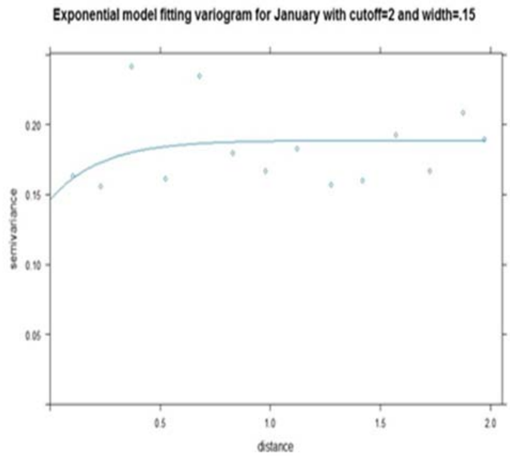

Figure 8b: Exponential model fit variogram

b) Precipitation prediction:

The dataset consists of precipitation data of twelve months in a calendar year. But in the study prediction has been performed on the basis of seasonality. Three different months across different seasons in a calendar year are taken into consideration and prediction of precipitation in un-sampled locations is performed. Here the three selected months are January, July and November. The prediction results are presented in the following figures.

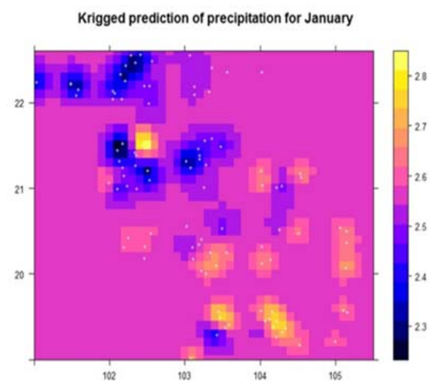

Figure 9a: Kriged prediction of January

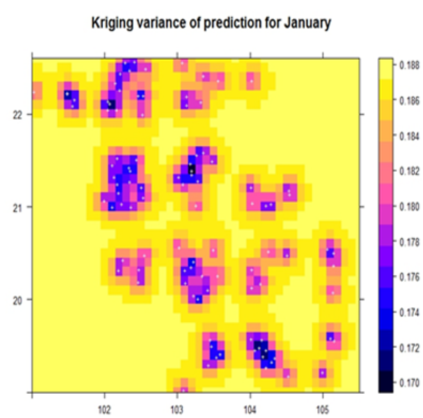

Figure 9b: Kriging variance of January

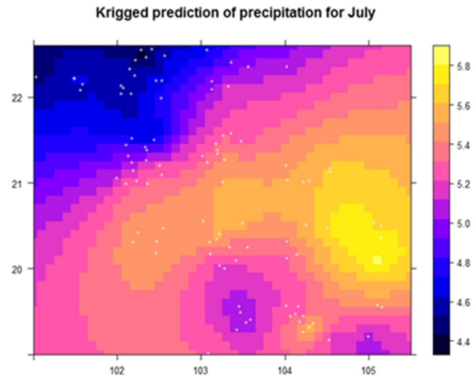

Figure 10a: Kriged prediction of July

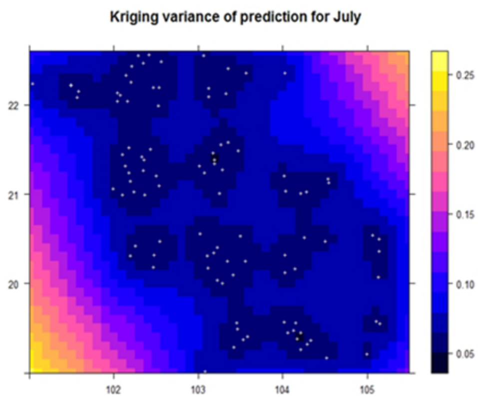

Figure 10b: Kriging variance of July

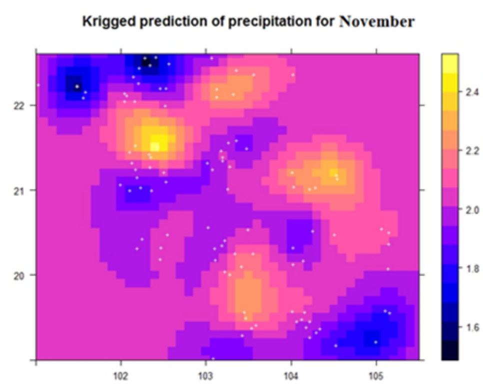

Figure 11a: Kriged prediction of November

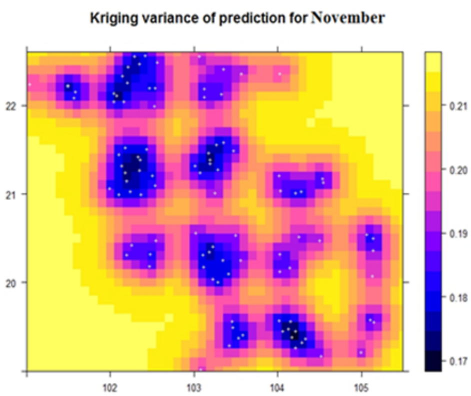

Figure 11b: Kriging variance of November

In the precipitation prediction maps (Figure 8a, 9a and 10a), the areas shown in orange and yellow are the areas with higher amount of predicted precipitation and the areas shown in blue are the areas with lower amount of precipitation. In the precipitation prediction variance maps (Figure 8b, 9b and

10b), yellow coloured areas are the areas with high kriging variance and the areas with blue colour represent areas with low kriging variance. The results show that the locations that are far from the observation points do have a higher kriging variance than the closer location to observation points.

The dataset also consists of elevation data of the observation points. Tests are performed to examine if this variable can be used as a co-variate in prediction. Regression is performed between precipitation amount data and corresponding elevation of the point to check the relationship between these two variables. The results of the month January as example are mentioned in the following figures (Figure 12a and 12b).

