**American Journal of Software Engineering and Applications**
2013; 2(2) : 40-48
Published online April 2, 2013 (http://www.sciencepublishinggroup.com/j/ajsea)
doi: 10.11648/j.ajsea.20130202.13

SciencePG
Science Publishing Group

# An approach to modeling domain-wide information, based on limited points' data – part II

**John Charlery, Chris D. Smith**

Dept. of Computer Science, Mathematics & Physics, Faculty of Science and Technology, University of the West Indies, Cave Hill Campus, Bridgetown, Barbados BB11000

**Email address:**
john.charlery@cavehill.uwi.edu (J. Charlery),chris.smith@mycavehill.uwi.edu (C. D. Smith)

**To cite this article:**
John Charlery, Chris D. Smith. An Approach to Modeling Domain-Wide Information, based on Limited Points' Data – Part II, *American Journal of Software Engineering and Applications.* Vol. 2, No. 2, 2013, pp. 40-48. doi:10.11648/j.ajsea.20130202.13

**Abstract:** Predicting values at data points in a specified region when only a few values are known is a perennial problem and many approaches have been developed in response. Interpolation schemes provide some success and are the most widely used among the approaches. However, none of those schemes incorporates historical aspects in their formulae. This study presents an approach to interpolation, which utilizes the historical relationships existing between the data points in a region of interest. By combining the historical relationships with the interpolation equations, an algorithm for making predictions over an entire domain area, where data is known only for some random parts of that area, is presented. A performance analysis of the algorithm indicates that even when provided with less than ten percent of the domain's data, the algorithm outperforms the other popular interpolation algorithms when more than fifty percent of the domain's data is provided to them.

**Keywords:** Data Modeling, Interpolation, Data Prediction, Sparse Data Analysis

## 1. Introduction

In Part I of this research, we presented the background and algorithms to a proposed interpolation methodology, whose primary objective was to develop an approach to the prediction of unknown stations values when some randomly selected other stations values are known. This was achieved by the implementation of an algorithm that established a historical statistical relationship between pairs of all the data stations in the region of interest. Since other algorithms do not incorporate historical relationships in the prediction of unknown data stations, this method is a "step-away" from the other popular established interpolations methods. In this second part of the research, we present an implementation of the methodology, through two very contrasting cases study.

## 2. Implementation

The algorithm is being implemented through the demonstration of two case studies. The performance of the algorithm is then analyzed through the usage of the two data sets. The first case study examines rainfall data (which represents data produced as a consequence of objective and physical rules,) and the second case study examines electoral ballots data at electoral polling stations (to represent data resulting from subjective processes). In both case studies the geographical domain is the country of Barbados, which was arbitrarily selected. The data from the stations of interest are scattered throughout the island. The map of the island is overlaid by a series of vertical and horizontal grid lines, which in turn creates a series of rectangular cells into which the data stations coincide. The virtual cells that house the data stations are used to calculate the Euclidean distances of the data stations from each other.

To determine the percentage deviations of the predictions, the error value must first be obtained. For simplicity, this is determined by simply taking the difference between the predicted values and the observed values for each station. The absolute values of the differences are used to eliminate the effects of having negative values after the subtractions. These absolute values are summed up and then divided by the sum of the observed values. Finally to get the percentage deviation, the error value is simply multiplied by 100. In other words.

Error Value is .
$$E = \left[ \frac{\sum_{1}^{i} D_i}{\sum_{1}^{i} A_i} \right] \qquad (1)$$

Deviation percentage is .

