The Philippine Statistician, 1999
Vol. 48, Nos. 1-4, pp. 1-16

A Cost-Effective Sampling Design For Multivariable Water Quality Monitoring

Emma P. Obalan[1] and Felino P. Lansigan[2]

ABSTRACT

A study was conducted to develop a method of determining the optimal sampling design parameters for a multivariable water quality monitoring program for a freshwater lake. The design parameters included the combination of frequency of sampling, number of stations, and number of replicates per sampling occasion and location. A geostatistical procedure called kriging, and an ordination method like correspondence analysis can be used together with cluster analysis to stratify the lake into homogeneous zones. The optimal combination of sampling design parameters was determined using a multivariable optimization approach and the principle of modified gradient search which yields the set of parameters with the highest minimum power attained by the design variables under a fixed budget, and also the combination that yields the least cost for a prescribed statistical power.

Keywords: Sampling design, kriging, correspondence analysis, cluster analysis, power, modified gradient search.

1. INTRODUCTION

Water quality monitoring is now an integral part of the natural resources management efforts to help save, prevent, or check deterioration of the quality of water resource systems such as lakes, rivers, and seas since these are often used as recipients of many industrial effluents, and domestic wastes. One objective of a water quality monitoring program for a freshwater lake is the detection of change in the water quality level. A water quality monitoring program requires a cost-effective sampling design for water quality data acquisition which considers the tradeoff between the cost of data collection and the statistical reliability. A cost-effective sampling design for the monitoring program is such that the two types of statistical errors (type I and type II errors) are minimized under a limited budget; alternatively, cost is minimized under maximum allowable statistical errors. Committing a type I error is expensive on the part of the monitoring agency because it requires action when none is actually needed. On the other hand, a type II error is detrimental to the environment because no action is taken when it is actually needed.

Two important activities in developing a cost-effective sampling design for multivariable water quality monitoring is the stratification of the lake into homogeneous zones, and the determination of the optimal sampling design parameters i.e. the combination of frequency of sampling, number of stations and number of replicates per sampling occasion and location.

2. DETERMINATION OF HOMOGENEOUS ZONES

In stratifying the lake into homogeneous zones in a situation with only few sampling points, one approach is to first interpolate values at other points of the lake using geostatistical methods like kriging. Keckler (1994) recommended the use of kriging with linear semi-variogram and zero nugget effect when sampling points are very few since kriging-based estimator is a linear unbiased estimator with minimum variance.

[1] Instructor, Department of Mathematics, Mindanao State University, 9700 Marawi City.
[2] Professor, Institute of Statistics (INSTAT), University of the Philippines Los Baños, 4031 College, Laguna.

The data set of measurements and interpolated values can then be transformed into a new set of independent variates by correspondence analysis (Lachance, et al, 1979). In order to make the analysis more stable, data on each water quality variable are first divided into equiprobable classes. The data are transformed into 0's or 1's depending on whether the observation of a station point belongs to that class or not. From the results of correspondence analysis, the scores obtained from the main factorial axes can be used in conducting cluster analysis to group the station points with similar characteristics. Transforming the data into a new set of independent variates makes the cluster analysis unbiased.

3. STATISTICAL MODEL FOR DETECTING SIGNIFICANT CHANGE

Assuming that observations on the water quality variables are taken at various sites or locations of the lake before and after the intervention into the ecosystem, a possible model for the data is (Millard and Lettenmaier, 1986)

$$Y_{jklmn} = \mu + A_j + B_k + C_l + AB_{jk} + AC_{jl} + BC_{kl} + ABC_{jkl} + \eta_{jklm} + \varepsilon_{jklmn} \qquad \text{Model 1}$$

where Y_{jklmn} = the observation for the n^{th} replicate on the m^{th} sampling occasion within season l at station j for event status k ; μ = the overall mean of observations over stations, event status, and seasons; A_j = main effect of station j (fixed); B_k = main effect of event status at level k (fixed); C_l = main effect of season l (fixed); AB_{jk} = station by event status interaction effect at level jk (fixed); AC_{jl} = station by season interaction effect at level jl (fixed); BC_{kl} = event status by season interaction effect at level kl (fixed); ABC_{jkl} = station by event status by season interaction effect at level jkl (fixed); η_{jklm} = effect due to sampling occasion m within season l for station j and event status k (random); ε_{jklmn} = random error in the n^{th} replicate on the m^{th} sampling occasion within season l at station j for event status k (random); $j = 1, 2, \ldots, J$ (stations); $k = 1, 2$ (event status index with 1 = pre-event and 2 = post-event); $l = 1, \ldots, L$ (seasons); $m = 1, \ldots M$ (sampling occasions); and $n = 1, \ldots N$ (replicates).

The model suggests that each observation is affected by the particular station and season, and whether the event or intervention has already occurred or not. Within each season, there are M sampling occasions that occur over all years of sampling for each level of event or project status. For convenience, let M_1 denotes the number of sampling occasions per season per year (fixed), and M_2 denotes the number of years of sampling for each level of event status (variable). Hence, $M = M_1 . M_2$.

The effect, denoted by η_{jklm}, represents the variability in the observations between sampling occasions in a season over all stations after consideration of event status. These are assumed to be independent random variables distributed normally with mean 0 and variance σ_n^2. The error term ε_{jklmn} accounts for the variability in the observations between replicates taken at the same station and on the same occasion. The ε_{jklmn}'s are also assumed to be independent with mean 0 and variance σ_ε^2. If the assumptions of normality and homoscedasticity of ε_{jklmn}'s and η_{jklm}'s cannot be satisfied, transformation of the raw data such as taking logarithms will usually make it possible to approximate these assumptions (Millard and Lettenmaier, 1986). However, Green (1979) discouraged resorting to nonparametric techniques since much information is lost and also discussed the robustness of the analysis of variance (ANOVA) to violations in the assumptions.

Before any test on the main effects can be done, all the interaction effects must be nonsignificant. However, when developing the sampling designs based on power calculations, Millard and Lettenmaier (1986) noted that it is simplest to assume that the interaction effects will be negligible, and to base the design on the power of the test on the main effect.

If MS(.) denotes the computed mean square, the null hypothesis of 'no change in the water quality variable after occurrence of the event', i.e. H_o: $B_1=B_2=0$ can be tested using the statistic

$$F = MS(B)/MS(\eta) \qquad (1)$$

which has an F-distribution with 1 and $2JL(M-1)$ degrees of freedom under the null hypothesis (Millard and Lettenmaier, 1986).