```
> summary(jan.lm)

Call:
lm(formula = log(Jan) ~ Elevation, data = precipitation)

Residuals:
     Min       1Q   Median       3Q      Max
-0.96386 -0.27150  0.03878  0.22354  1.89371

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.077e+00  1.085e-01  28.367  < 2e-16 ***
Elevation   -3.679e-04  6.773e-05  -5.431 4.01e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4205 on 99 degrees of freedom
Multiple R-squared:  0.2296,    Adjusted R-squared:  0.2218
F-statistic:  29.5 on 1 and 99 DF,  p-value: 4.007e-07
```

Figure 12a: Regression diagnostics

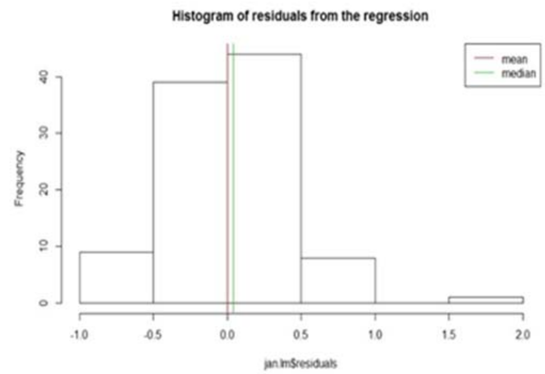

Figure 12b: Histogram of regression residuals

From the above figures (Figure 12a and 12b) it can be seen that there is a relationship to some extent between precipitation data and corresponding elevation. The coefficient corresponding to elevation (slope) is significantly different from zero (P value is very less, thus null hypothesis can be falsified). The low R-square value indicates that the data are not very close to the fitted regression line. It indicates to higher variation in the response variable. The histogram of the residuals show normal distribution but skewness is observed to be present. The mean and median are closely placed. So it can be inferred that elevation can be used as a covariate in precipitation prediction but it would not provide any drastically improved and accurate prediction than ordinary kriging.

Regression kriging is performed with the variogram consisting regression residuals (Exponential model fitted variogram of cutoff 2 and bin width. 15) and prediction map of precipitation for the month of January is generated along with the kriging variance map. The results are displayed in the following diagrams (Figure 13a and 13b).

