A Random Sampling Method in Estimating the Mean Areal Precipitation Using Kriging

Lee, Sang Il*

ABSTRACT/A new method to estimate the mean areal precipitation using kriging is developed. Unlike the conventional approach, points for double and quadruple numerical integrations in the kriging equation are selected randomly, given the boundary of area of interest. This feature eliminates the conventional approach's necessity of dividing the area into subareas and calculating the center of each subarea, which in turn makes the developed method more powerful in the case of complex boundaries. The algorithm to select random points within an arbitrary boundary, based on the theory of complex variables, is described. The results of Monte Carlo simulation showed that the error associated with estimation using randomly selected points is inversely proportional to the square root of the number of sampling points.

1. Introduction

The mean areal precipitation is defined as

$$z_A = \frac{1}{|A|} \int_A z(x, y)\, dxdy \tag{1}$$

Where $z(x, y)$ is a function defining the precipitation over the area of interest A and z_A is the mean areal precipitation. Since $z(x, y)$, a continuous function, is practically impossible to obtain, the usual procedure to evaluate equation (1) is to estimate the spatial average using point rainfall measurements at different locations (x_1, y_1), (x_2, y_2), $\cdots$, (x_n, y_n).

Various techniques have been developed, depending on how to give weights to measurements taken at different locations. Readers are referred to works by Whitmore et al. (1961), Singh and Birsoy (1986), and Yoon et al. (1991) for general description and comparison of techniques reported in the literature.

Kriging method [Matheron (1971) ; Rodriguez-Iturbe and Mejia (1974) ; Bastin et al. (1984)] has

* Assistant Research Professor, Center for Water Resources and Quality Management, Chung Buk National University, Cheong Ju, 360-763, Korea

advantages over other methods in that data are analysed in a systematic and objective way. Kriging also gives estimation error variance as well as estimation itself, which makes the method useful particularly in the design of monitoring networks and of sampling strategies.

In conventional applications of kriging to the problems of mean areal precipitation, the area of interest is discretized in a uniform fashion : The domain is subdivided into N elements (typically squares), and the coordinates of the center of each element are read for the calculation of the average rainfall estimate and the estimation error. The problems in this approach are as follows : 1) When the area of analysis is irregular in shape, discretization is not easy. 2) For different discretization either in numbers of the element or in shape, the coordinates of the center of each element must be read again, which is a time – consuming and laborious process.

In this paper, we present a new methodology for the calculation of mean areal precipitation. It is a kriging method based on the random sampling technique which requires only the coordinates of the boundary of study area. No discretization of the area, and thus the coordinates of the center of each element are required. Therefore, with the developed methodology, it is easy to automate the estimation procedure. The subject of estimation accuracy is discussed in the paper as well.

2. Estimation of Spatial Averages Using Kriging

Estimation is a procedure which uses data to infer the value of unknown quantity. Kriging may be defined as a linear minimum – variance unbiased estimation procedure. In it, the estimate of the unknown quantity is expressed as linear combinations of the measurements, i. e.,

$$\hat{z}_A = \sum_{i=1}^{n} \lambda_i z(x_i) \tag{2}$$

where $\hat{z}_A$ is the estimate of the mean areal precipitation, λ_i is the weighting coefficient, and $z(x_i)$ is the measurement at location x_i. This way, the problem is reduced to selection of a set of coefficients $\lambda_1, \cdots, \lambda_n$. The difference between the estimate $\hat{z}_A$ and the actual value z_A is the estimation error

$$\hat{z}_A - z_A = \sum_{i=1}^{n} \lambda_i z(x_i) - \frac{1}{|A|} \int_A z(x)\, dx \tag{3}$$

It is desired to select coefficients so that the estimator error becomes zero for any value of the mean m (unbiasedness), and the variance or mean square estimation error becomes minimum (minimum – variance).

