Comparing Web Scraped Establishment Survey Frames of Industrial Hemp Growers in Seven States:

Costs, Contact Data, and Accuracy of Frame

Samuel Garber, Mike Gerling,
Katherine Vande Pol & Tyler Wilson

National Agricultural Statistics Service,
Research and Development Division

Background

- Agricultural Marketing Service (AMS) wanted to know more about the hemp industry
  - Hemp industry was recognized by USDA and included in the 2018 Farm Bill

- AMS and NASS worked together to develop a survey for industrial hemp growers
  - Completely new commodity

- NASS compiled lists received from several USDA agencies such as Farm Service Agency (FSA) and Risk Management Agency (RMA)

Background

- NASS was also web crawling/scraping on its own for industrial hemp growers, processors, and transporters

- NASS decided to contract out with Multi Agency Collaboration Environment (MACE) to find every industrial hemp grower in the U.S.
  - MACE and NASS have worked previously on other web scraping projects
  - MACE's list also included marijuana growers, processors, medical marijuana facilities, dispensaries, chain stores (Home Depot), and gas stations

- NASS decided to use the USDA lists for the list frame part of the survey and use MACE's list for under-coverage adjustments

Background

- This collection of hemp producers resulted in 3 frames
  - MACE: contractor
  - RDD: internal
  - USDA/NASS lists: FSA, RMA, AMS

- Seven States
  - Colorado, Illinois, Missouri, Montana, Nevada, New York and Tennessee

Research Questions

1. What are the pros & cons of web scraping MACE vs RDD vs USDA/NASS lists?

2. How accurate are the new frames at identifying industrial hemp growers?

Analysis Plan

- Summary of descriptive statistics
- Frame overlap assessment
- Web scraped data quality
- Needed Resources
- Cost
- Contact data availability

Descriptive Statistics

- Compare the number and percent of records in each frame with information for:
  - Owner/Manager
  - Operation Name
  - Address, Operation Address
  - Phone Number(s)
  - Email
  - Website
  - And more…

Descriptive Statistics for Nevada

Number of records with individual or operation information by frame source, after removal of duplicates and data cleaning			
	Frame Source		
Item.	RDD	MACE	USDA/NASS
Total	361	105	175
Including:			
Whole Name	213	1	34
Address	311	104	24
Address (Other)	0	0	1
City	311	105	175
Zip	311	105	174
Phone	263	95	18
Phone (Other)	23	8	0
Email	163	22	82
Website	140	88	1
Operation Name	346	105	151
Operation Address	0	0	0
Operation City	0	0	0
Operation Zip	0	0	0
Operation Phone	0	0	0
Operation County	338	97	155
Operation Email	0	0	42
Hemp License Number	4	TBD	TBD

Frame Overlap

- Any 2 or 3-way Overlap/Non-Overlap across the 3 frames
  - Which records were captured by 2 or more frames?
  - How many of the records were unique to the frame and not captured by any other?

Frame Overlap for Nevada

Number and percent of records matching across 2 or more frames in Nevada.				
Type of match	MACE records on RDD frame	USDA/NASS records on RDD frame	MACE records on USDA/NASS frame	RDD records on both MACE and USDA/NASS frames
Number of records	74	92	3	2
Percent of records	70.5%	52.6%	2.9%	0.6%

Data Quality

- Determine the quality of records – how many are hemp growers
  - Number of matches with NASS's Agricultural Census Mail List (CML) Frame
  - Examine responses to NASS's recent Industrial Hemp Survey
  - Examine 3rd party numbers

Resources & Cost

- For each frame determine:
  - Time needed for frame preparation and data cleaning
  - Costs associated with obtaining and cleaning the data
  - Limitations or issues with the frame
    - E.g., contains mostly processers, dispensaries, etc. instead of hemp growers
    - Lack of contact information which would be needed for surveys

Next Steps

- Complete descriptive statistics, cost analysis, and frame overlap for all 7 states
- Conduct a phone survey of a sub-sample of records from each frame
  - Short 3-5 question survey to verify that the record:
    - Has a valid phone number
    - Is a hemp grower, not a dispensary, processer, etc.
  - Currently working on survey design, sample selection, and OMB clearance

References

- Link, M. W., M. P. Battaglia, M. R. Frankel, L. Osborn, and A. H. Mokdad (2006). "Address-Based versus Random-Digit-Dial Surveys: Comparison of Key Health and Risk Indicators", American Journal of Epidemiology, 164, 1019-1025. DOI: https://doi.org/10.1093/aje/kwj310.
- Young, L. J., M. Hyman, and B. R. Rater (2018). "Exploring a Big Data Approach to Building a List Frame for Urban Agriculture: A Pilot Study in the City of Baltimore", Journal of Official Statistics, 34(2), 323-340. DOI: http://dx.doi.org/10.2478/JOS-2018-0015.
- Lo, A., S. Srikukenthiran, M. Chen, K. N. Habib, and E. J. Miller (2020). "Impact of Multiple Sample Frames on Data Quality of Household Travel Surveys: The Case of the 2016 Transportation Tomorrow Survey", Transportation Planning and Technology, 43(6), 553-570. DOI: https://doi.org/10.1080/03081060.2020.1780707.
- Young, L. J., and M. Jacobsen (2021). "Sample Design and Estimation When Using a Web-Scraped List Frame and Capture-Recapture Methods", Journal of Agricultural, Biological and Environmental Statistics, 27, 261-279. DOI: https://doi.org/10.1007/s13253-021-00476-w.
- Hyman, M., L. Sartore, and L. J. Young (2021). "Capture-Recapture estimation of characteristics of U.S. Local Food Farms Using a Web-Scraped List Frame", Journal of Survey Statistics and Methodology, 00, 1-26. DOI: https://doi.org/10.1093/jssam/smab008.
- Kim, A. E., B. Loomis, B. Rhodes, M. E. Eggers, C. