CONF-880520--
(DE88013180)

QSAR 88

Proceedings of the
THIRD INTERNATIONAL WORKSHOP ON
QUANTITATIVE STRUCTURE-ACTIVITY RELATIONSHIPS
IN ENVIRONMENTAL TOXICOLOGY

May 22-26, 1988

Knoxville, Tennessee

Edited by: James E. Turner
M. Wendy Williams
T. Wayne Schultz
Norma J. Kwaak

SIMPLIFYING COMPLEX QSAR'S IN TOXICITY STUDIES WITH MULTIVARIATE STATISTICS

Gerald J. Niemi * and James M. McKim

Environmental Research Laboratory, Duluth
U.S. Environmental Protection Agency
6201 Congdon Boulevard
Duluth, MN 55804 USA

ABSTRACT

During the past several decades many quantitative structure-activity relationships (QSAR's) have been derived from relatively small data sets of chemicals in a homologous series and selected empirical observations. An alternative approach is to analyze large data sets consisting of heterogeneous groups of chemicals and to explore QSAR's among these chemicals for generalized patterns of chemical behavior. Exploratory analyses using multivariate statistical procedures in an iterative fashion have traditionally been a neglected tool in the effort to find relationships that can lead to testable hypotheses. Hence, statistical analysis does not need to be a device only to test a hypothesis. Moreover, multivariate statistical analyses (e.g., principal components analysis (PCA) and factor analyses) can simplify the complex relationships among variables. One of the major reasons for not considering multivariate statistical routines for "simplifying" complex relationships is a lack of understanding and routine use of these techniques by practicing QSAR scientists. The use of exploratory multivariate statistical techniques for simplifying complex QSAR problems is demonstrated through the use of research data on biodegradation and mode of toxic action. In these examples, a large number of explanatory variables were examined to explore which variables might best explain whether a chemical biodegrades or whether a toxic response by an organism can be used to identify a mode of toxic action. In both cases, the procedures reduced the number of potential explanatory variables and generated hypotheses about biodegradation and mode of toxic action for future research without explicitly testing an existing hypothesis.

* Present address: Center for Water and the Environment, Natural Resources Research Institute, University of Minnesota, Duluth, MN 55811 USA.

INTRODUCTION

The vast majority of QSAR's developed over the past several decades were largely derived from relatively small data sets of homologous series of chemicals. Furthermore, the "structural" variables used to make predictions about the "activity" variables in a "quantitative relationship" were primarily based on "secondary structural" variables such as the n-octanol water partition coefficient (log P). There is nothing inherently wrong with the development of these relationships except that certain limitations exist in their application. These include:

1) secondary structural variables such as log P (independent variables) are measured or calculated with error and hence these errors are propagated into predictions of the activity variables (dependent variables);

2) secondary structural variables are often impossible to calculate for some compounds, which limits their application; and

3) precise definitions on what constitutes a "homologous" series are often vague and, hence, the boundaries of a specific QSAR is also vague.

There clearly is no simple solution to these problems. However, the first two limitations can be overcome by considering primary structure activity variables or variables calculated directly from the structure of the chemical (e.g., chemical fragments or molecular connectivity indices). The third limitation can be overcome by building QSAR's in a more global context in which no subjective boundaries are placed on the realm of chemicals to which the QSAR will apply. Here our objectives are the following: 1) explore reasons why a more global perspective has not been pursued, and 2) present two examples of how a complex QSAR problem can be simplified through the use of multivariate statistics.

Limitations to a global perspective

The sheer magnitude of a QSAR problem from a global perspective is intimidating because of the large number of structures (e.g., hundreds of thousands) that can be considered, plus the number of potential structural variables that theoretically can be calculated for a structure. The magnitude of the problem has limited conceptual approaches and has immediately forced scientists into limiting the problem, usually by dealing with discrete homogeneous series of chemicals.

When one is working from an industrial perspective in attempts to design a new chemical or drug, this more focused approach might be feasible because a more limited number of solutions are possible given the availability of a lead structure. However, from a regulatory perspective it is not feasible because an initial evaluation must focus on an objective placement of the chemical into the proper group of chemicals and QSAR model, from which predictions can be made. Hence, the perspective in which scientists using QSAR techniques must examine a problem will partly determine the approach to appropriately bound the chemical within the global universe of potential chemical structures.