The disadvantage, however, in using Model 1 to determine the impacts of a specific agent is that the change may be the result of some event or events not associated with the specific agent. A change in the water quality variable can reasonably be attributed to the event only if it is known that no other change(s) occurred in the environment that might have also affected the water quality variable. Thus, a better approach is to utilize control stations (Palmer and Mackenzie, 1985; Millard and Lettenmaier, 1986).

If paired control stations are used in the design, Skalski and McKenzie (1982) replaced Y_{jklmn} and A_j in Model 1 with

$$Y_{jklmn} = \mu + A_j + B_k + C_l + AB_{jk} + AC_{jl} + BC_{kl} + ABC_{jkl} + \eta_{jklm} + \varepsilon_{jklmn} \qquad \text{Model 2}$$

where Y_{jklmn} = the observed difference between control and treatment stations for the n^{th} replicated on the m^{th} sampling occasion within season l for station pair j, and event status at level k; and A_j = main effect of station pair j (fixed); where $j = 1,\ldots\ldots, J$ (stations);

4. POWER OF THE TEST

The power of the test for the hypothesis that no change occurred in the water quality variable after the event (eqn.1) is given by (Millard and Lettenmaier, 1986) as $1-\beta = \Pr[F_{\nu_1,\nu_2}(\lambda) > F^{\alpha}_{\nu_1,\nu_2}]$ in which $F_{\nu_1,\nu_2}(\lambda)$ is a random variable that is distributed as non-central F with ν_1 and ν_2 degrees of freedom with non-centrality parameter λ; $F^{\alpha}_{\nu_1,\nu_2}$ = $(1-\alpha)^{th}$ fractile point of the central F-distribution with ν_1 =1 and $\nu_2 = 2JL(M-1)$ degrees of freedom.

The non-centrality parameter λ for Model 1 or Model 2 of detecting an abrupt change is given by (Millard and Lettenmaier, 1986) as

$$\lambda = \sqrt{\Delta^2 JLM/[2(\sigma_\eta^2 + \sigma_\varepsilon^2/N)]} \qquad (2)$$

where $\Delta = B_2 - B_1$ is an assumed step change in the water quality variable between pre- and post-event status..

5. COST MODEL

Evaluating the cost of the sampling program is just as important as evaluating the statistical power in developing a cost-effective sampling design. The cost model for multivariable water quality monitoring may be based on the simple linear equation developed by Wiens (1983). The total cost, C, for monitoring a single water quality parameter is given by

$$C = C_o + T \cdot C_t + T \cdot S \cdot C_s + T \cdot S \cdot N \cdot C_r$$

where C_o = the overhead cost, including program management which does not vary with the size of the program; C_t = the cost associated with a sampling occasion including travel to and from the site including boat rental; C_s = the cost per sampling station, including travel between stations and incremental time spent at each station; C_r = replication cost, the collection and laboratory analysis of each sample; T = the number of sampling occasion; S = the number of station sampled per occasion; N = the number of replicate taken at each station on each occasion.

For the multivariable water quality monitoring where the sampling design could be different for each of the p water quality variables, the computation for the total cost of the sampling program for Model 1 assuming a linear cost function would be as follows.

Let M_{2i} = the number of sampling years for water variable i; J_i = the number of stations for water variable i; N_i = the number of replications per station per sampling occasion for water variable i; and i = 1,2,..., p the total number of water quality variables under consideration. Thus, the total cost is the sum of the following cost components:

1. Overhead cost - fixed cost which includes program management and does not vary with the size of the program.

2. Total occasion cost - this considers the maximum number of years of sampling which is equal to 2 L M_1 M_2^* where M_2^* is the highest number of sampling years among the p water quality variables.

3. Total station cost – the total station cost in a year will be based on the highest number of stations that will be monitored in that year. The sum of these cost will constitute the overall station cost, or this is equal to $\sum_{m=1}^{M2^*} 2 \cdot L \cdot M1 \cdot J_m^* \cdot C_s$ where J_m^* = maximum number of stations to monitor in year m.

4. Total Replicate Cost – equals to the cumulative cost incurred in the total number of analysis required for each water variable, that is $\sum_{i=1}^{p} 2 \cdot L \cdot M1 \cdot M2_i \cdot J_i \cdot N_i \cdot C_{r_i}$ where C_{r_i} = cost per replicate of water variable i.

For Model 2, the total station cost and replication cost are doubled because of the use of control stations and these are given, respectively, by $\sum_{i=1}^{M2^*} 4 \cdot L \cdot M1 \cdot J_m^* \cdot C_s$ and $\sum_{i=1}^{p} 4 \cdot L \cdot M1 \cdot M2_i \cdot J_i \cdot N_i \cdot C_{r_i}$.

6. OPTIMIZATION MODEL

The mathematical programming technique can be used to come up with a cost-effective sampling design for water quality monitoring (Jagannathan, 1965). When the variables are independent, then sampling design for each water quality variable can be solved independently such that the design is within the budget constraint or minimum power requirement. When there is budget is limited, the objective function of the optimization model would be to find the combination of sampling design parameters (since each water quality variable may have different sampling designs i.e., number of sampling stations, number of sampling occasions, and number of replications), that would yield the highest minimum power in detecting a significant change. Alternatively, when there is no budgetary constraint, then the objective function is find a sampling design that would meet the minimum power requirements for detecting change at the minimum cost.

Solution to the mathematical programming problem can be done using a discrete approach known as the modified gradient search (Palmer and Mackenzie, 1985) which can be adapted for the multivariable case. This procedure is just one of the efficient combination techniques to find optimal discrete solutions for the number of sampling stations, sampling occasions and replications. The modified gradient search algorithm suggests to increase the value for the most cost-efficient activity (i.e. to increase either the number of sampling stations or the number of sampling occasions or the number of replications), and to decrease the value of the least cost-efficient activity. Cost-efficiency is measured by the ability of the sampling design to increase the power of the test per unit increase in cost.