$$Y = E \text{ x } 100 \qquad (2)$$

Where

E = Error Value

Y = Percentage Deviation

D = absolute difference between actual observations and predicted values

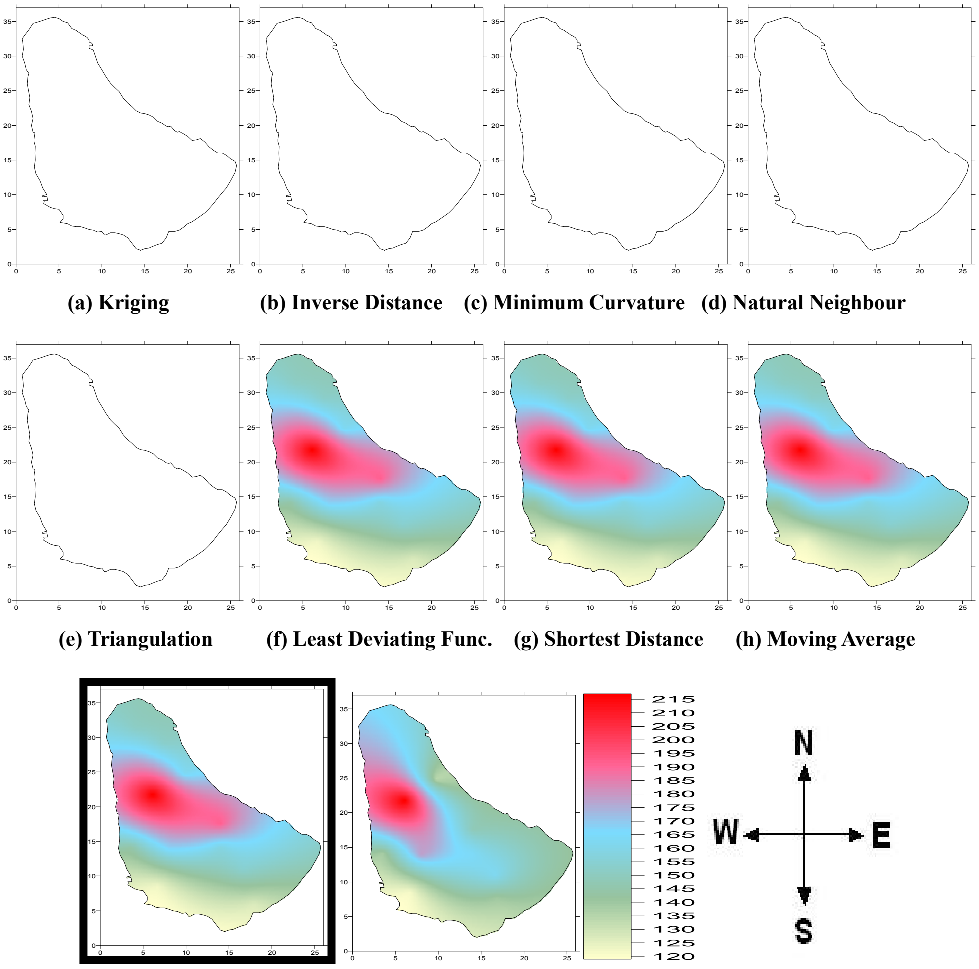
A = actual observation value

The prediction accuracy is given as 1 minus the error value. The maximum resulting value for the prediction accuracy is 1, which means perfect accuracy, or where the error value is zero.

### 2.1. Visualization of Established and Proposed Results for Case Study 1 – Rainfall Stations

The physical domain being used is 431 square kilometers and comprises of fourteen (14) data stations with monthly historical data values, which spans the period of 1960 to 2004. In the results, which are presented in Sections 6.1.1 to 6.1.3, we have arbitrarily and randomly selected one station with a known data value for January 2005, then three stations and then finally nine stations with known data values to perform the prediction for the entire domain for January 2005. A comparative analysis with some of the other popular interpolation algorithms is also presented.

#### 2.1.1. Analysis From Using One (1) Known Data Station Value



**(a) Kriging** **(b) Inverse Distance** **(c) Minimum Curvature** **(d) Natural Neighbour**

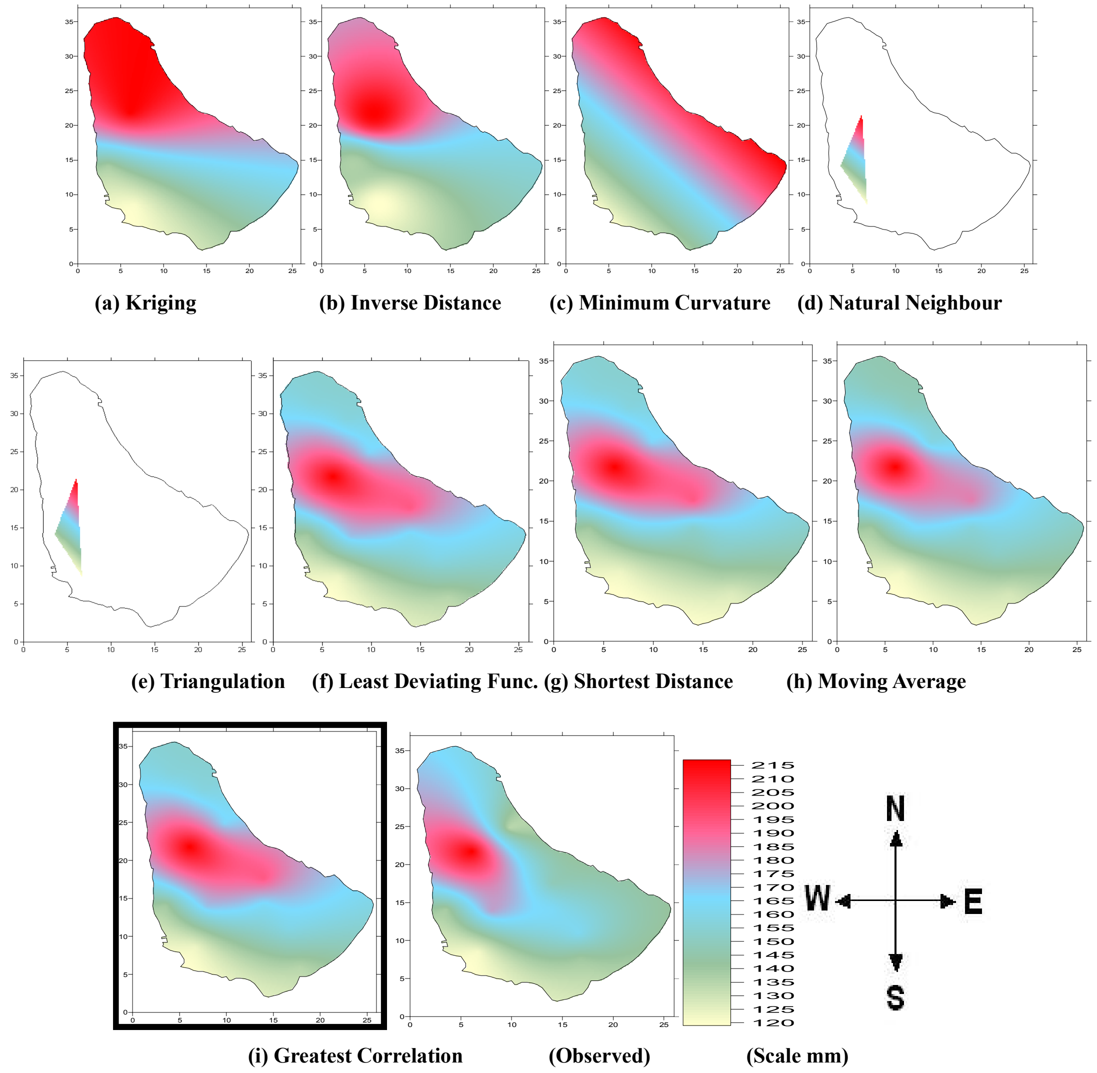
**(e) Triangulation** **(f) Least Deviating Func.** **(g) Shortest Distance** **(h) Moving Average**

