ENVIRONMENTAL STATISTICS, ASSESSMENT, AND FORECASTING

Chapter 3:
Patil, Gore and Sinha

EDITED BY

C. RICHARD COTHERN
CENTER FOR ENVIRONMENTAL STATISTICS DEVELOPMENT STAFF
U. S. ENVIRONMENTAL PROTECTION AGENCY
WASHINGTON, D. C.

N. PHILLIP ROSS
ENVIRONMENTAL STATISTICS AND INFORMATION DIVISION
U. S. ENVIRONMENTAL PROTECTION AGENCY
WASHINGTON, D. C.

LEWIS PUBLISHERS
Boca Raton Ann Arbor London Tokyo

Library of Congress Cataloging-in-Publication Data

Environmental statistics, assessment, and forecasting / edited by C. Richard Cothern and N. Phillip Ross.
p. cm.
Includes bibliographical references and index.
ISBN 0-87371-936-0
1. Environmental sciences—Statistical methods. 2. Environmental policy. I. Cothern, C. Richard. II. Ross, N. Phillip.
GE45.S75E58 1993
363.7'0072—dc20 93-25598
CIP

Lewis Publishers is an imprint of CRC Press

No claim to original U.S. Government works
International Standard Book Number 0-87371-936-0
Library of Congress Card Number 93-25598
Printed in the United States of America 1 2 3 4 5 6 7 8 9 0
Printed on acid-free paper

Environmen
sented, and in
different discip
in many differe
tivarious indivi
organized the s
the problems t
scribe the state
mental Statisti
aspects of the c
and discussion
ety (ACS) in W
symposium as
of the ACS. W
ing who acted
EPA; Daniel K
Nussbaum, U.

This volume
the fiel en
ume w o
shared by almo
Some call thi
There are as m
in environmen
describing thei
pleteness. Our
unfinished jig
is much work t
picture emerg
several steps tl
the environme

*The thoughts
not necessarily

CHAPTER 3

Environmental Chemistry, Statistical Modeling, and Observational Economy

G. P. Patil, S. D. Gore, and A. K. Sinha

ABSTRACT

Environmental sampling differs from classical theory of sampling in that the former may entail sampling of different types of material and the sampled materials often influence the sampling procedure. Therefore, while determining an optimal sampling design, the physical (chemical or biological) characteristics of the material to be sampled need to be taken into consideration. In this chapter, we discuss several environmental sampling and statistical modeling situations to illustrate how composite sampling and ranked set sampling facilitate observational economy while addressing substantive issues.

INTRODUCTION AND BACKGROUND

It is very important to recognize that there is no substitute for good data. Statistical thinking is an aid to the interpretation of data. The statistical approach is expected to contribute to the overall insight and perspective of the substantive issue and its resolution in the light of the evidence on hand. What makes the problem of environmental investigations different from studies in physical sciences is that, unlike in the hard sciences, we have a longer span of investigations depending on life stages and their age lengths. Also, the instrumentation changes are necessary in response to the advancing technology. That often puts us in a difficult cycle of no information, new information, and noninformation. When a question is asked of us, we promptly say, "Well, we don't have sufficient information. We need to collect new information." By the time new information is collected, we are ready to say that the information

0-87371-936-0/94/$0.00 + $.50

Table 11. Estimated Average Concentrations (pg/g) with Relative Standard Errors (%) for Selected Dioxins and Furans from FY87 NHATS Composite Samples

	Entire nation	Age group (yr)		
		0–14	15–55	45+
Population percentages	100	23	46	31
Compound				
Dioxins				
2,3,7,8-TCDD	5.38	1.98	4.37	9.40
	(6)	(41)	(12)	(4)
1,2,3,7,8-PECDD	10.7	3.30	9.33	18.2
	(4)	(22)	(7)	(4)
1,2,3,4,7,8/	75.1	23.4	70.9	120
1,2,3,6,7,8-HXCDD	(4)	(23)	(6)	(3)
1,2,3,7,8,9-HXCDD	11.7	6.13	10.8	17.1
	(4)	(18)	(7)	(4)
1,2,3,4,6,7,8-HPCDD	110	45.7	99.8	174
	(3)	(11)	(5)	(3)
1,2,3,4,6,7,8,9-OCDD	724	215	692	1150
	(4)	(17)	(7)	(5)
Furans				
2,3,7,8-TCDF	1.88	1.97	1.45	2.45
	(7)	(11)	(15)	(7)
2,3,4,6,7,8-PECDF	9.70	1.87	8.00	18.0
	(8)	(100)	(15)	(8)
1,2,3,6,7,8-HXCDF	5.78	1.80	4.59	10.5
	(13)	(83)	(26)	(13)

