

DEVELOPMENT OF A STATISTICALLY VALID PROGRAM FOR MONITORING PESTICIDES IN GROUNDWATER IN THE STATE OF INDIANA

by Greg A. Olyphant¹ and Denver Harper²

¹ Dept. of Geological Sciences, Indiana University, Bloomington, IN, 812-855-5154

² Indiana Geological Survey, 611 N. Walnut Grove, Bloomington, IN, 812-855-1369

GOALS AND APPROACH OF THE MONITORING NETWORK

The purpose of the proposed project is to develop a statistically valid basis for extrapolating pesticide data from wells in the Indiana Baseline Monitoring Program (IBMP) to aquifers throughout the state. The relevant chemical analytes have been determined by the State Chemist's Office, and a formal working group is developing standards for the inclusion of existing wells in the IBMP, as well as specifications for installing new wells.

The working group that is charged with locating the monitoring sites has concluded that approximately $1,600 \pm 400$ water samples could be analyzed each year. Thus, because quarterly sampling of each site is necessary in order to account for seasonal variability, the monitoring network will consist of a total of approximately 400 wells.

Hydrogeologic conditions within Indiana are extremely varied. The working group has concluded that random sampling should be conducted within the aquifers of eight principal categories of hydrogeologic settings:

- (1) Unconsolidated sediment underlying till plains
- (2) Moraine environments
- (3) Outwash fans
- (4) Outwash plains and valleys
- (5) Bedrock containing a thin cover of glacial drift
- (6) Tunnel valleys immediately beyond moraine-belts
- (7) Karst areas
- (8) Pennsylvanian bedrock of the driftless area.

The aquifers within each hydrogeologic setting are considered to be separate entities with hydrologic and chemical attributes that cannot be extrapolated beyond their physical boundaries. Some aquifers are comprised, however, of significant subentities that should be treated separately. For example, aquifers in outwash fans and morainal areas are extremely heterogeneous, so that monitoring wells within each significant subentity should be treated as independent samples representing different statistical populations.

Within each hydrogeologic setting, all of the monitoring wells will be in row-crop areas, so that varying land use does not strongly influence the data.

The statistical analysis of data will be targeted at detecting trends in the occurrence and quantity of pesticides in groundwater. In particular, the following questions will be addressed:

- (1) In any given year, which categories and (or) subcategories of aquifers have occurrences and concentrations of pesticides that are above a specified level of tolerance? The level of tolerance may be specified by the State Chemist or any other individual or

working group authorized to set a limit on occurrences or concentrations of pesticides in groundwater.

(2) Has there been a statistically significant (95 percent confidence level) increase or decrease in the occurrence and (or) concentration of pesticides in any of the categories or subcategories of aquifers? If so are the differences associated with a particular season?

STATISTICAL METHODS

The primary statistical procedure to be employed in the modelling program is Analysis of Variance (ANOVA), and its variant the t-test. This method of data analysis has been studied and refined by theoretical and applied statisticians for decades and has been shown to be robust in the face of problems associated with water quality variables (e.g., non-normality of population distribution and inequality of sample variances; Montgomery and Loftis, 1987). Any parametric statistical procedure will be highly sensitive to spatial and temporal autocorrelation, however. Random sampling of monitoring locations can prevent problems associated with spatial autocorrelation, but complications due to seasonality (a temporal factor) are more difficult to deal with.

In ANOVA the data collected from an individual sampling point has the following basic structure:

$$Y_{ij} = \mu_i + e_{ij} \quad (1)$$

where, in the context of the present study, Y_{ij} represents a measure of pesticide occurrence or concentration in the j th monitoring well of the i th population (monitoring subunit), μ_i is the mean of the i th population and e_{ij} is the deviation of Y_{ij} from its population mean (assumed to be a random variable having a mean of zero and a variance of s^2). An estimate of the population mean is given by:

$$Y_i = \sum Y_j / n_i \quad (2)$$

where Y_i is an estimate of μ_i , and an estimate of the population variance is given by:

$$s^2_i = \sum e_j^2 / (n_i - 1) \quad (3)$$

where s^2_i is an estimate of s^2_i . These calculations are made for each of the i populations contained in the monitoring network.

To test whether a particular monitoring subunit has an occurrence or concentration of pesticides above some specified level of tolerance, the following t-ratio will be computed:

$$t = (Y_i - mtol) / SEY \quad (4)$$

where $mtol$ is a specified average occurrence or concentration that is considered to be a maximum tolerable amount by the State Chemist (or other authority), and SEY is an estimate of the standard deviation of the sampling distribution of the means (a function of the sample size n_i and variance s^2_i). The computed t-ratios are compared to ordinates of Student's t-distribution with $n_i - 1$ degrees of freedom and a specified confidence level.

To compare the occurrences and or concentrations of pesticides in two monitoring subunits during the same time period, the t-ratio takes the following form:

$$t = (Y_2 - Y_1)/SEY_p \quad (5)$$

where Y_1 and Y_2 are means of two monitoring-subarea samples, and SEY_p is a pooled estimate of the standard deviation of the sampling distribution of the means (a function of the variances and sample sizes of the monitoring subunits being compared). The computed t-values can be compared to ordinates of Student's t-distribution with $n_1 + n_2 - 2$ degrees of freedom and a specified confidence level.