**(i) Greatest Correlation** **(Observed)** **(Scale mm)**

*Fig.. 1. Images using one known data station's value.*

With only one data point is used as input data, the interpolation algorithms of kriging, inverse distance, minimum curvature, natural neighbour and triangulation are unable to generate predictions for the domain. However, the algorithms of least deviating function, shortest distance, moving average and greatest correlation, which this work has adapted to work as interpolation techniques, have been able to capture the general distribution of the rainfall over the domain. Even with only one data point provided, the analyses from these techniques mimic the observed distribution. The principal and common difference between the observed analysis and the predicted analyses is in the orientation of the distribution. For each of the three methods, the rainfall's ridge axis (reddish color) is oriented in a somewhat east/west direction while the observed rainfall's ridge axis assumes a more northwest/southeast orientation. Notwithstanding this, the area of maximum rainfall (brightest red) corresponds almost perfectly with the actual observation.

### *2.1.2. Analysis From Using Three (3) Known Data Stations' Values*



**(a) Kriging** **(b) Inverse Distance** **(c) Minimum Curvature** **(d) Natural Neighbour**

**(e) Triangulation** **(f) Least Deviating Func.** **(g) Shortest Distance** **(h) Moving Average**

**(i) Greatest Correlation** **(Observed)** **(Scale mm)**

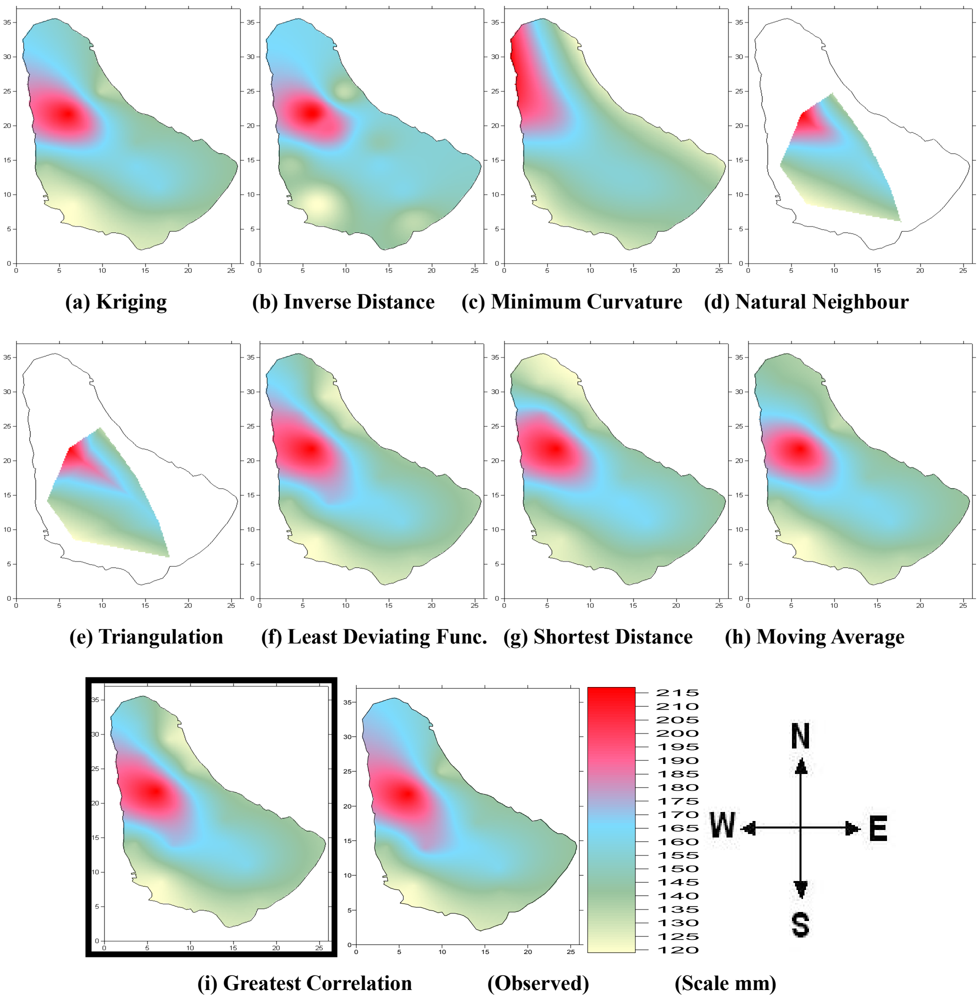
***Fig. 2.*** *Images using 3 known data stations' values.*