Since the objective is to maximize the minimum power, then the improvement of power always focus on the variable with the lowest power. Beginning with the design with the lowest possible statistical power, the variables are sorted in ascending order of their powers. The variable with the lowest power is chosen and power is improved by increasing its most cost-efficient activity. The least cost-efficient activity is decreased by that amount necessary to meet the budget constraint; and the median cost-efficient activity is decreased if required to meet the budget constraint. Then the variables are sorted according to their powers and the process is repeated again until the cost exceeds the given budget (Figure 1).

7. RESULTS AND DISCUSSION

The methodology was applied to the water quality data of Tadlac Lake in Los Baños, Laguna for the period July to December 1996. The lake has an area of 25 hectares, and the monitoring activity involved monthly sampling at six stations. Three water quality variables were considered, namely: total nitrogen (TN), phosphorus (P) and total suspended solids (TSS) due to their ability to provide information on the quality of the lake water for potential uses.

A rectangular grid of 17 rows by 15 columns was superimposed on the map of the lake and the value of each node (only 216 nodes are inside the lake area) was estimated by kriging. The input values to the kriging technique are the station coordinates and the mean of each water quality variable in each station. The estimated means were then divided into 4 equiprobable classes. The choice of 4 classes is arbitrarily chosen considering the compromise on the acceptable loss of information (resulting from the grouping of observations) and the desired level of resolution (Lachance et al., 1979). The range of each class formed is shown in Table 1.

Results of the correspondence analysis showed that the first five main axes have a total variance of 77.4%. The sum of the relative contributions of each class of each water quality variable on the first five axes has also reached 50% which indicates that the variables are well represented by these five factors (Myers, 1993). The first five factor scores were subjected to an average linkage cluster analysis and a partial tree diagram of the results showing the last clusters formed at a minimum junction level of 0.806 as shown in Figure 2. The tree can be cut-off at the level where the difference from the previous junction is highest, and this can be between 0.806 to 0.940 resulting to five clusters only. The resulting stratification of the lake is shown in Figure 3.

A characterization of each zone is done by counting the number of stations that fall in each class of each water quality variable (Table 2). Since these classes are equiprobable then a test is made whether the number of stations is significantly higher than what is expected. If it does, then that class is a major characteristic of that zone. Hence, zone 1 is mainly characterized by P2 and P4 as well as TSS3. Zone 5 has the lowest amount of pollutant present since the significant classes found are N1, P1, P2 and TSS1. The other zones have intermediate to high amount of pollutants present. The summary of this characterization is shown in Table 3.

The algorithm shown in Figure 1 is used in the search for a cost-effective sampling design that maximizes the minimum power that can be attained by the three water quality variables under an assumed total budget of P700,000. Estimates of the coefficients of variation (CV) of the observations among sampling occasions for the same season are found to be 37%, 73% and 56.5% for nitrogen, phosphorus, and total suspended solids, respectively. On the other hand, the CVs between replicates in the same station in every occasion are, respectively, 6.7%, 11.6% and 32.2%. Replicate cost for nitrogen is P330, phosphorus costs P210, and only P160 for total suspended solids based on the April 1996 price list of the UPLB Institute of Chemistry. Overhead cost is set to P100,000, occasion cost to P400 and station cost to P20. The number of seasons per year is 2 (wet and dry seasons) with 6 sampling occasions per season; maximum years of sampling is set to 15, and maximum number of replicates is set to 5 (based from the results of cluster analysis).

With a 0.05 probability of committing a type I error and the magnitude of change to be detected set to 30% from the mean value, the maximum of the minimum power that can be attained by the three water quality variables was found to be 0.730. The design parameters, i.e. number of sampling years, number of stations and number of replicates (M_2, J, N) for nitrogen is (6,2,1) with a power of 0.739; for phosphorus (9, 5, 1) with a power of 0.730; and for total suspended solids (8,5,1) with a power of 0.754 and total cost of P683,440. Phosphorus has the most extensive monitoring activity because it also has the highest CV while nitrogen has the least monitoring activity. Since the inputs are approximations, a sensitivity analysis maybe conducted to examine the robustness of the sampling design's effectiveness when some of the inputs are changed

7.1 Sensitivity Analysis on the Power

Figure 4 shows the effect of budget on the power attained by the water resource system. After reaching a high budget, the marginal rate of increase generally decreases. A satisfactory power of 0.70 can already be attained even at P600,000. This means that additional budget should be evaluated according to the amount of additional information that it can give.

Since the change to be detected is an abrupt change, then the usual tendency of the sampling design is to increase its number of stations (Millard and Lettenmaier, 1986). If this value is set at a small value, then the number of sampling occasions will increase. But if both have reached its maximum value, then it will naturally increase the number of replicates which is usually more expensive and usually has the least effect on power. Hence, when the number of sampling stations is set at a low value, the power attained by the design under the same budget is also lower. This increases for every step increase in the number of stations. But after setting it to a much higher value the power do not significantly change as shown in Figure 5.

The combination of the upper bound on the number of sampling years and sampling stations is a critical factor in finding a cost-effective sampling design. Figure 6 shows that with few sampling stations, more years are required before the test of no change can be declared satisfactory. For example, with 8 stations, a minimum of five years is required for the power to reach 0.70. However, with 4 stations, at least 10 years is required.

For this set of data, since station cost is much lower than occasion cost, then the design is more sensitive to a change in the occasion cost. Results have shown that even with a 100% increase in occasion cost, the powers attained were still above 0.70.

As can be inferred from the non-centrality parameter (eqn.2), the parameter is far more sensitive to a unit change in the natural variance than to a unit change in the measurement error. The power has almost remained constant even with a 25% change in the measurement error. However, with slight changes in the natural variance, the power can vary considerably (see Figure 7). Hence, in finding a cost-effective sampling design, allowance for errors especially the natural variance should be taken into consideration. Millard and Lettenmaier (1986) found out that the effectiveness of the sampling design is even more sensitive to a change in the natural variance than with changes in the number of stations, sampling years or replications.

For the test of detecting change, both the size α of the type I error and the power are crucial factors for the decision-makers because committing a type I or a type II error are both not welcomed. Hence, α is always set low, say 0.05, and power is set high to at least 0.70. Figure 8 shows that if α is further decreased, cost sharply increases. This means that at a lower value of α, the additional cost required to increase the power by one unit is higher.

