Exploring Sampling Techniques to Reduce Respondent Burden

Wei, Yijun(Frank)
USDA-NASS
Yijun.Wei@nass.usda.gov

Bejleri, Valbona
USDA-NASS
Valbona.Bejleri@nass.usda.gov
Outline

• Purpose of research
• Sampling procedures
• Coordination function
• Three simulation studies
• The results of simulations
• Some concluding remarks
National Agricultural Statistics Service (NASS)

• **Agricultural Estimates**
  - Conducts more than 100 surveys annually
  - Produces more than 400 reports
  - Publishes 7 federal principal economic indicators
  - Provide information for the commodity markets
  - Tight timelines

• **Samples are drawn prior to start of growing season**

• **Desire to spread the response burden to the extent possible**
  - Very large farms must be included in sample to get precise estimates
  - Potential to spread among other farms
Purpose of Research

Exploring for a sampling design that will allow:

- Optimal coordination of surveys
  - Small respondent burden (small number of appearances of an operation across sample surveys)

- Efficient estimators
  - Consistent
  - Unbiased
  - Efficient with respect to the variance

- Fixed sample size

- Simple implementation
Sequential Interval Poisson Sampling (SIP) at NASS

• Employed in Agricultural Resource Management Survey (ARMS)

• Controlling overlap between ARMS from previous year and Crop APS sample for the current year

• Poisson sampling is used with Probability Proportional to Size (PPS) scheme (Ohlsson, 1992)

• Poisson sampling yields fixed sampling fraction but not a fixed sample size

• Each element of population may have different probability of being included in the sample
Coordination Function

**Purpose**: Spreading respondent burden among multiple samples

**Steps**
1. Select sample $S_1$ using Permanent Random Numbers $U \sim Unif(0,1)$
2. For each chosen unit $k$, compute cumulative respondent burden $\Gamma_{k,t}(w)$ as a function of the number of times a unit $k$ is selected to participate (appears) in sample 1 through $t$
3. Use cumulative respondent burden as a criteria to construct coordination function

$$g_{k,t}(w_k) = \Gamma_{k,t}(w_k) + \sum_i 1_{A_i}(w_k) \int 1_{A_i \cap [0,w_k]} u \, du$$

where $w_k$ is the random number for unit $k$
Coordination Function (Continued)

Steps
4. Update the random number for each unit to the current value of the coordination function $g_{k,t}(w_k)$
5. Select a unit based on its “new” updated random number
6. Repeat $n$ times steps 1 – 5 to select samples $S_1, S_2, ..., S_n$
Previous Studies

• First study:
  o Simulated population of 100
  o Sampling rate is 25% for each of 10 samples
  o Coordination function led to reduced respondent burden compared to SRS, PPS, or a combination of SRS and PPS

• Second study:
  o NASS Agricultural Yield Row Crops, Agricultural Yield Small Grains, and Crop APS Survey data
  o Sampling rate is about 10% for each of 3 samples
  o Coordination function led to marginal reduction in respondent burden compared to SIP
Third Study

• Simulated farm population: 100,000 farms
  • Farm simulated data
    o ID: farm number
    o Farm size (continuous) in acreage
    o Farm size category (12 categories)
    o Farm type: crop or livestock
  • Frequencies proportional to 2012 Census of Agriculture data frequencies
    o Stratification: Farm size category x Farm type
  • Acreage (continuous)
    o Random Uniform [min acreage, max acreage] within stratum (except for largest size category)
    o For largest size category: Allow for long-tail distribution of size
Third Study

• Nine survey sequences with varied sampling fractions
• 3 sampling schemes
  o No stratification
  o 1-way stratification by Farm size (categorical)
  o 2-way stratification by Farm size x Farm type
• Sampling approaches to compare:
  o SIP
  o Coordination function
• 200 runs for each configuration
• Units’ number of appearances in series of samples is reported
Table 3: Sampling rate (percentage of total population)

<table>
<thead>
<tr>
<th>Sample</th>
<th>First Simulation</th>
<th>Second Simulation</th>
<th>Third Simulation</th>
<th>Fourth Simulation</th>
<th>Fifth Simulation</th>
<th>Sixth Simulation</th>
<th>Seventh Simulation</th>
<th>Eighth Simulation</th>
<th>Ninth Simulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st</td>
<td>25</td>
<td>75</td>
<td>75</td>
<td>90</td>
<td>90</td>
<td>25</td>
<td>10</td>
<td>25</td>
<td>10</td>
</tr>
<tr>
<td>2nd</td>
<td>25</td>
<td>25</td>
<td>25</td>
<td>25</td>
<td>25</td>
<td>10</td>
<td>25</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>3rd</td>
<td>25</td>
<td>25</td>
<td>25</td>
<td>25</td>
<td>10</td>
<td>25</td>
<td>10</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>4th</td>
<td>25</td>
<td>25</td>
<td>25</td>
<td>25</td>
<td>25</td>
<td>10</td>
<td>25</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>5th</td>
<td>25</td>
<td>25</td>
<td>25</td>
<td>25</td>
<td>25</td>
<td>10</td>
<td>25</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>7th</td>
<td>25</td>
<td>25</td>
<td>10</td>
<td>25</td>
<td>10</td>
<td>25</td>
<td>25</td>
<td>25</td>
<td>25</td>
</tr>
<tr>
<td>8th</td>
<td>25</td>
<td>25</td>
<td>10</td>
<td>25</td>
<td>10</td>
<td>25</td>
<td>25</td>
<td>25</td>
<td>25</td>
</tr>
<tr>
<td>10th</td>
<td>25</td>
<td>25</td>
<td>25</td>
<td>10</td>
<td>75</td>
<td>75</td>
<td>90</td>
<td>90</td>
<td></td>
</tr>
</tbody>
</table>
Difference of proportion (CF% – SIP%) using SIP and Coordination function. No stratification
Difference of proportion (CF% – SIP%) using SIP and Coordination function. One way stratification by size
Difference of proportion (CF% – SIP%) using SIP and Coordination function. Two-way stratification by size and type.
Summary

• Third study
  • Sampling rate varies for 10 scenarios and for three different sampling schemes
  • With no stratification: Coordination function led to higher respondent burden compared to SIP
  • With 1-way and 2-ways stratification: Coordination function led to reduced respondent burden compared to SIP

• As sampling rate increases, respondent burden over multiple samples increases
• Coordination function is more effective at reducing respondent burden among stratified samples, as sampling rate increases
• We have to think about when to use coordination function
• This work is preliminary—more studies are needed
Acknowledgements

This work has been done under Strategic Optimization Sample Selection Team of Response Rate Research Team at USDA-NASS

Team members: Wendy Barboza, Franklin Duan, Jonathan Lisc, Brian Richards, Shareefah Williams, Valbona Bejleri, Yijun Wei

Nathan Cruze helped us with the graph
References


Guggemos, Fabien, and Olivier Sautory (2012). "Sampling coordination of business surveys conducted by Insee." In proceedings of the Fourth International Conference of Establishment Surveys.
References


Ohlsson, E. (1992). SAMU, The system for Co-ordination of Samples from the Business Register at Statistics Sweden. 18


Any Questions?

Thank you!