Bull. Environm. Contam. Toxicol. 24,211-216 (1980)

Abundance of Organic Compounds in Water

David J. Schaeffer and Konanur G. Janardan*

*Illinois Environmental Protection Agency, 2200 Churchill Road, Springfield, Illinois 62706,
Sangamon State University, Springfield, Illinois 62708*

SHACKELFORD and KEITH (1976) have tabulated the occurrence of 1,258 organic compounds appearing in 5,720 observations on 33 water types from 1970-1976. The development of monitoring requirements under the Toxic Substances Control Act (TSCA 1976), and the questions of the applicability of the "priority pollutant" list and the nature and number of future additions to it, makes it imperative to evaluate the available data and the pending regulations systematically and jointly. Work is underway in Illinois to characterize as fully as possible the organic compounds contained in the discharges of candidate industries and in the receiving waters (SOMANI et al. 1979). This effort has found few priority pollutants in these discharges, and has identified over 150 compounds not in SHACKELFORD and KEITH's list. These findings are consistent with recent extensive tabulations which have also noted that many of the so-called "priority pollutants" are conspicuously absent in industrial discharges (ANON. 1978). These findings raise the question of the number of compounds which are actually in the environment, and the mechanisms which affect their discovery.

SHACKELFORD and KEITH (1976) have compiled 175 lists which give the identities and frequencies of occurrence of organic compounds. Such data constitute the type of information which will be obtained by broad-based monitoring programs. Some of the environmentally significant questions which arise from examining such lists are:

- What is the chance that a specific compound appears in a given list?
- What is the maximum number of times we could expect a compound to appear in a list?
- Is there any significance to compounds occurring more frequently than expected?
- Assuming that we wish to identify one compound from a list as a marker, what is the expected distribution of the frequency of the occurrence of this compound in such lists?
- If a number of lists are merged, how many compounds will appear once, twice, ... n times? Table 1, for example, presents this information for the merged lists in SHACKELFORD and KEITH.

0007-4861/80/0024-0211 $01.20

- Can we use the information from such merged lists to estimate the number of compounds which have not yet been observed?

TABLE 1. Frequency of Occurrence of Organic Compounds

Frequency of* occurrence	Observed no. of compounds	Expected no. of compounds
k	n(k)	m(k)
1	503	544
2	238	212
3	133	117
4	80	75
5	56	53
6	46	39
7	20	30
8	14	24
9	15	20
10	18	16
11	15	14
12	16	12
13	10	10
14	10	9
15	9	8
16	4	7
17	12	6
18	6	5
19	7	5
≥20	46	52

*Statistics are for 28 classes, prior to merging classes 20-28.

Mean = 4.5469 Variance = 63.3681 p = 0.1358 b = 5.7423

Chi Square = 30.093 $\chi^2(22,.95) = 33.924$

In this paper we examine these questions and develop a theoretical model which provides answers to them.

METHODS

Suppose a given compound occurs at some level in all samples of interest. The actual identification of this compound in a given sample depends on several factors such as the level of the compound in the sample, the effectiveness of the recovery method, and the sensitivity and specificity of the analysis (JANARDAN and SCHAEFFER 1979a). Thus, for any given sample, we will either not detect or will detect a compound which is actually present. Then the compound either will not appear or will appear one or more times in the lists.

Let p be the probability that a given compound appears once in a given list; 1-p is the probability that it will not appear. Let X_i be the number of times this compound appears in the i-th list (i=1,2,...,n lists). Let b denote the maximum number of times the compound is expected to appear in the i-th list. Recall that a binomial (BERNOULLI) process is characterized by the conditions that each analysis results either in detection or non-detection; the probability of detection is denoted by p, and that of nondetection by 1-p; each analysis is independent, and the outcome of any particular analysis is not affected by the outcome of any other analysis. It is clear that the mechanism by which compounds appear on the i-th list is described by the binomial distribution (HAHN and SHAPIRO 1968) (Eq. 1):

$$(1)\quad P(X_i=k) = \binom{b}{k} p^k (1-p)^{b-k}$$

$$k = 0,1,2...b, \text{ and } i = 1,2,..., n \text{ (lists).}$$

Let $X=X_1+X_2+,...,+X_n$ be the number of times a specified compound appears in the merged list. Since the sum of binomial variables is also a binomial variable, the probability of occurrence of this compound in n merged lists is given by:

$$(2)\quad P(X=m) = \binom{bn}{m} p^m (1-p)^{bn-m} \qquad m = 0,1,2,...,bn.$$

Equations 1 and 2 provide the probabilities that a given compound will appear in the ith and the merged lists, respectively.

The real intent in merging lists is to describe the overall environmental distribution of all the compounds on the list. Conceptually, a difficulty now arises because Eqs. 1 and 2 include those experiments in which the ith compound has not occurred (i.e., where k or m=0). In forming the merged lists all compounds are considered but some compound, j, which was actually present in the environment may not be observed in these lists. Considered over all compounds in the merged list, m=0 (Eq. 2) now implies that we have knowledge of the number of such unobserved compounds! Since this is not possible, Eq. 2 is modified by letting Y equal the number of (nonzero) times a given compound has actually been observed. Then

$$(3)\quad P(Y=k) = \frac{1}{k!} \left(\frac{d}{dt}\right)^{k-1} \{(1-p+pt)^b\}^k \Big|_{t=o}$$

$$P(Y=k) = \frac{1}{k} \binom{bk}{k-1} p^{k-1} (1-p)^{bk-k+1}.$$

If the total frequency of occurrence of all compounds in the merged list is N, the number of compounds occurring exactly k (=1,2,...) times is:

(4) $n_k = N\ P(Y=k)$.

