

TEMPORALLY STRATIFIED SAMPLING PROGRAMS
FOR ESTIMATION OF FISH IMPINGEMENT

MASTED

MASTER

NOTICE
This report was prepared as an account of work sponsored by the United States Government. Neither the United States nor the United States Department of Energy, nor any of their employees, nor any of their contractors, subcontractors, or their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness or usefulness of any information, apparatus, product or process disclosed, or represents that its use would not infringe privately owned rights.

K. D. Kumar and J. S. Griffith*

Environmental Sciences Division
Oak Ridge National Laboratory
Oak Ridge, Tennessee 37830

ABSTRACT

Impingement monitoring programs often expend valuable and limited resources and fail to provide a dependable estimate of either total annual impingement or those biological and physicochemical factors affecting impingement. In situations where initial monitoring has identified "problem" fish species and the periodicity of their impingement, intensive sampling during periods of high impingement will maximize information obtained. We use data gathered at two nuclear generating facilities in the southeastern United States to discuss techniques of designing such temporally stratified monitoring programs and their benefits and drawbacks.

Of the possible temporal patterns in environmental factors within a calendar year, differences among seasons are most influential in the impingement of freshwater fishes in the Southeast. Data on the threadfin shad (Dorosoma petenense) and the role of seasonal temperature changes are utilized as an example to demonstrate ways of most efficiently and accurately estimating impingement of the species.

*Present address: Idaho State University, Pocatello, Idaho 83201.

By acceptance of this article, the publisher or recipient acknowledges the U.S. Government's right to retain a non-exclusive, royalty-free license in and to any copyright covering the article.

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

CONF-771231-2

TEMPORALLY STRATIFIED SAMPLING PROGRAMS
FOR ESTIMATION OF FISH IMPINGEMENT

K. D. Kumar and J. S. Griffith*

Environmental Sciences Division
Oak Ridge National Laboratory
Oak Ridge, Tennessee 37830

I. INTRODUCTION

Some of the major purposes of impingement monitoring programs at power plants are to:

- (a) estimate the annual rate of impingement,
- (b) describe the species composition of fish impinged,
- (c) describe the temporal patterns of fish impingement, and
- (d) determine the factors causing significant impingement (if any)

so that proper measures can be taken to reduce impingement.

This paper is concerned with the estimation of annual rate of impingement when impingement counts exhibit temporal patterns. A temporally stratified sampling program using the Dalenius-Hodges (1951) cum \sqrt{f} rule and the Neyman optimal allocation policy will be developed (Sec. III). A possible approach to designing an efficient sampling program when no impingement data are available, but is known to be correlated with some known environmental parameter, will be discussed in Sec. IV.

*Present address: Idaho State University, Pocatello, Idaho 83201.

In a study of fish impingement in the Southeast United States (see J. M. Loar et al., this volume) it was found that impingement counts at several power plants exhibited distinct temporal patterns. It was reported that there were periods when the daily rate of impingement was greater than 100,000 fish and periods when the daily impingement rate was less than 10 fish. These patterns could be due to factors such as low temperatures, low dissolved oxygen, high turbidity, presence of young of the year near the intake, and/or migration patterns. It was reported that the sampling programs at these power plants ignored these patterns and used inefficient systematic sampling schemes leading to poor estimates of annual impingement rates.

An optimal sampling program must give more weight to periods of high impingement than to periods of low impingement, i.e., a temporally stratified sampling program (TSSP) must be used. When the impingement counts are more uniform throughout the year the stratification may not result in a smaller variance. When more than one species is of interest and these species exhibit different temporal patterns the best sampling program will be a compromise between the policies for the individual species. In this paper only the single species situation is considered. Furthermore, it is assumed that the impingement counts exhibit distinct temporal patterns. Due to the stochastic nature of impingement, it is not possible to stratify the year by day. Instead strata at a more gross level of months are formed. Certain months of the year will exhibit high impingement counts, some will have low impingement, while

others may have medium impingement. This paper discusses how such strata may be formed and how a sampling program may be implemented.