With three data point used as input data, the interpolation algorithms of kriging, inverse distance, minimum curvature, natural neighbour and triangulation are able to generate some predictions for the domain. However, for all three methods, when compared with the observed, their predictions are very misleading.

Kriging and inverse distance concentrate the maximum rainfall in the northern half of the island while minimum curvature concentrates the maximum rainfall in the eastern half of the island. With those three methods, neither rainfall distribution nor quantity reflects much similarity with what is actually observed. The natural neighbour and triangulation algorithms have fared better, but their predictions are limited to the zone bounded within the three data points while in both cases providing no information for the remainder of the domain.

The algorithms of least deviating function, shortest distance and moving average, maintain a similar distribution to the previous case in Section 6.1.1, although there are now some adjustments to the quantities being predicted. The analysis for greatest correlation is reflecting this tweaking and continues to generate the closest match to the observed.

### *2.1.2. Analysis From Using Nine (9) Known Data Stations' Values*



(a) Kriging (b) Inverse Distance (c) Minimum Curvature (d) Natural Neighbour

(e) Triangulation (f) Least Deviating Func. (g) Shortest Distance (h) Moving Average

(i) Greatest Correlation (Observed) (Scale mm)

***Fig. 3.*** *Images using 9 known data stations' values.*

With 64% of the data now used as input in the simulation, the analyses from the interpolation algorithms of kriging and inverse distance are now beginning to show some similarity to the observed. Compared to the previous case, the results from both the kriging and the inverse distance algorithm's rainfall's ridge axis (reddish color) have contracted and shifted from a northerly direction towards a more northwest/southeast orientation. However, with the minimum curvature, the rainfall's ridge axis has shifted from its location in the extreme northern part of the island to become oriented in a north/south direction, but still locked in the northwestern part of the island. However, neither rainfall distribution nor quantity reflects much similarity with what is actually observed.

The algorithms of natural neighbour and triangulation continues to increase its coverage region with the nine known data input. The predicted values for the unknown stations in their coverage area compare favorably with the observed. However their predictions are still limited only to the zone within the nine data points, while in both cases providing no information for the remainder of the domain.

The algorithms of least deviating function, shortest distance, moving average and greatest correlation continues to maintain a similar distribution to the previous case with further minor adjustments to the quantities being predicted.

The results from the greatest correlation algorithm continue to reflect further minor tweaking in the quantities being predicted for the unknown stations and also continues to provide the closest match to the observed in terms of both rainfall quantity and distribution.

### *2.1.4. Summary Conclusion for Case Study 1*

Although this is only a single case, some aspects of the comparison are very conspicuous and are noted as following.

**Kriging, Inverse Distance and Minimum Curvature.**

- No prediction was possible until at least 3 data stations were provided as input.
- With 3 data stations, predictions were made but these predictions turned out to be very misleading.
- Predictions began to show some similarity with the observed when at least 9 data stations (64% of all the stations) were used as

input. For kriging and inverse distance the predictions continued to improve steadily after more data stations were added. Minimum curvature displayed a significantly much slower rate of improvement than all the other methods.

**Natural Neighbour and Triangulation.**

- No prediction was possible until at least 3 data stations were provided as input.
- With 3 or more data stations used, predictions comparable to the observed were made, but these predictions were limited to the inner area created by those data points alone.

**Least Deviating Function, Shortest Distance, Moving Average and Greatest Correlation.**

- Captured general rainfall distribution pattern over the island with only one data point used as input.
- Predictions continued to improve with small tweaks, as more data were added as input.
- Consistently produced very good results with the greatest correlation method producing the best results at each stage.

