Estimation of Missing Rainfall Data in Northeast Region of Thailand
Using Kriging Methods: A Comparison Study

Jesada Kajornrit¹, Kok Wai Wong², Chun Che Fung³
School of Information Technology, Murdoch University
South Street, Murdoch, Western Australia, 6150
Email: j_kajornrit@hotmail.com, k.wong@murdoch.edu.au², l.fung@murdoch.edu.au³

Abstract. Ground-based rainfall observations are the primary sources of precipitation data used in most developing countries. However, those observations are frequently damaged or incomplete, thus missing data is always a problem. This comparison study examines a number of kriging methods used to estimate missing monthly rainfall data in the northeast region of Thailand. It was found that the characteristics of the datasets have significant effect on the estimation performance. This study recommends using the kurtosis value of observations’ histogram and nugget-sill ratio of fitted semivariogram models as a guideline to select between ordinary kriging and universal kriging methods. Since the study area is a large plateau, in which there is low correlation between rainfall and altitude, ordinary co-kriging method cannot make use of the altitude as a supplementary feature to improve the estimation performance.

Key words: Missing Rainfall Data, Spatial Interpolation, Kriging, Northeast of Thailand

1. Introduction

In developing countries historical rainfall data play an important role in hydrological management systems. A large number of rain gauge stations are installed throughout the study area to record the rainfall data. Those rainfall data are essential in creating the rain maps for the region. However, in practice, rainfall records often contain missing data values due to malfunctioning of equipment [1] and severe environmental conditions. Furthermore, in some cases, a large number of stations could be down simultaneously, thus creating many inaccurate readings or missing data. Such imperfect rainfall data could affect the accuracy of the rain maps. Therefore, estimating missing rainfall data is an important task in hydrology. To overcome the problem, most of the estimation can be achieved by spatial interpolation.

Spatial interpolation is a method used to estimate the value at unsampled points by using the values from neighbouring sampled points [2]. Such method is commonly used to estimate spatial data or create statistical surfaces passing through a region. Up to date, many spatial interpolation methods have been developed. They can be divided into two main groups: deterministic and geostatistical methods. Each method has its own specific assumptions and features. Due to the variation of features in a region, estimation results could be varied, even for the same neighbouring data. This study examined the estimation performance of some common spatial interpolation methods using kriging to estimate missing monthly rainfall data in the northeast region of Thailand. The objectives of our study are, firstly, to compare the estimation performance of the common methods, and secondly, to investigate the characteristics of the data that affect the estimation performance. This study also acts as a preliminary investigation that allows the understanding of the data obtained in this region of Thailand.

The rest of paper is organized as follows. In section 2, issues of spatial interpolation in climatology are introduced. Section 3 presents a brief discussion on the kriging spatial interpolation methods used in this paper. The case study is mentioned in section 4. Section 5 shows the comparison of the experimental results and provides some analysis on the data and methods used. Finally, conclusion will be discussed in section 6.

2. Spatial Interpolation Issues in Climatology

In general, spatial interpolation is widely used for various objectives in climatology. A number of works used spatial interpolation methods directly to estimate missing rainfall data. For examples, Teegavarapu and Chandramoul [3] applied Inverse Distance Weighting (IDW) method and its variants to estimate missing rainfall data in the state of Kentucky, USA. Jeffrey et al. [4] used Thin Plate Spline (TPS) to estimate missing daily climate variables to construct a comprehensive
archive of Australian climate data. As the characteristics of climate data change from region to region, comparisons of the methods are needed to find an appropriate solution. Luo et al. [5] compared seven spatial interpolation methods to estimate daily mean wind speed across England and Wales. Hartkamp et al. [6] compared several interpolation methods for climate data in Jalisco, Mexico. Cao et al. [7] compared five interpolation methods for climate data in China. In some works, the estimation accuracy of the spatial interpolation has been improved by adding supplementary features. Naelder and Wein [8] developed a novel method, Gradient plus-Inverse Distance Squared (GIDS) which combines multiple linear regression and distance weighting to interpolate monthly temperature and precipitation in western Canada. Price and McKenney et al. [9] applied two elevation-dependent interpolators, ANUSPLIN and GIDS in Canadian monthly mean climate data. Hong and Nix et al. [10] use TPS cooperating with Digital Elevation Model (DEM) to interpolate monthly mean climate data in China; similarly to the work of Goovaerts [11], which applied geostatistical interpolation methods incorporating DEM into the spatial interpolation of rainfall in Portugal. Hancock and Hutchison [12] used Bivariate TPS to interpolate large two climate datasets in Africa and Australia continent. Yu [13] purposed Geographically Weight Regression as an alternative of spatial interpolation. Up to this point, several works related to the applications of spatial interpolation have been discussed. One can realize that there is no one method which works well on every dataset. In order to select an appropriate method, comparison and analysis are necessary. In the initial study on the data set obtained from the northeast region of Thailand, Kriging is selected for this purpose. In the next section, kriging spatial interpolation methods used in our experiments will be discussed.