The statistical tools needed to design a TSSP for estimating annual impingement rate already exist. In the following sections these methods will be used to develop sampling programs which will greatly assist biologists in understanding the factors which cause impingement. A properly designed program may also reduce the cost of monitoring.

II. TEMPORAL STRATIFICATION

Temporal stratification is the allocation of the months of the year into L groups such that the annual rate of impingement can be estimated with the desired accuracy. In designing temporally stratified sampling programs the following parameters must either be specified or estimated:

- (a) Desired accuracy: This is usually specified as $\pm d\%$ of the true mean at some α level. This requires prior knowledge of the magnitude of the true mean (a crude estimate will suffice).
- (b) Number of strata: L .
- (c) Strata boundaries: $Y = \langle Y_0, Y_1, \dots, Y_L \rangle$. The i^{th} stratum has impingement values between Y_{i+1} and Y_i .
- (d) The total number of observations: n .
- (e) The allocation of samples to the strata: $(n_1, n_2, \dots, n_L, \text{ where } \sum_{i=1}^L n_i = n)$.

Items (b) through (e) are usually estimated from prior data and they will be a function of item (a) and the underlying distribution function of the impingement rate.

In designing a TSSP a variable, over which the stratification can be done, must be specified. An ideal situation occurs when historical data on impingement are available. If such data are not available a variable that is highly correlated with impingement may be used. For example, a high negative correlation between temperature and threadfin shad impingement exists at several power plants in the Southeast. We will discuss the former case in Section III and the latter in Section IV.

An efficient stratified sampling design will result in low within-stratum variability and high between-strata variability (i.e., the means of the strata will be as different as possible). Cochran (1977) gives a detailed discussion of various stratified sampling policies. In this paper the Neyman optimal allocation policy and some variants of it will be utilized. The Neyman allocation policy is given by

$$\frac{n_i}{n} = \frac{N_i S_i}{\sum_{j=1}^L N_j S_j} = W_i S_i, \quad (1)$$

where n_i = no. of samples in stratum i ,

$$n = \text{total number of samples} = \sum_{j=1}^L n_j,$$

N_i = no. of days in stratum i ,

S_i = standard deviation of stratum i , and

$$W_i = \frac{N_i}{\sum_{j=1}^L N_j S_j}$$

As S_i increases, the proportion of samples allocated to that stratum increases. There are other criteria for allocating samples. For example, we can replace S_i by Y_i , the i^{th} stratum mean, i.e.,

$$\frac{n_i}{n} = \frac{N_i Y_i}{\sum_{j=1}^L N_j Y_j} \quad (2)$$

If $S_i^2 \propto Y_i^2$, then (2) follows immediately from (1).

Another possibility is to equalize the coefficient of variation of the means in all classes (Deming 1950, pp. 233-238) resulting in

$$\frac{n_i}{n} = \frac{C_i^2}{\sum_{i=1}^L C_i^2}, \quad (3)$$

where C_i is the coefficient of variation of daily impingement rate in the i^{th} stratum. This criterion might be useful if the estimates of the stratum means are also important.

III. TSSP WHEN IMPINGEMENT DATA ARE AVAILABLE

In this section, two case studies to illustrate the TSSP are presented. In Fig. 1 the average daily rate of impingement of threadfin

shad at Arkansas Power Station is plotted for the various months in 1975. There is a distinct seasonal pattern with extremely high impingement during winter months and very low intensity during the rest of the year. At Browns Ferry Nuclear Plant in Alabama, threadfin shad exhibit high impingement rate during the winter months (Fig. 2). There is a second peak in August 1974 and in November 1975, which is probably related to factors other than temperature.

To design the TSSP, the strata must be formed and the strata boundaries must be defined. Dalenius and Hodges (1959) suggested the "cum \sqrt{f} " rule which can be used to define the strata boundaries if the number of strata L is known. This is an approximation to the strata boundaries obtained by solving iteratively a complicated set of equations (Dalenius et al. 1951). Cochran (1961) and Anderson et al (1976) have shown that the cum \sqrt{f} rule gives extremely good approximations to the optimal strata boundaries even when the distribution is skewed. For theoretical details refer to Dalenius and Hodges (1957, 1959) and Serfling (1968).