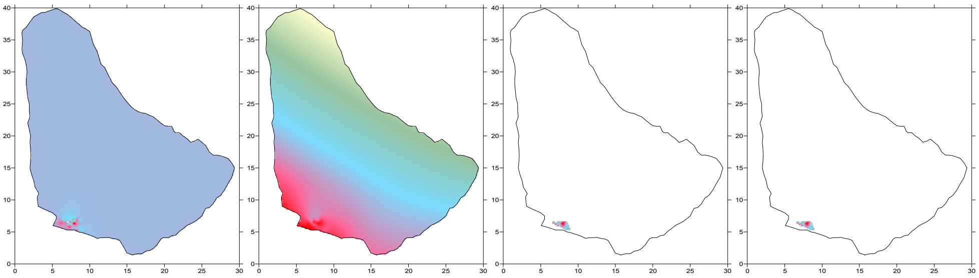
### *2.2. Visualization of Established and Proposed Results for Case Study 2 – Electoral Polling Stations*

For this case study, the algorithm is applied to data, which have been derived through a less rigid process, where the physical laws which determine the values are not as clearly defined. Electoral polling results for political election events in the country of Barbados are used. Far from being able to define the objective laws, which determine the casting of ballots in a democratic election, it may be more appropriate to describe this process as the result of subjective motivation. Nonetheless, it is expected that some polling stations will still have a positive relationship – even if such a relationship is weak.

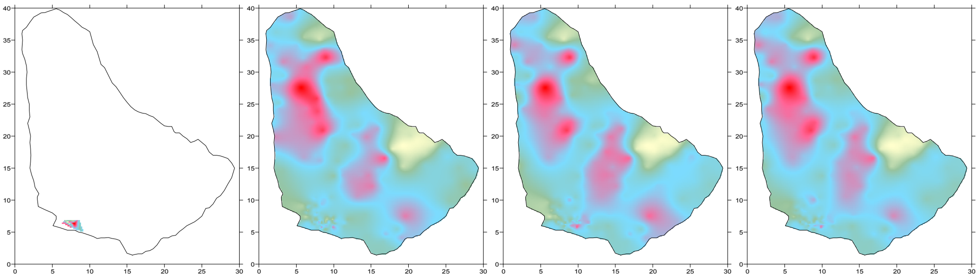
The polling stations' data were provided by 225 polling stations, which span the period from 1976 to 2003 and constituted seven (7) national electoral events. For this case study, the number of stations is relatively large. Hence, in the analysis, we will observe the proposed model's performance by randomly sampling 8%, 21% and 64% respectively of known data values, and noting how "well" the model is predicting the values for the remaining unknown polling stations.

Here the choice of 8%, 21% and 64% of known data values are not a requirement for the algorithm but have been deliberate for this case. These choices will allow a percentage comparison with the rainfall data in **Section 2.1.** The historical data from 1976 to 1999 will therefore be used to derive the relationships between the stations and the observed data of 2003 will be used as the "testing values". The data presented, represent the percentage of the ballot cast for one political organization (the Barbados Democratic Party).

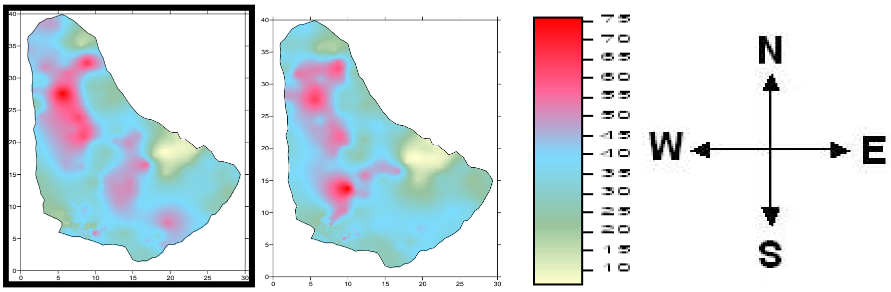
### *2.2.1. Analysis From Using 8% Of Known Data Stations' Values*