3. Background Theory

Kriging perform interpolation like IDW, but the method uses spatially dependent variance of data instead of spatial distance. Kriging not only predict data at unsampled points, but also assesses the quality of estimation. The assumption of kriging method is that the spatial variation of data is neither totally random nor deterministic. Instead, the spatial variation consists of three components, namely, (i) a spatial correlation component which represents the variation of the regionalized variable; (ii) a drift or structure which represents a trend; and (iii) random error term [2]. In kriging method, the key factor is semivariogram. Semivariogram is the model representing the spatial correlation presented as:

$$\gamma(h) = \frac{1}{2n} \sum_{i=1}^{n}[z(x_i) - z(x_i + h)]^2$$  \hspace{1cm} (1)

where $\gamma(h)$ is the average semivariance between sampled points separated by lag $h$, $n$ is the number of pairs of sample points, and $z$ is the attribute value. In another word, semivariogram is a relation between lag distance and semivariance which is depicted in Fig 1. In the figure, several important features are displayed: (i) the nugget is the semivariance at the distance of 0, representing the sampling error and/or spatial variance at shorter distance than the minimum sample space. (ii) The range is distance at which the semivariance starts to level off. Beyond the range, the semi-variance becomes a relatively constant value. (iii) The sill is semivariance at which the leveling takes place. Sill, in turn, consists of the partial sill ($C_1$) and nugget ($C_0$). In practice, such as experimental semivariogram, is not suitable for computation. So, mathematical models are used instead. Usually, four types of model are preferred, namely, spherical, exponential, Gaussian and linear models. However, the most commonly used model is the spherical model and it is therefore adopted in this study.

![Fig. 1. A General model of semivariogram shows the several important features: nugget, range, sill and partial sill.](image)

Ordinary Kriging (OKG) interpolates data by using fitted semivariogram. OKG focuses only on the spatial correlation and absence drift in interpolation. The general equation for estimating is
\[ z_0 = \sum_{i=1}^{s} z_i W_i \]  

(2)

where \( z_0 \) is the estimated value, \( z_i \) is the sampled points and \( W_i \) is the weight associated with point \( i \), and \( s \) is the number of sampled points. The weight can be derived from solving a set of simultaneous equations. One constraint for weight is that the sum of all weight is equal to 1. The variation estimation can be calculated by

\[ \sigma_0^2 = \sum_{i=1}^{s} W_i \gamma_{i,0} + \lambda \]  

(3)

where \( \sigma^2 \) is the variance, \( W_i \) is the weight at point \( i \), \( \gamma_{i,0} \) is the semivariance between point \( i \) and point to be estimated, and \( \lambda \) is the Lagrange multiplier added to ensure that the estimation results produce minimum error.

Universal Kriging (UKG) interpolates data in the same manner as OKG except the drift of data is taken into account. UKG has the general equation as

\[ z_0 = \sum_{i=1}^{s} z_i W_i + M \]  

(4)

where \( z_0 \) is the estimated value, \( z_i \) is the sampled points, \( W_i \) is the weight associated with point \( i \), \( s \) is the number of sampled points and \( M \) is a linear or quadratic polynomial equation. In this study linear equation is adopted and expressed as

\[ M = b_1 x_i + b_2 y_i \]  

(5)

where \( b_i \) is the drift coefficient. A higher-order poly-nominal are usually not recommended because it would leave little variation in the residual used to assess uncertainty and a large number of \( b_i \) would have an effect to the computation efficiency.

Cokriging (CKG) uses one or more secondary variables, which are correlated with primary variable of interest, in interpolation. It assumes that the correlation between the variables can be used to improve the prediction value of the primary variable. In our experiment, the secondary variable is assigned to the altitude. This is because moist air is normally present in higher altitude, where it could reach the dew point in producing rain. Thus the altitude information is important feature which may affect the amount of rain in the area. Cokriging requires much more estimation. It not only estimates the autocorrelation of the rainfall and the altitude, but also estimates the cross-correlation between them. Usually, this concept can be applied to both ordinary kriging and universal kriging. However, in this study, it is applied only to ordinary kriging. The equation of ordinary cokriging is as follows:

\[ z_0^{(1)} = \sum_{i=1}^{s_{(1)}} z_i^{(1)} W_i^{(1)} + \sum_{j=1}^{s_{(2)}} z_j^{(2)} W_j^{(2)} \]  

(6)

where \( z_0^{(1)} \) is the estimated value, \( z_i^{(1)} \) is the sampled points, \( W_i^{(1)} \) is the weight associated with point \( i \) of primary valuable, \( z_j^{(2)} \) is the sampled points and \( W_j^{(2)} \) is the weight associated with point \( j \) of secondary valuable. (This study uses the same sampled points for the primary and secondary variables, thus \( s_{(1)} = s_{(2)} \)). The weight can be derived from solving a set of simultaneous equations with two conditional constrains, the sum of all primary weight is equal to 1 and the sum of all secondary weight is equal to 0. This study uses spherical model to estimate rainfall correlation, elevation correlation and cross-correlation between them.

