Understanding the PUR – the good, the bad, and the ugly

Larry Wilhoit and Minghua Zhang
Department of Pesticide Regulation
Sacramento, CA 95814

Topics of Discussion

- How to get the PUR data
- The data and issues to watch out for (the ugly)
- Technical Tools (the good)
- Dealing with errors in the data (the bad)
- Conclusion

How to get the PUR data

- The PUR annual report (http://www.cdpr.ca.gov/docs/pur/purmain.htm)
- The full database on CD-ROM as fixed field or comma-delimited text files for each county
- Full database in one tab-delimited file per year
- Oracle export files

How to get the PUR data

- MKTBLDR: commercial access www.mktbldr.com
- Interactive web queries:
  - DPR: calpip.cdpr.ca.gov/cfdocs/calpip/prod/main.cfm
  - UCIPM: www.ipm.ucdavis.edu/PUSE/puse1.html
  - PAN: www.pesticideinfo.org/Search_Use.jsp

The Data

- PUR vs. UDC tables
  - PUR: one record per product application
  - UDC: one record per active ingredient application
- **Use_no**: uniquely identifies each record in the PUR

The Data

- **Record_id**: represents 3 things:
  - Whether the record is production agricultural (values 1, 4, A, B, E, F) or other kind of application (2, C, G)
  - Whether report was a 7-day (1, A, E) or monthly (4, B, F) production agricultural report
  - Whether data was entered by county staff (A, B, C), DPR staff (1, 2, 4), or Prison Industries Authority (E, F, G)

The Data

- **Prodno** uniquely identifies the pesticide product used but sometimes use report does not specify exact product
- **Chem_code** uniquely identifies the active ingredient in the pesticide product and can be very specific
- **Lbs_prd_used** and **lbs_chm_used** are the pounds of product and pounds of AI used
- **Site_code** indicates the target site to which a pesticide product was applied

The Data

- **Acre_treated** is not always acres; it is the area or other measure that was treated: need to check unit_treated
- **Acre_planted** is the size of the planted field: need to check unit_planted
- **Applic_cnt** is not very useful: for ag applications it is always one and for other applications a value is not always given

The Data

- The six **fields base_ln_mer, township, tship_dir, range, range_dir, section** indicate the geographic location
- **Grower_id** does not always mean grower: it identifies the "operator" of the field as well as other things
- **Site_loc_id** identifies distinct agricultural fields, but is used by growers who they want

Technical Tools

- Spreadsheets: Excel
  - Calculate basic statistics such as subtotals, means, variances
  - Make many kinds of graphs
  - Pivot tables and graphs
  - Limited to 65,536 rows
  - Limited querying of data

Technical Tools

- Databases
  - Oracle: powerful but expensive and difficult to use
  - Access, FileMaker Pro: easier to use but limitations
  - Open source: MySQL, PostGreSQL: some limitations but cheap
- ODBC: Open Database Connectivity
  - Provides access from any ODBC compliant application to data in most database systems

Technical Tools

- SQL: language for querying databases
  - Simple to use but powerful
  - For example, get the total pounds of each AI

```
SELECT    chem_code, SUM(lbs_chm_used)
FROM      udc2002
GROUP BY  chem_code;
```

Technical Tools

- SQL: language for querying databases
  - Easy to make mistakes
  - Dealing with NULL values can be tricky

```
SELECT      chem_code, SUM(lbs_chm_used),
            SUM(acre_treated), unit_treated
FROM        udc2002
WHERE       unit_treated != 'A'
GROUP BY    chem_code, unit_treated;
```

Technical Tools

- The first query will not return uses where unit is NULL; the second will find NULLs

```
SELECT      chem_code, SUM(lbs_chm_used),
            SUM(acre_treated), unit_treated
FROM        udc2002
WHERE       unit_treated != 'A'
GROUP BY    chem_code, unit_treated;

SELECT      chem_code, SUM(lbs_chm_used),
            SUM(acre_treated), unit_treated
FROM        udc2002
WHERE       unit_treated != 'A' or unit_treated IS NULL
GROUP BY    chem_code, unit_treated;
```

