Journal of Hazardous Materials, 13 (1986) 207—216
Elsevier Science Publishers B.V., Amsterdam — Printed in The Netherlands

STATISTICAL APPROACHES TO GROUNDWATER MONITORING

CARL A. SILVER

Environmental Institute for Waste Management Studies of the University of Alabama and

Department of Quantitative Methods, Drexel University, Philadelphia, PA 19104 (U.S.A.)

(Received May 23, 1985; accepted in revised form October 20, 1985)

Summary

Current and proposed regulations require that some form of Student's *t*-test be used for evaluation of groundwater pollution detection parameters. Several problems arising from that requirement lead to an inflated false alarm rate. Among these are the use of variability among replicates as the estimate of random sampling error, and the failure to take proper account of spatial and temporal sources of variation. In addition, conditions essential for the valid application of the *t*-test are usually lacking. The requirement that a separate *t*-test be conducted for each detection parameter, each reporting period, causes an additional severe increase in the likelihood of a false alarm. A dummy variable analysis of covariance is suggested as a desirable alternative. After statistically removing temporal variation from the data the averages of the upgradient and downgradient wells are compared. Examples of the use of this form of analysis are given.

Introduction

One purpose of a groundwater monitoring program is to protect the environment by measuring concentrations or detecting changes in significant indicators of contamination. To achieve this purpose in a valid, efficient, and effective manner a program must be designed so as to insure that environmental impacts will not go undetected while, at the same time, minimizing the likelihood of false detections. Because measurements of indicators vary from time to time, place to place, and include laboratory errors, it is widely held that detection programs must be statistically based. A minority opinion holds that a non-statistical (medical) model is more appropriate. Those who support such an approach point out that random sampling, a requirement for valid statistical analysis, seldom, if ever, obtains in field situations. They liken the assessment of the safety of a waste disposal site to the assessment of the health of a patient. The author is not unsympathetic to that point of view. The current regulatory climate, however, makes some form of statistical analysis mandatory. Accordingly, the present paper addresses the problems in statistical monitoring of hazardous waste disposal sites.

0304-3894/86/$03.50

Several approaches to statistical analysis of data from groundwater monitoring programs have been proposed or are in use. The author of the present report addresses some of these approaches and recommends use of more satisfactory statistical methods.

Detection of contaminants in groundwater may be attempted by either of two processes: (1) Detection of a contaminant occurs whenever a concentration significantly different from zero is discovered in groundwater. Statistical consideration is restricted to determination of uncertainty in any analytic measures, especially at very low concentrations. Statistical methods are used to decide whether an apparent detection represents the real presence of the material in question, or random fluctuations (laboratory error) in the measurements. (2) Detection is attempted against a background in which significant non-zero values of parameters can be expected in groundwater, whether contaminated or not. Specific conductance, pH, and total organic carbon are typical of such parameters. The logic of detection of groundwater contamination is based upon the comparison of groundwater sampled at site known to be free of contamination from a given source (an up-gradient site) with groundwater sampled from a site potentially exposed to contamination (a down-gradient site). When, a "significant" difference exists between up-gradient and down-gradient groundwater samples one concludes that the source has contaminated the groundwater. Fundamental to this analysis is the concept that, absent any impact, the up-gradient and down-gradient samples would not differ. This essential assumption is seldom, if ever, tested and is questionable in view of the fact that the transit time through an aquifer from an up-gradient to a down-gradient site may range from months to millenia.

Assumptions of random sampling and independence of observations are central to almost all statistical methods. Statistical procedures assume that, at some stage, there has been an element of randomness in the way the data have been selected. Randomness of analytical data from groundwater analyses may result from spatial or temporal factors. However, no provision for obtaining data through random sampling is included in any approved regulation or statute dealing with monitoring of groundwater. At the present time there is no agreement among statistical experts as to the best way to incorporate necessary randomness in sampling procedures for groundwater monitoring.

