Correlated versus Uncorrelated Hydrologic Samples

By Hugo A. Loaiciga[1] and Paul F. Hudak[2]

Introduction

The application of statistical analysis to spatial data collection and experimental design is an area of growing interest in engineers and earth scientists (Gilbert 1987; Loaiciga and Mariño 1988). The language and practice of statistics continues to gain acceptance among hydrologists and earth scientists given the uncertainty and at times, the inherent statistical nature of natural phenomena. For example, in subsurface environmental problems it is now common practice to report field survey results in terms of confidence intervals, tests of hypotheses, or probabilistic statements rather than using point, deterministic descriptions (McBean et al. 1988). One parameter of widespread interest in spatial data is the mean or average, examples of which are the mean precipitation in a watershed or the mean concentration of a solute in groundwater.

The classical statistical inference about the mean assumes independent and identically distributed data (Cochran 1977). However, when dealing with spatial data, it is important to consider the possibility of spatial correlation and its effects on statistical inferences about the mean. One interesting result of such consideration is the relationship between the optimal sample sizes needed to estimate the population mean of a spatial variable under the independent and the correlated assumptions. The effect of spatial correlation on statistical inference and optimal sample size for population-mean determination is addressed next.

Sample Sizes for Mean Estimation

Suppose that there are n spatially distributed sampling locations where observations z_i, $i = 1, 2, 3, \ldots, n$, are collected. Assume that the observations are independent and identically distributed (IID) and drawn from a normal population. The population mean and variance are μ and ν^2, respectively. A 100(1-α)% confidence interval (CI) for the population mean is given by

$$CI = \frac{\bar{z} \pm z_{\alpha/2}\nu}{n^{1/2}} \quad (1)$$

where α = the significance level (α is usually set equal to 0.05); $\bar{z}_{\alpha/2}$ = a standard normal variate (when $\alpha = 0.05$, $z_{\alpha/2} = 1.96$); and $\bar{z}$ = the arithmetic sample mean. According to Eq. 1, the tolerable error, E, committed in es-

[1]Asst. Prof., Dept. of Geography, Univ. of California, Santa Barbara, CA 93106.
[2]Grad. Student, Dept. of Geography, Univ. of California, Santa Barbara, CA.

Note. Discussion open until February 1, 1990. To extend the closing date one month, a written request must be filed with the ASCE Manager of Journals. The manuscript for this paper was submitted for review and possible publication on January 16, 1989. This paper is part of the ***Journal of Water Resources Planning and Management***, Vol. 115, No. 5, September, 1989. ©ASCE, ISSN 0733-9496/89/0005-0699/$1.00 + $.15 per page. Paper No. 23897.

J. Water Resour. Plann. Manage., 1989, 115(5): 699-705

timating the population mean μ by $\bar{z}$ is

$$E = \frac{z_{\alpha/2}\nu}{n^{1/2}} \quad \ldots\ldots (2)$$

If E is specified, then, given ν^2 and α, the sample size needed to achieve the specified error is

$$n^* = \frac{z^2_{\alpha/2}\nu^2}{E} \quad \ldots\ldots (3)$$

The applicability of Eqs. 1–3 within a hydrologic context is easily illustrated. Suppose that a large storm of a cyclonic, frontal nature affects a relatively flat basin (a situation typically found along coastal valleys in California). Under such hydroclimatic conditions, and given that precipitation measurements are uncorrelated (i.e., distances between gauges exceed the correlation scale of the precipitation field), the average basin precipitation may be characterized statistically by Eqs. 1 and 2. Another example is the mean solute concentration obtained by sampling wells that are separated by distances that exceed the correlation scale of the spatial concentration distribution. The statistical description of the mean concentration is then given by Eqs. 1 and 2 and the optimal sample size by Eq. 3.