The first example that will be considered is the impingement of threadfin snad at Arkansas Unit One during 1975. The data base consists of 24 hour counts on the screen collected during 1975. The sampling program was a systematic sampling program with two or three samples per week. In the following paragraphs we will describe in detail the various steps in the design of a TSSP. The first step is the development of Table 1. The daily impingement rates are grouped into cells that have equal width in the logarithmic scale (column 1 in Table 1). In the

second column of Table 1 the observed frequency f (number of samples falling into the class) is given. The third column is the square root of the observed frequency. The last column, the cum \sqrt{f} column, is the running sum of the \sqrt{f} column. For example, the entry in the second row (9.9) is the sum of 5.9 and 4.0, the first two entries in the \sqrt{f} column.

The Dalenius-Hodges rule states that if the number of strata, L , is known, then the strata boundaries are obtained by partitioning the cum \sqrt{f} column into L equal groups.

The results for $L=3$ and $L=4$ are shown in Table 2 (designs I and II, respectively). For example, when $L = 3$, the cell size of the cum \sqrt{f} column is $23.5/3 \approx 7.8$. The strata boundaries are shown in Table 2. The months are allocated to these strata according to the average daily rate of impingement during the month (Fig. 1). The result of such an allocation is shown in the fourth column of Table 2. For comparison, two additional designs (III and IV in Table 2) are presented. In design III, the year is stratified into three strata, where the strata are Jan-April, May-August, etc. Similarly, design IV has four strata with the year partitioned into groups of three sequential months. These strata are formed without regard to the historical data base.

The optimal allocation policies [equation (1)] for the four designs are given in the last column of Table 2. Due to the very large differences in the strata means, the allocation is concentrated during the last three months and the first three months of the year for designs I and II.

Since the Dalenius-Hodges rule is based on the assumption that $w_i S_i = \text{constant}$, the sample allocation is given by

$$n_i = \frac{n}{L} .$$

This will be satisfactory for large L . However, in the examples we are considering, L is usually 3 or 4. Moreover, given the historical data and the cum \sqrt{F} rule, we can estimate the strata means and variances and obtain the Neyman allocation policy directly as shown in Table 2. We still have to determine the total number of samples (n) and the optimal number of strata (L). These two can be determined simultaneously.

Assume that we wish to determine (L, n) such that the annual daily average impingement is estimated within an average $\pm d\%$ at some level. This is equivalent to specifying the half-width of the $(1 - \alpha)\%$ confidence interval. If an estimate of the true mean is available, the minimum variance V is given by

$$V = \frac{(d \bar{Y})^2}{t_{\alpha, v}^2} , \quad (5)$$

where \bar{Y} = true mean, and

$t_{\alpha, v}$ = Student-t statistic with v degrees of freedom.

Cochran (1977) has shown that

$$n = \frac{(\sum W_h S_h)^2}{V + \frac{1}{N} \sum W_h S_h^2} , \quad (6)$$

where $W_h = \frac{N_h}{N}$,

$N = 365$ (no. of days in the year),

and $S_h =$ standard deviation of the h^{th} stratum.

For a given value of L we can obtain a table of n for different values of d from equation (6) and Table 2 (Table 3). The best design is one which dominates the other, in the sense that it has the smallest sample size. In this example one could choose either design I or II, since they give similar results.

In the preceding discussion we started with a historical data base and developed a TSSP. Since the TSSP recommends that no sampling be conducted during six months of the year, this design might not be politically acceptable. Under such circumstances it might be advisable to set up the separate sampling programs for the six "winter" months and for the remaining six months. Moreover, if one of the aims of monitoring is to determine factors affecting impingement, it might be worthwhile to sample throughout the year. Clearly one must pay the penalty for the additional constraints by sampling at a higher rate during certain periods.