Source: Orban et al.[27]

chemicals should be analyzed. First, a chemical must be detected in at least 50% of the composites. Second, a minimum of 30 measurements were considered necessary to achieve sufficient precision of the estimates. Thus, of the 16 chemicals, there were 9 that met both criteria for performing the analyses. For each of the nine chemicals analyzed, Table 11 lists the estimated average concentration in the entire population and in the three age groups.

RANKED SET SAMPLING

Ranked set sampling (RSS) is a method of sampling which is mainly used for estimating a population mean. It utilizes prior information about the characteristic of interest for ranking the randomly selected sampling units from a population before resorting to quantification of some of the units so drawn. It is, in fact, useful when the quantification of a sampling unit is difficult,but the randomly drawn units could be ranked by a visual inspection or some other crude method without knowing their exact measurements. Interestingly, it combines the convenience of purposive sampling and the control of simple random sampling (SRS). As the SRS estimator of a population mean, the corresponding RSS estimator provides an unbiased estimator of the population mean; however, the estimator is more efficient than that of the SRS estimator in almost all situations. In the worst circumstances when a ranking is equivalent to a random ordering, its performance reduces to that of SRS. In view of these facts, it has tremendous potential for considerably economical investigation when we need to estimate the mean in environmental problems. In this section, we have made an attempt to illustrate the method and examine its superiority over SRS for estimating the average PCB concentration in the surface soil along the gas pipeline of the Texas Eastern Gas Pipeline Company. We have also tried to explore its application in evaluating the effectiveness of insecticides and pesticides to protect and preserve the environment. For its application in other areas, see Patil et al.[10]

Method

The selection of a ranked set sample involves the drawing of m random samples each of size m from a population. Having drawn the random samples, the units of each sample are ranked by a visual inspection or any other method not involving the exact measurements of the variable of interest. The unit with the smallest rank is quantified from the first sample; the unit having the second smallest rank is measured from the second sample, and so on until the unit with the highest rank is used for the determination of the magnitude of the characteristic of interest from the m^{th} sample. This yields m measurements corresponding to the quantification of m units out of m^2 randomly selected units, each representing a specific rank. The whole procedure is repeated r times which, in turn, gives mr quantified values in such a way that every rank has r quantified values. It is important to note that m^2r units are randomly selected from the population and used for ranking, but only the mr units are utilized for the determination of the magnitude of the characteristic. These measurements constitute a ranked set sample. If N and n denote the population and the sample size respectively, then:

$$N \geq m^2r \quad \text{and} \quad n = mr$$

In order to illustrate the method of drawing an RSS sample let us take $m = 3$ and $r = 2$. The scheme may be diagrammed following Stokes[28] and Muttlak,[29] as shown above.

In this diagram each row denotes a judgment ordered sample and the circled units indicate the units which are to be quantified. It means that 9 units are selected randomly from the population but only 3 units are quantified to get the required ranked set sample in each cycle.

r \ m	1	2	3
1	⊙	•	•
	•	⊙	•
	•	•	⊙
2	⊙	•	•
	•	⊙	•
	•	•	⊙

In general, let $X_{11}, X_{12}, \ldots, X_{1m}; X_{21}, X_{22}, \ldots, X_{2m}; \ldots; X_{m1}, X_{m2}, \ldots, X_{mm}$ be independent random variables all having the same cumulative distribution function (*cdf*)F. Then the i^{th} order statistic from the i^{th} sample is shown by $X_{(i:m)}$. In case the procedure of drawing random samples is repeated r times then the i^{th} order statistic from the i^{th} sample in the j^{th} cycle is denoted by $X_{(i:m)j}$.

