

# Pesticide Persistence in the Environment - Collected Data and Structure-Based Analysis

Sokratis ALIKHANIDI and Yoshimasa TAKAHASHI\*

Department of Knowledge-based Information Engineering, Toyohashi University of Technology,  
Tempaku-cho, Toyohashi 441-8580, Japan

\**e-mail*: taka@mis.tutkie.tut.ac.jp

(Received: October 23, 2003; Accepted for publication: January 23, 2004; Published on Web: March 22, 2004)

A data set of 420 pesticide persistences in the environment was collected as field half-life (HL) using several on-line databases. Due to the fuzziness of observed values, the compounds were grouped into three major categories: class 1 when a pesticide has  $HL \leq 30$  days; class 2, if  $30 < HL \leq 100$  days; and class 3, if  $HL > 100$  days. The Quantitative Structure-Biodegradation Relationship (QSBR) analysis was worked out on the training set of 315 pesticides. Thirty one topological substructural descriptors were used and the decision tree approach was employed for the modeling. Estimation results were as follows: for the train set, 5 compounds of two-unity (class 1/class 3) misclassification, 38 compounds of unity (class 1/class 2 and class 2/class 3) misclassification, and 272 compounds (86.3%) were correctly classified; for the test set, there were 3, 20, and 82 compounds (78.1%) respectively. The computer expert system EKeeper was developed on the basis of the QSBR model.

**Keywords:** QSAR, QSBR, Data mining, Estimation of biodegradation, Expert system, Pesticide persistence

## 1 Introduction

A great number of new chemical substances are employed by different present-day industries, such as chemical, pharmaceutical, and agricultural. The disposability of a new compound is a key feature of a chemical to be applied anywhere, together with its target functionality. This fact especially applies to the pesticides, which are normally released onto the ground.

The experimental evaluation of chemical degradation in the environment is highly complicated due to multiple reasons. These include the variation of moisture, temperature, chemical and microbiological composition of soil, ability of a chemical to volatilize and photo-degrade. The involvement of some longer-term processes, primarily related to changes in the population of microorganism species in soil or natural waters, can not be excluded. The assessment test for disposability is expensive and usually time-consuming. In some cases, when a chemical is not mineralized or broken down to nontoxic products, there is a possibility of extensive spoilage creating health hazards to fauna and humans.

Thus, the theoretical estimation of disposability is very valuable. Unfortunately, its modeling is extremely

complex, due to a quantity of reaction mechanisms for environmental degradation. On the other hand, the usage of the Quantitative Structure-Biodegradation Relationship (QSBR) [1, 2] techniques is a reasonable alternative way for developing of so-called "Expert Systems" for estimation of the environmental degradability. QSBR approaches are based on the postulation that the degradation ability of the compound can be described by some kind of modeling function of numerical descriptors representing the molecular structure. These numerical descriptors may be of many types: calculable physicochemical properties (octanol-water partition coefficient, surface area, refractivity, polarizability), various spatial or topological molecular indices, charge-distribution-related parameters, quantum chemical and molecular field parameters, and occurrences of certain structural features. Multiple QSBR models have been suggested for the estimation of degradability for some homogeneous series of molecules belonging to specific chemical classes [3–9]; most of them operate with short and congeneric training sets up to 40 compounds, and may have only a limited application. In a newer model [10], the size of a training set is not stated but the model has been validated by a homologous set of 177 mono benzene derivatives and 168

acyclic compounds only. Some models have been created with very short data sets of 18 compounds [11] and 36 compounds [12].

Models dealing with the larger sets of heterogeneous molecules are of more practical importance. Howard and coworkers developed linear and nonlinear models to estimate the probability of aerobic biodegradation of 264 compounds from the BIODEG database [13, 14]. To classify the compounds to be rapidly or slowly biodegradable, they used 35 fragmental descriptors achieving accuracies for the training set of 90.5 and 89.8% and for the test set (27 chemicals) of 81.5 and 88.8% for linear and nonlinear models, respectively. Klopman and coworkers created the program CASE [15] which automatically identifies and analyzes molecular fragments of a data set to create the discriminant function between rapid and slow biodegradability. They used 283 aliphatic and aromatic compounds from the BIODEG database and found 37 important substructures. For a test set of 27 compounds the correct prediction was 74%. Devillers and colleagues developed a new database of aerobic biodegradability of 184 chemicals with the help of 17 experts [16]. They associated the experts' 'hours', 'days', 'weeks', 'months', and 'longer' responses with 1 to 5 integers. As the next step, they used 66 autocorrelation indices encoding the hydrophobicity of molecules followed by the extraction of nine first principal components. Biodegradability was modeled by back-propagation neural network resulting in a squared correlation coefficient of 0.76 for the training set of 172 molecules and 0.49 for the test set of 12 molecules [17].

The newer studies deal with the more comprehensive MITI I data set for biodegradability of about 900 diverse compounds [18]. Loonen *et al.* used PLS analysis processing 127 predefined structural fragments and gained 85% and 83% of correct predictions for the total set and on average for four test sets, respectively [19]. On the other hand, Sabljic and coworkers applied the inductive machine learning method to the whole MITI I data set achieving 84% of correctly classified chemicals [20]. Those published results are in fact significant and may have a potential for practical use.

Our study was specifically directed at the pesticide area. In this work, (i) the collection of pesticide degradability database and (ii) the development of an estimating scheme on its basis were carried out.

## 2 Methodology and Experimental Part

### 2.1 Data set

As a first approximation, environmental disappear rate of a chemical is proportional to its concentration and the

first-order reaction may be assumed:

$$[A] = [A]_0 \cdot e^{(-kt)} \quad (1)$$

where  $[A]_0$  and  $[A]$  are initial and remained after time  $t$  concentrations of the chemical and  $k$  is a time constant. The convenient half-life period (HL) of a chemical is the time needed to decrease the concentration by factor 2. Then the equation (1) can be transformed to

$$\text{HL} = t = \ln 2/k. \quad (2)$$

Unfortunately, the environmental HL of a chemical compound is a highly fuzzy value, due to a number of reasons. Also, there may be different order degradation mechanisms, as well as some other peculiarities like the accumulation of hazardous but stable decay products.

Because of such uncertainty, the use of a discrete value may be a quite reasonable way to describe the degradation ability of a chemical compound. Several scales have been used, such as 2-level (*i.e.* a chemical degrades rapidly or not) [18], 3-level [21], and 4 (and 5)-level [16]. We consider that 3-level scale (*i.e.* a chemical degrades rapidly, moderately, or slowly) is a reasonable choice because the 2-level scale is too rough, whereas in scales with many levels it is often more difficult to find the correspondence for a chemical with an appropriate single level. 3-Level scale [21] categorizes the pesticides into the following classes:

$$\left\{ \begin{array}{l} \text{class 1, if HL} \leq 30 \text{ days,} \\ \text{class 2, if } 30 \text{ days} < \text{HL} \leq 100 \text{ days,} \\ \text{class 3, if HL} > 100 \text{ days.} \end{array} \right. \quad (3)$$

Compilation of the degradation rates of many compounds is a significant problem because test conditions ought to be uniform. Only the MITI (Japanese Ministry of International Trade and Industry) [18] assessment of 894 chemicals satisfies this normal requirement. However, it includes the 2-level valuation.