Independence of observations means that knowledge of any one data point does not decrease one's uncertainty about any other data point. This assumption, essential to many commonly used statistical tests, is violated in the extreme by methods of analysis required by regulations. One measure of independence is provided by the correlation coefficient, a number ranging between -1.0 and +1.0. A correlation coefficient, r, of 0.0 indicates linear independence among observations. Analysis of numerous data sets using the indicator parameters (pH, specific conductance, total organic carbon (TOC), and total organic halogen (TOX)) show correlation coefficients of 0.99 and

higher among replicate measures. Nevertheless, these measures are, under current regulations, such as those specified in 40 CFR Part 264, treated as independent observations. The results may be described as statistical chaos.

t-Tests

Several forms of Student's *t*-test have been employed in the analysis of data collected in groundwater monitoring programs. While forms of *t*-tests differ in detail, they share a number of similarities. (In the discussion that follows the Greek letter μ denotes a population mean, and the symbol $\overline{X}$ represents a sample mean.) *t*-Tests are statistical procedures for determining whether two sample means ($\overline{X}$s) differ by more than could be expected from mere sampling errors when, in fact, the population means (μs) are equal.

When conducting statistical tests on differences between $\overline{X}$s one usually begins by assuming the *null hypothesis*, i.e., that the two μs estimated by the two $\overline{X}$s are equal. If the $\overline{X}$s differ by more than can be attributed to variation in random samples, one concludes that the population means differ. There is, of course, a probability that the null hypothesis will be rejected even though the population means do not, in fact, differ. This error is known as an error of Type I and is usually signified by the Greek letter alpha (α). One can, in principle, choose any desired value of α. The practice in groundwater monitoring has been to use $\alpha = 0.01$.

Another error results from failure to reject the null hypothesis when, in fact, it is false. The likelihood of such an error, a Type II error, is generally indicated by the Greek letter beta (β). For any given value of α the value of β will be a function of sample size, n, and the true difference between the μs. The greater the difference between the μs and the larger the sample sizes the smaller β will be. The *power* of the statistical test is $1-\beta$. Unlike the case with α, there does not appear to be a generally accepted value for β, or for the difference between μs.

A *t*-test, when applicable, is a desirable statistical test. Not only is it widely known, easy to calculate, and available in most statistical software packages; it is also, when properly used, among the most powerful statistical techniques. Like many other statistical tests, *t*-tests are ratios between observed differences in $\overline{X}$s and estimates of sampling error. Several different methods of estimating sampling error are in current use in the field of groundwater monitoring. Some of these are discussed below.

Present regulations (40 CFR 246) stipulate that groundwater monitoring data, obtained in a single quarter from four down-gradient wells, be compared with the average of data obtained over the period of a year from one or more up-gradient wells. For example, during a control year one or more up-gradient wells are sampled quarterly. A number of determinations, typically four, are made on a sample from each well each quarter. The mean and standard deviation of the resulting data set are used to test whether the mean of each down-gradient well, measured during some later time period (quar-

ter), differs from the up-gradient mean. The mean and standard deviation of measurements from the down-gradient well are based upon multiple determinations (again, typically, four) of water sampled from the well in question during the quarter being evaluated. Some form of Student's *t*-test is used to determine statistical "significance". The most general form is shown below

$$t = \frac{\bar{X}_1 - \bar{X}_2}{[(DF_1 S_1^2 + DF_2 S_2^2)(1/n_1 + 1/n_2)]^{0.5}}$$

where subscripts 1 and 2 represent up-gradient and down-gradient, respectively, S^2 is the unbiased variance estimator (mean-square), n is the sample size, and DF is degrees of freedom (n-1).

In this approach the numerator of the *t*-test is the difference of the up-gradient and down-gradient means. The denominator is based upon the variability (standard deviation) of the up-gradient and down-gradient data. It is important to consider the meaning of that denominator. Clearly, the down-gradient variability contains nothing but laboratory error. All of the values upon which it is based were taken on a single occasion from a single well. Neither temporal (quarter-to-quarter) nor spatial (well-to-well) variation is present.

In the up-gradient data set, however, both laboratory error and temporal variability are present. Usually, laboratory error is such a small proportion of the total variability that it may be ignored (data illustrating this phenomenon are presented later in this discussion). A *t*-test based upon such data, therefore, compares the differences of the averages of wells in different locations (numerator) with the amount of variation arising from temporal and laboratory sources. This comparison is misleading. Laboratory error can be made arbitrarily small by improving laboratory techniques. Thus, any observed difference between means could, at least theoretically, be made statistically "significant".

