A Bayesian Approach to Determining the "Paternity" of Environmental Contamination

Douglas E. Splitstone
SPLITSTONE & ASSOCIATES
4530 William Penn Hwy. #110
Murrysville, Pennsylvania 15668
Phone (724) 325-7421 Fax (724) 327-5958

Michael E. Ginevan
Exponent
1730 Rode Island Ave., NW
Washington, DC 20036
Phone (202) 441-6484 Fax (703) 834-2707

Most evaluations of the contribution of different sources to environmental contamination at different locations begin with a set of measurements of chemical species at different locations. One then calculates either a variance-covariance or correlation matrix and performs some type of factor analysis. The resulting multivariate patterns shown in the factor solution are then used to identify the contribution of different sources to contamination and proportion the cleanup cost. The present discussion takes a different approach that derives from the Bayesian analysis of genetic data to determine the likely paternity of a given child given a set of data from potential fathers, the mother and the child. We demonstrate the use of two variants this approach. One assumes independence of chemical contaminant production while the other considers the specific multi-contaminant production. In both we begin with a non-informative prior that assumes that all sources are equally likely to have contributed to the site contamination and use the data to calculate a posterior probability of each source being responsible for the contamination.

1.Introduction

Determining the origins of environmental contamination at a site usually begins with a number, K, of sample points which have been characterized with respect to the concentrations of a number of chemical species. That is we have a number of samples each of which is characterized by a vector, $\mathbf{y}$, whose elements, $y_1..y_N$ consist of the measured concentrations of the N chemical species of interest. There are two broad approaches to using data like these. In the first, which can be termed factor approaches, the measurements are used to calculate the NxN correlation matrix, R, of the concentrations of the chemical species of interest. This matrix is then used in a principal component or factor analysis which yields a series of coefficient vectors, β. The relative magnitude of scores of a given observation on each β vector can be interpreted as the relative contribution of a given source to the chemical contaminants in the sample. Examples of this approach include Paterson et al. (1999), Rachdawong and Christensen (1997) and Lee et al. (2003).

Alternatively one can use the data to calculate a K x K similarity matrix, S, whose elements are the distances between the various samples using some measure of similarity such as euclidian or Mahalanobis distance. Because the relative concentrations of the chemical species in a sample are of greater interest than the absolute concentrations (which can change due to simple dilution) concentrations are usually normalized according to some criterion. Once one has calculated a similarity matrix, one can cluster the samples using one of a number of clustering algorithms to identify groups of similar samples, which one might infer have a common origin. This approach is discussed in Beebe et al (1998).

These approaches are useful in identifying interesting patterns of environmental contamination, but do not directly address questions which are often central to environmental contamination investigations. "What is the probability that the contamination was generated by a particular source?" And, "Who contributes how much to paying the cost of cleanup?"

It has been the author's experience that most often none of the above mentioned statistical techniques are employed to seek answers to the salient questions. Instead, one or more "experts" review the production and sales history of the designated principal responsible parties (PRPs) and apportion costs assuming that the difference between production and sales winds up as site contamination.

If we consider the apportioning of cleanup cost as one of determining paternity of the site contamination then one can, with certain modifications use a forensic approach (e.g. Aitken, 1995) to determine the desired probabilities. This approach utilizes the "evidence" provided by
the assessment and description of site contamination to assist in proportioning cleanup cost. Traditionally this evidence is used only to determine the total cost of cleanup.

Determining Paternity

To put things in perspective let us consider the determination of the likelihood of paternity. We rely heavily on the discussion of this problem as presented by Berry and Geisser (1986).

Assume that six men (M_i, I=1,...,6) are the only possibilities as the father of the child in question. All are deemed equally likely on the basis of evidence other than blood type, in other words the probability of paternity for each man, $Pr(M_i)$, is the same and equal to 1/6. This is the "prior" probability.

The "evidence," E, consists of the blood-type information given for the child, the mother and each of the six potential fathers. This is presented in the following Table 1:

Table 1. The Evidence - Blood Type

Person	Child	Mother	Mr. 1	Mr. 2	Mr. 3	Mr. 4	Mr. 5	Mr. 6
Phenotype	O	O	O	A	B	AB	O	A

The likelihood of the father being M_i based upon the evidence E is the conditional probability that the evidence would have resulted if M_i were the father, $Pr(E \mid M_i)$. What is of interest is the "posterior" probability that M_i is the father given the evidence, $Pr(M_i \mid E)$. This posterior probability is given by Bayes' theorem:

$$Pr(M_i \mid E) = Pr(E \mid M_i)\, Pr(M_i)/K.$$

The constant K is determined by the requirement that the total probability is 1:

$$K = Pr(E) = Pr(E \mid M_1)\, Pr(M_1) + Pr(E \mid M_2)\, Pr(M_2) + \ldots .$$

The information required to calculate $Pr(E \mid M_i)$ comes from Table 1 and the genotypic and phenotypic frequencies for the white California population. The results are as follows:

Table2. Posterior Probability of Being the Father

Possible Father	M_1	M_2	M_3	M_4	M_5	M_6
Prior $Pr(M_i)$	1/6	1/6	1/6	1/6	1/6	1/6
$Pr(E \mid M_i)$	1	0.431	0.472	0	1	0.431
Posterior	0.300	0.129	0.141	0	0.300	0.129

Obviously, the likelihood of being the father represented by the posterior probability is quite different from the prior probability.

