Using PUR Data to Estimate Crop Acreage

Stephan Orme
Pesticide Action Network North America
49 Powell Street, Suite 500, San Francisco, CA 94102

Because the State of California requires that pesticide applications to farmlands be reported, and because pesticides are applied to nearly all farm acreage, the PUR data set contains at least some record of nearly all farm land in California. Each PUR record contains a grower ID, site locator ID, site (commodity) code and the acres planted in that commodity. With this data, farm acreage can be estimated.

Building a Unique Location Identifier

The California Department of Pesticide Regulation (DPR) PUR data user documentation indicates that locations can be uniquely identified by concatenating the grower ID and site locator ID. However, due to variations in how farmers report pesticide use, a more reliable location identifier is a concatenation of the grower id, site locator ID, and site code, a value we call the **location ID**. If each of the component data fields is correctly reported, it is possible to calculate farm acreage for a specific crop. However, since the unique location identifier is made up of three components, errors or changes in any of these three elements results in double-reporting of fields. For example, if grower 5003309, growing almonds (site code 3001) on 20 acres, reports a site locator ID of 'A 1' in one pesticide use report and 'A-1' in a second pesticide use report, this appears as two different locations:

Location #1: 5003309 3001 A 1
Location #2: 5003309 3001 A-1

In similar fashion, errors or changes in the grower ID or site code result in double-reporting of locations. Examining the three location ID elements it is possible to identify some potential problems and solutions.

Grower ID Errors

A Grower ID for a location can change either because of a data entry error or because the field is sold or leased to a new grower. Errors and changes in grower IDs are difficult to detect because there are currently no independent data sources for comparison. In addition, grower IDs are not formatted for error detection. An example of a identification number containing error detection information is the CAS number, which contains a checksum. Incorrect CAS numbers can be easily identified by running a simple math routine. While grower IDs cannot be compared to independent data, they can be compared to data from a previous year. Grower IDs that only occur once are likely to be data entry errors. In a single year, the PUR data contain records from approximately 28,000 growers. For the 8-year period from 1991 through 1998, the PUR data contain 61,357 unique grower IDs. Of these, 4,958 have only one entry in the PUR dataset.

It is also possible to note the sale or lease of a field by examining the PUR data. For example, if a 55–acre field of grapes is listed under one grower ID in a selected MTRS

block (the MTRS block is a geographic identifier – one MTRS block is one square mile) for several years, then that grower ID disappears and a second field of 55 acres of grapes appears under a new grower ID, it is a reasonable assumption that the field has been sold. In the year that the field was sold, it will appear twice in the PUR data, once under the old owner and a second time under the new owner.

Site Code Errors

Errors in site codes are less common than other errors because there are only a limited number of site codes in use (typically 220 in a given year). The small number makes it difficult to accidentally enter an incorrect value. A notable exception to this is the incorrect use of a site code to identify a crop. For example, there are two site codes for tomatoes in the PUR data, one for fresh market tomatoes and a second for processing tomatoes. In the early years of PUR reporting, the distinction between these two types of tomatoes was not clear, and many growers used incorrect site codes.

Site Locator ID Errors

Site locator IDs are the most unreliable part of the location identifier in the PUR data. Because site locator IDs are not consistently formatted, it is very common for site locator IDs to be entered in several ways. One solution is to strip out common mistakes. In our data we remove all spaces, dashes, and hyphens in the site locator ID field. In addition, we convert all G's to 6's, Z's to 2's, and D's to 0's. This data cleanup technique was developed by Professor Lynn Epstein at UC Davis.

Using the Location ID to Calculate Crop Acreage

Once grower IDs, site codes, and the site locator IDs are corrected, these three fields can be concatenated to create a location ID. Each year, there are approximately 180,000 unique locations reported in the PUR data. Total acreage by crop can be calculated by grouping location IDs, removing duplicates, then summing the "acres planted" field.

While this technique works, it does not address errors in the "acres planted" data field. If the "acres planted" are reported correctly, the "acres planted" values in all records of pesticide use for that location should agree. In practice, about 96% of all records within a location set have a consistent "acres planted" value. To determine "acres planted" for the remaining set of records for that location, we calculate the mode (the most common value) of "acres planted" for the location set and use this value for all acreage calculations.

As a final step, we flag all location IDs with inconsistent "acres planted," all locations that are reported only once (this is a technique developed more extensively for use with a GIS spatial identifier by Minghua Zhang's group at UC Davis), and locations with unreasonably large acreage for the crop listed. These records are then assessed on an individual basis by comparing them to data from previous years and to location descriptions provided by the county agricultural commissioner offices.

Calculating acreage using these techniques works for crops that have one crop per year and are planted and harvested in the same year. When multiple crops of the same

commodity are grown in one year, it appears as if just a single crop was grown. As a result, this technique under-reports acreage. For crops with a growing season that spans two calendar years, "acres planted" calculated from the PUR data is generally higher than actual acreage, thus over-reporting acreage. Both of these problems can only be addressed by knowing the planting and harvest dates for each crop, data that are not currently collected under the PUR system. Fortunately, most crops are grown once per season and are planted and harvested in the same year. Some notable exceptions are carrots, garlic, cauliflower, celery, cabbage, lettuce (head and leaf), broccoli, onions, asparagus, strawberries, and spinach.