The probability function given by (3) is known as the Generalized Geometric Distribution (GGD) (PLUNKETT and JAIN 1975).

DISCUSSION

The GGD can be used to examine the questions raised before. In order to explore these questions, the model must be fitted to the actual data (such as in Table 1) of SHACKELFORD and KEITH (1976). In this case we wish to estimate p and b using the data in Table 1. These are obtained by the method of moments (HAHN and SHAPIRO 1968) as:

(5) $\hat{p} = 1-(S^2/\bar{X}^2(\bar{X}-1))$

$\hat{b} = \bar{X}(\bar{X}-1)^2/(\bar{X}^2(\bar{X}-1)-S^2)$

where $\bar{X}$ and S^2 are the sample mean and variance, respectively.

For the data in Table 1, $\bar{X}$=4.54, S=63.37, $\hat{p}$=0.135846, $\hat{b}$=5.74231. The theoretical frequencies (Eq. 4) are given in the last column of Table 1. The computed chi-square (30.09) is less than the theoretical value of 33.92 at the 95% probability level with 22 degrees of freedom. Thus, the chance that a given compound will appear in a particular list is given by the estimated value of p (=0.1358).

For the 175 lists comprising Table 1, any given compound appears in approximately 14% (25) of the lists. Thus, as states begin to develop lists of the occurrence of compounds in their waters, there is only a 14% chance of a specific compound being found in a given list. Assuming that the priority pollutants are members of SHACKELFORD and KEITH's (1976) compilation, we infer that the probability of observing any one of these compounds in a given state's list is 14% and the probability of observing t such pollutants in the same list is $(0.14)^t$.

If a number of lists are available, the maximum number of times a given compound is expected to appear is given by b (=6). Thus, even if a compound is actually present frequently in the environment, it will probably not appear in any given list more than 6 times. If a given compound appears more frequently than this, it suggests that its absolute rate of occurrence and detection is greater than that anticipated based on Table 1. Such changes could result from improved detection procedures, higher levels of discharge which favor increased detection, a different data base (i.e., selection of sample types), etc. The low probability (14%) of observing a given compound and the low mean occurrence (4.5) mitigate against developing inferential procedures for quality assurance or compliance which rely on a decreased frequency of observation (less pollution) than heretofore obtained.

It has been seen so far that the GGD (or the considerations leading to it) provides valuable environmental information. This is perhaps most important in the estimate it provides of the number of compounds already in the environment which we are likely to observe using present analytical techniques. Thus, the number of unobserved compounds is given as the frequency of the zeroth class, n_0, by:

$$(6) \quad n_0 = \sum_k |1-n_k/m_k|$$

where n_k is the observed number of compounds and m_k is the expected number compounds appearing exactly k times (in Table 1).

The estimated number of new compounds which could be found in another list similar in scope and size to SHACKELFORD and KEITH's is 351. This estimate compares exactly with those obtained by the authors using other techniques (JANARDAN and SCHAEFFER 1979b), where it is shown that in a single data base equal to SHACKELFORD and KEITH's, and in 24 such compilations, 348 and 1,200 new compounds respectively are likely to be identified.

CONCLUSION

Simple considerations lead to a powerful but easily utilized model for predicting the behavior of trace organics in the environment. Obviously, this technique can be applied to many similar problems such as air pollution data, trace inorganics, etc. Further, a data base such as SHACKELFORD and KEITH's could be analyzed in subsets comprising various sample types such as finished drinking waters, specific industrial categories, etc. For each of these, relevant estimates can be obtained to guide the collection and interpretation of new data.

REFERENCES

ANON., "EPA Priority Pollutant Rules Taking Shape," Chem. and Eng. News, September 25, p. 41 (1978).

HAHN, G.J., and S.S. SHAPIRO, Statistical Models in Engineering, New York, John Wiley (1968).

JANARDAN, K.G., and D.J. SCHAEFFER, "Propagation of Random Errors In Estimating the Levels of Trace Organics in Environmental Sources," Anal. Chem. 51, 1024 (1979a).

JANARDAN, K.G., and D.J. SCHAEFFER, "Some Models for Predicting Organic Pollutants in the Aquatic Environment," Tenth Annual Pittsburgh Conference on Modeling and Simulation, University of Pittsburgh, Pittsburgh PA April 25-27, 1979b.

PLUNKETT, I.G., and G.C. JAIN, "Three Generalized Negative Binomial Distributions," Biom. Z. 17, 286-302 (1975).

SHACKELFORD, W.M., and L.H. KEITH, Frequency of Organic Compounds Identified in Water. United States Environmental Protection Agency EPA 600/4-76-062 (1976).

SOMANI, S.M., J. JOHNSTON, K.G. JANARDAN, and D.J. SCHAEFFER, Assessment of Trace Organics in Illinois Discharges, Report of Work October 1, 1977 - December 31, 1978. Illinois Environmenal Protection Agency (1979).

Toxic Substances Control Act, PL94-469 15 USC 2601 et. seq., October 12, 1976.