A massive compilation of degradation data for about 240 pesticides was made by the U.S. Environmental Protection Agency [22]. The data were received from many references, thus the test conditions varied appreciably. The project is continuous and the renewed database is presented on the Internet [23] (we have processed 334 pesticides).

Nevertheless, some problems related to few observations for certain chemicals and in general with the absence of comments for a given half-life data, often arose while using Refs. [22, 23]. A good solution for these problems involved using the rich Hazardous Substances Data Bank of the U.S. National Library of Medicine (HSDB) [24]. It offers the explanation of degradation details and frequently provides new additional references.

The Pesticide Management Education Program at Cornell University was another and often complementary database [25]. For several compounds with

unreliable or too scanty persistence reports, it offered data unlisted in Refs. [22–24]. Those chemicals are dienochlor, dodine (cyprex), mepiquat chloride, propamocarb, propoxur, pyriithiobac, quizalofop-ethyl, temephos, terbufos, triadimefon, triflumizole, and trimethacarb. For trichloroacetic acid (TCA) a review of its environmental behavior has been published recently [26].

Gathering all the available data from Refs. [22, 23], with the help of the other above-mentioned sources, the data set of the persistence of 315 pesticides in field conditions was collected and presented in the Table 1 [27]. It consists of organic and ‘organic-like’ (carbon disulfide, ammonium sulfamidate, chloropicrin) compounds, where HL data exists. Several pesticides include nontoxic aluminium, iron, and silicon. Compounds with toxic arsenic and tin were excluded as these elements are dangerous for the environment, and the usage of such pesticides is specially regulated by law.

On the other hand, it was recognized that HSDB still has many pesticides not included in our data set. To take them, we used the comprehensive Compendium of Pesticide Common Names [28] (1086 compounds, up to October 2002) to extract all the new CAS numbers, followed by the extraction of corresponding information on environmental degradability from HSDB. As the last step, we cleaned the data removing the compounds with insufficient degradability information, substances which include toxic elements (As, Sn, Hg), or too complex mixtures of compounds under a single trade name (camphechlor - mixture of at least 177 chlorinated camphenes). Often, we have found specific relevant facts in the Pesticide Management Education Program at Cornell University [25]. As a result, the new data set has 105 pesticides, and is presented at the Table 2.

The pesticides were associated with the persistence class according to the rules (3). The persistence data were critically evaluated employing additional rules: (i) for the wide data ranges, a greater half-life period was decided to be the cheaper (*i.e.* safer) mistake for the environment; (ii) for the ranges of HL, geometric mean was assumed (as for log-normally distributed data). For instance, the pesticide aldrin has the following HL values collected: 28, 43-63, 10, 183, 273-365, 21-584, and 20-100 days. Definitely, the class 1 is not the case (rule (i)), and the smallest values (10, 20, 21, and 28) were removed to retain the greater values only. For the rest (43, 63, 183, 273, 365, 30, 584, 30, and 100 (30 were used twice to replace 21 and 20)), its geometric mean (rule (ii)) is 111 days associating aldrin with the class 3. Such serious data inconsistency problems arose frequently during the preparation of the data set. Often compounds had either too wide range of HL values (as for aldrin), which led to the determination of a fuzzy persistence class, or only 1 or 2 observations, which resulted in an unreliable conclusion; such cases were marked in Table 1 by ‘☺’ and ‘↓’

respectively [27].

For each compound the structural information was prepared. The ChemIDPlus database [29] of the National Library of Medicine (350K structures) was used as a convenient source for acquiring molecular structures. For other compounds, molecular structures were constructed from the chemical names.

Preprocessing of structures before QSBR analysis included:

(i) Compounds with weak ionic bonds were processed as a protonated free acid and tertiary free amine (or quaternary cation, if the case), for organic acid salts and organic ammonium salts respectively.

(ii) The aromaticity of benzene-like and heterocyclic fragments was carefully checked by the Hueckel rule. In particular, ring compounds with the exocyclic carbonyl group or with atoms of nitrogen and oxygen in the same ring were considered as non-aromatic. As an exception, only pyrimidine-2,4(1H,3H)-diones with an atom of chlorine or bromine in 5th position (for terbacil and bromacil) were assumed as aromatic, due to the positive mesomeric effect of the halogen in this case. Nevertheless, the problem of the unambiguous marking of aromaticity in rings is a subject for further clarification [30, 31].

## 2.2 QSBR Analysis

In classical QSAR/QSBR analysis, the predictive model is built by using many kinds of descriptors: physicochemical properties, molecular indices, and molecular fragments [32]. However, many authors noted the lack of the first two descriptor types for QSBR purposes, especially for diverse molecules. Redundancy and low discriminative power of those descriptors determine the situation, because often a small structure modification of a chemical can change its degradation ability appreciably [2, 33]. Only for sets of homogeneous compounds some success has been obtained.

Fragmental approaches, where the presence or absence of particular molecular functional groups and/or substructure in the molecule has influence on the model’s output, are especially designed for sets of diverse molecules. The better quality of descriptive models in this case is generally accepted [33]. Even the very large and heterogeneous data set of the MITI test (2-level scale) [18] was well explained [19, 20].

A number of learning approaches for determination of the relationship between degradability and molecular fragments were applied - classical MLR [11, 14, 15], PLS [19], rule-based [10, 20] as well as other methods including neural networks [12, 16, 34]; the newer data-mining arsenal is available from specific literature [35].

In this work, the decision tree approach [36] has been employed for developing the QSBR model. It has the following advantages: (i) simplicity of results - in most cases, the interpretation of results summarized in a tree

is simple. (ii) Tree methods are nonparametric and non-linear; final result for use of a tree method can be summarized in a series of (usually few) logical if-then rules (tree nodes). So, there is no implicit assumption that the underlying relationships between the predictor variables and the response (here the degradability class) are linear, follow some specific non-linear link function, or that they are even monotonic in nature. On the other hand, the sophisticated learning methods like variants of neural networks usually behave as the “black-boxes” with tangled inner interrelations and lack of transparency. This in turn leads to the serious complexity in interpretation of the structure of solution and the danger of unexpected “surprises” in predictions.

However, processing of a huge amount of possible substructural descriptors even for the algorithms especially designed for handling of a number of predictors is a challenge [37]. The development of the decision tree was carried out manually, and most efforts were directed to design and selection of appropriate fragments. The “manual” way has been chosen as it provides flexibility in generating and combining rules taking care of the chemically believable results. Testing of the hypotheses was done by supporting programming in the Statistica’s Basic [38] and the Perl languages.

The train set of 315 compounds for creation of the model is presented at the Table 1. The model was finally validated by the test set of 105 compounds from the Table 2.

The general way for construction of the decision tree was as follows: a compound is assumed as stable for degradation unless it contains some fragments associated with a quicker break-up tendency. First process was to find certain proper features that have an ability to discriminate diverse pesticides of the class 1 against more persistent ones. It was found that phosphorus compounds, derivatives of carbamic acid, amides of dicarbonic acids, amides of  $\alpha$ -chlorocarbonic acids, and some other types of compounds are generally of the class 1 (low persistence), while the structures with the aromatic nitrogen have special behavior and other parts of a molecule must be analyzed. As the next step, the decision tree was grown and optimized to decrease the total number of decision rules achieving the better generalization of the model.

