FINAL REPORT TO THE CALIFORNIA DEPARTMENT OF FOOD AND AGRICULTURE FOR CONTRACT AGREEMENT NO. 98-0241

PART I. DATA QUALITY OF CALIFORNIA'S PESTICIDE USE REPORT

By
Larry Wilhoit, Minghua Zhang, Lisa Ross

August 2001

STATE OF CALIFORNIA
Environmental Protection Agency
Department of Pesticide Regulation
Pest Management and Licensing Branch
Pest Management Analysis and Planning Program
Sacramento, California 95812-4015

PM 01-VT

EXECUTIVE SUMMARY
Of Report PM 01-XX Entitled
"Final Report to the California Department of Food and Agriculture
For Contract Agreement No. 98-0241
PART I. Data Quality of California's Pesticide Use Report"

Environmental Protection Agency
Department of Pesticide Regulation
Pest Management and Licensing Branch

BACKGROUND

The California Department of Pesticide Regulation's (DPR's) Pesticide Use Report (PUR) is probably the largest and most complete database on pesticide use in the world. A system to collect information on pesticide use in California has been in operation in some form for over 50 years, with the current use reporting system started in 1990. The PUR contains information on nearly all production agricultural pesticide use and some nonagricultural use in California. The data collected include the pesticide product used, the date it was applied, the particular field treated, and application location to a square-mile section. Production agricultural use includes applications to growing crops, agricultural fields, and most applications to forest trees and ornamental turf. Other pesticide uses reported to DPR include post-harvest commodity treatments, right of way, landscapes, structural use, and other nonagricultural uses by commercial applicators.

DPR expanded pesticide use collection in 1990 primarily to more accurately assess dietary risk as well as exposure and potential risk to workers. However, the PUR is also used for a wide variety of environmental and public health purposes. To ensure the accuracy of results from these uses and any regulatory decisions based on such assessments, PUR data must adequately represent actual pesticide use in the state.

Because of the importance of the PUR for many groups and individuals, it is critical that the database be as accurate and complete as possible. People who use the data need to feel confident that the data are sufficiently accurate for their purposes. Bad data are worse than no data. If a system fails to minimize errors, two serious problems will arise. Some people will use the data not knowing about the errors and, therefore, make wrong conclusions. If laws and regulations are based on these conclusions, there is the potential for serious consequences. Other people will not trust the data and therefore ignore it. In either case, the huge effort at data collection is wasted.

PURPOSE

In an effort to minimize errors and improve the quality of the data in the PUR, DPR's Pest Management Analysis and Planning Program entered into a contract with the Office of Pesticide Consultation and Analysis of the California Department of Food and Agriculture (CDFA). Improvement in data quality was focused on four main goals: (1) identify specific PUR data fields and information in need of improvement, (2) assess the quality of spatial data in the PUR, (3) improve error-checking procedures used when loading data into the PUR database, and (4) conduct a survey of county staff to identify the variation in data-field definitions and procedures used for data entry into the PUR. In Part I of this final contract report, the first three goals are described. Goal four will be completed later this year and results related to the acres-planted questions in the survey will be described in Part II.

IDENTIFICATION OF DATA FIELDS AND OTHER INFORMATION

At the onset of the contract, staff from DPR and CDFA met to discuss the data fields most important to CDFA that needed improvement. The most critical data fields identified were acres treated, acres planted, and rate of pesticide use. These needs were incorporated into the contract work as described in more detail below.

In addition to meeting with CDFA staff, DPR also established a committee of various PUR users to identify their concerns about the PUR database as it impacts their assessments. This committee met from September 1999 to May 2000 and discussed many issues related to data quality and produced a list of issues they hoped would be addressed by DPR (Appendix A). This committee also helped organize and plan a conference on the PUR from which an expanded issues list was developed (Appendix A). The conference drew over 200 attendees who listened to seminars on various uses for PUR data. These uses included topics such as PUR data role in policy and regulatory decision-making, human health and environmental quality assessments, exposure and epidemiological studies, economic analyses, and integrated pest management studies. In addition, attendees participated in concurrent sessions designed to reveal problems encountered when using PUR data. Attendance at the conference demonstrated the importance of the PUR database to a wide variety of people and organizations. It also illustrated the need for accurate PUR data given the various links made between pesticide use, environmental impacts, human health assessments, and policy/regulatory decisions.

In response to the conference and the list of issues developed over the prior months, DPR staff developed a work plan for addressing these issues (Appendix B). Updates to the work plan can also be found on DPR's web page at www.cdpr.ca.gov under "Programs and Services". In the coming months and years, DPR hopes to address these various issues.

ASSESSMENT OF DATA QUALITY OF SPATIAL ATTRIBUTES IN THE PUR

Various data fields in the PUR database were assessed for their potential rate of error from 1990 through 1997 (Appendix C). Most of the data fields assessed relate to spatial attributes of pesticide applications. These include county code, meridian/township/range/section (MTRS) designation, grower identification number and site location identification. Also assessed were the data fields for crop planted, area planted, and area treated. In addition, agricultural records (rows in the PUR database pertaining to agricultural use) were checked for potential duplicates. Errors in these data fields could lead to erroneous assessments about the amount of pesticide used and location of use relative to sensitive sites. Errors in these types of data fields could have consequences for community, environmental, and worker health assessments as well as for the regulated community. Therefore, it is important to understand the rate of error found in these data fields so the data can be used with a certain degree of confidence. It is also important when planning for future improvements to the PUR database.

