Evaluating the Use of Web-Scraped List Frames to Assess Undercoverage in Surveys: Lessons from Local Foods Marketing

Linda J. Young
National Agricultural Statistics Service
March 7, 2018
2015 Local Foods Marketing Survey

• Fall 2015: NASS became aware of the need to conduct a local foods study with results to be published in 2016

• 2015 was chosen as the reference year

• Publication was slated for December, 2016
Key Definition for Local Foods

Operation: A farm having at least $1,000 in sales or potential sales, which in 2015 produced and sold food for humans to eat or drink directly to
» consumers
» retail markets
» institutions
» intermediary businesses marketing the food as being locally produced
Available History on Local Foods

• 2012 Census of Agriculture
  – 144,530 Local Foods Farms
  – $1,309,827,000 in Sales

• 2007 Census of Agriculture
  – 136,817 Local Foods Farms
  – $1,211,270,000 in Sales
Emerging Sectors of Agriculture

- Emerging sectors
  - Urban agriculture
  - Organics
  - Horticulture
  - Local Foods

- These tend to be
  - Smaller
  - More diverse
  - More transient
  - More dispersed than the more traditional farms in rural areas

- Hard to Quantify
Challenges with Identifying Small Farms

• NASS list frame
  – List of all known farms and potential farms
  – Known to be incomplete, especially for small farms
  – In 2012 Census of Agriculture, a 12.3% adjustment in the number of farms was due to undercoverage
Challenges with Identifying Small Farms

• Need to be able to assess undercoverage on NASS list frame
  – Sampling from NASS area frame not cost-effective when farms are dispersed
  – June Agricultural Survey (JAS) sample from NASS area frame—Insufficient number of small farms
  – Need a new approach

Big Idea: Create an independent list frame using web scraping
A Web-Scraped List Frame for the Local Foods Marketing Practices Survey

• Create a web-scraped list frame of all US local foods farms
  – Farm Name
  – Farm Type (Crops, Livestock, Poultry)
  – Farm Address
  – Farm State
  – Farm Latitude
  – Farm Longitude
  – POC Name
  – POC Address
  – POC State
  – POC Phone
  – POC E-mail
Timeline for Creating the Web-Scraped List

Consequence: Incomplete harvesting of potential open source data
Capture-Recapture: The Big Idea

How many bass are in your pond?

• Catch some bass (say 100)
• Tag each one and return to pond
• Next day catch some more (say 50, 25 are tagged)
• Half in second group have a tag so estimate half in pond have a tag

\[
\frac{25}{50} = \frac{100}{N}
\]

• Solve to find \( N = 200 \)
List Frames Available for the Survey

• 2,007,110 on NASS List Frame
  – Includes all (not just local foods) operations
  – Consists both of confirmed farms and potential local foods farms

• 33,394 on Web-Scraped List Frame, which only has potential local foods
  – Are not confirmed to be farms
  – In urban ag pilot study about half had agricultural activity
Local Foods Sampling Design

- NASS list – Stratified Sample Design (24,907)
  - Four groups
    - A: Census and Organic respondents + Value of Sales for food
    - B: Local Foods indicator – No Value of Sales
    - C: Potential local foods entities
    - D: All others – stratified by likelihood of local foods
  - Sample Allocation: Target CV’s (Value of Sales)
    - US level  2.0 – 3.0
    - Regional   8.0 – 10.0
    - State Level 10.0 – 12.0

- Web-Scraped (WS) list – Systematic Sample (19,365)
  - Ordered by state and web-scraped farm type

1,466 records were in both NASS and Web-Scraped list samples
Primary Assumptions for Analysis

- Two Independent Samples:
  - NASS List Frame
  - Web-Scraped List Frame
- Proportion of web-scraped local foods farms captured in the NASS list frame sample is equal to the proportion of the US local foods farms captured by the NASS list frame sample

U.S. Agricultural Operations

Sampled Units on NASS List Frame

Sampled Units on Both NASS List Frame and Web-Scraped Samples

Local Farms

Sampled Units on Web-Scraped List

23,441

17,899

1,466
Signal of a Challenge Ahead

Response Rates

A: Census and Organic respondents with Value of Sales for food
B: Local Foods indicator – No Value of Sales

In-Scope Rates

C: Potential local foods entities
D: All others – stratified by likelihood of local foods
Responding Local Foods Farms for Capture-Recapture

U.S. Agricultural Operations

- Responded as Local Farms and in NASS List Frame Sample
- Local Farms and in Both NASS List Frame and Web-Scraped Samples
- Local Farms in Web-Scraped Sample

5,641
4,373
368

15
Operations Selling Directly: Count and Sales Through all Marketing Channels, 2015

- 167,009 ± 5845 operations used direct marketing practices to sell food in the US.

- $8,747 million ± $892 million of food was sold through direct marketing practices, including value-added products at the first point of sale.
  - $4.8 billion were direct food sales of raw commodities.
  - $3.9 billion were food sales of value-added commodities.
Local Food Marketing Practices
Publication Levels

Levels of Publication: US, Regional, and 30 States

Count of Published items by level

<table>
<thead>
<tr>
<th>Level</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>US</td>
<td>393</td>
</tr>
<tr>
<td>7 Regions</td>
<td>33</td>
</tr>
<tr>
<td>30 States</td>
<td>15</td>
</tr>
</tbody>
</table>

=States with published data.
A Closer Look at the Assumptions

• The population is closed (no “births” or “deaths” during the time between the two samples)
  – Samples collected during the same timeframe

• The two lists are independent
  – Web sources used in developing the NASS list frame
  – Lack of independence introduces bias

• All farms are equally likely to be captured in each sample
  – Tried to control for this using logistic regression or by forming categories
  – Heterogeneity tends to cause downward bias
A Closer Look at the Assumptions

• Capturing a farm in one sample does not affect its catchability in the other sample
  – Operations in both samples only receive one questionnaire

• Farms caught in the first sample can be identified if they are caught in the second sample
  – Assumes perfect record matching
Discussion: List Comparison

Response Rates

In-Scope Rates

A: Census and Organic respondents with Value of Sales for food
B: Local Foods indicator – No Value of Sales
C: Potential local foods entities
D: All others – stratified by likelihood of local foods
Discussion

• Web scraping for list building
  – More thorough web scraping
  – Prescreening to determine farm status
  – Coverage

• Capture-recapture modeling
  – Same population for both lists?
  – Should sample design emphasize records not on NASS list frame?
  – Probability of capture
Local Foods Team Members

Mark Apodaca
Adam Cline
Jeff Bailey
James Barham—RD
Jeremy Beach
Jeff Beranek
Doug Boline
Kara Daniel
Saira Farooq
Ginger Harris
Andrew Brosier
Donald Buysse
Vincent Davis
Sarah Goodale
Pat Gregory
Brandon Hopkins
Mike Hyman
Troy Joshua
Doug Kilburg
Tom Kruchten
Megan Lipke
Christy Meyer
Peter Quan
Scott Shimmin
Dominique Sims
Holly Smith
Elanor Starmer – AMS
Danielle Tarpley
Nate Vandermeer
Stephen Vogel – ERS
Krissy Young


Selected References


Thank you!

Linda.Young@nass.usda.gov