The key criterion for selection of a decision rule for the model was its high discrimination. No additional rule or fragment have been added for explanation of only 1 or 2 compounds, but the extension of generalization of some existing fragment or rule was allowed. The generalization of rules was done by using of the Boolean operators combining several descriptors into the one rule. Special attention was paid to the design of the fuzzy fragmental descriptors with multiple atom types or bond types, as can be seen in Figure 2 (fragments N-CA-X, N-CA-C, Ph-COO, and others) for better generalization. Some de-

scriptors (N-CO, CN, NO<sub>2</sub>, CYC\_3C, nBO, nDB) were used several times in different places in the decision tree to compress the descriptor dimensionality of the solution.

### 3 Results and Discussion

The constructed tree of decisions and respective molecular fragments are shown in Figure 1 and Figure 2. There are 12 decision nodes and one algorithmic rule of 7 steps. The model employs 31 topological descriptors totally. Contributions of the core rules are listed in Table 3, classification results for the training set and the test set are presented in Table 4 and Table 5, respectively.

Discrimination power of descriptors is presented in terms of the total separated compounds in the Table 3. The table shows that the highest discrimination is achieved by UNSATW and N-CO-O (carbamic acid’s derivatives) descriptors. UNSATW parameter, as some kind of the weighted unsaturation index, summarizes the instability of a compound determined mostly by the number of presenting CC, CN, or CO double bonds normalized by the number of total bonds excluding hydrogen (nBO parameter). UNSATW along with HETERO value (absence of sulfur or halogens except iodine) and nBO were introduced to detach heterogeneous structures of weak persistence. Fragments O-C (single or aromatic bond between carbon and oxygen in molecule with amide group), O-CO (esters), and Ph-NHC (aromatic secondary or tertiary amines) have a high significance as well. Another interesting fact is that the presence of a 3-membered ring in a molecule noticeably increases its persistence (leaves 8 and 12 in Figure 1).

Generally, the opening nodes of the decision tree move compounds into classes of less persistence and separation is easier and more numerous. Taking into account Figure 1 and Table 3, the opening nodes (leaves 1-4) may be believed as the most important, together with the special algorithmic part for processing of the aromatic nitrogen-bearing heterocycles. Compounds with an aromatic atom of nitrogen were separated from the decision tree because no fragment was detected to be able to discriminate a particular degradability class. Presence of fragment N-CA-X decreases the pesticide stability, possibly by coupling effect, but presence of fragment N-CA-C does oppositely. Other fragments of the algorithmic part decrease the estimated degradability class (it is ‘I’ in the algorithm).

Table 4 and Table 5 have summarized the classification results for training and test sets. For the extremely heterogeneous training data set of 315 compounds including almost all types of pesticides, there were only 31 descriptors and in total 19 rules employed producing 86.3% of correctly calculated values among three classes (see Table 4). For the test set of 105 compounds, there were 78.1% of correctly predicted values. Many mis-

classified pesticides have either very wide range of half-life values or few observations, and may not belong unambiguously to one specific class; such compounds are marked in Table 1 and Table 2. There are also many unique structural classes where no reliable rule may be produced.

The case of two-unity misclassifications should be specially discussed. (i) Negative value means that a compound can be broken down very quickly, unexpectedly with the model output; benodanil and bentazon have unique fragments (iodine atom and sulfuric diamide chain, respectively), while the instability of propanil in soil is caused by known specific sensitivity to microbial hydrolysis [39]. (ii) Positive value of the mistake has a dangerous potential for the environment. Both fenac and chloroneb are persistent but were falsely marked as

quickly degradable by UNSATW descriptor. For the test set there were only positive two-unity misclassifications also incorrectly marked by UNSATW descriptor (leaf 4): 2,3,6-tba acid, clofencet (rare phenyl-hydrazide), and cycloheximide (heterogeneous and very limited persistence information). The UNSATW parameter was designed for discrimination of molecules with weak persistence of really various structural classes and has the highest discriminating power for the training set (see Table 3). As such, this high generalization may lead to mistakes in specific cases.

Comparing this result with the former works, we should notice that the prediction rate is on a comparable level with the best models handling two classes (rapidly/slowly degradable) [19, 20], while we processed three classes (low, moderate, or high persistence).

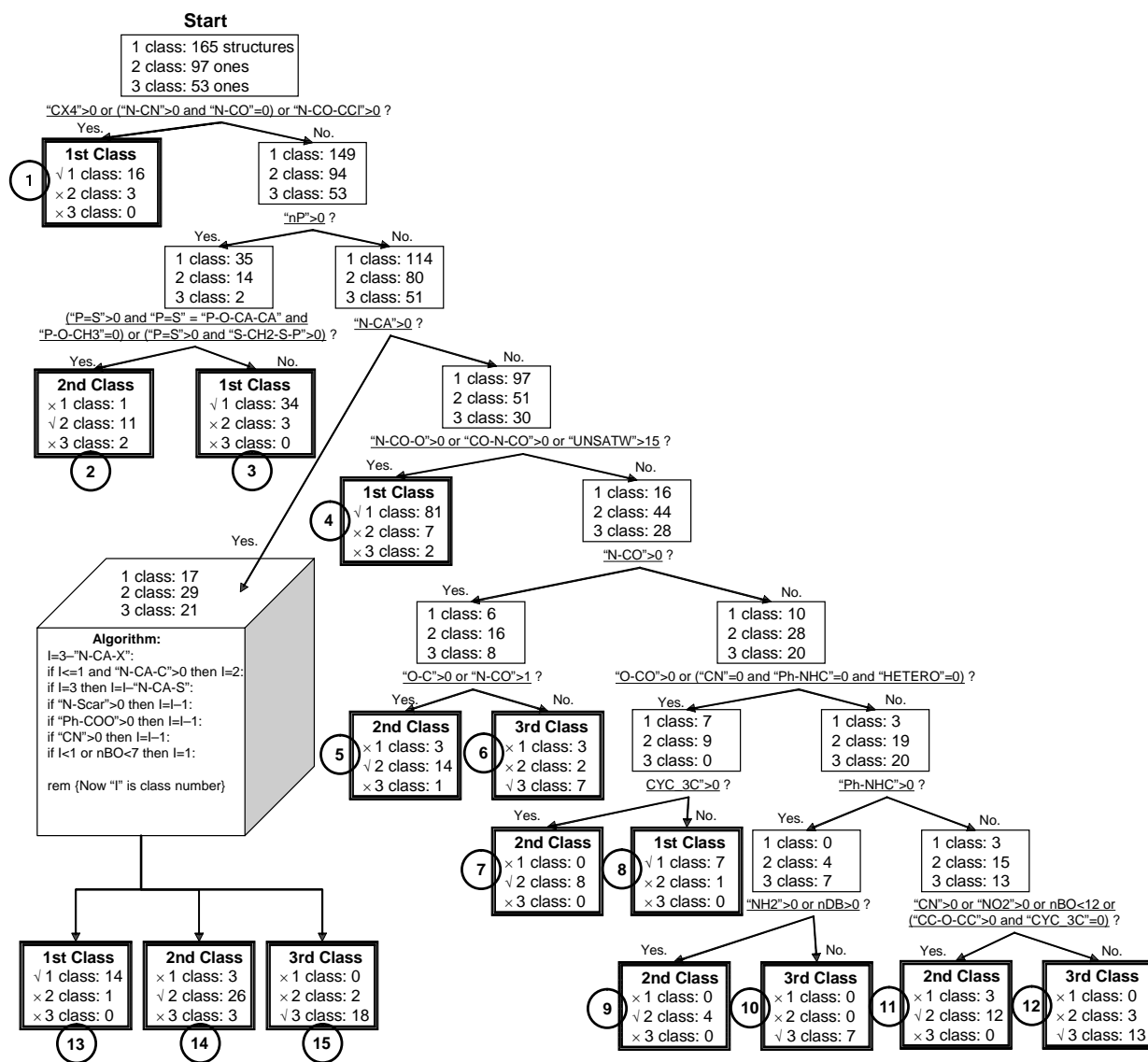


Figure 1. Decision tree model for classification of pesticide persistence in environment. Each decision node is accompanied by the numbers of compounds that arrive at the node and flow away. Terminal leaves are marked by double border; their index numbers are given in circles. Algorithmic part is shown in the BASIC-like style.