4. Case study

The datasets used in our experiments are monthly rainfall data from May to October in the year 2001. The numbers of observations are collected from 294 stations. Half of them are randomly selected to be sampled points used for estimation and the other half are used as unsampled points for validation. In general, sampled and unsampled points are normally allocated in 70:30 ratio. However, in this case study, the sampling ratio is based on 50:50. This is because a practical assumption that at least half of the observations must be working. Another reason is that the unsampled points will be used for further experiments, which need enough data for further selection. The statistics of data are shown in Table 1 and the study area is shown in Fig 2.

The R value in the table is the correlation between monthly rainfall and altitude. It is evident that in such large plateau area, rainfall and altitude is relatively independent as reflected by small value of R. In the experiment, the accuracy of
estimation methods are mainly validated by two measures, that is, Relative Mean Error (RME), Relative Root Mean Square Error (RRMSE). Besides, coefficient of fit ($R^2$) is another additional measure.

**Fig. 2.** The figure shown above is the case study area, northeast region of Thailand. The altitude data in this study can be found at http://www.gdem.aster.ersdac.or.jp.

5. Experimental Results and Analysis

The estimation results, namely, RME, RRMSE and $R^2$, are shown in Fig 3. Taking RME into account, it is obvious that all interpolator give positive estimation bias on nearly all datasets and give negative estimation bias only in September. OKG and CKG give very similar results in all datasets, whereas UKG provides higher estimation bias form May to July. This can be concluded that, based on the RME measures, OKG and CKG provide better estimation performance than UKG because they give close to zero RME value when compare to the result of UKG in the first four datasets.

When taking the RRMSE measure into account based on the data from May to August, OKG gives the lowest RRMSE and UKG provides highest RRMSE, whereas CKG gives the estimation error close to OKG. In contrast, for the case of the September and October data, UKG provides lower error and CKG provides the highest error, whereas OKG provides estimation result close to UKG. It is possible that linear trend appears on the last two months, which causes UKG to provide better results than the first four months. If one takes the statistics of the sampled data into consideration, one measure that reflects the linear trend is kurtosis.

In Table 1 the kurtosis measures fall between -0.1 and 1.8 approximately from May to August. However, in September and October, the kurtosis measures increase almost two times, that is, 3.09 and 4.35. Coupled with the lowest RMSE of UKG in these two months, it is high possibility that linear trend appears on the datasets, which causes UKG outperforms OKG. So, one suggestion is that if the kurtosis is less than 2.0 (in this study area), OKG should be considered. In contrast if kurtosis is more than 3.0, UKG is more appropriate.

<table>
<thead>
<tr>
<th>Statistics</th>
<th>May</th>
<th>Jun</th>
<th>Jul</th>
<th>Aug</th>
<th>Sep</th>
<th>Oct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>1970.35</td>
<td>2306.39</td>
<td>1613.56</td>
<td>3701.21</td>
<td>2369.25</td>
<td>1624.34</td>
</tr>
<tr>
<td>Median</td>
<td>1875</td>
<td>2164.5</td>
<td>1573</td>
<td>3357</td>
<td>2177.5</td>
<td>1515</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>1000.27</td>
<td>1275.26</td>
<td>845.83</td>
<td>1840.11</td>
<td>1145.95</td>
<td>953.22</td>
</tr>
<tr>
<td>Relative Standard Deviation</td>
<td>0.55</td>
<td>0.63</td>
<td>0.59</td>
<td>0.55</td>
<td>0.53</td>
<td>0.59</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>0.547</td>
<td>1.792</td>
<td>1.160</td>
<td>-0.130</td>
<td>3.090</td>
<td>4.351</td>
</tr>
<tr>
<td>Skewness</td>
<td>0.823</td>
<td>0.939</td>
<td>0.736</td>
<td>0.634</td>
<td>1.169</td>
<td>1.707</td>
</tr>
<tr>
<td>Minimum</td>
<td>198</td>
<td>108</td>
<td>206</td>
<td>341</td>
<td>287</td>
<td>0</td>
</tr>
<tr>
<td>Maximum</td>
<td>5486</td>
<td>7476</td>
<td>5100</td>
<td>8334</td>
<td>7497</td>
<td>5399</td>
</tr>
<tr>
<td>R</td>
<td>0.016</td>
<td>-0.327</td>
<td>-0.258</td>
<td>-0.007</td>
<td>-0.104</td>
<td>0.309</td>
</tr>
</tbody>
</table>
Considering into the CKG, because of the low correlation between altitude and rainfall, it causes CKG to provide the estimation never better than OKG or UKG in all datasets because the cross-semivariogram between altitude and rainfall cannot reflect the relation between these features, which results in the lower estimation performance comparing to the other methods.