**(a) Kriging** **(b) Inverse Distance** **(c) Minimum Curvature** **(d) Natural Neighbour**



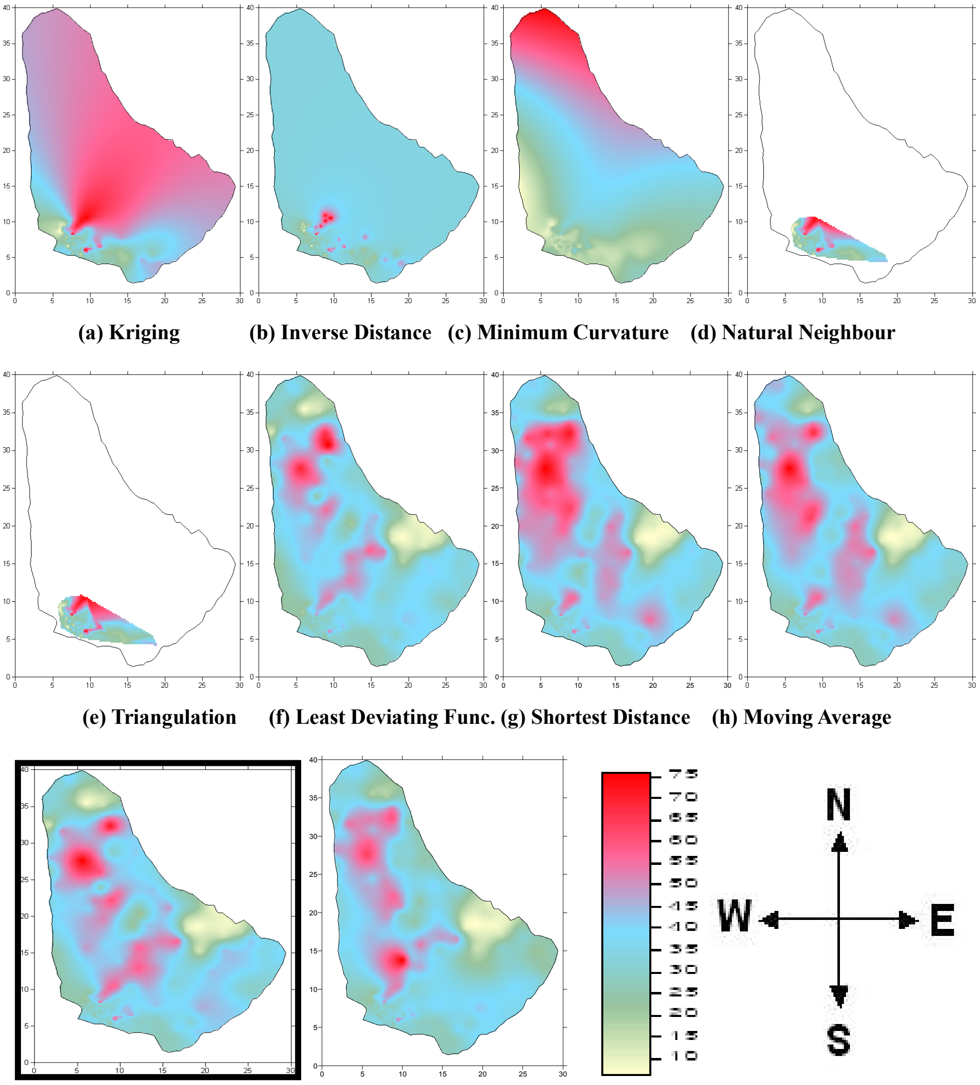
**(e) Triangulation** **(f) Least Deviating Func.** **(g) Shortest Distance** **(h) Moving Average**



**(i) Greatest Correlation** **(Observed)** **(Scale %)**

*Fig. 4. Images using 8% of known stations.*

### *2.2.2. Analysis From Using 48% Of Known Data Stations' Values*



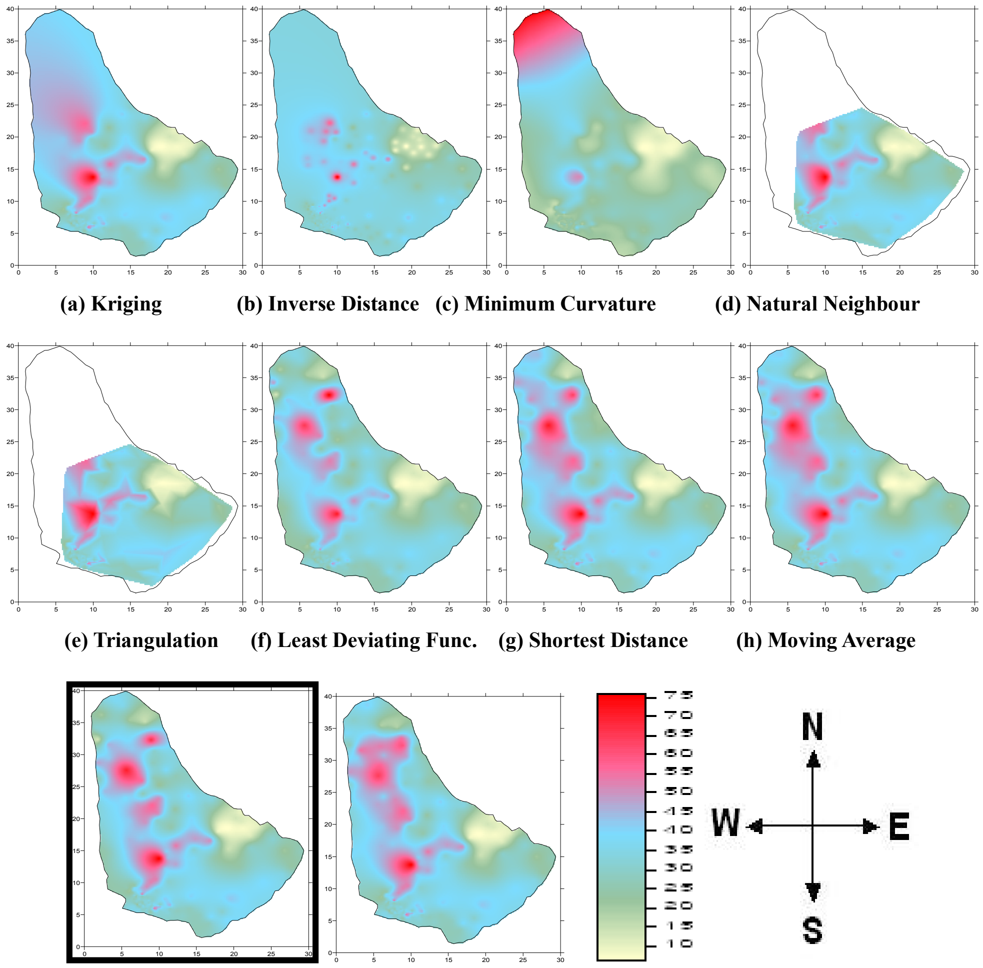
**(a) Kriging** **(b) Inverse Distance** **(c) Minimum Curvature** **(d) Natural Neighbour**