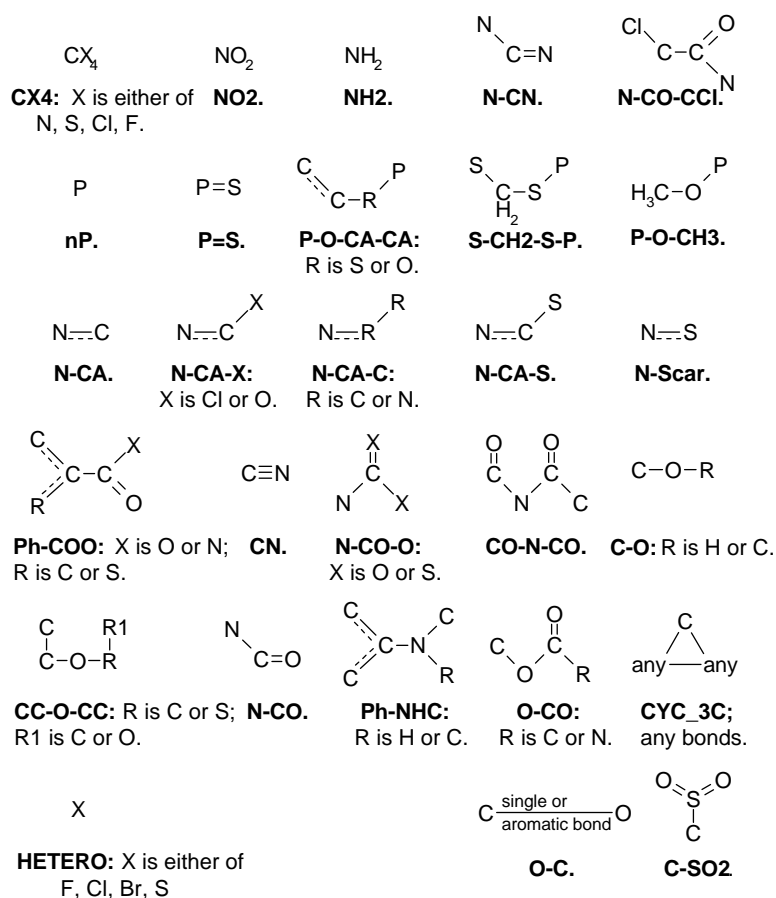
## 4 Conclusion

We developed a data set of pesticide persistence and explained it by the decision tree model. The structure of QSBR solution is visual and simple. For every pesticide from Table 1 and Table 2, the terminal leaf is presented; therefore, the classification path through the tree can be reproduced with ease. This is an important achievement because our model does not only allow prediction of the field persistence, but also assists in the construction of new compounds with desired degradation ability.

On the basis of this research, a computer expert sys-

tem **EKeeper** has been developed and made available for downloading [40]. Future investigations will be directed to the completely automated QSBR analysis of the presented data set using the new concept of fuzzy fragments.

This work was partially supported by Japan Chemical Industry Association. The authors are also thankful to the U.S. Environmental Protection Agency, the U.S. National Library of Medicine, and the Pesticide Management Education Program at Cornell University for free access to corresponding databases.



$$nDB = \text{NumberOfDoubleBonds}; nBO = \text{NumberOfBonds\_HydrogenSuppressed};$$

$$UNSATW = 100 * (nDB - 1.5 * "N-CO" + "C-O" + "NO2" - 3 * "C-SO2") / nBO;$$

Figure 2. Chart of topological descriptors. Parameters nDB and nBO are the constitutional descriptors; UNSATW is the empirical parameter. Each other predictor is the number of non-overlapped occurrences of the corresponding fragment in a molecular structure.

Table 1. Classified Observed and Predicted persistence of pesticides in environment, and predicting Terminal Leaves (see Figure 1) for the training set of 315 compounds.