Figure 9 shows that the value of delta (magnitude of change to be detected), which is often inadvertently neglected in monitoring activities is very critical in determining a cost-effective sampling design. The cost increases with increasing rate as the amount to be detected decreases. Also, the difference in cost under 0.70 and 0.80 minimum power requirement is significant. Further analysis also showed that with different upper bound on the number of stations (5, 10 or 15 stations), the cost requirement is generally least with less than 15 stations.

Figure 10 shows that for every step increase in the power requirement, the cost also increases. The marginal rate of increase generally increases as power requirement is set higher. It can also be seen that under 4 stations, a power greater than 0.75 cannot be attained, and under 5 stations, the highest power that can be attained is only less than 0.85. All these designs are set under a maximum of 15 years of sampling. The graph also showed that cost is generally lower under 8 stations.

8. CONCLUDING REMARKS

As lake water quality monitoring requires assessing the levels of several variables, a cost-effective sampling design is needed that considers the trade-off between statistical reliability and cost of monitoring. This involves determining the optimal combination of the sampling design parameters (number of sampling occasions, number of stations, and number of replicates) that maximizes the power of monitoring program given a fixed budget, or one that minimizes the cost of monitoring for a prescribed level of statistical power of detecting change in the level of water quality variables.

In the search for cost-effective sampling design, it is important to distinguish the cost components associated with the monitoring program and to analyze the sensitivity of the design to the estimated natural variability and measurement error. Relative comparisons of the effect of cost of the data collection, variability of data, and statistical power can be made to determine the best sampling design for water quality monitoring among the alternative sampling plans.

9. Acknowledgments

The authors would like to give special thanks to Department of Science and Technology (DOST) through the Engineering and Science Education Project (ESEP) for the scholarship support extended to the first author; to the Institute of Statistics at UPLB; to the Mathematics

Department of MSU-Marawi City; and to Dr. Maxima E. Flavier, ESEP Coordinator, Mr. and Mrs. Gregorio Santos, Jr. and Ms. Sonny Tababa for their invaluable contributions to this study.

10. References

GREEN, R.H. 1979. *Sampling Design and Statistical Methods for Environmental Biologists.* John Wiley and Sons, Inc., New York.

JAGANNATHAN, R. 1965. *The Programming Approach in Multiple Character Studies.* Econometrica. 33(1):236-237.

KECKLER, D. 1994. User's Guide: SURFER for Windows: Contouring and 3D Surface Mapping. Colorado: Golden Software, Inc.

LACHANCE, M., BOBEE, B., AND D. GOUIN. 1979. *Characterization of the Water Quality in the Saint Lawrence River: Determination of Homogeneous Zones by Correspondence Analysis.* Water Resources Research. 15:1451-1462.

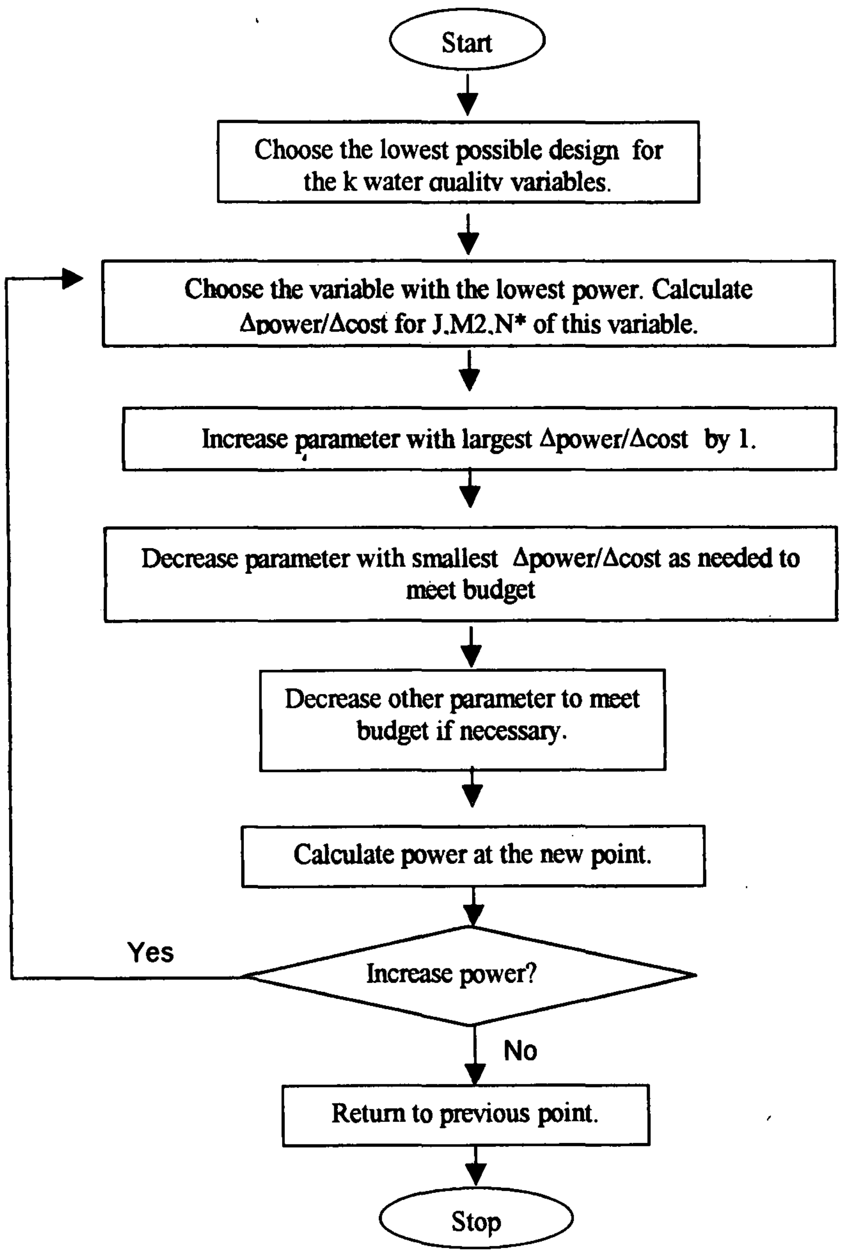
MILLARD, S.P., AND D.P. LETTENMAIER. 1986. *Optimal Design of Biological Sampling Programs Using the Analysis of Variance.* Estuarine, Coastal and Shelf Science. 22:637-656.

MYERS, D.E. 1993. *Correspondence Analysis Applied to Environmental Data Sets.* Multivariate Environmental Statistics. Patil, G.P., and C.R. Rao (editors). Elsevier Science Publication.