Another measure in this study is $R^2$. After considering the relationship between $R^2$ and RRMSE (Fig 3c), it is obvious that $R^2$ is inversely related to RRMSE when comparing the datasets. However, in July, even if RRMSEs are higher than both June and August, $R^2$ is still the highest. This is because in July rainfall distributed more evenly over the study area than June and August, i.e., does not have much peak values. This confirms the work of Li and Heap, which criticised that $R^2$ should not be used as a model performance measure because it is often misleading [14].
The estimation results, (a) Relative Mean Error (RME), (b) Relative Root Mean Square Error (RRMSE) and (c) Coefficient of Fit ($R^2$). 

![Graph showing RME, RRMSE, and $R^2$ for different months.]

Fig. 3. The estimation results, (a) Relative Mean Error (RME), (b) Relative Root Mean Square Error (RRMSE) and (c) Coefficient of Fit ($R^2$).

Table 2. The fitted Semivariogram’s Parameters of OKG.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>May</th>
<th>June</th>
<th>July</th>
<th>August</th>
<th>Sep</th>
<th>Oct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Range</td>
<td>4.29681</td>
<td>4.29681</td>
<td>4.32976</td>
<td>4.32976</td>
<td>1.22947</td>
<td>1.50752</td>
</tr>
<tr>
<td>Partial Sill</td>
<td>1106100</td>
<td>2390800</td>
<td>1161000</td>
<td>3913000</td>
<td>192290</td>
<td>209810</td>
</tr>
<tr>
<td>Nugget</td>
<td>414960</td>
<td>441390</td>
<td>160340</td>
<td>977960</td>
<td>1107600</td>
<td>718050</td>
</tr>
<tr>
<td>Lag</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>Lag Size</td>
<td>0.3625</td>
<td>0.3625</td>
<td>0.36528</td>
<td>0.36528</td>
<td>0.21353</td>
<td>0.36528</td>
</tr>
<tr>
<td>Nugget/Sill</td>
<td>0.2728098</td>
<td>0.1558476</td>
<td>0.1213465</td>
<td>0.1999526</td>
<td>0.8520721</td>
<td>0.7738775</td>
</tr>
</tbody>
</table>

The kurtosis measure has been used as a consideration in examining the trend in sampled data. Another way to describe the characteristics of sampled data is the ratio of nugget to sill of fitted semivariogram model. Table 2 shows the parameters used in the fitted semivariogram models of OKG method in each dataset. From this table, the ratio of nugget to sill falls between 0.12 and 0.27 approximately in first four datasets. However, this ratio increases to more than 0.7 in last two datasets. Due to the better estimation performance of UKG to OKG in last two datasets, it is a possibility that this strong values of nugget to sill ratio are caused by appearing of trend in sampled data. Therefore, this ratio can be used as another consideration in examining the trend in sampled data. Such behavior suggests that (for this case study) if the nugget to sill ratio is greater than 0.5, it is worth investigating to use the UKG instead of OKG.

6. Conclusion

This study examines the performance of kriging spatial interpolation for estimating missing rainfall data in the northeast region of Thailand. The datasets used in this study reflect various characteristics during the raining season. The comparison results are that ordinary kriging provided the best estimation in the first four datasets (from May to August); whereas universal kriging provided the best estimation in the last two datasets (September and October). This study investigated into the characteristics of datasets affecting the estimation results and found that the kurtosis of observation’s histogram and the ratio of nugget to sill of fitted semivariogram model can be used as guideline to detect the trend on the datasets, which results in the appropriate selection of interpolators. This study also suggested that the relationship between altitude and rainfall in this region could not improve the performance of ordinary kriging. Furthermore, the experiments confirmed the problem of using coefficient of fit ($R^2$) to measure estimation performance since it is often misleading.
Acknowledgements

The author would like to thank Dr. Marc Souris, Computer Sciences Researcher and GIS Developer, IRD, Asian Institute of Technology (AIT), which kindly provides valuable datasets for the public at website http://www.rsgis.ait.ac.th/~souris/Thailand.htm. Also the authors would like to thank all GIS staffs at the Hydro and Agro Informatics Institute (HAII), Thailand, who have provided the expert consultations and advise on this work.

References