Next, suppose that n' observations are identically distributed with mean μ and variance ν^2 drawn from a normal population, as before, but now assuming that they are spatially correlated with the covariance between the observations at locations i and j, v_{ij}, given by

$$v_{ij} = \nu^2 r_{ij} \quad \ldots\ldots (4)$$

where r_{ij} denotes the correlation between the observations at locations i and j. We seek to determine the 100(1-α)% confidence interval for the population mean taking into account the correlation between sampling sites. The variance of the sample mean when the observations are correlated is given by

$$v(\bar{z}) = \left(\frac{1}{n'}\right)^2 \nu^2 \left(n' + 2\sum_{i=1}^{n'}\sum_{j>i}^{n'} r_{ij}\right) \quad \ldots\ldots (5)$$

from which it follows that the 100(1-α)% confidence interval for the population mean is given by

$$CI = \bar{z} \pm z_{\alpha/2} v(\bar{z})^{1/2} \quad \ldots\ldots (6)$$

The tolerable error incurred in estimating the population mean by the sample mean is given by

$$E' = z_{\alpha/2} v(\bar{z})^{1/2} \quad \ldots\ldots (7)$$

Eqs. 6 and 7 are of general applicability to characterize hydrologic phenomena with a spatial structure. The hydrologic examples previously cited can be extended to the more general use of spatially correlated precipitation and solute concentration measurements and represent important applications of the statistical theory embodied by Eqs. 6 and 7.

For a given variance ν^2, correlation r_{ij}, significance level α, and specified tolerable error E', there is a unique sample size n' that would achieve the target error E'. That sample size is determined by Eq. 7. Suppose that the

errors E and E' of Eqs. 2 and 7, respectively, are set equal to each other. What is the relationship between the sample sizes n^* (Eq. 3) and n' (implied by Eq. 7) that would achieve the same tolerable error, E? The relationship is obtained by equating Eqs. 2 and 7, using Eq. 5, to yield

$$n^* = \frac{(n')^2}{n' + 2\sum_{i=1}^{n'}\sum_{j>i}^{n'} r_{ij}} \quad \ldots\ldots (8)$$

Eq. 8 means that if n^* normal IID observations are used to estimate the population mean by the sample mean, then the tolerable error would be the same as that obtained from n' normal correlated variates. It is important to understand the assumptions behind Eq. 8: (1) The observations, independent or correlated, must be normally distributed; and (2) the population variance or covariance must be known, rather than estimated, for Eq. 8 to hold. In the case of independent samples, assumption 2 could be relaxed when deriving Eq. 2 and the estimated standard deviation would replace v (see Eq. 2) in that equation whereas the t statistic with n-1 degrees of freedom would replace the standard normal variate. However, neither assumptions 1 nor 2 can be relaxed in obtaining Eq. 7, thereby restricting the conditions under which Eq. 8 is applicable. [The reader is referred to Loaiciga et al. (1988) for a treatment of covariance estimation.]

Further Analysis of the Correlation Effect

It is not obvious from Eq. 8 whether n^* is less or greater than n'. Intuitively, one may expect that the "effective" information content of n^* IID observations should exceed that of n^* correlated variables, because in the latter each observation is partly explained by the other variables that are correlated to it. In contrast, each IID observation is a separate information unit by itself, which neither explains nor is explained by other independent variates. Therefore, this heuristic analysis indicates that in order to achieve the same estimation error, the size of the correlated sample must exceed that of the independent sample:

$$n' > n^* \quad \ldots\ldots (9)$$

The previous analysis obviously applies only when the spatial correlation is positive, which is typically the case of interest in spatial variables describing geophysical phenomena. However, there is no need to speculate about the relative magnitudes of n^* and n'. Their relative magnitudes can be obtained from Eq. 8 by setting r_{ij} equal to zero or equal to one, which correspond to the uncorrelated and perfectly correlated cases, respectively. The result is

$$r_{ij} = 0, \quad \text{implies } n^* = n' \quad \ldots\ldots (10a)$$

$$r_{ij} = 1, \quad \text{implies } n^* = 1 \quad \ldots\ldots (10b)$$

Eqs. 10*a–b* indicate that if there is no correlation, n^* and n' are equal—an obvious result. On the other hand, if there is perfect correlation, then knowing one observation value means that all other values are also known (and equal to the observed one), meaning that the effective sample size is one regardless of the total number of observations. For r_{ij} between zero and one,

