Errors in the PUR database

Before reported pesticide use data is stored in the PUR database at the Department of Pesticide Regulation (DPR), a computer program (called the "loader" program) checks the values of all data fields for several possible kinds of errors. Another program is run after all data for each year have been entered into the database to determine if any values for rate of use are unusually high relative to other reported uses during that year. When possible DPR or county staff attempt to determine correct values and correct them. However, not all errors are corrected.

A table of all errors found in the PUR is given in the file "errorsYYYY.txt" (where "YYYY" represents the PUR year) and a list of changes made to erroneous values is given in file "changesYYYY.txt". Erroneous values were either estimated or replaced with a character that represents no value, which is called the "null" character. The records in this table can be linked back to the PUR data in the UDC files using the use_no, which uniquely identify each application by a product for a particular year. Also, each record in the UDC table with one or more errors will have an "X" in its error_flag field.

The file errorsYYYY.txt contains the following information:

1. use_no: uniquely identifies each application of a pesticide product in the PUR for a particular year
2. error_id: uniquely identifies each error
3. error_code: identifies which of about 40 different errors occurred
4. error_type: indicates whether the error is "invalid", "possible", or "inconsistent"
5. duplicate_set: if this record is an erroneous duplicate of another record, this field will have a number identifying the set of duplicate records it belongs to
6. error_description: description of type of error
7. comments: provides additional information about some error codes, especially error codes 37, 39, 72, and 75

The file changesYYYY.txt contains the following information for each value that was estimated or replaced by a "null" character:

1. use_no: uniquely identifies each application of a pesticide product in the PUR for a particular year
2. field_name: the name of the column or field in the data with an error
3. old_value: the original value reported for this data field
4. new_value: the estimated value, or "null" character, that was used to replace the erroneous value
5. error_id: identifies the corresponding record in the ERRORS table.

A brief explanation of each kind of error is given in the file error_descriptions.txt, but more complete explanations are provided in the DPR reports "Pesticide Use Report Loading and Error-Handling Processes", Larry Wilhoit, January 2002, Report # PM 02-01 and "A Computer Program to Identify Outliers in the Pesticide Use Report Database", Larry Wilhoit, April 1998, Report # 98-01. These reports can be found on DPR's website at www.cdpr.ca.gov.

A record is flagged with error 37 if the reported 4-part registration number does not match any pesticide in DPR's pesticide label database. If a match is not found for the 4-part registration number then the program looks for a product that matches the 3-part or 2-part registration number. If one is found then that product is chosen; if more that one is found one is chosen using procedures described in Report PM 02-01 mentioned above. If a 3-part registration number is found the error_type is returned as "possible"; if only a 2-part registration number is found error_type is set to "invalid"; the comments field gives reported and chosen registration number of the product and why that product was chosen. If no product is found, the error_type is set to "invalid" and prodno is set = -1.

The program will flag records with a rate of use (pounds of product or AI per unit treated) that are so large they are likely errors. We use three different criteria to identify rate outliers by comparing each use rate with an estimate of the maximum rate for that type of use. These records will have either error code 72 or 75 depending on which criteria were exceeded; for records with error code 75 the program will make estimates for either amt_prd_used, lbs_prd_used, acre_treated, or unit_treated.

Records with error code 75 have rates higher than limits determined by at least one of two criteria. The criterion 1 limit is 200 pounds of active ingredient per acre (or 1000 pounds per acre for fumigants). The criterion 2 limit is 50 times the median rate for all uses with the same pesticide product, crop treated, unit treated, and record type (that is, production agriculture or monthly summary report). Records with error code 72 have rates higher than the criterion 3 limit, which is a value determined by a neural network procedure. These criteria are explained in more detail below and in Report PM 98-01 mentioned above.

Identifying outliers using the 2nd and 3rd criteria require a distribution of actual rates of use. Rate outliers are identified at two different times each year: when the data are first loaded into DPR's database and after all the data for a year have been loaded. When the data are first loaded, outliers are identified by comparing the rates with the outlier limits determined from the previous year's data; after all the data for the current year are loaded, the outlier program determines the distribution of rates based on that year's data and then uses that distribution to once again identify outliers. If a rate is greater than either the criterion 1 or 2 limit, the value used in the PUR is estimated as the median rate of use for all applications of that product on that site in the previous or current year.