Sample means will differ across both time and space because of naturally occurring "background" (non-polution source related) factors. Such factors may include seasonal, temperature, microgeological, or other sources of variability. Each of these factors may generate much larger variation than will laboratory errors.

The *t*-test uses both laboratory error and temporal variation in the denominator. The temporal (seasonal) variation may be expected to be much larger than the laboratory component, typically by one or more orders of magnitude. The *t*-test compares groundwater data from down-gradient well, measured at one time, with the average of up-gradient wells measurements made at different times. In the numerator variation may be due to four factors: laboratory error, naturally occurring temporal variation, naturally occurring spatial variation, and contamination from a population source. The denominator contains variation from laboratory error and seasons. When more than one up-gradient well is used the denominator also contains spatial variation. When a single up-gradient well's data are tested against data from a single

down-gradient well the denominator is clearly inappropriate because only the numerator includes spatial variation. When the denominator contains spatial variation (more than one up-gradient well) a problem remains, but is less obvious.

Temporal variation in many groundwater parameters, e.g., pH, specific conductance, is cyclic. Testing results from a given quarter against an annual average may result in testing the peak of a cycle against the mean value of that cyclic system. Such comparisons will often result in a significant *t*-test, demonstrating that quarters differ from each other by more than random sampling error. This kind of inappropriate *t*-test often leads to a large number of "significant" results in the "wrong" direction, i.e., up-gradient wells showed more contamination than did the corresponding down-gradient wells. One such instance resulted in discovering 24 "significant" tests at the $\alpha = 0.01$ level among 40 *t*-tests. Of those 24, 14 were in the "wrong" direction. More appropriate analysis showed no significant differences among the wells for any of the measured parameters. The apparent differences between up- and down-gradient wells were due, in part, from failure to consider seasonal differences and, in part, from lack of independence among the replicate samples. (So little information is gained from replicate samples in most instances that, for a constant budget, it would be better to spend the costs of replicates in some other way, such as installing more wells, or more frequent sampling. Little statistical justification exists for more than one replicate (two measures) for each parameter.)

In another approach *t*-tests are used to compare each down-gradient well against the average of one or more up-gradient wells using only data from the time period (quarter) under investigation. Use of one up-gradient well is equivalent to the use of laboratory error as the sole basis of comparison. In such cases almost every comparison turns out to be "significant" because virtually all wells can be expected to differ by more than laboratory error. Data from multiple up-gradient wells, however, will reduce the number of false positives. The problem is not completely solved by increasing the number of up-gradient wells because the variability of the down-gradient measures still includes only laboratory error. Furthermore, such a procedure lacks statistical power because only a small part of the available data is being used in each comparison. Even when multiple up-gradient wells exist the *t*-test does not have as much statistical power as procedures which use all available data.

If there are several down-gradient wells a number of *t*-tests must be performed, one for each down-gradient well. If the chosen value of Type I error (α) is 0.01 the probability that at least one test will be significant by chance is $1-(1-\alpha)^n$, where n is the number of *t*-tests being conducted. For $n = 6$, that probability is 0.0585. When $n = 10$ the probability of at least one false positive rises to 0.0956, almost ten times greater than the acceptable value. The use of many *t*-tests requires that some additional precaution be taken to insure that the nominal α and the achieved α be the same. *A posteriori*

tests have been designed for this purpose, but their use is complex and often unsatisfactory. The problem can be avoided by using a single F-test instead of multiple t-tests. For this, and other reasons, analysis of covariance deserves special consideration.

Analysis of covariance

Important considerations for a statistical test of analytical data include:

(1) use of all of the available relevant data
(2) removing the effects of temporal variation
(3) taking proper account of spatial variation
(4) maintaining the desired Type I error rate
(5) maximizing statistical power

The analysis of covariance provides all of these desirable characteristics. Furthermore, the procedure is well documented in widely distributed textbooks, and is available in many statistical analysis packages for computers, e.g., SAS, BMDP, SPSS, etc. The examples used below were computed using the SAS package [1].