**(e) Triangulation** **(f) Least Deviating Func.** **(g) Shortest Distance** **(h) Moving Average**

**(i) Greatest Correlation** **(Observed)** **(Scale %)**

*Fig. 5. Images using 48% of known stations.*

### *2.2.3. Analysis from Using 80% of Known Data Stations' Values*



(a) Kriging (b) Inverse Distance (c) Minimum Curvature (d) Natural Neighbour

(e) Triangulation (f) Least Deviating Func. (g) Shortest Distance (h) Moving Average

(i) Greatest Correlation (Observed) (Scale %)

***Fig. 6.*** *Images using 80% of known stations.*

### *2.2.4. Summary Conclusion for Case Study 2*

**Kriging, Inverse Distance and Minimum Curvature.**

- With almost half of the data used as input, the simulations produce results, which bear absolutely no relation to the observed.
- Between 55% and 75% of the data used as input, the values of the maxima and minima areas slowly begin to look somewhat like the observed. However, the distributions for the three models are widely varying and still not truly capturing the distribution indicated by the observed.
- Not until 80% of the data are used as input do the generated images begin to mimic a similar distribution and orientation to the observed. Even with 90% of the data used as input, there still remained many areas where the quantities are either being significantly under- or over-predicted.

**Natural Neighbour and Triangulation.**

- The algorithms for natural neighbour and triangulation produce a steady increase in the domain's area for which predictions are made. However, upon close examination, the maxima and minima areas did not correspond very well to those of the observed when less than 50% of the data were used as input.
- The values predicted became progressively more reliable as the input data grew beyond 50%. Even then the results provided information for only a subset of the domain.
- With over 80% of the data used as input, the images produced started to reflect the distribution of maxima and minima areas, which corresponded to the observed.

**Least Deviating Function, Shortest Distance, Moving Average and Greatest Correlation.**

- With as little as 8% of the data provided as input, both distributions and orientations of the analyses matched the observed fairly closely.
- Starting with using 8% of the data as input, the predictions for the entire domain continued to improve steadily as more data were added progressively to the simulations.

- Consistently produced very good results with the greatest correlation method producing the best results at each stage.

## 3. Conclusion

The primary objective of this work was to develop an approach to the prediction of unknown stations values when some randomly selected stations values are known. This was achieved by the implementation of an algorithm that established a historical statistical relationship between pairs of all the data stations in the region of interest. The use of a historical relationship in the prediction of unknown data stations is a “step-away” from the established interpolations methods such as Kriging, Inverse Distance and Minimum Curvature (to name a few). These established methods do not factor in any historical relationships in their formulae.

This different approach to the interpolation of unknown data stations have yielded some results that are impressive, in that it shows that irrespective of the type of data being studied, whether objective or subjective, there is a strong correlation between historical trends and the future values of the data stations.

This observation is evident in the presentation of the color distribution maps provided in the proposed algorithm’s results in Section 6. These maps show that when even as few as only one station’s data is used as input in the simulation of the domain’s data, irrespective of either objective or subjective data, the predictions generated for the unknown stations by the proposed algorithm, have still managed to capture the quantities, distributions and general orientations of the domain’s data as evidenced by the actual image map for the observed data. The differences between the predicted values for the unknown stations (as well as the analysis for the entire domain) and the actual observations became increasingly and progressively smaller as the number of known stations to be used as input increases.

The results show that the algorithm has a higher degree of accuracy when applied to objective data as opposed to subjective data. In looking at the deviation from the observed values, it can be seen that the deviation starts at a much higher comparative value for the subjective data than that for the objective data. This behaviour was expected since the engine which drives the generation of objective data is better understood and therefore more consistent over time. This consistency is therefore an ideal factor when searching for any historical significance between the data stations. On the other hand, subjective data by its own definition, is derived through a less rigid process, where the physical laws which determine the values are not as clearly defined and can therefore be inconsistent and periodically difficult to explain in quantified scientific terms.

A secondary objective of this thesis was to compare the results of some of the more popular interpolation methods with the proposed algorithm. This comparison was implemented through the detailed examinations of the two cases studied in Section 6. In the image distribution maps for both the objective and subjective data, the Natural Neighbour and Triangulation methods both showed some shortcoming, in that they were not well equipped in their interpolation analyses to generate predictions for the entire domain. Both methods only provided values within the areal bounds created by the known stations and did not extrapolate beyond these stations. Despite their skills within the bounded areas, they were unable to provide the domain-wide analyses which were desired.