Cost Allocation

In constructing the parallel example for cost allocation consider a site where the soil is contaminated by several analytes. These are arsenic, total BNA's, total DDT, dioxins/furans lead and other metals. Production began at the site in1900 and ceased in the late 1950's. During that time various operations were active on the site including the distillation of spirits for human consumption, the manufacture of pesticides and chemical compound as well as the repackaging and distribution of these and other products.

The "evidence," E, is the volume of site soil contaminated by each of the 63 contaminant combinations (See Appendix Table). This is the volume of soil requiring treatment according to the combinations of contaminants found (the soil either requires treatment because of contaminant X or it does not), and $63 = 2^6 - 1$. The one is obviously the case of soil which is not contaminated and can be left in place.

Assume there are four possible responsible parties Company A, Company B, Company C and Company D which are successor companies to one or more of those operating at the site. The initial assessment of relative responsibility is based upon production and is shown in the following Table 3.

Table 3. Relative Responsibility Based On Production

PRP Company (M_i)	Arsenic	BNA	DDT	Dioxins & Furans	Lead	Other Metals
A	0.949	0.914	0.005	0.610	0.944	0.0
B	0.0	0.086	0.995	0.059	0.055	0.0
C	0.001	0.0	0.0	0.331	0.001	1.0
D	0.050	0.0	0.0	0.0	0.0	0.0

The conditional probability of the evidence given the company, $Pr(E \mid M_i)$, may be calculated in two ways. The first is to assume that a company must have produced <u>each</u> constituent identified as a contaminant in one of the 63 types of contaminated soil to be considered as a contributor for that type. That is, if a company produced or used potential contaminants X <u>and</u> Y and the soil was identified as contaminated with X <u>and</u> Y, the probability that the company contributed is

$$Pr(X \text{ and } Y) = Pr(X)\, P(Y).$$

We have assumed the statistical independence among the production of various potential contaminants. We have also weighted the company's contribution to each of the 63 soil types by the relative soil volume within each type. The desired value of $Pr(E \mid M_i)$ is the accumulated probability of the company's contribution summed over the 63 soil types.

Assuming a uniform prior distribution, ie. each company is initially equally responsible, the following results:

Table 4. Bayes Allocation for "And" Contribution

	Potentially Responsible Company			
	A	B	C	D
Prior Probability	0.25	0.25	0.25	0.25
$Pr(E \mid M_i)$	0.438	0.208	0.060	0.037
Posterior Probability, $Pr(M_i \mid E)$	0.589	0.280	0.081	0.050

The second way of calculating the conditional probability of the evidence given the company $Pr(E \mid M_i)$ is to assume that if a company produced or used any constituent identified as a contaminant in one of the 63 types of contaminated soil, it is to be considered as a contributor for that type. That is, if a company produced or used potential contaminants X or Y and the soil was identified as contaminated with X or Y, the probability that the company contributed is

$$Pr(X \text{ or } Y) = Pr(X) + Pr(Y) - Pr(X)\,P(Y).$$

The "or" as used here is the inclusive "or."

As the number of contaminants increases, the number of terms required for the calculation of the correct probability increases rapidly. Seven terms are required for calculating Pr(X or Y or Z). Sixty three are required in the calculation that any one of the six identified contaminants are present. This makes the calculation of $Pr(E \mid M_i)$ somewhat tedious.

As in the "And", case above, we have assumed the statistical independence among the production or use of various potential contaminants for this "Or" case. I have again weighted the company's contribution to each of the 63 soil types by the relative soil volume within each type.

Table 5. Bayes Allocation for "Or" Contribution

	Potentially Responsible Company			
	A	B	C	D
Prior Probability	0.25	0.25	0.25	0.25
$Pr(E \mid M_i)$	0.810	0.478	0.182	0.037
Posterior Probability, $Pr(M_i \mid E)$	0.537	0.317	0.121	0.025

Note that the posterior probabilities of company responsibility given the evidence are somewhat different for the "And" and "Or" cases. In addition, a uniform prior probability may not be a correct assumption in view of the site history. A uniform prior is frequently used in the absence of any additional information. The prior may be adjusted to take into account additional information such as a qualitative assessment of the housekeeping practices of each company.

Conclusion

Section 113(f)(1) of the Comprehensive Environmental Response, Compensation and Liability Act ("CERCLA") provides:

> Any person may seek contribution from any other person who is liable or potentially liable under Section 9607(a) of this title, during of following any civil action under Section 9606 of this title. . . Nothing in this subsection shall diminish the right of any person to bring an action for contribution in the absence of a civil action under Section 9606 of this title of Section 9607 of this title.