Name	CAS#	Obs.	Est.	Leaf	Name	CAS#	Obs.	Est.	Leaf
1,2-dichloropropane	78-87-5	2	2	11	chlorpropham	101-21-3	1	1	4
1,3-dichloropropene	542-75-6	1	1	4	chlorpyrifos-ethyl	2921-88-2	2	2	2
2,4,5-t acid	93-76-5	1	1	4	chlorpyrifos-methyl	5598-13-0	1	1	3
2,4-d acid	94-75-7	1	1	4	chlorsulfuron	64902-72-3	2 ☉	2	14
2,4-db acid	10433-59-7	1	1	4	chlozolinate	72391-46-9	1	1	4
3cpa	101-10-0	1	1	4	cinmethylin	87818-31-3	2 ↓	1	8
abamectin	65195-56-4	1	1	4	clofentezine	74115-24-5	2	3	15
acephate	30560-19-1	1	1	3	clomazone	81777-89-1	2 ☉	2	5
acifluorfen	62476-59-9	1	1	4	clopyralid	1702-17-6	2	2	14
acrolein	107-02-8	1	1	4	cyanazine	21725-46-2	1	1	13
alachlor	15972-60-8	1	1	1	cycloate	1134-23-2	1	1	4
aldicarb	116-06-3	1	1	4	cyfluthrin	68359-37-5	2	2	7
aldoxycarb	1646-88-4	1	1	4	cymoxanil	57966-95-7	1	1	4
aldrin	309-00-2	3 ☉ <sup>a</sup>	3	12	cypermethrin	52315-07-8	2	2	7
ametryn	834-12-8	2 ☉	2	14	cyromazine	66215-27-8	3	3	15
aminocarb	2032-59-9	1	1	4	dalapon	127-20-8	1	1	4
amitraz	33089-61-1	1	1	1	daminozide	1596-84-5	1	1	4
amitrole	61-82-5	1	1	13	dazomet	533-74-4	1	1	4
ams	7773-06-0	1	1	4	dbcp	96-12-8	2	2	11
ancymidol	12771-68-5	3 ↓ <sup>b</sup>	3	15	dcna (dicloran)	99-30-9	2 ☉	2	11
anilazine	101-05-3	1	2 <sup>c</sup>	14	dcpa	1861-32-1	2	1	4
asulam	3337-71-1	1	1	4	ddd (tde)	72-54-8	3	3	12
atrazine	1912-24-9	2 ☉	2	14	dde	3424-82-6	3	3	12
azimsulfuron	120162-55-2	2	2	14	ddt	50-29-3	3	3	12
azinhos-methyl	86-50-0	1	1	3	demeton-o	298-03-3	1	1	3
barban	101-27-9	1	1	4	desmedipham	13684-56-5	1	1	4
benalaxyl	71626-11-4	2	2	5	di-allate	2303-16-4	1 ☉	1	4
bendiocarb	22781-23-3	1	1	4	diazinon	333-41-5	2	2	2
benefin	1861-40-1	3	3	10	dicamba	1918-00-9	1	1	4
benodanil	15310-01-7	1	3	6	dichlobenil	1194-65-6	2 ☉	2	11
benomyl	17804-35-2	3	3	15	dichlone	117-80-6	1 ☉	1	4
bensulfuron methyl	83055-99-6	1	1	13	dichlormid	37764-25-3	1	1	1
bensulide	741-58-2	2	1	3	dichlorprop	120-36-5	1	1	4
bentazon	50723-80-3	1	3	6	dichlorvos (ddvp)	62-73-7	1	1	3
bifenox	42576-02-3	1	1	4	diclofop-methyl	51338-27-3	1	1	4
bifenthrin	82657-04-3	2	2	7	dicofol	115-32-2	2	3	11
bromacil acid	314-40-9	3	3	15	dicrotophos	141-66-2	1	1	3
bromoxynil	1689-84-5	1	2	11	dieldrin	60-57-1	3	3	12
butachlor	23184-66-9	1	1	1	dienochlor	2227-17-0	1 ↓	1	4
butylate	2008-41-5	1	1	4	diethyl-ethyl	38727-55-8	1	1	1
captafol	2425-06-1	1	1	4	difenzoquat (as amine)	43222-48-6	3	3	15
captan	133-06-2	1	1	1	diflubenzuron	35367-38-5	1	1	4
carbaryl	63-25-2	1	1	4	dimethipin	55290-64-7	3	3	12
carbendazim	10605-21-7	3 ↓	3	15	dimethirimol	5221-53-4	3 ↓	2	14
carbofuran	1563-66-2	2	1	4	dimethoate	60-51-5	1	1	3
carbon disulfide	75-15-0	1	1	4	dinitramine	29091-05-2	2 ☉	2	9
carbophenothion	786-19-6	2	2	2	dinocap	39300-45-3	1	1	4
carboxin	5234-68-4	1	2	5	dinoseb	88-85-7	1	1	4
cdaa	93-71-0	1	1	1	dioxacarb	6988-21-2	1	1	4
chloramben	133-90-4	1	1	4	diphenamid	957-51-7	3	3	6
chlorbromuron	13360-45-7	2	2	5	dipropetryn	4147-51-7	2 ☉	2	14
chlordane	57-74-9	3	3	12	diquat	85-00-7	3	3	15
chlordimeform	19750-95-9	2	1	1	disulfoton	298-04-4	1	1	3
chlorethoxyfos	54593-83-8	1	1	3	diuron	330-54-1	3	3	6
chlorimuron ethyl	90982-32-4	1	1	13	dnoc	534-52-1	1	1	4
chlorobenzilate	510-15-6	1	1	8	dodine (cyprex)	2439-10-3	1	1	1
chloroneb	2675-77-6	3	1	4	endosulfan	115-29-7	2	2	11
chloropicrin	76-06-2	1	1	1	endothall	145-73-3	1	1	4
chlorothalonil	1897-45-6	2	2	11	epn	2104-64-5	2 ☉	2	2
chloroxuron	1982-47-4	2	2	5	epte	759-94-4	1	1	4

Name	CAS#	Obs.	Est.	Leaf	Name	CAS#	Obs.	Est.	Leaf
esfenvalerate	66230-04-4	1	1	8	maneb	12427-38-2	1	1	4
ethalfluralin	55283-68-4	2	2	9	mcpa	94-74-6	1	1	4
ethametsulfuron methyl	97780-06-8	2 ↓	1	13	mcpb	94-81-5	1	1	4
ethephon	16672-87-0	1	1	3	mecoprop	93-65-2	1	1	4
ethion	563-12-2	3 ☉	2	2	mefluidide	53780-34-0	1	1	1
ethofumesate	26225-79-6	2	2	11	mepiquat chloride	24307-26-4	1 ↓	1	8
ethoprop	13194-48-4	1	1	3	metalaxyl	57837-19-1	2	2	5
ethylene dibromide	106-93-4	2	2	11	metalddehyde	108-62-3	1	1	4
etridiazole	2593-15-9	1	1	13	metham acid	137-42-8	1	1	4
fenac (chlorfenac)	2439-00-1	3	1	4	methamidophos	10265-92-6	1	1	3
fenaminosulf	140-56-7	1	1	4	methazole	20354-26-1	1	1	4
fenamiphos	22224-92-6	1	1	3	methidathion	950-37-8	1	1	3
fenarimol	60168-88-9	3	3	15	methiocarb	2032-65-7	1	1	4
fenfuram	24691-80-3	2 ↓	2	5	methomyl	16752-77-5	1	1	4
fenitrothion	122-14-5	1	1	3	methoxychlor	72-43-5	3	3	12
fenoprop	93-72-1	1	1	4	methyl bromide	74-83-9	1	2	11
fenoxaprop-ethyl	66441-23-4	1	1	4	methyl isothiocyanate	556-61-6	1	1	4
fenoxycarb	72490-01-8	1	1	4	methyl parathion	298-00-0	1	1	3
fenpropathrin	64257-84-7	2 ☉	2	7	metolachlor	51218-45-2	2	1	1
fensulfthion	115-90-2	2 ☉	2	2	metribuzin	21087-64-9	3 ☉	3	6
fenthion	55-38-9	2 ☉	1	3	metsulfuron methyl	74223-64-6	1	1	13
fenuron	101-42-8	3	3	6	mevinphos	7786-34-7	1	1	3
ferbam	14484-64-1	1	1	4	mexacarbate	315-18-4	1	1	4
fluazifop-butyl	69335-91-7	2	2	14	mirex	2385-85-5	3	3	12
fluchloralin	33245-39-5	3	3	10	molinate	2212-67-1	1	1	4
flucythrinate	70124-77-5	1	1	8	monocrotophos	6923-22-4	1	1	3
flumetralin	62924-70-3	3	3	10	monolinuron	1746-81-2	2	2	5
flumetsulam	98967-40-9	2	2	14	monuron	150-68-5	3	3	6
fluometuron	2164-17-2	2	3	6	myclobutanil	88671-89-0	2 ↓	2	14
flupyrsulfuron (acid)	144740-54-5	1	1	13	naled	300-76-5	1	1	3
fluridone	59756-60-4	3	3	12	naphthalene	91-20-3	1	1	8
flusilazole	85509-19-9	3	3	15	napropamide	15299-99-7	2	2	5
fluvalinate	69409-94-5	1	1	8	naptalam	132-66-1	1	2	5
fomesafen	72178-02-0	3 ☉	2	5	neburon	555-37-3	3	3	6
fonofos	944-22-9	2	2	2	nicosulfuron	111991-09-4	1	1	13
formetanate	23422-53-9	1	1	4	nitrapyrin	1929-82-4	2	2	14
fosamine	25954-13-6	1	1	3	nitrofen	1836-75-5	2 ☉	2	11
fosetyl-aluminum	39148-24-8	1	1	3	norflurazon	27314-13-2	2 ☉	1	1
glufosinate	77182-82-2	1	1	3	oryzalin	19044-88-3	2 ☉	2	9
glyphosate	1071-83-6	1	1	3	oxadiazon	19666-30-9	2	1	4
haloxyfop-methyl	69806-40-2	2	2	14	oxamyl	23135-22-0	1	1	4
heptachlor	76-44-8	3	3	12	oxycarboxin	5259-88-1	2 ↓	2	5
hexachlorobenzene	118-74-1	3	3	12	oxydemeton-methyl	301-12-2	1	1	3
hexazinone	51235-04-2	2 ☉	2	5	oxyfluorfen	42874-03-3	2 ☉	2	11
hexythiazox	78587-05-0	1	1	4	oxythioquinox	2439-01-2	1	1	13
hydramethylnon	67485-29-4	1	1	1	paclobutrazol	76738-62-0	3	3	15
imazalil	35554-44-0	3	3	15	paraquat	1910-42-5	3	3	15
imazamethabenz methyl	81405-85-8	2	2	5	parathion	56-38-2	2	2	2
imazapyr	81334-34-1	2	2	14	penb	82-68-8	2	2	11
imazaquin	81335-37-7	2	2	14	pebulate	1114-71-2	1	1	4
imazethapyr	81335-77-5	2	2	14	pendimethalin	40487-42-1	3	3	10
iprodione	36734-19-7	1	1	4	pentachlorophenol	87-86-5	2	3	11
isazofos	42509-80-8	1 ☉	1	3	perfluidone	37924-13-3	1	1	1
isofenphos	25311-71-1	2 ☉	2	2	permethrin	52645-53-1	2	2	7
isopropalin	33820-53-0	3 ☉	3	10	phenmedipham	13684-63-4	2 ☉	1	4
isoxaben	82558-50-7	2	2	5	phenthoate	2597-03-7	2 ☉	1	3
lactofen	77501-63-4	1	1	4	phorate	298-02-2	2	2	2
lambda-cyhalothrin	91465-08-6	2	2	7	phosalone	2310-17-0	1	1	3
lenacil	2164-08-1	2 ↓	1	4	phosmet	732-11-6	1	1	3
lindane	58-89-9	3	3	12	phosphamidon	13171-21-6	1	1	1
linuron	330-55-2	2	2	5	picloram	1918-02-1	2	2	14
malathion	121-75-5	1	1	3	piperalin	3478-94-2	1 ↓	1	8
maleic acid hydrazide	123-33-1	2 ☉	2	5	pirimicarb	23103-98-3	2	2	14