TABLE 1. Values of n^* for Spherical Correlation Model and Square Sample Pattern

	L^a											
n' (1)	0.10 (2)	0.50 (3)	1.00 (4)	2.00 (5)	3.00 (6)	4.00 (7)	5.00 (8)	6.00 (9)	7.00 (10)	8.00 (11)	9.00 (12)	10.00 (13)
4	4.00	4.00	4.00	2.30	1.68	1.45	1.34	1.27	1.22	1.19	1.16	1.15
9	9.00	9.00	9.00	4.41	2.68	2.01	1.71	1.54	1.44	1.36	1.31	1.27
16	16.00	16.00	16.00	7.28	4.10	2.81	2.21	1.89	1.70	1.57	1.48	1.42
25	25.00	25.00	25.00	10.88	5.89	3.86	2.87	2.33	2.02	1.82	1.68	1.58
36	36.00	36.00	36.00	15.23	8.02	5.12	3.69	2.90	2.43	2.12	1.92	1.78
49	49.00	49.00	49.00	20.31	10.51	6.58	4.66	3.58	2.92	2.49	2.21	2.01
64	64.00	64.00	64.00	26.13	13.35	8.25	5.75	4.35	3.49	2.93	2.54	2.27
81	81.00	81.00	81.00	32.69	16.53	10.11	6.98	5.22	4.14	3.43	2.94	2.59
100	100.00	100.00	100.00	39.98	20.06	12.17	8.34	6.19	4.86	3.98	3.38	2.94

[a]Correlation length scale (L) in multiples of a, the grid separation in a square sampling pattern.

n^* varies from n' to one. The exact correspondence between n^* and n' depends on (1) The correlation model; and (2) the geometrical distribution of sampling sites (e.g., rectangular, triangular, random spatial arrangement of sampling sites). Consider, for example, an isotropic, spherical correlation model:

$$r_{ij}(h_{ij}) \doteq 1 - \frac{3}{2}\frac{h_{ij}}{L} + \frac{1}{2}\frac{h_{ij}^3}{L^3}, \quad h_{ij} < L \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (11a)$$

$$r_{ij}(h_{ij}) = 0, \quad h_{ij} \geq L \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (11b)$$

where L = the correlation length scale; and h_{ij} = the distance between any two sampling points, i and j. For a given number of sampling sites n', and assuming that the spatial correlation follows Eqs. 11*a–b*, Eq. 8 would yield the independent sample size n^*. Table 1 lists n^* values resulting from values of n' ranging from 4 to 100, arranged in a square sample pattern (see Fig. 1). For $L > a$ (a being the minimum unit in the square grid), n^* decreases as L increases.

The relationship between n^* and n' is not continuous for a square configuration of points (e.g., one cannot lay out a square grid with three or five points). However, by slightly rearranging the information contained in Table 1, the continuous dependence of n^* on the correlation scale (L) for a given number of grid points (n') can be depicted (see Fig. 2), revealing interesting behavior of n^*. It is seen in Fig. 2 that regardless of the value of n', as the correlation scale becomes large (i.e., in Fig. 2, $L = 10a$), n^* converges rapidly to the asymptotic value of one. Fig. 2 can be useful for network design. Suppose that a given error of estimation, E, of the population (e.g., precipitation) mean is specified for a square grid, and that a spherical correlation is adequate. Assuming independent observations, Eq. 3 would provide the number of independent sites n^* corresponding to the error E. Suppose that, for the sake of argument, $n^* = 40$. Entering Fig. 2 with that value of n^*, it is seen that if the network of sampling points (i.e., the grid) is designed so that $L = 2a$, then exactly 100 sampling (correlated) sites are