Industry threatened with a Section 113 action has traditionally relied upon an ad hoc approach to apportioning clean up costs among PRPs. This approach is usually based upon production and sales records and ignored the evidence provided by actual site contamination. The alternate approach suggested here can take into account production, maintenance and the evidence provided by the site contamination assessment.

REFERENCES

Aitken, C.G.G. 1995. *Statistics and the Evaluation of Evidence for Forensic Scientists*. John Wiley, NY.

Berry, D.A. and S. Geisser 1986. Inference in Cases of Disputed Paternity in *Statistics and the Law*, eds M. H. DeGroot, S. E Feinberg & J. B. Kadane, John Wiley.

Beebe, K.R., R.J. Pell, and M.B. Seasoltz. 1998. Chemometrics: A Practical Guide. John Wiley, NY.

Lee, P. K. H., J. R. Brook, E. Dabek-Zlotorzynska, and S. A. Mabury. 2003. Identification of the Major Sources Contributing to PM 2.5 Observed in Toronto. Environ. Sci. Technol. 37(21): 4831-4840.

Paterson, K. G.; J. L Sagady, D. L Hooper, S. B. Bertman, M. A. Carroll, and P. B. Shepson. 1999. Analysis of Air Quality Data Using Positive Matrix Factorization. Environ. Sci. Technol. 33(4): 635-641.

Rachdawong, P. and Christensen, E. R. 1997. Determination of PCB Sources by a Principal Component Method with Nonnegative Constraints. Environ. Sci. Technol. 31(9): 2686-2691.

Appendix Table. Example Site Soil Volume Distribution.

Case	Arsenic	BNAs	DDT	TCDD	Other Metals	Lead	Relative Volume	Volume in yd^3
1	1	0	0	0	0	0	0.333	18090
2	0	1	0	0	0	0	0.001	46
3	1	1	0	0	0	0	0.002	93
4	0	0	1	0	0	0	0.106	5752
5	1	0	1	0	0	0	0.074	4016
6	0	1	1	0	0	0	0.016	868
7	1	1	1	0	0	0	0.018	984
8	0	0	0	1	0	0	0.056	3056
9	1	0	0	1	0	0	0.035	1921
10	0	1	0	1	0	0	0	0
11	1	1	0	1	0	0	0.02	1088
12	0	0	1	1	0	0	0.022	1204
13	1	0	1	1	0	0	0.046	2512
14	0	1	1	1	0	0	0.012	660
15	1	1	1	1	0	0	0.035	1910
16	0	0	0	0	1	0	0.02	1111
17	1	0	0	0	1	0	0.011	579
18	0	1	0	0	1	0	0.001	35
19	1	1	0	0	1	0	0	0
20	0	0	1	0	1	0	0.008	440
21	1	0	1	0	1	0	0.004	197
22	0	1	1	0	1	0	0	12
23	1	1	1	0	1	0	0.006	313
24	0	0	0	1	1	0	0	0
25	1	0	0	1	1	0	0	0
26	0	1	0	1	1	0	0	0
27	1	1	0	1	1	0	0	0
28	0	0	1	1	1	0	0	0
29	1	0	1	1	1	0	0	0
30	0	1	1	1	1	0	0.001	35
31	1	1	1	1	1	0	0.005	278
32	0	0	0	0	0	1	0.008	417
33	1	0	0	0	0	1	0.024	1285
34	0	1	0	0	0	1	0	0
35	1	1	0	0	0	1	0	0
36	0	0	1	0	0	1	0	0
37	1	0	1	0	0	1	0.026	1424
38	0	1	1	0	0	1	0	0

Appendix Table (Continued). Example Site Soil Volume Distribution.

Case	Arsenic	BNAs	DDT	TCDD	Other Metals	Lead	Relative Volume	Volume in yd^3
39	1	1	1	0	0	1	0.008	440
40	0	0	0	1	0	1	0	0
41	1	0	0	1	0	1	0.004	231
42	0	1	0	1	0	1	0	0
43	1	1	0	1	0	1	0.012	637
44	0	0	1	1	0	1	0.001	35
45	1	0	1	1	0	1	0.017	949
46	0	1	1	1	0	1	0	0
47	1	1	1	1	0	1	0.042	2292
48	0	0	0	0	1	1	0	0
49	1	0	0	0	1	1	0.003	162
50	0	1	0	0	1	1	0	0
51	1	1	0	0	1	1	0.001	46
52	0	0	1	0	1	1	0	0
53	1	0	1	0	1	1	0.004	197
54	0	1	1	0	1	1	0	0
55	1	1	1	0	1	1	0.003	162
56	0	0	0	1	1	1	0	0
57	1	0	0	1	1	1	0.001	58
58	0	1	0	1	1	1	0	0
59	1	1	0	1	1	1	0.001	69
60	0	0	1	1	1	1	0	0
61	1	0	1	1	1	1	0.003	185
62	0	1	1	1	1	1	0	0
63	1	1	1	1	1	1	0.009	475
Totals							1.000	54264