Two additional reasons for an inhibited global perspective in understanding QSAR are the lack of training of scientists in statistical analysis, especially multivariate statistics, and the relatively rapid and expanding development of computer capabilities. Regarding the former, scientists using QSAR techniques are generally either chemists, biologists, or biochemists. Most of these scientists have formal training in mathematics including calculus because most undergraduate and graduate curricula require some mathematical training. Some have formal training in one or two elementary statistics courses in which at most two variables (e.g., correlation or regression) are considered in statistical tests. Few have any training in multivariate statistical techniques, training which is necessary to consider a multivariate, global perspective to QSAR.

In regard to the latter limitation, phenomenal progress has been made during the past 30 years in the development, design, and use of computer hardware and software. Yet, despite this progress, the capacity and cost to use many computers and the availability of software capable of handling hundreds of variables for thousands of chemical cases is still limited. For example, we can calculate literally hundreds of potential primary structure-activity variables based on various mathematical routines (Basak et al. 1987), yet one of the most commonly available statistical packages, the Statistical Package for the Social Sciences (SPSS, Nie et al. 1975) is limited to analysis of < 100 independent variables (Niemi et al. 1985). Therefore, we have powerful computer capabilities today, but they may not yet be as powerful as we desire nor do enough scientists have the proper technical training to fully utilize statistical routines or existing computers. Progress in developing QSAR, especially global-multivariate relationships, will be inhibited until a larger critical mass of QSAR scientists are educated, hardware and software capabilities of computers are improved, and costs to obtain and use these computer capabilities are reasonable.

QSAR in biodegradation research

Development of QSAR models to predict whether chemicals microbially degrade in aquatic environments or to determine the rate at which a chemical degrades can be difficult because of the many interacting factors that contribute to biodegradability (e.g., see Alexander 1981). Niemi et al. (1987) attempted an objective multivariate statistical approach to this problem by using a data base of 287 compounds with available 5-day BOD values (BOD_5). BOD_5 was used as an approximate measure of the inherent ability of a chemical to microbially degrade in a modern sewage treatment facility. For each of these compounds 54 molecular connectivity indices and five physicochemical parameters including log P (Leo and Weininger 1984) were calculated and used as potential explanatory variables for assessing whether the BOD_5 value was relatively high or low (e.g., biodegradable or persistent respectively).

The first manipulation of these data was to separate the compounds into biodegradable and persistent groups based on a natural division in the BOD_5 values. Discriminant function analysis (DFA) with the molecular connectivity indices and the five physicochemical factors were used as explanatory variables in an attempt to separate these two groups. In general, DFA is a multivariate statistical technique that identifies whether differences in a set of explanatory variables exist between two or more groups. Although two previous papers reported some success with this technique (Geating 1981, Enslein et al. 1984), only 50 % of the 287 compounds could be correctly discriminated in this exercise.

From the perspective of chemical structure, it is likely that there are many different factors that contribute to the persistence or degradability of a chemical and, hence, the chemicals need to be assessed in smaller groups. Because there was no a priori rationale for defining these groups, an objective multivariate technique, K-means clustering (Dixon 1981), was used. Prior to the use of clustering, a principal components analysis (PCA) was calculated on 45 of the molecular connectivity indices. PCA is a technique used to reduce the number of variables to be considered in a problem and here it was used to reduce the molecular connectivity indices to less than 10 variables that still explained > 90 % of the variation in the original variables. PCA was a necessary step here because the K-means clustering software of the Biomedical Computer Program (BMDP, Dixon 1981) and the PDP-11/70 computer system used at the time was limited to a maximum of nine variables for eight clusters that could be defined for 287 cases.

Two additional problems were encountered in this analysis. First, there was no a priori rationale for defining how many clusters should be identified to improve the predictions. Secondly, compounds that were outliers in the principal components space were often identified as single compound clusters. To avoid the latter problem, any compounds that were > 2 standard deviations from the mean for any of the principal components used were identified as belonging to an "outer" space and analyzed separately from those compounds within 2 s.d.'s for all principal components. The former problem was solved by iterating the number of clusters to be formed over a range of clusters and identifying the number of clusters that produced the best discrimination of biodegradable from persistent chemicals. Hence, the statistical process consisted of the following:

(1) PCA of 45 molecular connectivity indices that described the structure of the compounds,
(2) identification of an "outer" and "inner" space, and
(3) iterative clustering of the outer and inner space followed by DFA of biodegradable and persistent groups within each iterative cluster.