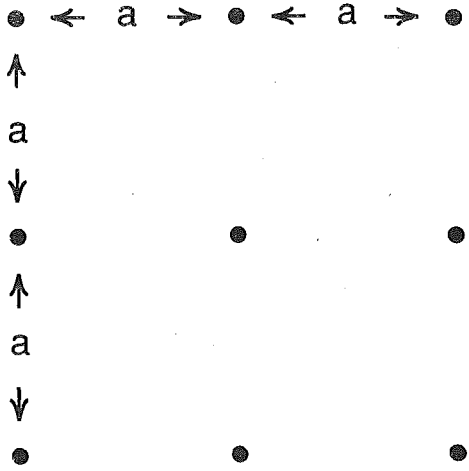

FIG. 1. Square Sample Pattern Layout; Dots Represent Sampling Locations

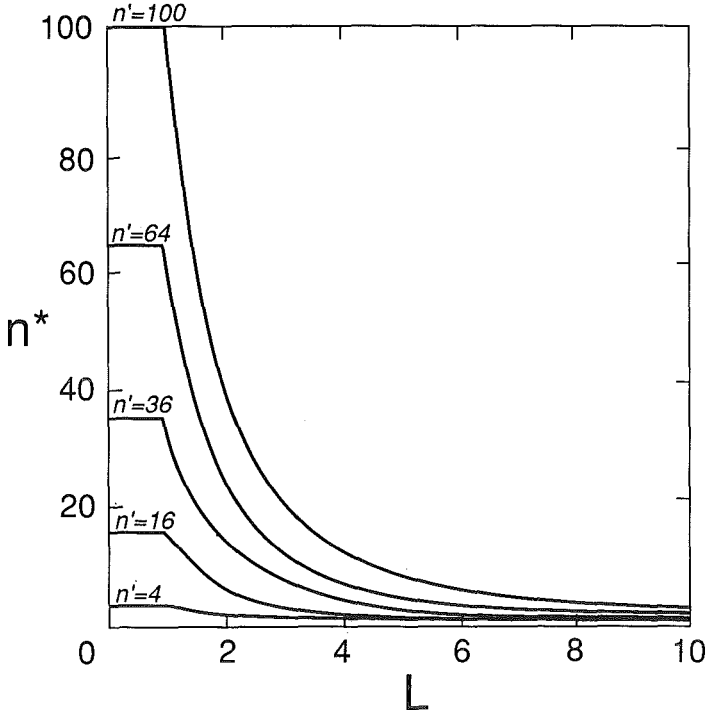

FIG. 2. Graphs of n^* versus L for $n' = 4$, 16, 36, 64, and 100

needed. Additional simulations for other network configurations (triangular, hexagonal) and correlation models (exponential) indicated the same qualitative behavior.

A simplification of Eq. 8 occurs when the correlation r_{ij} is replaced by a constant, average correlation coefficient $\bar{r}$ defined as follows:

$$\bar{r} = \left(\frac{1}{N}\right) \sum_{i=1}^{n'} \sum_{j>i}^{n'} r_{ij} \quad \ldots\ldots (12)$$

in which

$$N = \frac{n'(n' - 1)}{2} \quad \ldots\ldots (13)$$

is the number of distances between pairs of sampling sites (e.g., if there are four sampling sites, then there will be six connecting distances among points). Substitution of Eqs. 12 into Eq. 8 yields

$$n^* = \frac{n'}{1 + \bar{r}(n' - 1)} \quad \ldots\ldots (14)$$

in terms of an average correlation. Eq. 14 was first reported by Matalas and Benson (1961) in the hydrologic literature. Those authors, however, did not present a derivation of Eq. 14 nor did they elaborate on it's relationship to the more general case of Eq. 8, the parent equation. This note provides the statistical context of Eqs. 8 and 14 and indicates their relevance to hydrologic applications.

Conclusions

This work has derived a general relationship describing the equivalence of independent and correlated samples used to estimate the population mean. The relationship is useful in experimental design of data collection of spatial data. Eq. 8 indicates that ignoring spatial correlation would lead to the use of a smaller sample size than that actually required to achieve a specified level of accuracy, thus leading to err on the unsafe side. On the other hand, if the correlation length scale (L) of a spatial variable is known, one could design a sampling scheme that would yield an IID sample and thus introduce a substantial simplification in statistical inference about the mean. A simplification of Eq. 8 in terms of the "average" correlation is given by Eq. 14. The derivation of Eqs. 8 and 14 appears to be novel. This note provides the derivation and expands on the relevance of correlated versus uncorrelated samples in hydrologic statistical inference and sampling network design.

Acknowledgment

The motivation for writing this paper derived from a conversation with John L. Wilson at the 1988 American Geophysical Union Spring Meeting.

Appendix. References

Cochran, W. G. (1977). *Sampling techniques,* 3rd ed., John Wiley, New York, N.Y.
Gilbert, R. O. (1987). *Statistical methods for environmental pollution monitoring.*

Van Nostrand–Reinhold, New York, N.Y.
Loaiciga, H. A., and Mariño, M. A. (1988). "Fitting minima of flows via maximum likelihood." *J. Water Resour. Plng. and Mgmt.*, ASCE, 114(1), 78–90.
Loaiciga, H. A., Shumway, R. H., and Yeh, W. (1988). "Linear spatial interpolation analysis with an application to the San Joaquin Valley, California." *Stochastic Hydrol. and Hydr.*, 2(2), 113–136.
Matalas, N. C., and Benson, M. A. (1961). "Effect of interstation correlation on regression analysis." *J. Geophys. Res.*, 66(10), 3285–3293.
McBean, E. A., Kompter, M., and Rovers, F. (1988). "A critical examination of approximations implicit in Cochran's Procedure." *Ground Water Monitoring Rev.*, winter, 83–87.