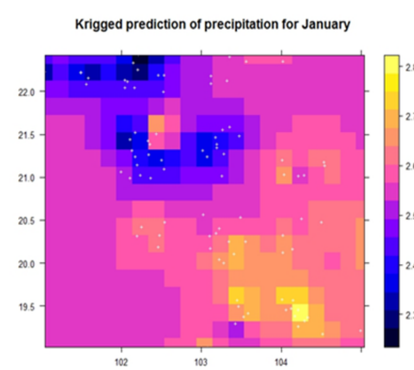

Figure 13a: Regression kriged prediction of January

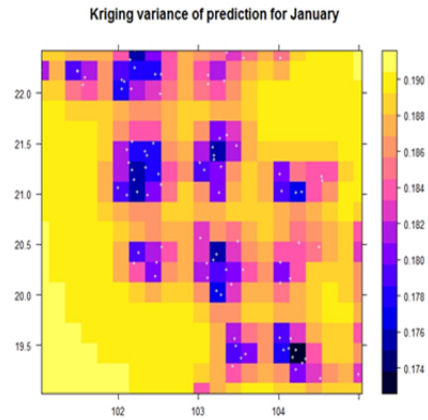

Figure 13b: Regression kriging variance of January

In the above precipitation prediction map (Figure 12a), the areas shown in orange and yellow are the areas with a higher amount of predicted precipitation and the areas shown in blue are the areas with lower amount of precipitation. In the precipitation prediction variance maps (Figure 12b), yellow coloured areas are the areas with high kriging variance and the areas with blue colour represent areas with low kriging variance. It is observed that maximum variance is more than of the ordinary kriging which justifies the regression diagnostics stated previously. The variance did not differ drastically than the ordinary kriging because kriging variance depends on the location of the points, not on the predicted value.

Cross-validation is performed for variograms incorporated in both the kriging techniques to find out the more accurate one. The results are tabulated below (Table 3). The results differ in terms of cross-validation parameter such as ME and RMSE values.

Table 3: Cross-validation results of kriging techniques

Kriging Technique	**Mean Error**	**RMSE**
Ordinary kriging	0.0003	0.01116
Regression kriging	0.0021	0.02193

In the table above (Table 3) it can be seen that mean error is less in the case of ordinary kriging whereas root mean square error (RMSE) is more than ordinary kriging. Regression kriging is useful when a trend is observed in the data. For January there was no such trend in the data. So, it can be inferred that for the month of January regression kriging does not provide any major advantage over ordinary kriging. However, searching for trend in rest of the months followed by regression kriging can be a future extension of this study.

6. DISCUSSION

In this study, for the provided dataset, variogram with a cutoff of 2 km. and bin width of. 15 km is found to be suitable. Spherical model fits the obtained variogram better than exponential model in predicting precipitation at un-sampled locations. Kriging using the selected model shows that more distant the un-sampled points are to observation points, higher the kriging variance is observed. It can be explained due to the less correlation with the distant points than of the closer ones to observation points. The observations (Figure 8a, 9a and 10a) show that the precipitation varies throughout the months. Seasonality of precipitation can be obtained from these results. Among the tested months, July shows the highest amount of precipitation prediction and November shows the minimum amount of precipitation. It can also be justified to the seasonal characteristics of these months. July is considered to be in the rainy season and prediction shows similar results. The dataset also supports this prediction. In the dataset, precipitation is observed to be maximum in July among the three months considered in the study. Thus it can be inferred that the predictions obtained from the study are valid and justified.

Regression kriging is also performed in the study as elevation can be considered as a covariate in the dataset and relationship is found with precipitation. Regression kriging is performed for January but no major difference is noted. Regression kriging is useful when there is a trend in the data. But in January no such trend can be observed, thus regression kriging does not provide any significantly different results. The co-kriging technique also requires another variable that may affect the precipitation, which was absent, so further analysis was not possible.

7. CONCLUSION

i. Spherical variogram is best suitable model to fit the variogram for the given dataset as it has a lower SSErr.

ii. The observation shows that precipitation is not constant through the calendar year. It varies from month to month. Thus, precipitation can be considered as a variable seasonal process.

iii. The predicted values obtained from kriging can be used to identify the months when the precipitation will be high.

iv. Regression kriging by incorporating elevation as a covariate will be useful when there will be a trend in the dataset.

REFERENCES

[1] Diggle, P. J., & Ribeiro, P. (2007). *Model-based Geostatistics*. New York, NY: Springer New York. https://doi.org/10.1007/978-0-387-98135-2

[2] Hengl, T., Heuvelink, G. B. M., & Rossiter, D. G. (2007). About regression-kriging: From equations to case studies. *Computers and Geosciences*, *33*(10),1301–1315. https://doi.org/10.1016/j.cageo.2007.05.001

[3] Van Groenigen, J. W. (2000). The influence of variogram parameters on optimal sampling schemes for mapping by kriging. *Geoderma*, *97*(3-4), 223–236. https://doi.org/10.1016/S0016-7061(00)00040-9

[4] Webster, R., & Oliver, M. A. (2008). *Geostatistics for Environmental Scientists*. (S. Senn, M. Scott, & V. Barnett, Eds.), *John Wiley & Sons, Ltd* (Second). John Wiley & Sons, Ltd. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/2483231

[5] Western, A. W., & Blöschl, G. (1999). On the spatial scaling of soil moisture. *Journal of Hydrology*, *217*(3-4), 203–224. https://doi.org/10.1016/S0022-1694(98)00232-7

[6] Wilson, E. M. (1983). *Engineering hydrology*. *Macmillan civil engineering hydraulics*. London: Macmillan Education UK. https://doi.org/10.1007/978-1-349-00329-7