A second example is the impingement of shad at Browns Ferry Nuclear Station (Fig. 2). Tables 4 through 6 show the results similar to the ones we derived for the Arkansas Unit One. Due to the more even impingement counts throughout the year (as compared to the previous example) the sampling is conducted throughout the year. As before, Table 6 can be extended for several values of L and pick the dominant design.

IV. TSSP WHEN NO IMPINGEMENT DATA ARE AVAILABLE

Consider the impingement at Arkansas Unit One (Fig. 1). Observe that there is a high negative correlation between water temperature and threadfin shad impingement counts. We wish to design a stratified sampling program at another plant where shad are known to be abundant but no impingement data are available.

If water temperature data are available (possible sources are state and federal agencies, pre-operational monitoring program), one could use the correlation between impingement counts and temperature to design a TSSP. In this section, the mechanics of designing a TSSP under each circumstance are discussed.

Anderson et al. (1976), have developed some tables relating the number of strata (L), the increase in efficiency due to stratification (as compared to a simple random sample), and the correlation when one can treat the response (impingement counts) and the stratification variable (temperature) as having a bivariate normal distribution. By assuming a value for the correlation, the optimal number of strata can be determined. For example, when the correlation is 0.8, Anderson et al. (1976) report a gain in efficiency (over a random sample) of 0.57. The final decision is based on the desired increase in efficiency. In Table 7 the cum \sqrt{f} rule is applied to temperature data from Arkansas power plant. In Table 8 the designs when L is 3 and 4 are shown.

The next step is the determination of the percent allocation and the total sample size. One approach would be to use a proportional allocation policy. Let us consider another approach. In this example,

it has been assumed that the correlation between temperature and impingement is negative. Hence, one would wish to give more weight to the stratum with lower temperature. This is accomplished by writing

$$y_i \propto \frac{1}{T_i} \quad , \quad (7)$$

where T_i is the average temperature of the i^{th} stratum. T_i could also be the midpoint of the stratum. By substituting this in equation (2), we get

$$\frac{n_i}{n} = \frac{N_i/T_i}{\sum_{j=1}^L N_j/T_j} \quad . \quad (8)$$

The allocation based on this rule is shown in the last column of Table 8.

The design will be complete when we estimate the total sample size, n . This can be approximated by assuming a value for the coefficient of variation (say 200%) and using the Stein estimator (Cochran 1977) for the sample size for finite populations. For example, if α is 0.1 and d is 20%, the value of n is 155 (see Sect. 4.6 of Cochran 1977).

If it is suspected that more than one factor influences impingement (e.g., temperature, dissolved oxygen, turbidity, etc.), one could use the first principal component as the stratification variable. The advantage of using a covariate like temperature to form the strata is to make sure that the sampling is conducted during all major configurations of temperature. In Table 8 when L is equal to 4, November forms a

separate stratum. During this month, the daily temperature exhibits a very rapid drop ($0.5^{\circ}\text{C}/\text{day}$). Hence, this month must be treated separately.

SUMMARY

A practical approach to the designing of impingement monitoring programs was discussed under the two situations, namely (a) when prior data on impingement were available and (b) when no impingement data were available but the impingement counts were known to be correlated with some other parameter like water temperature. It was shown that a temporally stratified sampling program (TSSP) can be used to obtain reliable estimates of annual impingement rates. Some areas for future studies are the design of sampling programs for several species simultaneously, subsampling of the fish impinged on the screen when the impingement rate is in thousands of fish per day and the effect of estimating the strata variances from prior samples on the optimal properties of the TSSP.

ACKNOWLEDGEMENTS

The authors wish to thank their colleagues J. J. Beauchamp and R. B. McLean for their many helpful comments on this manuscript. This research was sponsored by the Division of Biomedical and Environmental Research, U. S. Department of Energy, under contract W-7405-eng-26 with Union Carbide Corporation. Publication No. 1174, Environmental Sciences Division, ORNL.