The results of this iterative analysis process improved the correct prediction of biodegradability to an overall 88 % (85 % for biodegradable compounds and 94 % for persistent compounds). To identify the types of structural features associated with biodegradability or persistence, the discrimination within each of the clusters was examined and summarized into a set of heuristic rules. When possible, each of the heuristic rules was related with previous knowledge published on structural relationships associated with biodegradability. After some obvious misclassifications based on DFA were translated into the set of heuristic rules, the set of heuristic rules correctly classified 93 % of compounds into the appropriate biodegradability group (91 % for degradable chemicals and 96 % for persistent chemicals).

In summary, the iterative multivariate statistical procedures described above allowed for an eventual simplification of structural features associated with the complex process of biodegradability of chemicals into a set of heuristic rules. These rules can be viewed as tentative hypotheses to be tested in future experimentation and modified as a result of those subsequent experiments. Admittedly, the statistical procedures are complex, especially to those unfamiliar with these techniques, but the eventual results led to a simplification in understanding potential structural features associated with biodegradability.

QSAR in mode of toxic action research

Over the past five years scientists at U.S. EPA's Environmental Research Laboratory in Duluth have studied eight xenobiotic chemicals from the perspective of four different biological disciplines; two chemicals for each of four different known modes of toxic action. The major objective of this research was to identify effective, but cost-efficient sets of toxic responses in fish that would correctly identify specific modes of action. These response sets were termed fish acute toxicity syndromes or FATS (McKim et al. 1987a). The basic premise of this research was based on the idea that if an appropriate FATS could be identified for a chemical, then a reasonable prediction of mode of action could be made for that chemical. A QSAR equation could then be used for that mode of action and subsequently a prediction about its toxicity (McKim et al. 1987a).

The four biological disciplines and number of parameters included in the analysis were the following:

(1) 17 physiological variables measured on four individual rainbow trout (Salmo gairdneri) exposed to each of the test chemicals (primarily respiratory-cardiovascular) (McKim et al. 1987b, c);

(2) 14 behavioral variables measured on fathead minnows (Pimephales promelas) exposed to each of the test chemicals in standard 96-h LC_{50} assays (Drummond et al. 1986);

(3) 25 hematological variables measured on individual trout exposed to each of the test chemicals (Snarski and Stokes, pers. comm.); and

(4) 14 biochemistry variables measured on individual trout exposed to each of the test chemicals (Christensen, pers. comm.).

These data represent a substantial multivariate problem and one in which substantial violations of statistical assumptions are possible as well as a situation in which spurious results are expected. For example, the common denominator that links the observations for each variable for each discipline are the eight chemicals. Thus, the multivariate situation is that there are 70 variables for eight cases; a reversal from the ideal situation in which one would like 70 cases for each of eight variables. However, from a biological perspective, it is seldom that information of this detail is available across disciplines and we argue that despite the statistical problems these data are worthy of exploratory analysis. These data are especially worthy from the perspective of using multivariate statistical analysis to simplify future analyses of FATS predictions.

The initial major question of interest here is what variables can best discriminate between the four FATS groups (each reflective of a mode of toxic action) which are represented by the eight chemicals. The four modes of toxic action and the associated chemicals studied were:

(1) nonpolar narcosis (tricaine methanesulfonate and 1-octanol);
(2) acetylcholinesterase inhibitors (malathion and carbaryl);
(3) uncouplers of oxidative phosphorylation (pentachlorophenol and 2,4-dinitrophenol); and
(4) mucous membrane irritants (acrolein and benzaldehyde).

Selection of the most useful variables for discriminating between FATS was based on the following steps: (1) identification of those variables that provided the best discrimination (lowest alpha values) of all four FATS, (2) identification of those variables that best discriminated between two FATS groups, and (3) elimination of one variable from a pair of variables that were highly correlated within a biological discipline (here defined as $\underline{r} >$ 0.85). In steps one and two above, univariate F values and associated alpha values from an analysis of variance were used to determine the best discriminating variables. In step three Kendall's rank correlation was used because of the relatively low sample size. The final step in the analysis was to use some of the variables identified in the first three steps above in a DFA to identify the smallest number of variables that could discriminate the four FATS.

A total of 23 variables including six physiological, five behavioral, five hematological, and seven biochemical variables were highly significant ($\underline{p} < 0.001$) discriminators of the four FATS groups. In addition to these 23 variables, six physiological, three behavioral, three hematological, and one biochemical variable were significant ($\underline{p} < 0.01$) discriminators of the four FATS groups (Table 1). Therefore, a total of 36 of the 70 variables (51 %) considered here were significant discriminators of the four groups.