When comparisons are made over time (i.e. to determine any temporal trends) seasonality may influence the results. Ideally, the comparisons will be made using data collected during the same seasons, in which case the same procedure described above would be appropriate for determining if a temporal change in the pesticide occurrence or concentration in groundwater has occurred. If, on the other hand, the samples to be compared were not collected during the same seasons then the model changes from (1) to:

$$Y_{ijk} = \mu_i + \alpha_k + \epsilon_{ijk} \quad (6)$$

where Y_{ijk} is a measure of pesticide occurrence or concentration in the j th well of the i th monitoring unit during the k th season, μ_i is the mean of the i th monitoring unit, α_k is an unknown parameter representing the effect of seasonality on the sample mean, and ϵ_{ijk} is a random variable with a mean of zero and a variance of s^2 . In this situation an estimate of α_k will need to be made (this will involve an analysis of data collected over a sufficiently long period of time; e.g. 2 years) and the appropriate tests would involve calculation of F-ratios and comparison with the ordinates of Snedecor's F- distribution (the details are not presented here but are available in Dixon and Massey, 1957).

SAMPLE SIZES

Appropriate sample sizes are a critical factor in the monitoring program, because too small a sample will leave questions regarding statistical validity of inferences while too large a sample will either limit the number of subunits that can be monitored or cause the cost (both money and man hours) to be prohibitive. If the distribution of means is assumed to be normal (a safe assumption to make) then the appropriate sample size from which to compute the mean of each subarea is (Cochran, 1963):

$$n_i = (s_i t_m / d)^2 \quad (7)$$

where s_i is an estimate of the standard deviation (square-root of population variance), d is a specified margin of error, and t_m is the ordinate of Student's t-distribution for a specified confidence interval and m degrees of freedom. In a practical application, s_i and m are determined from an initial sample of about 15 randomly selected wells. Thus, the additional number of wells to be chosen in the particular monitoring subunit would be $n_i - m + 1$.

PROJECT SCHEDULE

The project will take place in four primary phases. The execution of each phase will depend on the progress of other subgroups that are operating separately from ours, but we expect that two years will be sufficient to deliver a final product.

Phase 1 The exact boundaries of each monitoring subunit must be specified before any analyses can be performed. Much progress has been made concerning the identification of the units (by the monitoring network working group) but geographical boundaries have not been determined. Once the boundaries have been agreed upon, they will be digitized as polygons in a GIS coverage.

Phase 2 Random samples of 10-to-15 monitoring wells will be selected from each monitoring subunit. Information regarding each well that will be entered into the GIS database will include: (1) well identification number(s), (2) location (Universal Transverse Mercator (UTM) coordinates, digitized from a 1:24,000-scale map), and (3) a subunit identification code. Existing wells that meet the criteria of the monitoring-well working group can be used only if they do not exhibit strong spatial clustering (which may result in biased estimates of statistical parameter values). Clustering of wells will be evaluated using Nearest Neighbor Analysis (Theakstone and Harrison, 1970). If a spatially random sample cannot be achieved using existing wells, then new wells will need to be installed at the closest accessible points to randomly selected UTM coordinates.

Phase 3 From the initial samples, the exact number of sample points in each monitoring subunit will be determined using the method described above, and any additional well locations will be determined by random sampling. The need for new sampling subunits will be addressed during this phase, as the preliminary data will provide a basis for identifying significant subpopulations within the data. It is also possible that one or more of the a priori groupings can be eliminated on the basis of the preliminary statistical analysis.

Phase 4 Software will be written (in FORTRAN) to analyze the data using the methods briefly described above. An interface will be developed between the statistical software and a database management system. The operating system will be UNIX, the GIS will be ARC/INFO, and the database will be INFO. If desired, a set of ASCII files will be provided to be loaded into other databases, such as Oracle.

DELIVERABLES

At the end of the project, we will deliver the following:

(1) A report on the development of the statistically valid program for monitoring pesticides in Indiana groundwater. This will be published by the Indiana Geological Survey and will include a spatial and frequency-density analysis of monitoring-well locations and a summary of the preliminary data that were used to develop the permanent monitoring program.

(2) A set of GIS procedures that interface with a statistical analysis package, so that the user can view maps showing the status of pesticide contamination in monitoring subunits (i.e., above or below specified tolerances) at any given time, and maps showing changes of status through time, as well as graphs showing time trends of average pesticide concentrations in the monitoring subunits. Documentation will be provided in the form of a manual.

REFERENCES

Cochran, W.G., 1963: Sampling Techniques. Wiley and Sons, New York.

Dixon, W.J., and F.J. Massey, 1957: Introduction to Statistical Analysis. McGraw-Hill, New York.

Krumbein, W.C., and F.A. Graybill, 1965: An Introduction to Statistical Models in Geology. McGraw-Hill, New York.

Montgomery, R.H., and J.C. Loftis, 1987: The applicability of the t-test for detecting trends in water quality variables. Water Resources Bulletin, 23(4): 653-666.

Theakstone, W.H., and C. Harrison, 1970: The Analysis of Geographical Data. Heinemann Ltd., London.