The analysis of covariance is a procedure which combines features of analysis of variance and regression analysis. The procedure first removes the effects of a covariable (in the present case, time or "quarter") from the values of the dependent variable and then performs an analysis of variance on the residuals. These residuals represent values that would have been obtained if all of the data had been obtained at the same time, i.e., if there had been no seasonal variation.

The data in Table 1 for the variable specific conductance were gathered at a waste disposal site which had two up-gradient and three down-gradient wells.

In Table 1, the following abreviations have been used: SOURCE — the source of variation; DF — degrees of freedom; SS — sum of squares; MS — mean square (SS/DF); F — the F statistic; PROB — the probability of observing an F as large as, or larger than, the one obtained by chance alone; R-

TABLE 1

Analysis of variance for specific conductance

Source	DF	SS	MS	F	PROB
Model	9	6144329.49	682703.27	122.66	<0.0001
Error	158	879394.16	5565.78		
Total	167	7023723.66	R-square = 0.875		
Time	5	605746.06	121149.21	21.77	<0.0001
Loc	1	218904.41	218904.41	0.12	0.7486
Well(Loc)	3	5319679.03	1773226.34	318.59	<0.0001

square — the proportion of variance accounted for by the model; Model — the general linear model used in the analysis (in this case specific conductance = *f*(Time, Loc, Well(Loc)); Time — the quarter in which the data were taken; Loc — location of the well from which the data were taken (up-gradient of down-gradient); Well(Loc) — well within location.

Well(Loc) plays an important role in the analysis. In order to say that there is a real difference between up-gradient and down-gradient, there should be more difference between the up-gradient and down-gradient $\bar{X}$s than can be explained on the basis of variation among the up-gradient wells or the down-gradient wells. Well(Loc) provides a pooled estimate of the variation among $\bar{X}$s of wells in the same location. The ultimate statistical test in the procedure being described is to devide the mean square for Loc by the mean square for Well(Loc). When there is no real difference in the analytic data between locations that ratio should be about unity. Ratios sufficiently larger than unity lead to the conclusion that there is a real (non-random) difference in the parameter of interest between up-gradient and down-gradient wells.

The data set consisted of 168 observations as shown in Table 2. These 168 observations produced 167 degrees of freedom for the Total *SS* in Table 1. The six quarters provided in the five *DF* for Time. Between the two locations, "up" and "down", there is one *DF*. The two up-gradient wells (UP-1, UP-2) accounted for one *DF* while the three down-gradient wells have two *DF*, comprising the three *DF* for Well(Loc). The remaining 158 *DF* for Error arise mainly from laboratory error.

TABLE 2

Design matrix showing number of observations made for each well at each time period

Quarter	Well				
	Up-1	Up-2	DN-1	DN-2	DN-3
1	4	—	—	4	4
2	4	—	4	4	4
3	4	4	4	4	4
4	4	4	4	4	4
5	4	4	4	4	4
6	16	16	16	16	16

The *F*-test for the model (model *MS*/error *MS*) of 122.66 illustrates that not all of the observed differences can be attributed to chance (random sampling error). Examination of the components of the model reveals the source of the differences. The very small *F* for Loc (Loc *MS*/Well(Loc) *MS*) of 0.12 indicates that up-gradient wells do not differ from down-gradient wells. The *F* for Well(Loc) (Well(Loc) *MS*/Error *MS*) of 318.59 cannot be meaningly

interpreted because the Error *MS* used as the denominator of this test represents only laboratory error. The same problem occurs in the interpretation of the *F*-test for Time. However, those problems are of no consequence because the purpose of the monitoring program is to determine whether the down-gradient wells show any evidence of contamination. The *F*-test for Loc fulfills that function.

The value of *R*-square, 0.875, which shows that the model used was suitable for the data, was calculated by Model *SS*/Total *SS* and is thus the proportion of the total variation accounted for by the model. Of the remaining 12.5% of the variation about one percent is attributable to laboratory error.

Table 3 is an example of an analysis of four quarters of data from a waste disposal site with one up-gradient and six down-gradient wells. Groundwater from each well had four replicate measures in each of the four quarters for a total of 112 observations of pH. The analysis follows the same plan as the example above.

Locations (up-gradient vs. down-gradient) clearly do not differ. The large *R*-square shows that the model is excellent for these data. It may be instructive to examine Table 4 which presents the means of Time and Loc.