BIBLIOGRAPHY

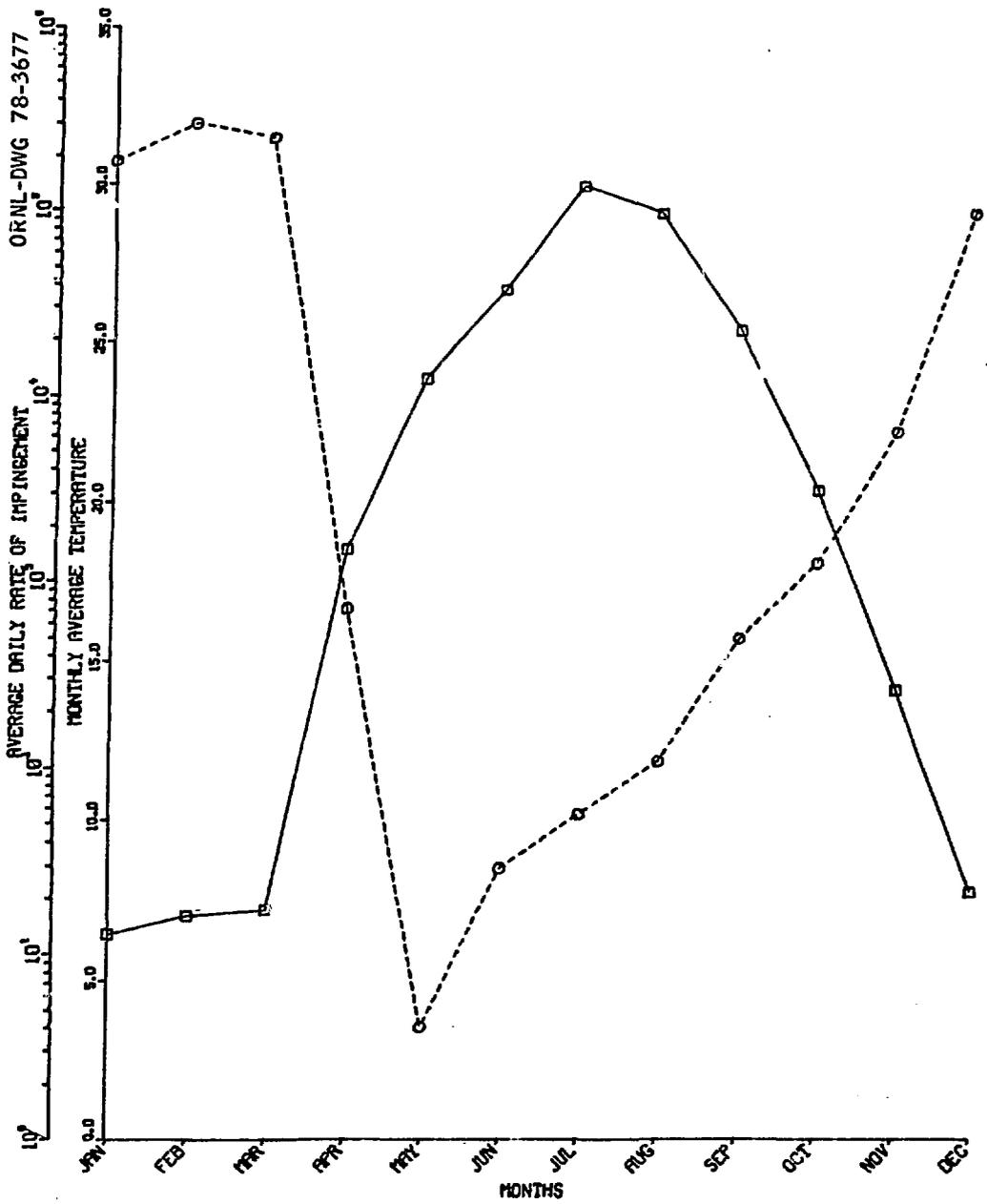
- Anderson, D. W., L. Kish, and R. C. Cornell. 1976. Quantifying gains from stratification for optimum and approximately optimum strata using a bivariate normal model. *J. Am. Stat. Assoc.* 71:887-892.
- Cochran, W. G. 1961. Comparison of methods for determining stratum boundaries. *Bull. Int. Stat. Inst.* 38:345-358.
- Cochran, W. G. 1977. *Sampling Techniques*. John Wiley and Sons, New York. 428 pp.
- Dalenius, T., and M. Gurney. 1951. The choices of stratification points, *Skandinavisk Akturietidskrift* 3-4:133-148.
- Dalenius, T., and J. L. Hodges, Jr. 1957. The choice of stratification points. *Skandinavisk Akturietidskrift* 1-2:198-203.
- Dalenius, T., and J. L. Hodges, Jr. 1959. Minimum variance stratification. *J. Am. Stat. Assoc.* 54:88-101.
- Deming, W. E. 1950. *Some theory of sampling*. Dover Publications Inc., New York. 602 pp.
- Serfling, R. J. 1968. Approximately optimal stratification. *J. Am. Stat. Assoc.* 63:1298-1309.