Unbiasedness requires that

$$E[\hat{z}_A - z_A] = \sum_{i=1}^{n} \lambda_i m - \frac{1}{|A|} \int_A m\, dx = \sum_{i=1}^{n} \lambda_i m - m = 0 \tag{4}$$

For this condition to hold for any value of m, it is required that

$$\sum_{i=1}^{n} \lambda_i - 1 = 0 \tag{5}$$

Now, the variance of the estmation error if a covariance function can be defined is

$$E[(\hat{z}_A - z_A)^2] = E[\{\sum_{i=1}^{n} \lambda_i (z(x_i) - m) - \frac{1}{|A|} \int_A (z(x) - m) dx\}^2]$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j R(x_i - x_j) - 2 \sum_{j=1}^{n} \lambda_i \frac{1}{|A|} \int_A R(x_i - u) du + \frac{1}{|A|^2} \int_A \int_A R(u - v) du dv \tag{6}$$

Note that R is the covariance function defined as

$$E[(z(x_1) - m(x_1))(z(x_2) - m(x_2))] = R(x_1, x_2) \tag{7}$$

A basic assumption behind the derivation is that mean value is constant and the covariance function is a function not of the spatial coordinates of the two points but of the coordinates of the separation, i. e., stationarity of the random function $z(x)$:

$$E[z(x)] = m$$
$$E[(z(x_1) - m)(z(x_2) - m)] = R(x_1 - x_2) \tag{8}$$

Integral notations in equation (6) are

$$\int_A R(x_i - u) du = \iint R(x_{i1} - u_1, x_{i2} - u_2) du_1 du_2 \tag{9}$$

and

$$\int_A \int_A R(u - v) dudv = \int\int\int\int R(u_1 - v_1, u_2 - v_2) du_1 du_2 dv_1 dv_2 \tag{10}$$

Thus the problem of best (minimum mean-squared error) unbiased estimation of the λ coefficients may be reduced to the constrained optimization problem : Select the values of $\lambda_1, \cdots, \lambda_n$ which minimizes equation (6) subject to the constraint (5).

Using Lagrange multipliers, the necessary conditions for the minimization are given by the linear "kriging system" equations :

$$\sum_{i=1}^{n} \lambda_j R(x_i - x_j) + \nu = \frac{1}{|A|} \int_A R(x_i - u) du, \qquad i = 1, \cdots, n \tag{11}$$

$$\sum_{i=1}^{n} \lambda_i = 1 \tag{12}$$

– where ν is the Lagrange multiplier. Once coefficients $\lambda_1, \cdots, \lambda_n$ are obtained, the mean areal precipitation is calculated through equation (2). The mean square estimation error is given as

$$E[(\hat{z}_A - z_A)^2] = -\nu - \sum_{i=1}^{n} \lambda_i \frac{1}{|A|} \int_A R(x_i - u)\,du + \frac{1}{|A|^2} \int_A \int_A R(u - v)\,dudv \tag{13}$$

If the function is intrinsic rather than stationary, a variogram is used instead of a covariance function. It is well known that the relations which are valid for a stationary case are also valid for a more general intrinsic case where we replace $R(h)$ by $-\gamma(h)$, where $\gamma(h)$ is the semi – variogram, in equations (11) and (13) [Matheron, 1971]. The definition of a semi – variogram is

$$\gamma(h) = \frac{1}{2} E[(z(x_1) - z(x_2))^2] \tag{14}$$

As seen above, the estimation of the kriging coefficients and the estimation error involves the calculation of the double and quadruple integrals. Conventional approach to calculate these integrals is to divide the total area A into subareas $A_1, \cdots, A_N$. Each area may be represented by a point u (e. g, the center of the area.) Then :

$$|A| = \sum_{k=1}^{N} A_K \tag{15}$$

$$\frac{1}{|A|} \int_A R(x_i - u)\,du = \frac{1}{|A|} \sum_{k=1}^{N} R(x_i - u_k) A_K \tag{16}$$

$$\frac{1}{|A|^2} \int_A \int_A R(u - v)\,dudv = \frac{1}{|A|^2} \sum_{k=1}^{N} \sum_{j=1}^{N} R(u_k - v_l) A_K A_l \tag{17}$$

3. Random Sampling Method

When subareas of equation (15) are represented by points u, one must discretize the domatin externally to the kriging algorithm. Even though each area can be any shape, squares are most frequently used for they are easy to calculate the location of the center. Different number and shape of subareas would result in different estimates of areal precipitation and estimation errors because representative points affect the calculation of integrals in equations (16) and (17).

A different approach to choose points representing the domain is to select points randomly. Given the geometry of the area, the issue is to determine whether a randomly selected point is within the boundary of the area or not. The only information we have about the geometry of the area is the coordinates of the vertices of the polygon which approximate the boundary of the domain. The problem of determining whether a point is within the boundary or not

is resolved using the concept of "branch" in the theory of complex variables [Hildebrand, 1976].