TABLE 3

Analysis of covariance for the variable pH

Source	*DF*	*SS*	*MS*	*F*	PROB
Model	9	34.77	3.86	116.9	<0.0001
Error	102	3.37	0.03		
Total	111	38.14	*R*-square = 0.91		
Time	3	21.79	7.26	219.8	<0.0001
Loc	1	1.42	1.42	0.6	0.4689
Well(Loc)	5	11.56	2.31	69.9	<0.0001

TABLE 4

Means for the analysis of covariance shown in Table 3

Time	Mean
1	5.16
2	4.08
3	4.45
4	4.08
Loc	**Mean**
Up	4.49
Down	4.17

These means show how large the temporal variation is when compared to the variation between locations.

Appendix 1 contains the SAS program statements used to perform the analysis shown in Table 1. Readers wishing a formal treatment of the analysis of covariance may consult Ott (Ref. [2], Chap. 17).

The analysis of covariance, although far superior to the *t*-test for analysis of groundwater monitoring data, must be used with a certain degree of caution. It shares with the *t*-test the requirement that several assumptions be met. The necessity for random sampling has already been discussed. Like other parametric statistical tests, both the analysis of covariance and the *t*-test require that errors (residuals) be normally distributed and that the population variances be homogeneous. Some data sets may meet neither of these assumptions. Another problem arises when the data contain many "less than" values because such censored data sets will violate the assumptions common to most parametric statistical procedures. A possible solution involves the use of specially designed non-parametric statistical tests. At present there are no widely available non-parametric procedures that can be substituted for the analysis of covariance in the analysis of groundwater data. Development and distribution of such tests should have a high priority.

It is doubtful whether the statistical procedures currently specified could, or should, be used in enforcement of environmental protection standards. The most recent guidelines for groundwater monitoring do nothing to improve the unsatisfactory situation with regard to statistical analysis of monitoring data [3]. Appendix B of that document specifies the use of the Fisher–Behrens *t*-test in a way that incorporates nearly every error discussed above. The present author believes that *t*-tests used as the basis of legal actions cannot withstand attacks on their scientific validity. The resulting vacuum in enforcement must be promptly filled. Legislation or regulations should specify statistical performance standards rather than particular statistical procedures. It would then be possible to optimize statistical procedures by assessing individual site characteristics while at the same time insuring that industry and the public recieve a satisfactory degree of protection.

References

1 SAS Institute Inc., SAS User's Guide, SAS Institute, Inc. Cary, NC 27511, U.S.A., 1982.

2 L. Ott, An Introduction to Statistical Methods and Data Analysis, second edn., Duxbury Press, Boston, MA, 1984.

3 United States Environmental Protection Agency, Groundwater Technical Enforcement Guidance Document (draft), Office of Solid Waste and Emergency Response, March 21, 1985.

Appendix 1

The data in Table 1 were entered as shown below:

```
DUG1 1 sc 1580 1580 1560 1560
DUG1 2 sc 1780 1760 1790 1800
.    . . .    .    .    .
.    . . .    .    .    .
```

The SAS programming statements were as follows:

```
Data One;                                    (Creates a data set called "one".)
Infile Sher;        (Reads a file called "Sher" which looks like the data above.)
Input Depth $ 1 Loc $ 2 Well-Num 4           (Reads the data from "Sher" into
@ 1 Well $ Time Chem $ X1—X4;                "one" according to the specification in
                                                          the input statement.)
Drop X1—X4;                          (X1—X4 will not be retained in the data set.)
Con = X1; Output;                                (These statements create separate
Con = X2; Output;                                observations for each of the four
Con = X3; Output;                                                replicate values.)
Con = X4; Output;
Cards;                    (Signals the end of the input portion of the program.)
Proc Sort; By Chem;                (Sorts the data by parameter, pH, TOC, etc.)
Proc GLM; By Chem; Classes Well Time Loc;                    (These three statements
Model Con = Time Loc Well(Loc);                  cause the analysis of covariance to
Test H = Loc E = Well(Loc)/Htype = 1 Etype = 1;                  be conducted.)
Title Analysis of Covariance of Groundwater Data;                (Self-explanatory.)
```