Figure 1

1975 Threadfin shad impingement counts (⊙----⊙) and water temperature (□—□) at Arkansas Unit One.

Figure 2

1974-1975 Shad impingement counts (⊙----⊙) and water temperature (□—□) at Browns Ferry Nuclear Station.



ORNL-DWG 78-3537

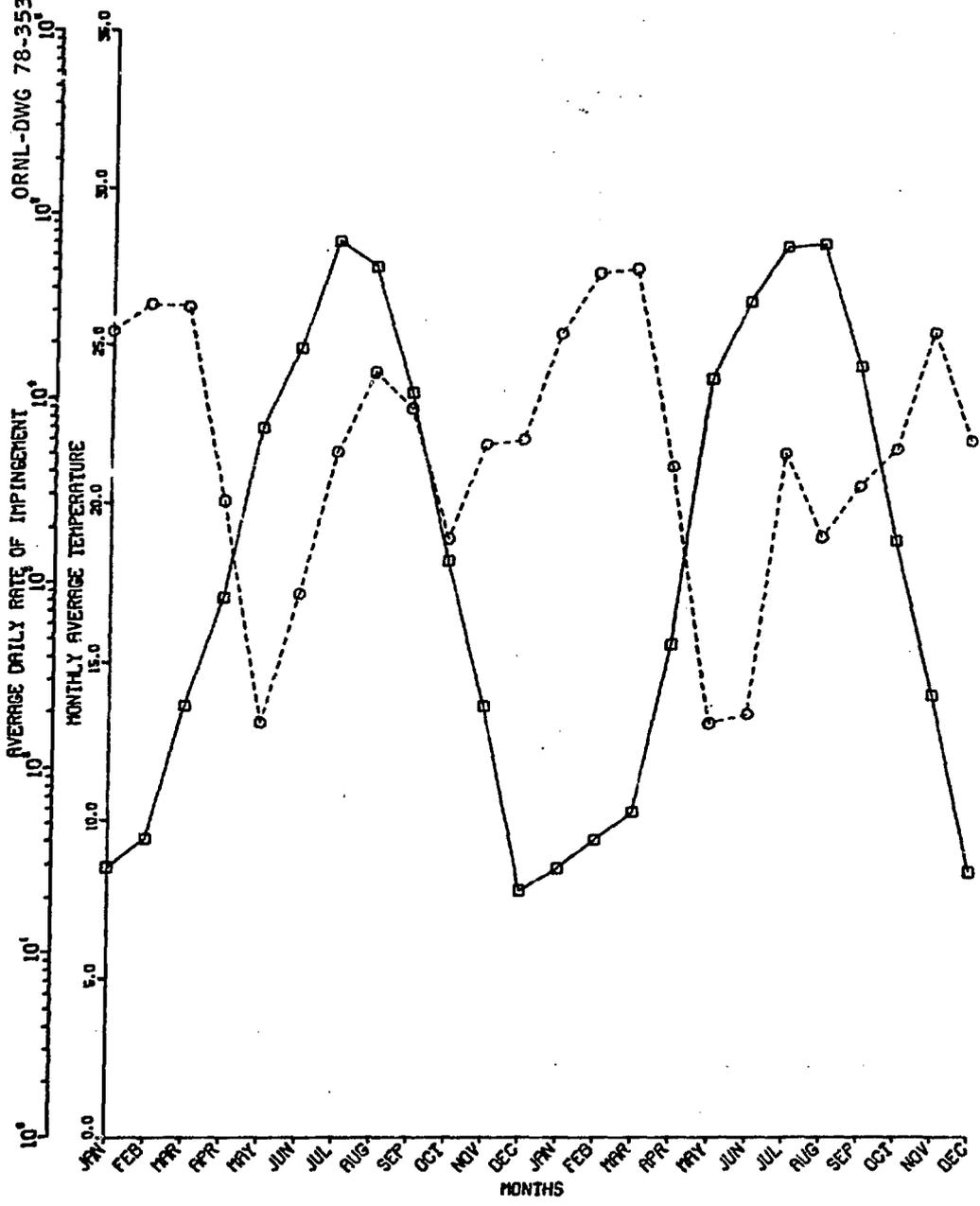


Table 1
 Cum \sqrt{f} Rule for Threadfin Shad at Arkansas Unit 1
 Based on the 1975 Impingement Counts

Class*	f_i	$\sqrt{f_i}$	cum $\sqrt{f_i}^\dagger$
$10^1 - 10^2$	35	5.9	5.9
$10^2 - 10^3$	16	4.0	9.9
$10^3 - 10^4$	19	4.4	14.3
$10^4 - 10^5$	13	3.6	17.9
$10^5 - 10^6$	31	5.6	23.5

* f_i = number of days with total (24 hour) observed impingement counts that fall into the class i .

† cum $\sqrt{f_i}$ = running sum of $\sqrt{f_i}$ column.

Table 2
Four Temporally Stratified Designs for Threadfin Shad at
Arkansas Unit 1 Based on 1975 Impingement Counts

L	Stratum Number	Boundaries*	Months†	$\bar{x}^{\dagger\dagger}$	S [§]	Percent Allocation**
<u>Design I</u>						
3	1	10 - 10 ²	May, Jun, July	30.0	53.5	§§
	2	10 ² - 10 ⁴	April, Aug, Sep, Oct, Nov	2364.0	6669.4	4.0
	3	10 ⁴ - 10 ⁶	Jan, Feb, Mar, Dec	216,529.0	200,061.2	96.0
<u>Design II</u>						
4	1	10 - 10 ²	May, Jun, July	30.0	53.5	§§