The comments give the reported rate of use, both as pounds of product per unit treated and pounds of AI per unit treated, and the median rate of use based on data from the current or previous year.

Further explanation of outliers

Rate of use is not one of the fields in the PUR table. Rates are calculated by dividing the pounds of pesticide used by the acres or other unit treated. Thus, an extremely high rate value could occur from either an extremely high weight or extremely low unit treated.

Only extremely large rates are flagged, not extremely small ones, because only large values will have a major influence on statistics involving pounds of pesticide use. What value to use for the

maximum rate in each criterion is somewhat arbitrary; the value determines how conservative one wants to be. We chose maximum rates to be close to what were considered obvious outliers by a group of scientists in a survey described below in the description of the neural network criterion.

There are many possible methods for determining if a value is an outlier. If we knew the maximum label rates for particular uses, then rates in the PUR could be compared to these maximum rates, but unfortunately this information is not available in the PUR or in the Pesticide Label Database. The other methods to identify outliers involve looking at the distribution of the actual use rates. If the values are normally distributed, then one can identify outliers using a number of statistical procedures. If the values have an unknown or nonstandard distribution, then there exist no standard statistical procedures for identifying outliers. Nevertheless, people can look at a distribution and usually say with different degrees of confidence whether some value is an outlier. This suggests there could be a procedure that can be developed to make similar judgments.

For most of the pesticide use data, distributions of rates are not even close to normal. They may have several different peaks (multi-modal). They can have either very broad distributions or very narrow distributions. None of the standard statistical measures of outliers are very useful for these data. The best single method is the one based on neural networks. However, each different criterion will catch different outlier values so it is usually best to use all three criteria. It should be noted that these criteria are not perfect. They are conservative, meaning a value must very extreme to be flagged, and they will miss some errors. On the other hand, they may occasionally flag an extreme value that is actually correct. Because the criteria are conservative these later kinds of errors are minimized.

Criterion 1: Pounds per acre of active ingredient is larger than 200 (for non-fumigants), or 1000 (for fumigants).

These limit values were chosen based on what is known about typical rates of use for most pesticides. Note that this criterion uses the pounds of active ingredient. Also, this criterion only applies to records where the unit treated is acres or square feet. The other criteria use pounds of pesticide product and apply to any unit treated, such as square feet or cubic feet.

Criterion 2: Pounds per unit treated of a product is larger than 50 times the median of all rates with similar types of use.

The median, like the mean (average), is a measure of the location of a set of values and is defined as the value in the set that has an equal number of values above and below it. It was used rather than the mean because it is not as likely to be affected be a few extreme outliers. The median was calculated from the set of all use rates of the same pesticide product and uses as that of each record being examined.

By the same uses, we mean the uses of a product on the same crop or site, same unit treated, and same record type. The record type is either a production agriculture report or a monthly summary report. The type of report is identified in the PUR by the field RECORD_ID. Production agricultural reports have RECORD_ID values of A, B, 1, or 4; monthly summary reports have RECORD_ID values of C or 2.

Criterion 3: Pounds per unit of product is larger than a value generated using a neural network.

A neural network is a function that maps a set of input values to a set of output values. This function has a large number of parameters that must be determined so that the function will give the correct outputs for every possible set of inputs. The values for these parameters are found by a training procedure that involves presenting to the neural network program data consisting of many sets of input and corresponding output values. The program then adjusts the parameters in the neural network function until it produces the correct output values for each input set. Once the neural network has been successfully trained, it can then be used to produce appropriate output values for any input data set provided to it.

The data used to train the neural network used in the PUR outlier program were generated from frequency distributions of the pounds of pesticide product per unit treated for a selected set of pesticide products and sites. Groups of `products` and sites were chosen that included a wide range of types of distributions, including many unusual distributions. Two hundred frequency distributions were plotted. These plots were examined independently by 12 scientists in DPR who marked rates on each plot they thought were outliers.

The results of this survey were summarized by determining an outlier limit for each distribution. The limit was set at a value where all 12 scientists thought higher rates were obvious outliers. These limits were used as the output values for training the neural network. The input values were a set of statistical measures that described the frequency distributions. These sets of input and output values were used to train the neural network. After the neural network was successfully trained, it was used to find the outlier limit for all sets of pesticide use types in the PUR.