Consider a function f(w) defined on the plane of complex variable w. Suppose further that w makes a complete circuit (counterclockwise) around the origin w = 0 starting from point S (see Figure 1). It is well known that when the function f(w) has the same values for the arguments $\theta = \theta_1 + 2k\pi, (k = 0, \pm 1, \pm 2, \cdots)$ we call the difference of phase angles for the function 2π. The positive axis behaves as a branch cut, and w = 0 a branch point.

Table 1 summarizes the algorithm to select a point and to determine whether it is whthin the boundary or not. The area within which the mean areal precipitation needs to be calculated can be modeled as a polygon and the boundary of the area as a circuit. Suppose a point u is selected randomly. Relative positions between the vertices of the polygon and the point u can be calculated from the absolute coordinate system. Those relative positions constitute a vector of complex variables. Argument vector consisting of the angles of complex variables are easily obtainable. Remember that the difference of arguments of a complex variable on two consecutive branches is 2π. Therefore, if the phase angle difference of a vector from u to a vertex on two consecutive branches (for example, k = 0 and k = 1) is 2π, the point u is a branch point, meaning that it is within the boundary. If the argument difference has the value other than 2π, the point u is outside of the boundary.

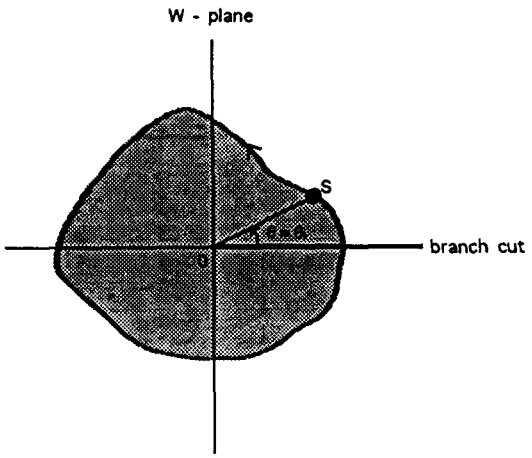

Figure 1. Phase angle and branch cut on a complex plane

Table 1. Algorithm for selecting a random point inside a boundary

1. Enter coordinates of vertices $x(i), y(i), i = 1, \cdots, n$
2. Pick a point u randomly.
3. Construct $\vec{v} = \begin{bmatrix} x(1)-x_u \\ \vdots \\ x(n)-x_u \\ \cdots\cdots \\ x(1)-x_u \end{bmatrix} + i \begin{bmatrix} y(1)-y_u \\ \vdots \\ y(n)-y_u \\ \cdots\cdots \\ y(1)-y_u \end{bmatrix}$
4. Calculate argument vector $\vec{p}$ of $\vec{v}$.
5. Find out min. $p(i)$ and let
$$\vec{q} = [p_{min}, \cdots, p_{min}]^T.$$
6. $\vec{r} = \text{mod}(\vec{p} - \vec{q}, 2\pi) + \vec{q}$
7. $\vec{b} = [r(1) : r(2) - r(1), \cdots, r(n+1) - r(n)]^T$
8. Construct $\vec{c}$ and $\vec{d}$ whose elements are
$$\vec{c} = \begin{cases} -1 \text{ if } b(i) > \pi \\ 0 \text{ if } b(i) \le \pi \end{cases}$$
$$\vec{d} = \begin{cases} 1 \text{ if } b(i) < -\pi \\ 0 \text{ if } b(i) \ge -\pi \end{cases}$$
9. $\vec{e} = (\vec{c} + \vec{d}) * 2\pi$
10. $\vec{f} = [e(1), e(1) + e(2), \cdots, e(1) + \cdots + e(n+1)]^T$
11. $\vec{g} = \vec{p} + \vec{f}$ and $\delta = g(n+1) - p(1)$
12. u is $\begin{cases} \text{inside if } \delta = 2\pi \\ \text{outside if } \delta \neq 2\pi \end{cases}$

4. Application

Let's consider an example studied in Kitanidis [1989] (see Figure 2). Absolute reference frame is located at the left bottom. The location of the reference frame is arbitrary. There are four raingage stations whose coordinates and point rainfall measurements are tabulated in Table 2. We consider that rainfall exhibits variability at a scale comparable to the distance between stations, h. Assume that

$$\gamma(h) = \begin{cases} 0 & \text{if } |h| = 0 \\ 1 + |h| & \text{if } |h| > 0 \end{cases} \tag{18}$$