Name	CAS#	Obs.	Est.	Leaf	Name	CAS#	Obs.	Est.	Leaf
pirimiphos-ethyl	23505-41-1	2	2	2	temephos	3383-96-8	1	1	3
pirimsulfuron-methyl	86209-51-0	1	1	13	terbacil	5902-51-2	3	3	15
prochloraz	67747-09-5	3 $\perp$	3	15	terbufos	13071-79-9	1	2	2
procymidone	32809-16-8	1 $\perp$	1	4	terbutryn	886-50-0	2 $\odot$	2	14
prodiamine	29091-21-2	2 $\perp$	2	9	tetrachlorvinphos	22248-79-9	1	1	3
profenofos	41198-08-7	1	1	3	thiabenzazole	148-79-8	3	3	15
profluralin	26399-36-0	3	3	10	thidiazuron	51707-55-2	2 $\perp$	2	14
promecarb	2631-37-0	1	1	4	thifensulfuron methyl	79277-27-3	1	1	13
prometon	1610-18-0	3	2	14	thiobencarb	28249-77-6	1	1	4
prometryn	7287-19-6	2 $\odot$	2	14	thiocyclam	31895-22-4	1 $\perp$	2	11
pronamide	23950-58-5	2	3	6	thiodicarb	59669-26-0	1	1	4
propachlor	1918-16-7	1	1	1	thiophanate-methyl	23564-05-8	1 $\perp$	1	4
propamocarb	25606-41-1	1	1	4	thiram	137-26-8	1	1	4
propanil	709-98-8	1	3	6	tolclofos-methyl	57018-04-9	1 $\perp$	1	3
propargite	2312-35-8	2	2	11	tralomethrin	66841-25-6	2	2	7
propazine	139-40-2	2	2	14	triadimefon	43121-43-3	2 $\odot$	3	15
propham	122-42-9	1	1	4	triadimenol	55219-65-3	3	3	15
propiconazole	60207-90-1	3	3	15	triallate	2303-17-5	2	/	4
propoxur	114-26-1	2 $\odot$	/	4	triasulfuron	82097-50-5	2 $\odot$	2	14
pyrazon (chloridazon)	1698-60-8	1	1	1	tribenuron methyl	101200-48-0	1	1	13
pyrethrin ii	121-29-9	1	1	4	tribufos	78-48-8	1 $\perp$ $\odot$	1	3
pyrithiobac	123343-16-8	1	1	13	trichlorfon	52-68-6	1	1	3
quizalofop-ethyl	76578-14-8	2 $\perp$	2	14	trichloronat	327-98-0	3	2	2
resmethrin	10453-86-8	2 $\perp$	2	7	tricyclpyr	55335-06-3	2	2	14
rimulfuron	122931-48-0	1 $\perp$	2	14	tricyclazole	41814-78-2	1 $\perp$	2	14
rotenone	83-79-4	1	1	4	tridiphane	58138-08-2	2 $\perp$	3	12
sebumeton	26259-45-0	3 $\perp$	2	14	triflumizole	99387-89-0	1 $\perp$	1	1
sethoxydim	74051-80-2	1	1	4	trifluralin	1582-09-8	3	3	10
siduron	1982-49-6	3	3	6	triflusulfuron methyl	126535-15-7	1 $\perp$	1	13
simazine	122-34-9	2 $\odot$	2	14	triforine	26644-46-2	1 $\perp$	2	5
simetryn	1014-70-6	2 $\perp$	2	14	trimethacarb	12407-86-2	1 $\perp$	1	4
sulfometuron-methyl	74222-97-2	2 $\odot$	2	14	vernolate	1929-77-7	1	1	4
sulprofos	35400-43-2	2 $\odot$	2	2	vinclozolin	50471-44-8	1	1	4
tca (trichloroacetic acid)	76-03-9	1	1	4	zineb	12122-67-7	1	1	4
tebuthiuron	34014-18-1	3	3	15					

- a) Symbol “ $\odot$ ” marks very wide range of half-life values; fuzzy persistence class.  
b) Symbol “ $\perp$ ” marks few and unclear observations for a compound; unreliable persistence class.  
c) Misclassification is marked by *Italic* font.

Table 2. Classified Observed and Predicted persistence of pesticides in environment, and predicting Terminal Leaves (see Figure 1) for the test set of 105 compounds.

