Errors and Oddities in total PUR Crop Acreages

Christopher DeMars
Minghua Zhang

UC Davis AGIS Lab
Errors and Oddities in total PUR Crop Acreages

Christopher DeMars
Minghua Zhang

UC Davis AGIS Lab
Problem

- The PUR database is very accurate if the question is “how much of what was applied when?”
- The database can be very inaccurate for a given crop if the question is “what is grown where?”
PUR Crop Accuracy is important because...

- PUR database is one of the only datasets that has yearly, statewide crop distribution at a sub-county spatial resolution.
  - It is the only such dataset that gets data directly from farmers
- Sub-county resolution is needed:
  - To determine risk to waterways, you need to know how close the application was to a waterway.
  - To predict possible contributors to measured surface water pollution, you need to know who applied that chemical upstream.
  - Statistical correlation between crop type and/or yield and environmental factors like soil, weather or DWR deliveries requires as much spatial information as possible.
  - All hydrologic modeling of California is dependent on a good crop layer.
Data Sources

- **PUR database:**
  ftp://pestreg.cdpr.ca.gov/pub/outgoing/pur_archives/

- **PUR Fields Layer**
  - created by Larry Wilhoit

- **County Agricultural Commissioners' Crop Reports:**

- **California Strawberry Commission:**
  http://www.calstrawberry.com/commission/asurvey.asp

- **DWR Land Use Surveys:**
  http://www.water.ca.gov/landwateruse/lusrvymain.cfm
Identifying Individual Fields

- Naïve method to identify individual fields:
  - select distinct year, county_cd, grower_id, license_no, site_loc_id, site_code, comtrs, acre_planted
  - License_no and comtrs can be ignored with little change.

- Problems:
  - Sometimes reported acreage changes during the season.
  - Reporting errors can count the same field twice: changing site_loc_id or grower_id, using different comtrs, etc
PUR-Fields by Larry Wilhoit

- Larry Wilhoit developed a SQL script to overcome many of the duplicate field problems.
  - Cleans grower_id & site_loc_id
  - Converts non-acre units to acres
  - Tries to identify distinct fields and their MTRS location
  - Collects and organizes multiple reported field sizes into new database fields
PUR-Fields MTRS detection

- If reported MTRS is within reported county:
  - If this is the only valid MTRS reported for this field:
    - Set MTRS to reported MTRS
  - Else if there are other valid MTRS reported, the MTRS reported most often is set as that field's MTRS
- Else
  - if there is another record using this field that has a valid MTRS in the reported county, set the field's MTRS to it
  - else if there is another record using this field that shares an MTRS and has a valid County/MTRS, set the field's county to that valid County/MTRS pair
  - else if there is another record using this field within the same county, set the field's MTRS to that
  - else create new field_id for field, keep county and MTRS, but tag MTRS as incorrect
Many historical anomalies are smoothed out.

Unfortunately, there is still a large discrepancy between the PUR Fields layer and the county crop report or strawberry commission.

This discrepancy is getting worse with time.

The error must be something fundamental with PUR, and not a trivial error.
Not all crops are problematic

- Most of the crops in the PUR database have very accurate totals.
- Some crop acre totals are getting more accurate over time.
Statewide, what crops acre totals are getting less accurate over time?

- For each crop, for each year find the absolute value of the difference between PUR-Field total and the Crop Report
- Fit a line to these residuals
- Plot the slope of that line:
  - (+) = getting worse
  - (-) = getting better
While the gap between PUR or PUR-Fields totals and the crop report is increasing slightly over time, it looks like the difference is decreasing in recent years.

Even so, given the stable nature of vineyards, the difference should be minuscule.
Lettuce & Spinach

**Statewide Lettuce Acreage**
- **PUR naive**
- **PUR fields**
- **Crop Report**

**Statewide Spinach Acreage**
- **PUR naive**
- **PUR fields**
- **Crop Report**
Strawberries & Broccoli

Statewide Strawberry Acreage

Statewide Broccoli Acreage

- PUR naive
- Crop Report
- PUR fields
- Strawberry Comm.
Spatially, where are the errors?