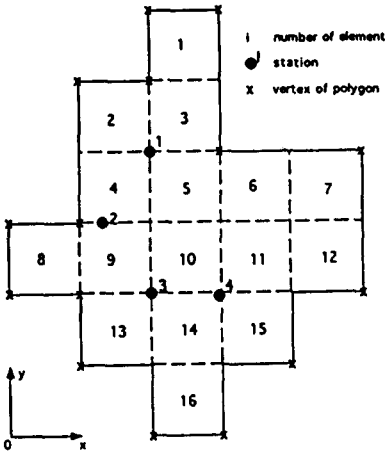

Figure 2. Study area with raingage stations

Table 2. Locations of raingage stations and rainfall measurements

Station	x (km)	y (km)	Rainfall(mm)
1	5	10	7.6
2	3.5	7.5	4.5
3	5	5	3.0
4	7.5	5	14.5

4.1 Uniform Sampling

In the conventional approach the area is divided into subareas. In this example, the area has 16 squares with sides equal to 2.5km. Thus $|A_i| = 6.25$. The coordinates of the center of each element are given in Table 3.

Table 3. The coordinates of the center of each element

Element i	u_1 (km)	u_2 (km)	Element i	u_1 (km)	u_2 (km)
1	6.25	13.75	9	3.75	6.25
2	3.75	11.25	10	6.25	6.25
3	6.25	11.25	11	8.75	6.25
4	3.75	8.75	12	11.25	6.25
5	6.25	8.75	13	3.75	3.75
6	8.75	8.75	14	6.25	3.75
7	11.25	8.75	15	8.75	3.75
8	1.25	6.25	16	6.25	1.25

First, compute the multiple integrals :

$$\frac{1}{|A|}\int_A \gamma(x_i-u)\,du = \frac{1}{16}\sum_{k=1}^{16}\gamma(x_i-u_k) = \begin{cases} 5.75 \\ 5.66 \\ 5.45 \\ 5.39 \end{cases}$$

$$\frac{1}{|A|^2}\int_A\int_A \gamma(u-v)\,dudv = \frac{1}{16^2}\sum_{k=1}^{16}\sum_{j=1}^{16}\gamma(u_k-v_j) = 6.21$$

Then form the kriging system of equations in equation (11).

$$\begin{pmatrix} 0. & -3.915 & -6.000 & -6.590 & 1. \\ -3.915 & 0. & -3.915 & -5.717 & 1. \\ -6.000 & -3.915 & 0. & -3.500 & 1. \\ -6.590 & -5.717 & -3.500 & 0. & 1. \\ 1. & 1. & 1. & 1. & 0. \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \lambda_4 \\ \nu \end{pmatrix} = \begin{pmatrix} -5.75 \\ -5.66 \\ -5.45 \\ -5.39 \\ 1. \end{pmatrix}$$

Solving the above matrix equation, we obtain the weighting coefficients as $\lambda_1 = 0.31$, $\lambda_2 = 0.16$, $\lambda_3 = 0.19$, $\lambda_4 = 0.34$. The Lagrange multiplier is $\nu = -1.76$. Equation (2) gives the best estimate as 8.596. Equation (11) yields the mean square estimation error as 1.1063.

4.2 Random Sampling

Now, let's apply the methodology developed in the previous section to select points needed for the calculation of integrals. Table 4 lists coordinates of vertices of the area in Figure 2 and one sample set of points selected randomly by computer. For comparison purposes, the number of points were chosen to be 16. One can enter the coordinates of vertices starting from any vertex as long as they are entered counterclockwise.

Table 4. The coordinates of vertices and a sample set of random points

	Vertex		Random points	
No.	$x(km)$	$y(km)$	$u_1(km)$	$u_2(km)$
1	5.0	0.0	6.99	9.78
2	7.5	0.0	7.48	2.32
3	7.5	2.5	5.12	8.93
4	10.0	2.5	8.39	9.20
5	10.0	5.0	7.21	10.89
6	12.5	5.0	7.02	6.42
7	12.5	10.0	5.70	4.99
8	7.5	10.0	4.44	9.99
9	7.5	15.0	6.22	4.19
10	5.0	15.0	7.16	3.19
11	5.0	12.5	6.64	14.58
12	2.5	10.0	5.27	8.58
13	2.5	7.5	8.87	2.72
14	0.0	7.5	3.86	8.60
15	0.0	5.0	8.31	8.44
16	2.5	5.0	9.19	3.88
17	2.5	2.5		
18	5.0	2.5		