Name	CAS#	Obs.	Est.	Leaf	Name	CAS#	Obs.	Est.	Leaf
1-naphthol	90-15-3	1	1	8	chloranil	118-75-2	1 $\perp$	1	4
2-phenylphenol	90-43-7	1	1	8	chlorazine	580-48-3	2 $\odot$	2	14
2,3,6-tba acid (tba)	50-31-7	3	/	4	chlorbufam	1967-16-4	1	1	4
4-aminopyridine	504-24-5	3	3	15	chlordecone	143-50-0	3	3	12
4cpa	122-88-3	1	1	4	chlordimeform	6164-98-3	2	/	1
8-hydroxyquinoline	134-31-6	3	3	15	chlormequat	999-81-5	2 $\perp$	2	11
acetochlor	34256-82-1	1 $\perp$	1	1	chloroacetic acid	79-11-8	1	1	4
acibenzolar-s-methyl	135158-54-2	1	1	4	chlorofenvinphos	470-90-6	2 $\odot$	/	3
acrylonitrile	107-13-1	1	1	4	chloroform	67-66-3	2	2	11
allyl alcohol	107-18-6	1	1	4	chloropropylate	5836-10-2	2 $\perp$ $\odot$	/	8
alpha-chlorohydrin	96-24-2	1 $\odot$	1	4	chlortoluron	15545-48-9	2 $\odot$	3	6
alphacypermethrin	67375-30-8	1	2	7	clodinafop	105512-06-9	1	2	14
anthraquinone	84-65-1	1	1	8	clofencet	129025-54-3	3	/	4
azobenzene	103-33-3	1	1	8	cloransulam	147150-35-4	1	1	13
biphenyl	92-52-4	2 $\odot$	/	8	coumaphos	56-72-4	3	2	2
bromophos	2104-96-3	2	1	3	cresol	1319-77-3	1	1	8
butylamine	13952-84-6	1	1	8	crotoxyphos	7700-17-6	1	1	3
carbon tetrachloride	56-23-5	1	1	1	cyanophos	2636-26-2	1	1	3

Name	CAS#	Obs.	Est.	Leaf	Name	CAS#	Obs.	Est.	Leaf
cyclanilide	113136-77-9	2	2	5	hydrogen cyanide	74-90-8	2	2	11
cycloheximide	66-81-9	3 ↓ ⊙	1	4	imazamox	114311-32-9	2	2	14
cycluron	2163-69-1	3	3	6	ioxynil	1689-83-4	1	2	11
cyhalothrin	68085-85-8	2	2	7	isobenzan	297-78-9	1 ⊙	2	11
cyprazine	22936-86-3	2	2	14	isodrin	465-73-6	3	3	12
cyprodinil	121552-61-2	3	3	15	jodofenphos	18181-70-9	2 ↓	1	3
dcip	108-60-1	2	2	11	kresoxim-methyl	143390-89-0	1	1	4
dehydroacetic acid	520-45-6	1	1	4	mecarbam	2595-54-2	1	1	3
deltamethrin	52918-63-5	2	2	7	menazon	78-57-9	1	1	3
demeton-s-methyl	919-86-8	1	1	3	methoprene	40596-69-8	1	1	4
desmetryn	1014-69-3	2	2	14	methylchloroform	71-55-6	2	2	11
dibutyl phthalate	84-74-2	1	1	4	methylene chloride	75-09-2	1	2	11
dichlofluanid	1085-98-9	1	1	1	methyleugenol	93-15-2	1	1	4
dichlorophene	97-23-4	3	3	12	metolcarb	1129-41-5	1	1	4
diethyltoluamide	134-62-3	3	3	6	nitralin	4726-14-1	2	2	9
diflufenzopyr	109293-97-2	1	2	14	o-dichlorobenzene	95-50-1	2	2	11
dimethylphthalate	131-11-3	1	1	4	omethoate	1113-02-6	1 ↓	1	3
dinobuton	973-21-7	1	1	4	p-dichlorobenzene	106-46-7	2	2	11
dioxathion	78-34-2	1	1	3	phenothrin	26002-80-2	3 ↓	2	7
disul	136-78-7	2	1	4	piperonyl butoxide	51-03-6	1 ↓	1	4
endrin	72-20-8	3	3	12	proprtamphos	31218-83-4	1	1	3
ethyl formate	109-94-4	1	1	4	pyrethrin i	121-21-1	1	1	4
ethylene	74-85-1	1	1	4	pyrimethanil	53112-28-0	2	3	15
ethylene_dichloride	107-06-2	2	2	11	strychnine	57-24-9	2	2	5
ethylene_oxide	75-21-8	1	1	4	sulfentrazone	122836-35-5	3	3	6
fenchlorphos	299-84-3	1	1	3	terbutylethylazine	5915-41-3	3	2	14
fluzifop	69806-50-4	2	2	14	tetrachloroethane	79-34-5	2	2	11
flufenacet	142459-58-3	2 ⊙	2	14	tetradifon	116-29-0	3	3	12
flumioxazin	103361-09-7	1	1	4	thiazopyr	117718-60-2	2	2	14
fluoroacetamide	640-19-7	2	2	5	thiometon	640-15-3	1 ↓	1	3
fluroxypyr	69377-81-7	1	2	14	thionazin	297-97-2	2	2	2
folpet	133-07-3	1	1	1	tralkoxydim	87820-88-0	1	1	4
formaldehyde	50-00-0	1	1	4	xmc	2275-23-2	2 ⊙	1	4
formothion	2540-82-1	1	1	3	xlenols	1300-71-6	1	1	8
hexachlorobutadiene	87-68-3	1	1	4					

Table 3. Contribution of the core rules.

Rule	Tree's leaves	Totally separated compounds	Incorrectly separated compounds
"CX4">0	1	4	0
"N-CN">0 and "N-CO"=0	1	5	1
"N-CO-CC1">0	1	10	2
"P=S">0 and "P-O-CH3"=0 and "P=S" = "P-O-CA-CA"	2	10	1 (0) <sup>a</sup>
"P=S">0 and "S-CH2-S-P">0	2	4	2 (1) <sup>a</sup>
"N-CO-O">0	4	44	4
"CO-N-CO">0	4	8	1
"UNSATW">15	4	45	3
"N-CO">0 and "O-C">0	5	15	3 (1) <sup>a</sup>
"N-CO">1 <sup>b</sup>	5	3	1 (0) <sup>a</sup>
"O-CO">0	7-8	13	0
"HETERO"=0	7-8	4	1
"Ph-NHC">0	9-10	11	0
"CN">0 <sup>b</sup>	11	3	1 (0) <sup>a</sup>
"NO2">0	11	4	0
"nBO"<12	11	7	3 (0) <sup>a</sup>
"CC-O-CC">0 and "CYC_3C"=0	11	3	0

a) For this terminal rule, the one-unity (class 1/class 2 and class 2/class 3) misclassification was not assumed to be a serious difficulty, if in the case of absence of the condition, the two-unity (class 1/class 3) misclassifications would occur. Number of misclassifications excluding such "safe" one-unity mistakes is given in parenthesis.

b) The descriptor is used several times in different rules of the model.

Table 4. Summary of classification results for the training set.

Interclass mistake (Observed - Predicted)	Count
-2	3
-1	17
0	272
1	21
2	2

Table 5. Summary of classification results for the test set.