Statewide, potential error rates averaged over the years 1990 to 1997 were about 5% or less of all agricultural records collected (Appendix C). One exception was the potential error rate calculated for acres planted. The statewide error rate averaged 8.1% of all agricultural fields and 17% of all agricultural records collected from the years 1990 to 1997. The specific reason for such a relatively high error rate is not certain but could relate to the inconsistent use of site location identifiers that should uniquely identify each agricultural field.

Examination of potential error rates over the eight-year period statewide and by county, indicated variation temporally and spatially. Three of seven error types indicate a decline in the rate of error averaged statewide. However, duplicate records, MTRS inconsistency, acres planted, and acres treated did not show a similar trend. By county, error rates typically declined over the years with some exceptions. Error-rate details by county and year are provided in Appendix C. The accuracy of any environmental or human health assessment will depend in part on the accuracy of the pesticide use data used for those assessments. Therefore, it is important to be aware of the error rates encountered in the PUR database, and any details on the data field, county, and year, as applicable to the assessment. Appendix C provides those details and could be used by researchers to help assess the degree of accuracy of their work.

THE PUR LOADER AND ERROR HANDLING PROCESSES

In concert with the assessment of error rates in selected data fields of the PUR (Appendix C), the loader program was modified to reduce the number of errors in the database. In addition, DPR recently decided to convert its PUR database from FOCUS to Oracle. Since this involved rewriting the program to load and error check the data, this conversion provided an opportunity to improve the error-checking procedures. We were able to add several new error checks and improve some of the previous checks thanks to the monetary support from CDFA.

The PUR loader program is a computer program that loads data from pesticide use reports into an Oracle database (Appendix D). These data are first entered into databases at each County Agricultural Commissioner (CAC) office and then sent to DPR where the loader program is run. In addition to loading the data, the program also searches for errors, records any errors found, and corrects errors where possible.

During collection and processing of the PUR data, several kinds of error checks occur, both at the CAC offices where the data are entered and during the loading of the data from the county into DPR's PUR database. Ideally, error checking should occur at the time the data are first entered, but until the data collection system is changed, DPR can make improvements to the data received from the county.

The improved loader program carries out several steps (see Appendix D for details):

1. Checks that data from each PUR county data file sent to DPR has not already been loaded.
2. Checks that the new data file has the correct structure and fixes certain kinds of errors.
3. Loads all the data from the new file directly into an Oracle table.
4. Logs the name and date of the county file, number of rows of data entered, and the date loaded.
5. Checks that each record in the new file is not an erroneous duplicate of records already loaded.
6. Checks each data field in this table for a series of possible errors.
7. Corrects errors or makes estimates to replace invalid data where possible, otherwise replaces the value with a blank or leaves it unchanged.
8. Records any uncorrected errors it finds including both the original value and new value, and the kind of error.
9. Records any changes made to the data, the date of the change, and whether it was corrected, estimated, or replaced with a null.
10. Identifies each agricultural field, assigns it a field identification code, determines the most likely acres planted and location of the field, and creates a record of this agricultural field.
11. If the agricultural field has records in the PUR with different values reported for its acres planted or location, the agricultural field is marked as inconsistent and a list of all inconsistent values of acres planted and location is made for this field.
12. Makes some calculations and conversions (such as getting the DPR product identification number from the pesticide registration number and the pounds of active ingredient used from the amount of product used).
13. Loads the valid and converted data into another table.
14. Creates and prints an error report with all the errors found in the PUR data; and
15. E-mails loader log files to the loader administrators; these report the data files successfully loaded, the files not loaded because of errors, and any database or operating system error messages that may have been produced during loading.

The new loader program includes several error-checking routines that were not previously performed. These error checks are listed below and described in more detail in Appendix D:

1. Product identification number.
2. Grower identification number.
3. Inconsistent values for an agricultural field.
4. Duplicate records.
5. High rates of use.

Once various errors are identified and stored in a series of data files, the counties receive a file containing the records in need of data improvement. Staff at the CAC office correct the records, wherever possible, and return the data to DPR. An error-correction program will be developed to enter corrected data into the PUR and will be used for the first time on the 1999 PUR preliminary data. Once this database is corrected, we plan to compare the rate of errors found in 1999 with prior years. This will help us identify where to focus future data improvement efforts.

CONCLUSIONS

Much time, effort, and resources are spent on collecting pesticide use data in California because of the need for such information. Many groups and individuals use PUR data for various human health and environmental assessments. These assessments may lead to policy and/or regulatory decisions that affect the agricultural community as well as others who use pesticides. Because of the importance of the PUR database, it is critical that these data be as accurate and complete as possible.

In our efforts here, we have made an initial attempt to assess the rate of error in certain data fields in the database as well as established an error-checking process to catch as many errors as possible. Although error rates in certain data fields have declined since 1990, these measures are by no means complete and efforts to improve data quality of the PUR are on going. Future efforts, in addition to those outlined in DPR's PUR Improvement Plan, will include assessing improvements made with our new PUR Loader Program and evaluating a county survey used to identify variation in use and definition of data fields in the PUR. Part II of this report will focus on results of the acres-planted questions in the county survey.

TABLE OF CONTENTS

Page