PALMER, R.N., AND N.C. MACKENZIE. 1985. *Optimization of Water Quality Monitoring Networks.* Journal of Water Resources Planning and Management. 34:15-30.

SKALSKI, J.R., AND D.H. McKENZIE. 1982. *A Design for Aquatic Monitoring Programs.* Journal of Environmental Management. 14(3):237-251.

WIENS, G.J. 1983. *Cost-Optimization of Aquatic Monitoring Programs.* Master's thesis, Dept. of Civil Engineering, University of Washington, Seattle, Washington.



*J,M2,N refers to J stations, M2 sampling years, and N replications per station per sampling occasion.

Figure 1. Flowchart for the algorithm of maximizing the minimum power under a fixed budget.

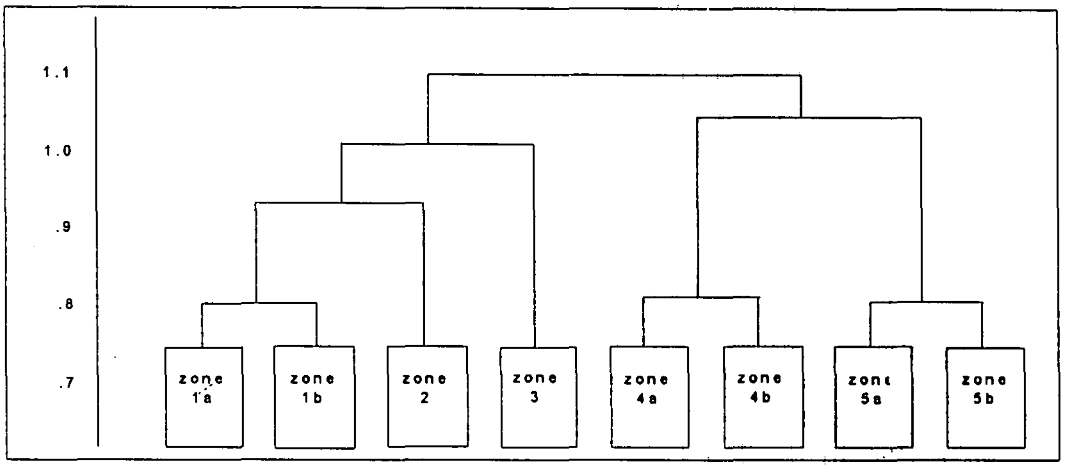


Figure 2. Tree diagram on the average cluster analysis of the stations' scores on the first five factorial axes showing the last clusters formed at a minimum junction level of 0.81.

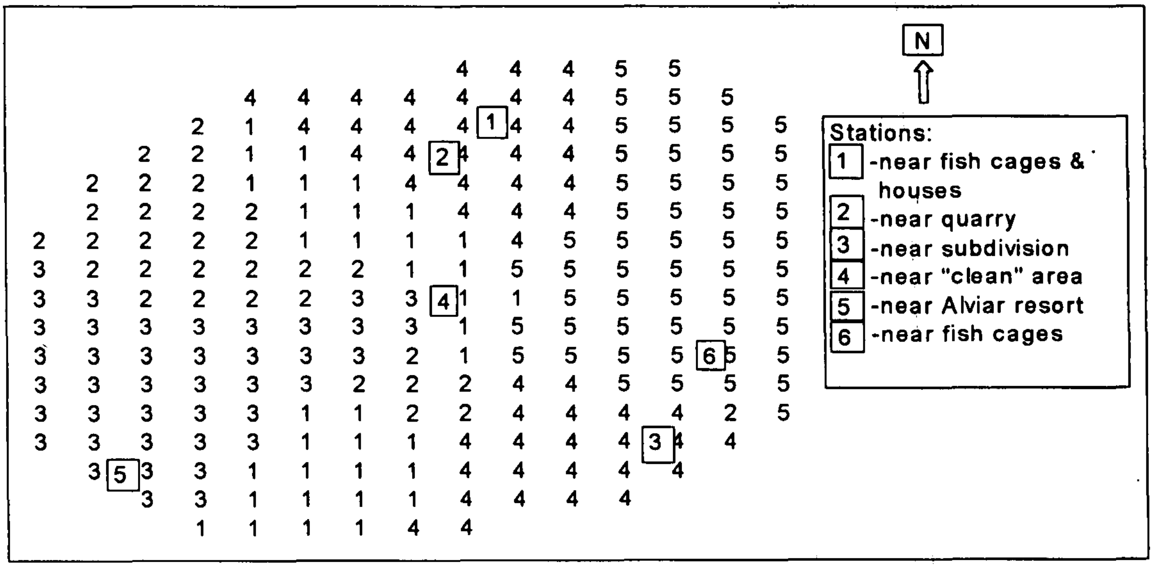


Figure 3. Stratification of Tadlac Lake into five homogeneous zones using the stations' coordinates on the first five axes.