Interclass mistake (Observed - Predicted)	Count
-2	0
-1	9
0	82
1	11
2	3

## References

- [1] We accepted the common tendency and used the QSBR abbreviation (Quantitative Structure-Biodegradation Relationship). In fact, the biodegradation is one of the major ways of chemical decay, and often is associated with the whole degradation process.
- [2] G. Klopman and M. Tu, *Encyclopedia of Computational Chemistry*, Wiley, Chichester (1998), pp. 128-135.
- [3] W. J. G. M. Peijnenburg, *Pure Appl. Chem.*, **66**, 1931 (1994).
- [4] J. R. Parsons and H. A. J. Govers, *Ecotoxicol. Environ. Safety*, **19**, 212 (1990).
- [5] G. J. Niemi, G. D. Veith, R. R. Regal, and D. D. Vaishnav, *Environ. Toxicol. Chem.*, **6**, 515 (1987).
- [6] R. S. Boethling, B. Gregg, F. R. Gabel, N. W. Campbell, and A. Sabljic, *Ecotoxicol. Environ. Safety*, **18**, 252 (1989).
- [7] S. M. Desai, R. Govind, and H. H. Tabak, *Environ. Toxicol. Chem.*, **9**, 473 (1990).
- [8] P. Bhagat, *Chem. Eng. Prog.*, **86**, 55 (1990).
- [9] G. Klopman and M. J. McGonigal, *J. Chem. Inf. Comput. Sci.*, **21**, 48 (1981).
- [10] K. Hiromatsu, Y. Yakabe, K. Katagiri, and Tsu. Nishihara, *Chemosphere*, **41**, 1749 (2000).
- [11] H. H. Tabak, C. Gao, S. Desai, and R. Govind, *Water Sci. Technol.*, **26**, 763 (1992).
- [12] H. H. Tabak and R. Govind, *Environ. Technol. Chem.*, **12**, 251 (1993).
- [13] BIODREG, Environmental Fate Database of Syracuse Research Corporation, Environmental Science Center division, 301 Plainfield Road, Syracuse, NY 13212 USA.  
URL: <http://esc.syrres.com/>
- [14] P. H. Howard, R. S. Boethling, W. M. Stiteler, W. M. Meylan, A. E. Hueber, H. A. Beauman, and M. E. Larosche, *Environ. Toxicol. Chem.*, **11**, 593 (1992).
- [15] G. Klopman, D. M. Balthasar, and H. S. Rosendranz, *Environ. Toxicol. Chem.*, **12**, 231 (1993).
- [16] J. Devillers, D. Domine, and R. S. Boethling, *Neural Networks in QSAR and Drug Design*, ed by J. Devillers, Academic Press, New York (1996), pp. 65-82.
- [17] Correlation coefficient for this test set had been absent in the original, but was calculated by ourselves just from the tabulated results of original's Table II.
- [18] Japan Chemical Industry Ecology-Toxicology & Information Center (JETOC), "Biodegradation and Bioaccumulation Data of Existing Chemicals Based on the Chemical Substances Control Law (CSCL Japan)," Tokyo (1992).
- [19] H. Loonen, F. Lindgren, B. Hansen, W. Karcher, J. Niemela, K. Hiromatsu, M. Takatsuki, W. Peijnenburg, E. Rorije, J. Struijs, *Environ. Toxicol. Chem.*, **18**, 1763 (1999).
- [20] A. Sabljic and W. Peijnenburg, *Pure Appl. Chem.*, **73**, 1331 (2001).
- [21] A. C. Waldron, "Pesticides and Groundwater Contamination," Ohio State University Extension Bulletin, 820, Columbus (Ohio) (1992). Available on the Internet at <http://ohioline.ag.ohio-state.edu/b820/index.html>.
- [22] A. G. Hornsby, R. Don Wauchope, and A. E. Herner, *Pesticide Properties in the Environment*, Springer, New York (1995).
- [23] Pesticide Property Database of the Alternate Crops and Systems Laboratory of Beltsville Agricultural Research Center. Available on the Internet at <http://wizard.arsusda.gov/acsl/ppdb.html>.

- [24] Hazardous Substances Data Bank of U.S. National Library of Medicine. Available on the Internet at <http://toxnet.nlm.nih.gov/cgi-bin/sis/htmlgen?HSDB>.
- [25] Pesticide Management Education Program at Cornell University. Available on the Internet at <http://pmep.cce.cornell.edu>.
- [26] A. McCulloch, *Chemosphere*, **47**, 667 (2002).
- [27] Very short record for each chemical is given in the Table 1, because of space limitation. Detailed information about all HL values and molecular structures is available from authors on request.
- [28] Compendium of Pesticide Common Names. Available on the Internet at <http://www.hclrss.demon.co.uk>.
- [29] ChemIDPlus Database of U.S. National Library of Medicine. Available on the Internet at <http://chem.sis.nlm.nih.gov/chemidplus/>.
- [30] J. March, *Advanced Organic Chemistry: Reactions, Mechanisms, and Structure*, Wiley, New York (1992).
- [31] M. K. Cyrański, T. M. Krygowski, A. R. Katritzky, and P. von R. Schleyer, *J. Org. Chem.*, **67**, 1333 (2002).
- [32] M. Karelson, *Molecular Descriptors in QSAR/QSPR*, Wiley, New York (2000).
- [33] J. Devillers, *Encyclopedia of Computational Chemistry*, Wiley, Chichester (1998), pp. 932-941.
- [34] J. W. Raymond, T. N. Rogers, D. R. Shonnard, and A. A. Kline, *J. Hazard. Mater.*, **84**, 189 (2001).
- [35] "KDnuggets News," the e-newsletter on Data Mining, Data Mining Books section. Available on the Internet at <http://www.kdnuggets.com/publications/books.html>.
- [36] J. R. Rose, *Encyclopedia of Computational Chemistry*, Wiley, Chichester (1998), pp. 1521-1525.
- [37] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*, Wadsworth, Belmont (1984).
- [38] Data analysis and statistical programming environment STATISTICA v. 5-6. Information is available on the Internet at <http://www.statsoft.com>.
- [39] R. Bartha, *J. Agr. Food Chem.*, **19**, 385 (1971).
- [40] EKeeper software for evaluation of the level of persistence of chemicals in environment. Available on Internet at <http://www.mis.tutkie.tut.ac.jp/>; go to "English" / "MIS-services".

## 農薬の環境残留性-データの収集と構造特徴解析

Sokratis ALIKHANIDI, 高橋 由雅\*

豊橋技術科学大学 知識情報工学系, 〒441-8580 愛知県豊橋市天伯町雲雀ヶ丘 1-1

\*e-mail: [taka@mis.tutkie.tut.ac.jp](mailto:taka@mis.tutkie.tut.ac.jp)

本研究では 420 種の農薬の土壌中での半減期 ( half-life, HL ) データを収集し構造活性相関を行った。解析に際しては結果の不確実性等を考慮し、これらデータを 3 つのクラス ( HL(30 日をクラス 1、30 日 < HL(100 日をクラス 2、100 日 < HL をクラス 3 ) に分けて用いた。315 化合物を訓練集合とし、31 種の構造記述子を用いて構造 生分解性相関のモデル化を行った。その結果 272 化合物 (86.3%) の活性クラスを正しく予測できた。一方、38 化合物は隣接クラスへの誤分類であり、非隣接クラスへの誤分類は 5 化合物であった。一方、予測集合 105 化合物に対しては 82 化合物 (78.1%) が正しく予測され、非隣接への誤分類は 3 化合物であった。また、得られたモデルをもとに生分解性予測システム EKeeper を作成した。

キーワード: QSAR, QSBR, データマイニング, 生分解性予測, エキスパートシステム, 農薬残留性