The multiple integrals using 16 randomly selected points u are

$$\frac{1}{|A|}\int_A \gamma(x_i-u)\,du = \frac{1}{16}\sum_{k=1}^{16}\gamma(x_i-u_k)$$

$$= \begin{Bmatrix} 5.20 \\ 5.54 \\ 5.24 \\ 4.89 \end{Bmatrix}$$

$$\frac{1}{|A|^2}\int_A\int_A \gamma(u-v)\,dudv = \frac{1}{16^2}\sum_{k=1}^{16}\sum_{j=1}^{16}$$

$$\gamma(u_k-v_j) = 6.39$$

The kriging system of equations becomes

$$\begin{bmatrix} 0. & -3.915 & -6.000 & -6.590 & 1. \\ -3.915 & 0. & -3.915 & -5.717 & 1. \\ -6.000 & -3.915 & 0. & -3.500 & 1. \\ -6.590 & -5.717 & -3.500 & 0. & 1. \\ 1. & 1. & 1. & 1. & 0. \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \lambda_4 \\ \nu \end{bmatrix} = \begin{bmatrix} -5.20 \\ -5.54 \\ -5.24 \\ -4.89 \\ 1. \end{bmatrix}$$

The weighting factors obtained are $\lambda_1 = 0.36$, $\lambda_2 = 0.10$, $\lambda_3 = 0.16$, $\lambda_4 = 0.38$ while the Lagrange multiplier $\nu = -1.36$ is resulted. Equations (2) and (13) give the best estimate as 9.13 and the mean square estimation error as 1.0871, respectively. It is interesting to note that the estimation error of the random sampling method is less than that of the uniform sampling. It is because the set of random samples chosen in this case represents a smaller area than the case of uniform sampling (see Figure 3). A different set of sampling points would result in different values of estimates and estimation error.

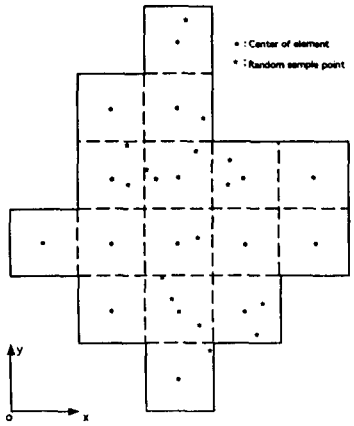

Figure 3. Two ways of selecting representative points : uniform sampling and random sampling

5. Error Analysis

The accuracy of estimation in the uniform sampling method is known to be inversely proportional to the number of subarea N, i. e., $\hat{z}_A - z_A \propto \frac{1}{N}$ [Journel and Huibregts, 1978]. In section 4.2 we have seen that the random sampling method results in different estimates and estimation errors depending on the locations of selected points. Therefore, a statistical approach must be made in order to investigate the

accuracy of estimation. We adopted Monte Carlo simulation approach.

In Monte Carlo simulation, different "realizations" of area are generated in computer, given the boundary and the number of points for integration. Outcomes of kriging using many realizations of random points, rather than a single realization, are interpreted in some average sense.

Simulations were conducted for the same example studied in the previous section. Twenty realizations were generated for a fixed number of random samples. The number of sample points were varied ranging from 5 to 55. The statistics of mean areal precipitation estimates and their errors are tabulated in Table 5.

The averages of estimation errors are plotted in Figure 4 as a function of the number of integration points. As seen in the figure, the mean square estimation error of random sampling method is inversely proportinal to the number of sample points. In other words, the error associated with estimation is inversely proportional to the square root of the number of sample points :

$$\hat{z}_A - z_A \propto \frac{1}{\sqrt{N_S}} \tag{19}$$

where N_S is the number of sampling points. This observation agrees with the basic theorem of Monte Carlo integration [Press et al., 1986]. The integral of a function f over the multidimensional volume V can be written as

$$\int f dV \approx V\langle f\rangle \pm V\sqrt{\frac{\langle f^2\rangle - \langle f\rangle^2}{N_S}} \tag{20}$$

where the angle brackets denote taking the arithmatic mean over the N_S sample points,

$$\langle f\rangle = \frac{1}{N_S}\sum_{i=1}^{N_S} f(x_i) \qquad \langle f^2\rangle = \frac{1}{N_S}\sum_{i=1}^{N_S} f^2(x_i) \tag{21}$$