Technical Tools

- Procedural languages needed to extend capabilities of SQL
- Statistics: SAS
  - SQL does not have many functions for doing statistical analyses
  - SAS has additional data handling capabilities

Technical Tools

- SAS: Find the median rates of use, number of records, and percentiles for each AI and year:

```
PROC MEANS NOPRINT NWAY DATA = ai_rates;
      CLASS chem_code year;
      VAR rate;
      OUTPUT OUT = ai_year_stats
            N(rate) = num_recs
            MEDIAN(rate) = med_rate
            MIN(rate) = min_rate
            MAX(rate) = max_rate
            P10(rate) = p10_rate
            P90(rate) = p90_rate;
```

Technical Tools

- SAS: PROC TABULATE: a more powerful pivot table

```
PROC TABULATE DATA = almond.ops_sac_valley FORMAT=comma7.;
      CLASS chemname crop year;
      VAR lbs_ai;
      WHERE crop in ('ASPARAGUS', 'ORANGE')
            and year > 1995;
      TABLE crop*chemname, SUM*lbs_ai*year / rts=22;
```

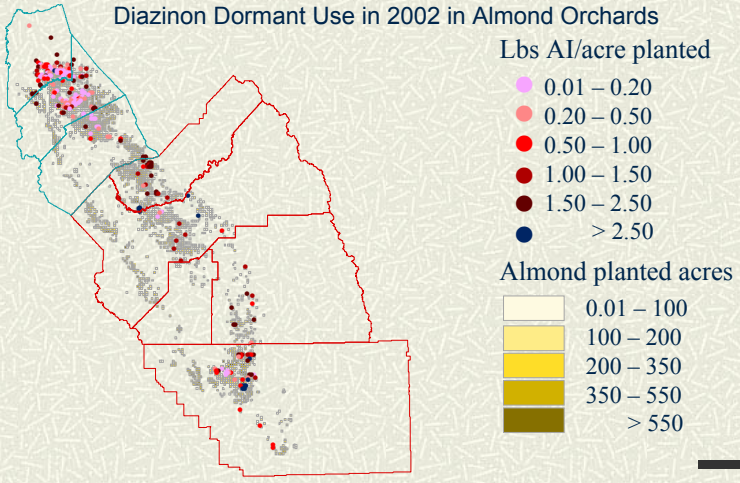
Technical Tools

- SAS: PROC TABULATE

		Sum					
		LBS_AI					
		YEAR					
		1996	1997	1998	1999	2000	2001
CROP	CHEMNAME						
ASPARAGUS	CHLORPYRI-FOS	4	82	16	22	690	15
	DISULFOTON	1,788	1,217	1,347	1,130	1,984	1,691
	FONOFOS	808	71	114	.	.	.
ORANGE	CHLORPYRI-FOS	1,106	832	735	763	365	346
	DIAZINON	.	3	1	.	2	1
	DIMETHOATE	773	510	111	97	111	1
	MALATHION	106	94	296	77	84	91
	METHIDATH-ION	62	50	3	1	.	11