With the Minimum Curvature, Inverse Distance Weighting and Kriging algorithms, the results were more suited for a comparison to the proposed method. In the images produced for the objective data, it was only after 63% of the stations were used as input to the simulation did the established methods began to resemble the observed domain’s analysis. With the subjective data, a similar comparison was only achieved after 88% of the stations’ data were used as input.

This work has therefore presented an algorithm which can be compared favorably with the other established algorithms of.

- Natural Neighbour
- Triangulation
- Minimum Curvature
- Inverse Distance Weighting
- Kriging.

## 4. Further Work

The strength of this work lies in the existence of historical information to determine the relationships between data points within the domain of interest. This dependence restricts the beneficial usage of the algorithm as frequently, such historical information may either not be available or not be sufficient to truly inform the relationships between points within the domain. The scope of this work has not addressed this limitation. However, it is believed that further investigations, possibly implementing some variance of the Bayesian strategy [16] could provide an appropriate support to those cases where historical information is not available.

The need for further refinement is also being highlighted as a path for further attention to be focused. This work dealt with the historical data relationships of static stations whose data can then be used as input for the simulation. This, of course poses another limitation on the system as one may not only want to use the value of an already existing station but may want to include the available values for stations that have been newly added to the region of interest. The challenge, therefore, will be to make the system of predictions into one that is dynamic in nature or possibly to also develop relationships between zones or geographical areas. This would add a further element of complexity to the system, but will also make it more robust.

Finally, the relationships, which have been used between stations, have been for the express purpose of making predictions for the data elements, which define the relationships. Investigations into one set of data elements at one location to produce information on a different element in another location may provide better relationships between data points than that provided by the same element. For example, the humidity value at point A and the wind speed value at point B could possibly provide a better prediction relationship for rainfall at point C than rainfall values at points A and B. It is believed that this approach, particularly for subjective data, may further enhance the strength of the proposed algorithm.

## References

[1] N. I. Fisher, T. Lewis, B. J. J. Embleton, Statistical Analysis of Spherical Data, Cambridge University Press, 1987.

[2] D. Dorsel, T. La Breche, Kriging. <http.//ewr.cee.vt.edu/environmental/teach/smprimer/kriging/kriging.html>, January 2009.

[3] R. V. Jesus, Kriging. An Accompanied Example in IDRISI, GIS Centrum, University of Lund for Oresund, Summer University, 2003.

[4] M.L. Stein, Interpolation of Spatial Data. Some Theory for Kriging, Springer, New York, 1999.

[5] W.C.M. Van Beers, J. P.C. Kleijnen, Kriging Interpolation in Simulation . A Survey, in. R .G. Ingalls, M. D. Rossetti, J. S. Smith, and B. A. Peters (Eds.), Proceedings of the 2004 Winter Simulation Conference,Washington, DC, 2004, pp. 113-121.

[6] D. T. Lee, B.J. Schachter, Two Algorithms for Constructing a Delaunay Triangulation. International Journal of Computer and information Sciences, Vol. 9, 1980, pp. 219-242.

[7] P-J. Laurent, Wavelets, Images, and Surface Fitting, in. A. Le Mehaute (Ed.), A.K Peters Ltd., 1994.

[8] W. H. F. Smith, P. Wessel, Gridding with Continuous Curvature Splines in Tension, Geophysics, 55, 1990.

[9] D. Shepard, A two-dimensional interpolation function for irregularly spaced data. Proceedings of the 23rd ACM National Conference (128), 1968, pp.517-524.

[10] R. Sierra, Rigid Registration. <http.//www.rsierra.com/DA/node10.html#SECTION001030000000000000000>, May 2009.

[11] J. Parag, Class Presentation. <http.//arcib.dowling.edu/~JainP/Research1/slide2.html>, May 2009.

[12] D. Kleinbaum, L. Kupper, K. Muller, Applied Regression Analysis and other Multivariable Method, Duxbury Press, 1987.

[13] Wasson J. Statistics in Educational Research - An Internet Based Course. <http.//www.mnstate.edu/wasson/ed602pearsoncorr.htm>, April 2009.

[14] J. Deacon, Correlation, and regression analysis for curve fitting. <http.//www.biology.ed.ac.uk/research/groups/jdeacon/statistics/tress11.html >, January 2009.

[15] R.J. Rummel, Understanding Correlation. <http.//www.mega.nu.8080/ampp/rummel/uc.htm>, December 2009.

[16] E. Yudkowsky, An Intuitive Explanation of Bayesian Reasoning, <http.//yudkowsky.net/bayes/bayes.html>, May 2009.

[17] P. E. Gill, W. Murray, Algorithms for the solution of the nonlinear least-squares problem. SIAM Journal of Numeral Analysis, 15 [5], 1978, pp. 977-992.

[18] W. R. Greco, M. T. Hakala. Evaluation of methods for estimating the dissociation constant of tight binding enzyme inhibitors, Journal of Biological Chemistry, (254), 1979, pp.12104-12109.

[19] D. F. Symancyk, Visualizing Gaussian Elimination, <http.//ola4.aacc.edu/dfsymancyk/vgetalk/VGEtalkexpanded.html>, May 2006.