Top five counties per crop in 2010, using Crop Report totals
County: Strawberries & Broccoli
What's up with Monterey?

- Of all the crops whose state totals diverge significantly from county crop report totals, Monterey has the largest error. And it is getting worse.
- Start with a visual comparison: using spatial crop data from 2006 acquired directly from Monterey County, compare those field locations with what sections PUR-Field says those crops are grown in.
Spinach
Strawberries
Possible Error Sources

- Unlike the central valley, most fields in the Salinas valley don't follow section boundaries.
  - Could result in the same field being reported in two different sections. But the PUR-Fields dataset should clean that up by choosing the most common reported MTRS.
- There are many ghost sections: MTRS sections that PUR claims contain a given crop but which lack a field in the county supplied spatial dataset.
Error sources continued...

- Sometimes an error is made when entering in site_loc_id

<table>
<thead>
<tr>
<th>operator_id</th>
<th>site_loc_id_fix</th>
<th>mtrs</th>
<th>acres_planted</th>
</tr>
</thead>
<tbody>
<tr>
<td>270P048</td>
<td>IOOOI</td>
<td>M13S02E05</td>
<td>8</td>
</tr>
<tr>
<td>270P048</td>
<td>IOOOOI</td>
<td>M13S02E05</td>
<td>8</td>
</tr>
</tbody>
</table>

- Some site_loc_id seem to refer to subsets of a larger field that is previously listed

<table>
<thead>
<tr>
<th>operator_id</th>
<th>site_loc_id_fix</th>
<th>mtrs</th>
<th>acres_planted</th>
</tr>
</thead>
<tbody>
<tr>
<td>270P043</td>
<td>600I2</td>
<td>M12S02E10</td>
<td>5.9</td>
</tr>
<tr>
<td>270P043</td>
<td>600I3</td>
<td>M12S02E10</td>
<td>9.8</td>
</tr>
<tr>
<td>270P043</td>
<td>600I4</td>
<td>M12S02E10</td>
<td>9.9</td>
</tr>
<tr>
<td>270P043</td>
<td>600II</td>
<td>M12S02E10</td>
<td>3.5</td>
</tr>
<tr>
<td>270P043</td>
<td>6000I</td>
<td>M12S02E10</td>
<td>31</td>
</tr>
<tr>
<td>270P043</td>
<td>70002</td>
<td>M12S03E18</td>
<td>5.6</td>
</tr>
<tr>
<td>270P043</td>
<td>70003</td>
<td>M12S03E18</td>
<td>16.9</td>
</tr>
<tr>
<td>270P043</td>
<td>7000I</td>
<td>M12S03E18</td>
<td>50</td>
</tr>
</tbody>
</table>
Conclusion

- The PUR database can be a tremendous resource for data mining to solve environmental, agricultural and economic questions.
- To be more effective, the PUR database needs accurate spatial data and field identification.
- There should be a standardized way of recording site_loc_id:
  - Only use alphanumeric characters
  - Never designate a subfield using a different site_loc_id. Either:
    - Use the original site_loc_id for all entries and when subfields are treated, have acres_treated less than acres_planted
    - Give all subfields a unique site_loc_id, but do not enter the unity of subfields as its own site_loc_id
Conclusion continued...

- Many of the existing errors in the database could be identified and fixed by extending Larry Wilhoit's field script.
- Future errors can be avoided by using a `site_loc_id` naming ruleset common to all counties.
  - This could be automated in the new online data entry app: error checking at the point of entry against previous entries could raise an exception to the user asking if they are really referring to a field already in the system.
In an ideal world....

- site_loc_id would be a statewide unique ID
- A separate table would associate the unique ID to a field boundary, or at least a lat/lon centroid point.