The RSS estimator ($X_{(mv)}$) of the population mean (μ) is computed by:

$$X_{(mv)} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{r} X_{(i:m)j}}{mr}$$

or

$$X_{(mv)} = \frac{1}{m}\sum_{i=1}^{m} X_{(i:m)}$$

since

$$X_{(i:m)} = \frac{1}{r}\sum_{j=1}^{r} X_{(i:m)j}$$

Further as $EX_{(i:m)} = \mu_{(i:m)}$ and $\mu = \frac{1}{m}\sum_{i=1}^{m}\mu_{(i:m)}$ we get $E(X_{(mv)}) = \mu$, where $\mu_{(i:m)}$ denotes the expected value of the i^{th} order statistic. This suggests that $X_{(mv)}$ is an unbiased estimator of the population mean (μ). The expression for the variance of $X_{(mv)}$ is given by:

$$Var(X_{(mv)}) = \frac{1}{m^2}\sum_{i=1}^{m}\frac{\sigma^2_{(i:m)}}{r}$$

where $\sigma^2_{(i:m)}$ represents the variance of the i^{th} order statistic. Also,

$$Var(X_{(mv)}) = \frac{1}{mr}\left\{\sigma^2 - \frac{1}{m}\sum_{i=1}^{m}(\mu_{(i:m)} - \mu)^2\right\}$$

where σ^2 denotes the population variance.

Comparison of the RSS Estimator with the SRS Estimator

In order to examine the performance of the two estimators under consideration we compare the precision of the RSS estimator relative to that of the SRS estimator with the same sample size. To compute the SRS estimator of the population mean, we need to draw a random sample of the size mr. For this, one unit is randomly selected from each sample in each cycle. The unit is then quantified. We denote the SRS estimator by $\bar{X}$. It is computed as shown below:

$$\bar{X} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{r} X_{ij}}{mr}$$

Its variance is obtained as follows:

$$Var(\bar{X}) = \frac{\sigma^2}{mr}$$

The relative precision (RP) is defined as mentioned below:

$$RP = \frac{Var(\bar{X})}{Var(X_{(mv)})}$$

or,

$$RP = \frac{1}{1 - \frac{1}{m}\sum_{i=1}^{m}\left\{\frac{\mu_{(i:m)} - \mu}{\sigma}\right\}^2}$$

Its limits are given by:

$$1 \le RP \le \frac{m+1}{2}$$

Takahasi and Wakimoto[10] gave the rigorous proof of the limits for all continuous distributions with finite variances. It means that almost ($m + 1/2$) times as many random samples are required to equal the precision of the RSS estimator, provided ranking is perfect. The relative cost (RC) and the relative savings (RS) are computed as shown below:

$$RC = \frac{1}{RP}$$

$$RS = 1 - RC$$

and

$$RS = \frac{1}{m}\sum_{i=1}^{m}\left\{\frac{\mu_{(i:m)}-\mu}{\sigma}\right\}^2$$

It means $RS \geq 0$.

Impact of Imperfect Ranking

As ranking of the randomly drawn units are carried out on the basis of some crude method in the absence of the exact magnitude of the characteristic of interest that a unit possesses, it may not always be feasible to perform the ordering correctly. Even in this situation of imperfect ranking Dell and Clutter[31] have shown that the RSS estimator of the population mean is unbiased and the $RP \geq 1$. However, the magnitude of RP gets reduced in this case. In fact, the higher the magnitude of ranking error the smaller is the magnitude of the RP. As a solution to the impasse David and Levine[32] and Stokes[33] have suggested to use some other variable which helps in ranking relatively more correctly and conveniently than the main variable of interest. The variable is known as a concomitant variable in the literature. It is supposed to be correlated with the variable of interest. If both the variables have the same cumulative distribution function and follow the bivariate normal distribution, then the RP of the RSS estimator with the ranking done on the basis of a concomitant variable, and the SRS estimator is given by:

$$RP = \frac{1}{1-\frac{\rho^2}{m}\sum_{i=1}^{m}\left\{\frac{\mu_{(i:m)}-\mu}{\sigma}\right\}^2}$$

where ρ denotes the correlation coefficient between the variable of interest and the concomitant variable. It is evident from the expression that the RP depends on the value of ρ^2.

Unequal Allocation of Sample Sizes

With a view to improve the magnitude of RP, McIntyre[34] and Takahasi and Wakimoto[30] have suggested allocating sample size to each group (i.e., rank order) proportional to its standard deviation.

Let r_i denote the number of times (i.e., the size of the sample of the i^{th} group) the units with rank i to be quantified. Then, it is computed as follows:

$$r_i = \frac{n\sigma_{(i:m)}}{\sum_{i=1}^{m}\sigma_{(i:m)}}$$

$$i = 1, 2, \ldots, m$$

and

$$r_1 + r_2, \ldots, r_m = n, r_i \geq 1$$

If T_i denotes the sum of the measurements for the units having the i^{th} rank, the RSS estimator ($\bar{X}_{(m)w}$) of the population mean is given by:

$$\bar{X}_{(m)w} = \frac{1}{m}\sum_{i=1}^{m}\frac{T_i}{r_i},\ E(X_{(m)w}) = \mu$$

and

$$Var(\bar{X}_{(m)w}) = \frac{1}{m^2}\sum_{i=1}^{m}\frac{\sigma^2_{(i:m)}}{r_i}$$

In order to estimate $Var(\bar{X}_{(m)w})$, r_i should be greater than or equal to two. The RP of the RSS estimator relative to the SRS estimator is defined as follows:

$$RP = \frac{\sigma^2/n}{\frac{1}{m^2}\sum_{i=1}^{m}\frac{\sigma^2_{(i:m)}}{r_i}}$$

and

$$0 \leq RP \leq m$$

See Takahasi and Wakimoto.[30] However, it is important to note that the RP under the equal allocation is less than or equal to that of the unequal allocation provided the allocation is carried out proportional to the standard deviation of each group. Further, if $r_1 = r_2 = \ldots = r_m$ then the RSS design is said to be balanced; otherwise it is unbalanced.

Illustration

With the aim to illustrate the effectiveness of RSS relative to SRS in estimating the level of concentration of polychlorinated biphenyls (PCB) at the Armagh site along the gas pipeline of the Texas Eastern Gas Company, we examine the schemes mentioned below:

1. balanced allocation of samples using all possible combinations of each set size
2. balanced allocation of samples for a specific sample
3. unbalanced allocation of samples

For this purpose we use the measurements of the contaminant obtained by using grids A and C each with phase I and II together. Table 12

Table 12. Number of Observations, Mean, SD, CV, Coefficient of Skewness, and Kurtosis of PCB Values in Grids A and C

Characteristics	Grid A	Grid C
Number of observation	184	68
Mean	200.9	600.2
SD	902.9	1585
CV	4.49	2.64
Coefficient of skewness	9.27	4.48
Coefficient of kurtosis	99.69	20.88

gives the number of observations, mean, standard deviation, coefficient of variation, coefficients of skewness, and kurtosis of PCB values for grids A and C separately. For applying RSS protocol we have considered set sizes two, three, and four. Using all possible combinations of the PCB values for each set size, the magnitude of the relative savings (RS) due to the RSS estimator relative to that of the SRS estimator has been computed using balanced allocation of samples. The findings are mentioned in Table 13. We observe that the value of RS increases with the set size, but its amount is higher for grid C than grid A because the values of PCB under the former are more homogeneous and symmetric than those of the latter. Also, the values of $X_{(m)r}$, RP, and RS have been computed for a specific sample. These results are mentioned in Table 14. It is evident that the results suffer from the sampling fluctuation. We find from Table 15 that the values of RS are quite substantial due to unequal allocation of samples. It is obvious that RSS with this allocation performs much better than with equal allocation. Though it is difficult to determine the proportion of samples in advance in the absence of the standard deviation at each rank order for an unknown population, one could take help of other surveys of similar nature conducted earlier or conduct a preliminary survey based on a small sample size.

Table 13. Relative Saving (RS) Considering All Possible Combinations of Each Set Size Under Perfect Ranking with Balanced Allocation

Set size (m)	Grid A RS	Grid C RS
2	4	9
3	7	16
4	10	22

Table 14. The Values of $X_{(m)r}$, RP, and RS Under Perfect Ranking with Balanced Allocation in the Case of a Specific Sample

Set size (m)	Grid A $X_{(m)r}$	A RP	A RS	Grid C $X_{(m)r}$	C RP	C RS
2	155.035	1.02	2	711.3	1.14	12
3	286.187	1.19	16	223.6	1.01	1
4	166.817	1.07	7	635.9	1.56	36

RSS may also be used for forming relatively more homogeneous composite samples compared to those based on random groupings. With m samples of size m we form m composite samples by physically mixing the units of the same rank before resorting to quantification. Thus, we get mr composite samples out of m^2r units drawn from the population. The standard deviation of these measurements is expected to have smaller variance than the same number of measurements obtained after quantifying the composite samples obtained by random groupings of the units. The results are summarized in Table 16 and 17 for grids A and C, respectively.