Technical Tools

GIS: Geographic Information Systems



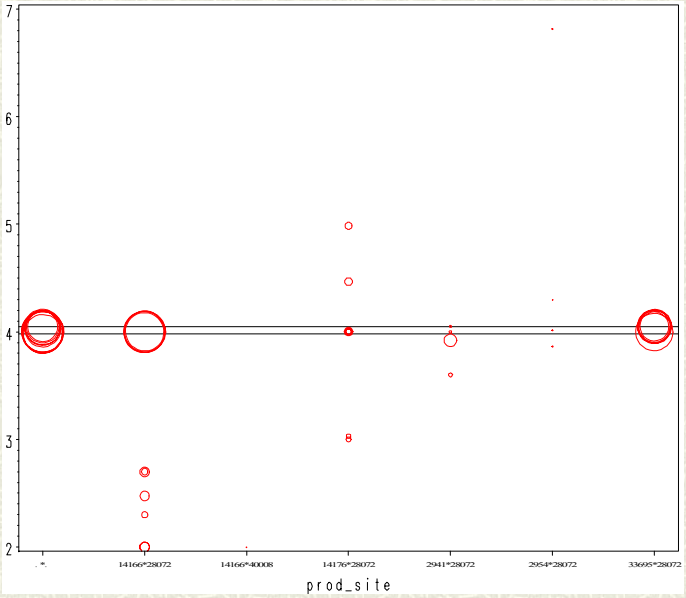
Dealing with errors in the data

- When analyzing data, it is important to check for possible errors
- The loader program looks for possible errors and enters them in an errors table.
- In the PUR the erroneous value is either left as it is, or replaced with an estimate, a blank, a “-1”, or a “?”

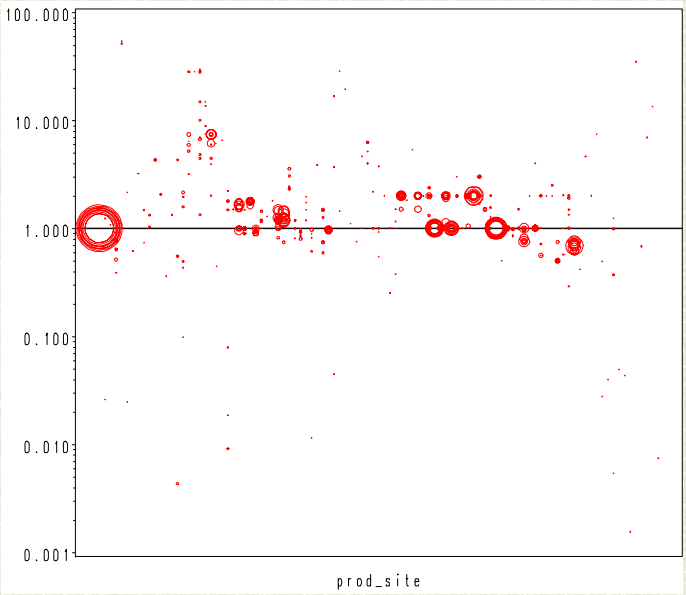
Dealing with errors in the data

- Errors in pounds of pesticide can be especially important
- A rates of use is flagged if it is greater than the median rate of use in other records with the same product and crop.
- However, there may be few records with this product and crop.
- Maybe better to get median rate from all records with the same AI.

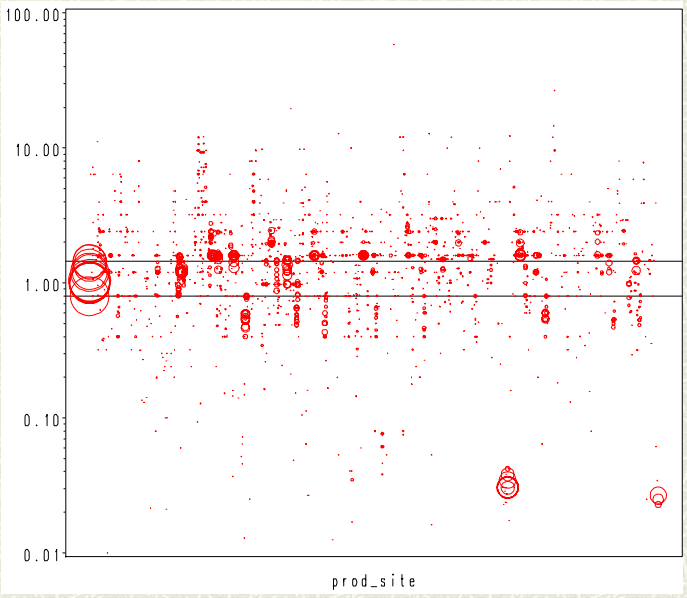
Molinate median rates by product and site



Disulfoton median rates by product and site



Diuron median rates by product and site



Conclusions

- There are issues that make analyzing the PUR confusing:
  - Poor database design
  - Confusing regulations
  - Errors, some of which can difficult to find
- There are many tools that can help
  - Databases, SQL, ODBC
  - Other procedural computer languages
  - Spreadsheets such as Excel
  - Statistics packages such as SAS