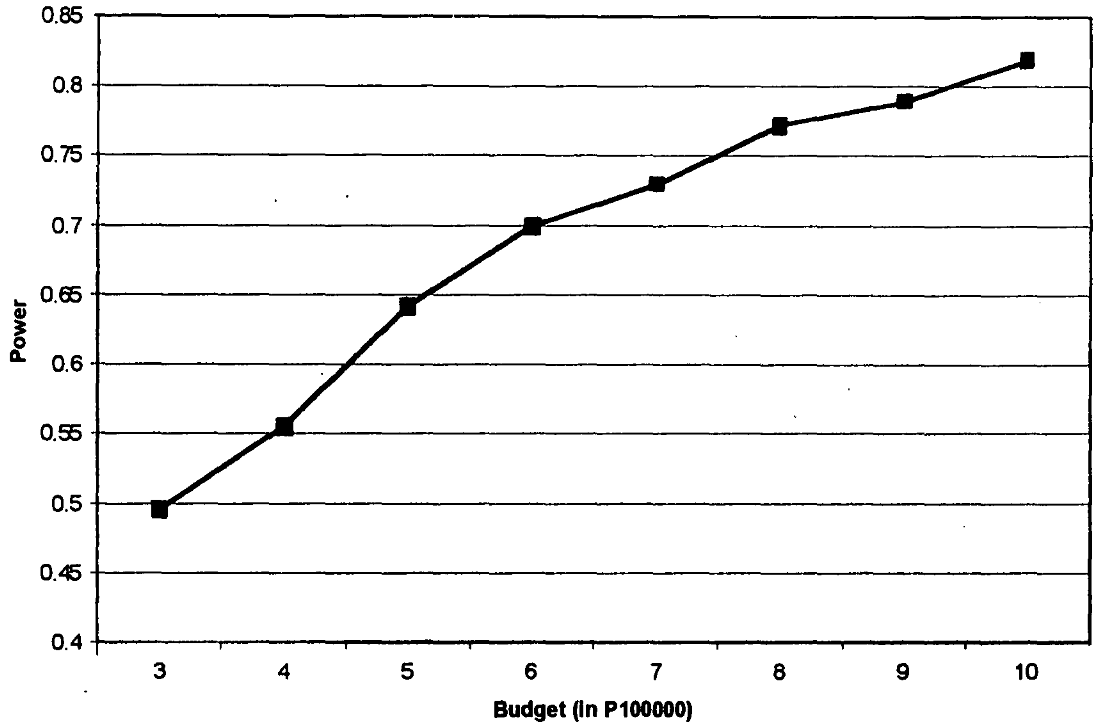


Figure 4. Trend in the minimum power as a function of budget.

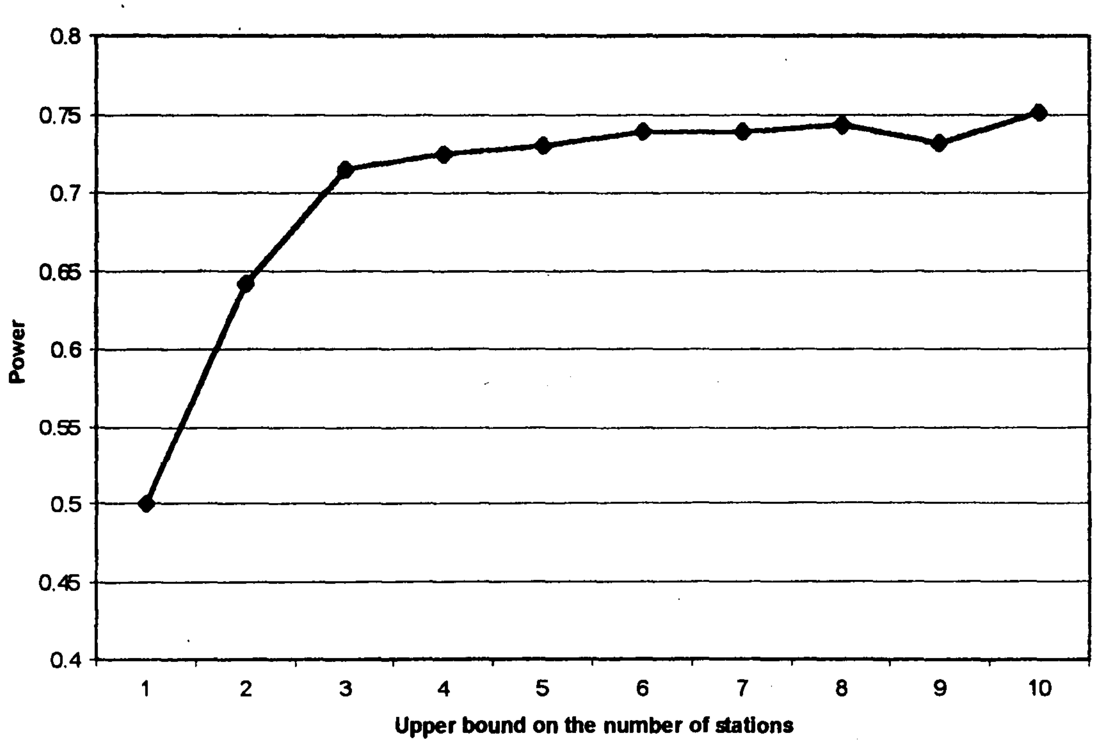


Figure 5. Trend in the minimum power as a function of the upper bound on the number of stations

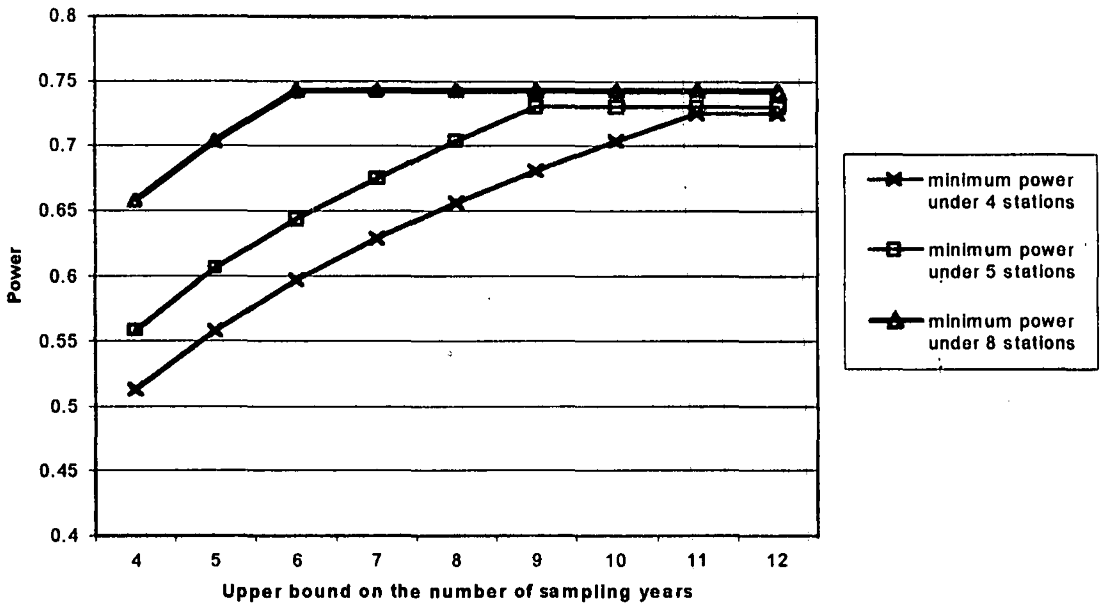


Figure 6. Trend in the minimum power as a function of the upper bound on the number of years of sampling and number of stations.