Table 1. Summary of three steps in reducing the number of potential explanatory variables in discriminating four FATS among four different biological disciplines (see text for details in reduction process).

Discipline	Original variables	Step-1	Step 2	Step 3	Total
Physiological	17	12	3	- 1	14
Behavioral	14	8	4	- 1	11
Hematological	25	8	3	0	11
Biochemical	14	8	1	- 1	8
Total	70	36	11	- 3	44

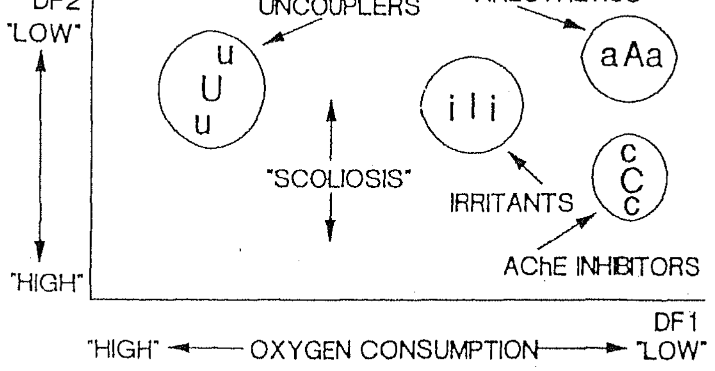
In considering all six pairwise combinations of the four FATS groups, three additional physiological, four behavioral, three hematological variables, and one biochemical variable were significant ($\underline{p} < 0.01$) discriminators of at least two FATS groups (Table 1). Hence, a cumulative total of 15 of 17 physiological variables (88 %), 12 of 14 behavioral variables (86 %), 11 of 25 hematological variables (44 %), and 9 of 14 biochemical variables (64 %) or 47 of 70 potential explanatory variables (67 %) of at least 2 FATS groups were significant at $\underline{p} < 0.01$.

Pairwise correlations between variables that were good discriminators within a discipline (Steps 1 and 2) showed that only six pairwise variables (Table 1) had correlation values greater than $\underline{r} > 0.85$ ($\underline{r}^2 > 0.72$). Therefore, only three variables could be eliminated at Step 3 and 26 of the original 70 variables (37 %) could be eliminated using this reduction process (70-44=26).

The final step in the analysis was to conduct a stepwise DFA to identify the best variables that could discriminate the eight chemicals among the four FATS groups. In this process, instead of including all 44 good discriminating variables of the four FATS groups, we selected the two best discriminating variables from each of the four biological disciplines. This process could still produce spurious relationships in the results because the number of variables is equal to the number of cases. However, this analysis is a better alternative as compared with including all 44 good discriminating variables and it is only being calculated to explore the best combination of variables among disciplines for potentially discriminating all four FATS groups.

The first variable selected was oxygen consumption, a physiological variable, (McKim et al. 1987b), which discriminated the narcosis and uncoupler FATS groups from the inhibitor and irritant FATS groups. After this step, six of the eight chemicals were correctly classified. The second variable selected was a behavioral variable, scoliosis (a morphological abnormality, Drummond et al. 1986), which correctly discriminated the remaining two chemicals and all four FATS.

Figure 1. Plot of first two discriminant functions from a DFA of eight chemicals in which the response of fish (as measured by two variables (oxygen consumption and scoliosis) were the best discriminators and correctly classified each chemical into one of four FATS groups.



In summary, we established a criteria for potentially reducing the number of variables to be considered for correctly classifying chemicals into a respective FATS group based on biological responses of fish exposed to those chemicals. By selecting the best discriminating variables and variables that were highly intercorrelated, 26 of 70 potential variables (37 %) were eliminated. The lack in our ability to reduce the dimensionality further is partly due to the good selection of discriminating variables by the scientists involved among the respective disciplines and partly due to our lack of knowledge regarding FATS. For example, one would not want to eliminate a variable that might prove to be a good discriminator of a FATS not yet tested with the model. In contrast, two variables were able to correctly discriminate eight chemicals into four FATS groups. This likely indicates the problem of discriminating FATS can be accomplished with a relatively small set of response variables and that the response of fish to chemical intoxication is manifested by a number of variables; each of which is measurable at a variety of levels (physiologically, behaviorally, hematologically, and biochemically). Discovery of the best combination of variables to use for screening a large number of chemicals will best be accomplished by an examination of the cost-effectiveness and the precision and accuracy of measuring the respective variables.