	2	10 ² - 10 ³	Apr, Aug, Sep	293.0	452.3	§§
	3	10 ³ - 10 ⁵	Oct, Nov	4372.0	9065.6	2.0
	4	10 ⁵ - 10 ⁶	Jan, Feb, Mar, Dec	204,077.0	200,628.9	98.0
<u>Design III</u>						
3	1	-	Jan, Apr	232,360.0	19,052.0	66.0
	2	-	May, Aug	58.0	97.0	§§
	3	-	Aug, Dec	30,172.0	100,847.0	34.0
<u>Design IV</u>						
4	1	-	Jan, March	250,648.0	195,293.0	63.0
	2	-	Apr, June	119.0	362.0	§§
	3	-	July, Sept	210.0	382.0	§§
	4	-	Oct, Dec	37,829.0	112,191.0	37.0

* Stratum boundaries (number of fish impinged per 24 hours).

† Months that form the stratum (i.e. temporal stratum).

†† \bar{x} = estimated average 24 hour impingement rate of the individual strata.

§ S = estimated standard deviations of the 24 hour impingement rate of the individual strata.

** Percent allocation as per Neyman allocation policy.

§§ Represents allocation less than 0.01%.

Table 3
 Sample Size Requirements for Four Temporally Stratified*
 Sampling Programs for Threadfin Shad at Arkansas Unit 1

d [†]	Design Number			
	I	II	III	IV
0.05	121 ^{††}	117	157	205
0.10	98	95	126	172
0.15	75	72	95	136
0.20	56	54	70	105
0.25	43	41	53	81
0.5	14	14	18	28

*See Table 2 for details on design.