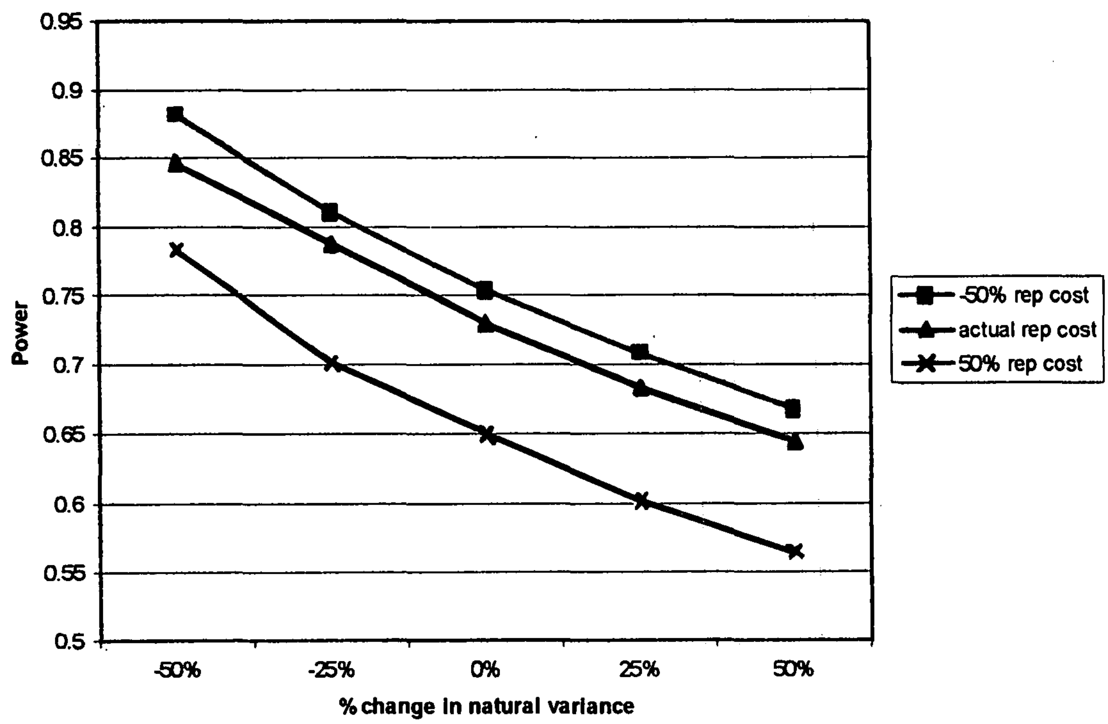


Figure 7. Trend in the minimum power as a function of the natural variance and replicate cost.

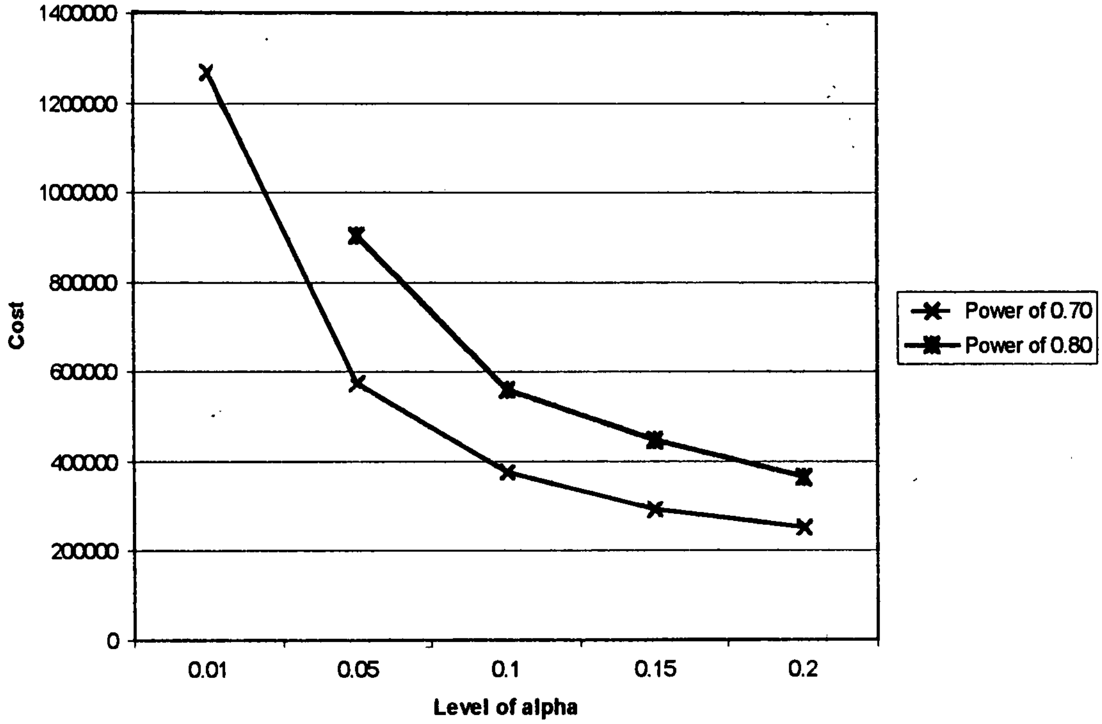


Figure 8. Trend in the minimum cost requirement as a function of alpha and power.