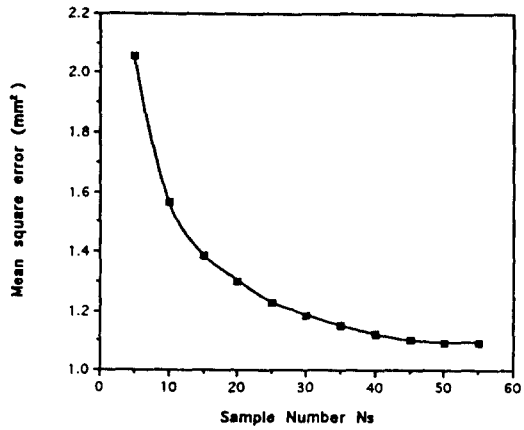

Figure 4. Mean square error decrease with the number of random sampling points

The "plus – or – minus" term in equation (20) is a one standard deviation error estimate for the integral.

Table 5. Statistics of mean areal precipitation estimates and mean square errors (MSE). Twenty Monte Carlo simulations were conducted using random sampling method.

Sample points	Average of estimates	Average of MSE
5	8.71	2.056
10	8.63	1.562
15	8.79	1.384
20	8.76	1.300
25	8.87	1.226
30	8.76	1.185
35	8.61	1.148
40	8.63	1.121
45	8.65	1.099
50	8.57	1.092
55	8.66	1.091

6. Conclusions

Various techniques are available to estimate the mean areal precipitation over the area of interest from point precipitation measurements. In this paper, a new methodology based on kriging was developed to estimate the mean areal precipitation and the associated estimation error. The followings are the major conclusions drawn from this work :

(1) Compared to the conventional approach where the points for integration in kriging equations are uniformly selected, the developed method selects points randomly.

(2) The developed method has advantages in that only the coordinates of the boundary are required, not the coordinates of the center of subarea as in the conventional kriging. This feature makes the method more useful for the area with complex boundaries in which subdivision of the area might be inefficient and inaccurate.

(3) The accuracy of estimation using random sampling is inversely proportional to the square root of the number of sample points whereas the accuracy of estimation using uniform sampling is inversely proportional to the number of subareas.

Acknowledgment

The author wishes to thank Professor Peter Kitanidis of Stanford University for his contributions to this research. The support from Hyundai Electronics Industries Co., Ltd. is gratefully acknowledged.

References

1. Batsin G., Lorent B., Duque C., and Gevers M., (1984) "Optimal estimation of the average areal rainfall and optimal selection of rain gage location," *Water Resources Research*, Vol. 20, No. 4, pp. 463 - 470.
2. Hildebrand, F. B., (1976) *Advanced calculus for applications* (2nd Ed.), Prentice - Hall, Inc., Englewood Cliffs, New Jersey.
3. Journel, A. G., and Huibregts C. J., (1978) *Mining Geostatistics*, Academic Press, London.
4. Kitanidis, P. K., (1989) *Estimation of spatial functions and predictive groundwater modeling*, Class Notes, Stanford University.
5. Matheron, G., (1971) *The theory of regionalized variables and its applications*, Les Cahiers du Centre de Morphologie Mathematique de Fontainebleau, France.
6. Press, W. H., Flannery B. P., Teukolsky, S. A., and Vetterling, W. T, (1986) *Numerical Recipes*, Cambridge University Press.
7. Rodriguez - Iturbe I . and Mejia J. M., (1974) "On the transform of point to areal rainfall," *Water Resources Research* Vol. 10, No. 4, pp. 729 - 735.
8. Singh V. P., and Birsoy Y. K., (1986) "Comparison of the methods of estimating mean areal rainfall," *Water Resources Bulletin*, Vol. 22, No. 2, pp. 275 - 282.
9. Whitmore J. S., Van Efden F. J., and Harvey K. J., (1961) "Assessment of average annual rainfall over large catchments," Inter - African Conference on Hydrology, C. C. T. A. Publication 66, pp. 100 - 107.
10. Yoon K. H., Kim W., and Cheong S. D., (1991) *An analysis of the areal interpolation techniques and regional variability of the rainfall by areal reduction factor*, Korea Institute of Construction Technology Report 91 - WR - 113.