In view of the findings we may conclude that it has enormous capability for application in evaluating the impact of chemical treatments, for example, on vegetation, much more economically. Its dependence on

Table 15. The Values of $X_{(m)u}$, RP, and RS Under Perfect Ranking with Unbalanced Allocation of Samples

Set size (m)	Grid A Proportion of samples (exact no.)	$X_{(m)u}$	RP	RS	Grid C Proportion of samples (exact no.)	$X_{(m)u}$	RP	RS
2	1:10 (8, 84)	205.9	1.724	42	1:10 (3, 31)	535.2	2.041	51
2	1:15 (6, 86)	203.1	1.818	45	1:15 (2, 32)	520.4	2.174	54
3	1:4:20 (2, 10, 48)	203.6	2.174	54	1:1.7:1.5 (5, 8, 8)	560.1	1.471	32
3	1:4:25 (2, 8, 50)	201.1	2.326	57	1:2:7 (2, 4, 15)	615.2	1.923	48
4	1:3:5:16 (8, 5, 9, 28)	247.1	1.695	41	1:2:3:4 (2, 3, 5, 6)	576.6	2.083	52
4	1:3:9:27 (2, 2, 10, 30)	226.1	1.316	24	1:1:3:5 (2, 2, 4, 8)	802.4	1.449	31

Table 16. Sample Size, Mean, and Standard Deviation for Individual Samples, Composite Samples, and Composites of Ranked Samples for Grid A

Set size	Item	Sample size	Mean	SD
2	Individual samples	184	200.72	902.9
	Composite samples	92	200.72	627.9
	Composites of ranked samples	92	200.72	618.4
3	Individual samples	180	183.8	870.7
	Composite samples	60	183.8	490.6
	Composites of ranked samples	60	183.8	470.4
4	Individual samples	176	187.8	880.2
	Composite samples	44	187.8	509.8
	Composites of ranked samples	44	187.8	321.5

Table 17. Sample Size, Mean, and Standard Deviation (SD) for Individual Samples, Composite Samples, and Composites of Ranked Samples for Grid C

Set size	Item	Sample size	Mean	SD
2	Individual samples	68	601	1585
	Composite samples	34	601	1067
	Composites of ranked samples	34	601	982.8
3	Individual samples	63	599	1630
	Composite samples	21	599	865
	Composites of ranked samples	21	599	663.3
4	Individual samples	64	590	1618
	Composite samples	16	590	761.0
	Composites of ranked samples	16	590	952.7

prior information of the characteristic of interest for ranking can be tackled effectively to a great extent by utilizing the experience and the expertise of the field personnel. For applications in other areas, see Patil et al.[10]

ACKNOWLEDGMENT

Adapted from the Keynote Address by G. P. Patil to the Symposium on Environmental Statistics of the American Chemical Society, 1992. It has been prepared with partial support from the Statistical Analysis and Computing Branch, Environmental Statistics and Information Division, Office of Policy Planning, and Evaluation, U.S. Environmental Protection Agency, Washington, D. C. under a Cooperative Agreement Number CR-815273. The contents have not been subjected to Agency review and, therefore, do not necessarily reflect the views of the Agency and no official endorsement should be inferred.

REFERENCES

1. G. P. Patil (1991). Encountered data, statistical ecology, environmental statistics, and weighted distribution methods. *Environmetrics*, 2(4), 377–423.
2. G. P. Patil, C. Taillie, and S. Talwalker (1992). Encounter sampling and modeling in ecological and environmental studies using weighted distribution methods. Technical Report 92-0402, Center for Statistical Ecology and Environmental Statistics, Department of Statistics, Pennsylvania State University, University Park, PA.
3. M. T. Boswell, S. D. Gore, and G. P. Patil (1990). Efficiency of various composite sample retesting schemes to classify samples with presence/absence measurements. Technical Report 90-0901, Center for Statistical Ecology and Environmental Statistics, Department of Statistics, Pennsylvania State University, University Park, PA.
4. M. T. Boswell, S. D. Gore, G. D. Johnson, and G. P. Patil (1992). Composite sampling protocols for site characterization and evaluation of cleanup attainment. Technical Report 92-0401, Center for Statistical Ecology and Environmental Statistics, Department of Statistics, Pennsylvania State University, University Park, PA.
5. M. T. Boswell, S. D. Gore, G. Lovison, and G. P. Patil (1992). Annotated bibliography of composite sampling. Technical Report 92-0802, Center for Statistical Ecology and Environmental Statistics, Department of Statistics, Pennsylvania State University, University Park, PA.