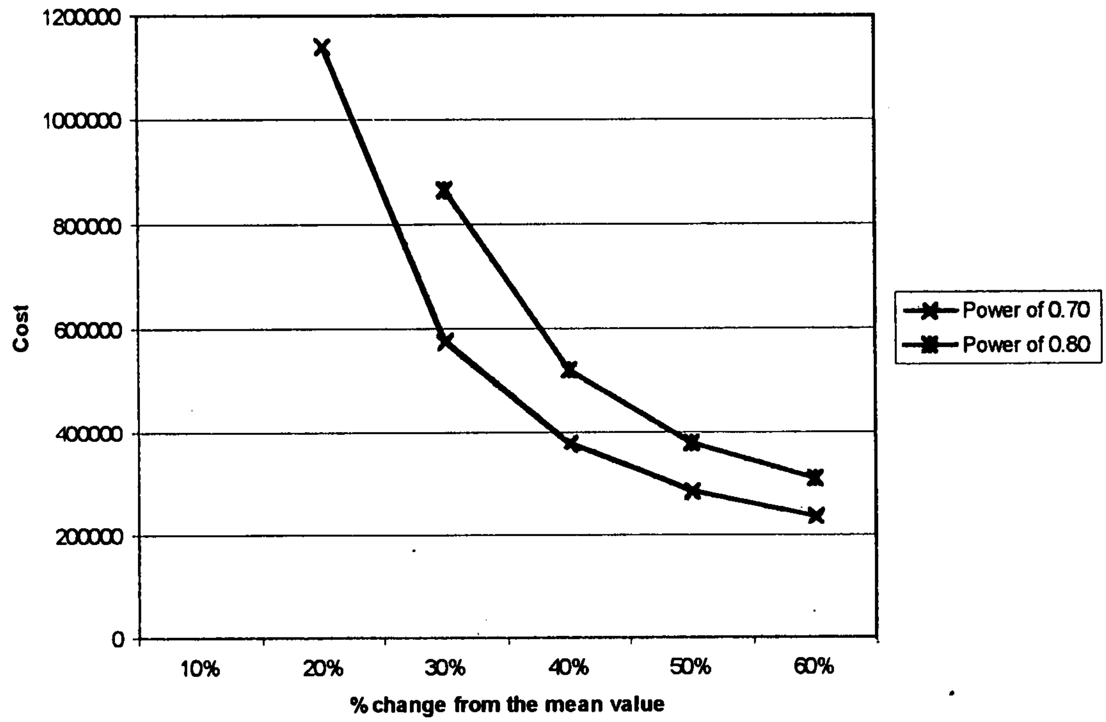


Figure 9. Trend in the minimum cost requirement as a function of delta and power.

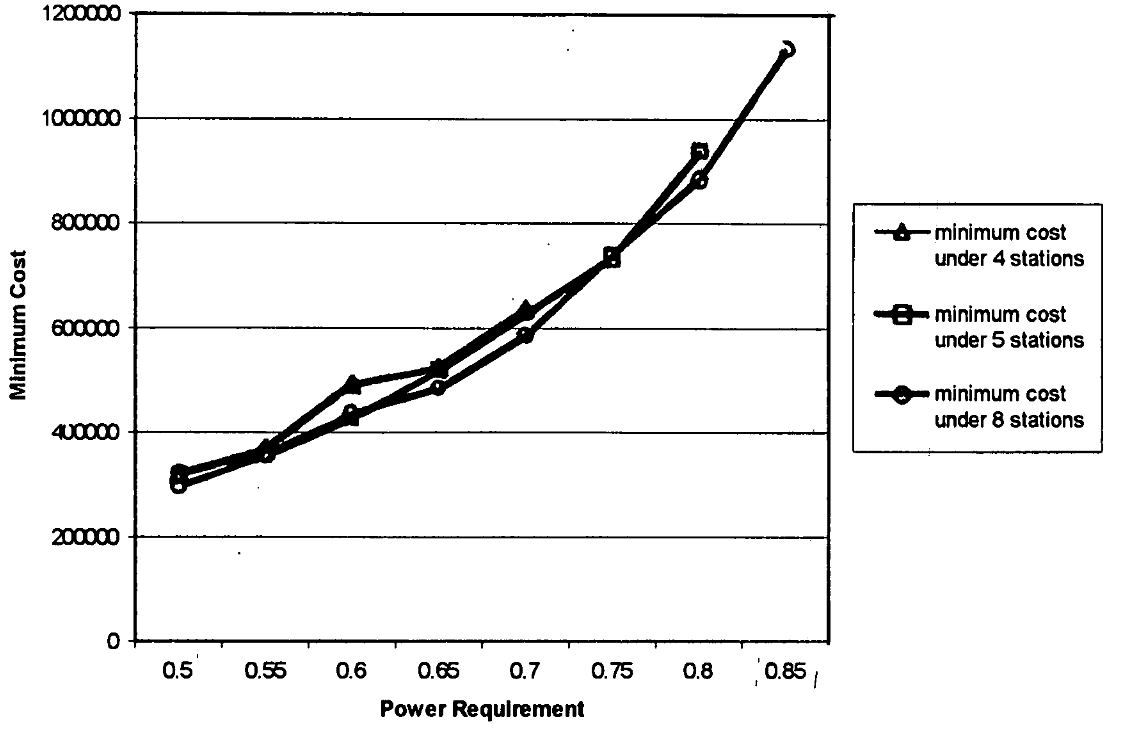


Figure 10. Trend in the minimum cost as a function of power requirement.

Table 1. Limits and size of the classes taken from measurements of nitrogen, and phosphorus and total suspended solids.

Class	Size	Nitrogen		Phosphorus		Total Suspended Solids	
		Limits	Class Name	Limits	Class Name	Limits	Class Name
1	54	1.75-2.08	N1	0.319-0.322	P1	9.48-10.55	TSS1
2	54	2.08-2.13	N2	0.323-0.335	P2	10.56-11.56	TSS2
3	54	2.13-2.20	N3	0.336-0.358	P3	11 57-12.26	TSS3
4	54	2.20-2.40	N4	0.359-0.400	P4	12.26-12 55	TSS4

Table 2. Characterization of the groups of stations obtained by counting the number of stations falling in each class of the three water quality variables.

ZONE	NUMBER OF STATIONS	NITROGEN				PHOSPHORUS				TOTAL SUSPENDED SOLIDS			
		1	2	3	4	1	2	3	4	1	2	3	4
1	36	3	9	13	11	4	15	0	17	0	0	36	0
2	32	0	4	23	5	0	1	31	0	1	0	18	13
3	41	9	29	3	0	0	3	5	33	0	0	0	41
4	52	0	4	10	38	18	12	18	4	4	46	0	0
5	55	42	8	5	0	32	23	0	0	47	8	0	0
Total	216	54	54	54	54	54	54	54	54	54	54	54	54

Table 3. Significant classes of the water quality variables that characterize the groups of stations.

Parameter	**Zone No.**				
	1	**2**	**3**	**4**	**5**
Nitrogen	-	N3	N2	N4	N1
Phosphorus	P2 & P4	P3	P4	-	P1 & P2
Total Suspended Solids	TSS3	TSS3 & TSS4	TSS4	TSS2	TSS1