[†]d = half width of the derived confidence interval expressed as a fraction of the true mean.

^{††}Sample size.

Table 4
 Cum \sqrt{f} Rule for Shad at Browns Ferry Nuclear Station
 Based on the 1974-1975 Impingement Data

Class	f_i^*	$\sqrt{f_i}$	cum $\sqrt{f_i}^\dagger$
$10^1 - 10^2$	12	3.5	3.5
$10^2 - 10^3$	39	6.2	9.7
$10^3 - 10^4$	61	7.8	17.5
$10^4 - 10^5$	38	6.2	23.7
$10^5 - 10^6$	4	2.0	25.7

* f_i = number of days with total (24 hour) observed impingement counts that fall into the class i .

† cum $\sqrt{f_i}$ = running sum of $\sqrt{f_i}$ column.

Table 5
Temporally Stratified Design for Shad at
Browns Ferry Nuclear Station

L*	Stratum Number	Boundaries [†]	Months ^{††}	\bar{X}^{\S}	S ^{**}	Percent Allocation ^{§§}
3	1	10 - 10 ³	May, June	184.0	138.0	***
	2	10 ³ - 10 ⁴	April, July- Oct, Dec	4,480.0	5,740.0	18.0
	3	10 ⁴ - 10 ⁶	Jan, Feb, Mar, Nov	36,264.0	40,677.0	82.0

*L = number of strata.

[†]Stratum boundaries (number of fish impinged per 24 hours).

^{††}Months that form the temporal stratum.

[§] \bar{X} = estimated stratum mean.

^{**}S = estimated stratum standard deviation.

^{§§}Percent allocation as per Neyman allocation policy.

^{***}Represents allocation < 0.01%.

Table 6
Sample Size Requirements for Shad at
Browns Ferry Nuclear Station*

d^\dagger	Sample Size
0.05	158
0.10	126
0.15	95
0.20	70
0.25	52
0.50	17

* See Table 5 for details on design.

$^\dagger d$ = half width of the desired confidence interval
expressed as a fraction of the time mean.

Table 7
Cum \sqrt{f} Rule Arkansas Unit 1 Using Temperature
as the Stratification Variable

Class*	f^{\dagger}	\sqrt{f}	cum $\sqrt{f}^{\dagger\dagger}$
0 - 2	1	1.0	1.0
2 - 4	2	1.4	2.4
4 - 6	9	3.0	5.4
6 - 8	44	6.6	12.0
8 - 10	7	2.7	14.7
10 - 12	4	2.0	16.7
12 - 14	5	2.2	18.9
14 - 16	2	1.4	20.3
16 - 18	5	2.2	22.5
18 - 20	9	3.0	25.6
20 - 22	8	2.8	28.4
22 - 24	6	2.5	30.9
24 - 26	6	2.5	33.3
26 - 28	6	2.5	35.8
28 - 30	21	4.6	40.4
30 - 32	2	1.4	41.8

* 2°C cells.

$^{\dagger}f$ = observed frequency of 24 hour (daily) impingement counts.

†† cum \sqrt{f} = running sum of \sqrt{f} column.

Table 8
Stratified Designs for Arkansas Unit } Using Temperature
as the Stratification Variable

Design No.	L	Boundaries *	Months [†]	T_i ^{††}	Percent Allocation [§]
I	3	0 - 10	Jan-March, Dec	5.0	68.0
		10 - 22	April, Oct, Nov	16.0	16.0
		22 - 32	May-Sept	27.0	16.0
II	4	0 - 8	Jan-March, Dec	4.0	68.2
		8 - 16	Nov	12.0	5.6
		16 - 24	April, May, Oct	20.0	10.4
		24 - 32	June, July Aug, Sept	28.0	15.8

* Stratum boundaries expressed in °C.

† Months that form the temporal stratification.

†† Mid-point of the stratum in °C.

§ Percent allocation as per equation (8).