Acknowledgments

We thank Mr. Robert Drummond, Ms. Virginia Snarski, Ms. Nan Stokes, and Mr. Glenn Christensen for access to their data on the biological responses of fish to the eight chemicals considered here. We also thank Drs. Steven Bradbury, Subhash Basak, and Gilman Veith for their comments on this manuscript. This paper has not been peer-reviewed by the U.S. Environmental Protection Agency and therefore the views expressed herein do not necessarily reflect the views of the Agency. This research was partly supported by Cooperative Agreement No. CR811981 to Dr. Ronald Regal of the University of Minnesota.

Literature Cited

Alexander, M. 1981. Biodegradation of chemicals of environmental concern. Science 211: 132-139.

Basak, S.C., V.R. Magnuson, G.J. Niemi, R.R. Regal. and G.D. Veith. 1987. Topological indices: their nature, mutual relatedness, and applications. Pages 300-305 in X.J.R. Abulah, G. Leitmann, C.D. Mote, and E. Y. Rodin, eds. Proceedings, Fifth International Conference on Mathematical Modelling, Berkeley, CA. Pergamon Press, New York, NY, USA.

Dixon, W.J. Ed. 1981. BMDP Statistical Software, 1981. University of California Press, Berkeley, CA, USA.

Drummond, R.A., C.L. Russom, D.L. Geiger, and D.L. DeFoe. 1986. Behavioral and morphological changes in fathead minnows, Pimephales promelas, as diagnostic endpoints for screening chemicals according to mode of action. Pages 415-435 in Aquatic Toxicology. Ninth Aquatic Toxicology Symposium, American Society for Testing and Materials, Philadelphia, PA, USA.

Enslein, K., M.E. Tomb, and T.R. Lander. 1984. Structure-activity models of biological oxygen demand. Pages 89-109 in K.L.E. Kaiser, ed., QSAR in Environmental Toxicology. D. Reidel, New York, NY, USA.

Geating, J. 1981. Literature study of the biodegradability of chemicals in water. Vols. 1 and 2. EPA/600/2-81-175/176. U.S. Environmental Protection Agency, Office of Research and Development, Cincinnati, OH.

Leo, A. and D. Weininger. 1984. CLOGP version 3.2 user reference manual. Medicinal Chemistry Project, Pomona College, Claremont, CA, USA.

McKim, J.M., S.P. Bradbury, and G. J. Niemi. 1987a. Fish acute toxicity syndromes and their use in the QSAR approach to hazard assessment. Environmental Health Perspectives 71: 171-186.

McKim, J.M., P.K. Schmieder, R.W. Carlson, E.P. Hunt, and G.J. Niemi. 1987b. Use of respiratory-cardiovascular responses of rainbow trout (Salmo gairdneri) in identifying acute toxicity syndromes in fish: part 1. pentachlorophenol, 2,4-dinitrophenol, tricaine methanesulfonate, and 1-octanol. Environmental Toxicology and Chemistry 6: 295-312.

McKim, J.M., P.K. Schmieder, G.J. Niemi, R.W. Carlson, and T.R. Henry. 1987c. Use of respiratory-cardiovascular responses of rainbow trout (Salmo gairdneri) in identifying acute toxicity syndromes in fish: part 2. malathion, carbaryl, acrolein, and benzaldehyde. Environmental Toxicology and Chemistry 6: 313-328.

Nie, N.H., C.H. Hull, J.G. Jenkins, K. Steinbrenner, and D.H. Bent. 1975. SPSS, statistical package for the social sciences. McGraw-Hill Book Company, New York, NY, USA.

Niemi, G.J., R.R. Regal, and G.D. Veith. 1985. Applications of molecular connectivity indexes and multivariate analysis in environmental chemistry. Pages 148-159 in J.J. Breen and P.E. Robinson, eds., Environmental applications of chemometrics. ACS symposium series No. 292. American Chemical Society, Washington D.C., USA.

Niemi, G.J., G.D. Veith, R.R. Regal, and D.D. Vaishnav. 1987. Structural features associated with degradable and persistent chemicals. Environmental Toxicology and Chemistry 6: 515-527.

Veith, G.D., D.J. Call, and L.T. Brooke. Structure-toxicity relationships for the fathead minnow: narcotic industrial chemicals. Canadian Journal of Fisheries and Aquatic Sciences 40: 743-748.