Use of Record Linkage to build a Brazilian List Frame for Agricultural Statistics

Andrea Diniz da Silva
Programa de Pós-graduação em População, Território e Estatísticas Públicas
Use of Record Linkage to build a Brazilian List Frame for Agricultural Statistics

Andrea, Diniz da Silva  
National School of Statistical Science – ENCE/IBGE  
Rua Andre Cavalcanti, 106  
Rio de Janeiro, Brazil  
andrea.diniz100@gmail.com

Flavio, Pinto Bolliger  
The Food and Agriculture Organization of the United Nations – FAO  
Rome, Italy  
flavio.bolliger@fao.org

Jose Andre, de Moura Brito  
National School of Statistical Science – ENCE/IBGE  
Rua Andre Cavalcanti, 106  
Rio de Janeiro, Brazil  
jambrito@gmail.com
Summary

1. Context

2. Building a Brazilian List Frame
   2.1 Data Sources
   2.2 Available information
   2.3 Record Linkage Strategy
   2.4 Record Linkage Methods

3. Discussion
1. Context

- Most of agricultural statistics are produced by IBGE, under National Statistical System.

- Census since 1920 and subjective surveys conducted by IBGE since 1938 (before it was produced by the Ministry of Agriculture).

- Decennial census and subjective surveys are on the pillar of official agricultural statistics.

- To run away from this model a probabilistic surveys system is on the way.

- Problem: it still misses an important component: Master Sampling Frame.
1. Context

- Many available frames allow using *Multiple Frame Approach*.
  - Area: IBGE
  - List: IBGE, State Tax Files, Ministry of Agrarian Development.

- Many Record Linkage methods: none of them is considered best regardless data.

- Comparison of methods on real data is necessary.
2. Building a Brazilian List Frame

• An extension of previous work (2012)
  • More than 50 thousand matches, linking census of agriculture, address file of population census and business register.

• Addition of sources
  • Five sources: IBGE: 2 survey and 1 register; State Tax Administration: 1 register; and Ministry of Agrarian Development: 1 register.

• Other methods: tree based and SVM methods.
### 2.1 Data Sources

**Table 1: Number of records available for linkage by source**

<table>
<thead>
<tr>
<th>State</th>
<th>Source</th>
<th>IBGE</th>
<th>State</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Census of Agriculture (2006)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ceará</td>
<td>381,017</td>
<td>84,954</td>
<td>979</td>
<td>509</td>
</tr>
<tr>
<td>Mato Grosso do Sul</td>
<td>64,864</td>
<td>68,453</td>
<td>1,262</td>
<td>123,104</td>
</tr>
<tr>
<td>Paraíba</td>
<td>167,286</td>
<td>88,092</td>
<td>298</td>
<td>649</td>
</tr>
<tr>
<td>Maranhão</td>
<td>287,039</td>
<td>35,761</td>
<td>647</td>
<td>13,475</td>
</tr>
<tr>
<td>Santa Catarina</td>
<td>170,913</td>
<td>150,043</td>
<td>1,803</td>
<td>92,338</td>
</tr>
</tbody>
</table>

- Ag. Census has better coverage than Pop. Census.
- Business register is poor for agriculture.
- Tax file: may have office address (other State).
- MDA shows applications (more than one for same unit).
2.2 Available information

<table>
<thead>
<tr>
<th>Variable</th>
<th>State Tax File</th>
<th>IBGE</th>
<th>MDA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Ceara</td>
<td>Mato Grosso do Sul</td>
<td>Paraíba</td>
</tr>
<tr>
<td>Name</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>ID on Tax File*</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Activity Code</td>
<td>x</td>
<td>-</td>
<td>x</td>
</tr>
<tr>
<td>Activity Description</td>
<td>x</td>
<td>-</td>
<td>x</td>
</tr>
<tr>
<td>Address (street, number etc.)</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Owners: name</td>
<td>x</td>
<td>-</td>
<td>x</td>
</tr>
<tr>
<td>Owners: ID on Tax File</td>
<td>x</td>
<td>-</td>
<td>x</td>
</tr>
<tr>
<td>Owners: address</td>
<td>x</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

* CPF ou CNPJ

- Info available include basically names and address.
- Sources hold much more information but there is confidentiality issues.
2.3 Record Linkage Strategy

- Three independent procedures:
  - Register <-> register.
  - Survey <-> survey.
  - Survey <-> register.
2.4 Record Linkage Methods and Tools

- Two classes of theoretical model.
  - Stochastic: FS decision model.
  - Non-stochastic: Tree based and SVM.
- Blocking/ Indexing: municipalities.
- String comparator: Jaro-Winkler.
- R programming language.
3. Discussion

- Use of regular government sources: may ensure maintenance but may not be that easy obtaining.

- A number of method for record linkage are now available: a compromise has to be done based on the conditions of record linkage project.

- Extension of (ad hoc) model: other years, other regions, other sources.
THANK YOU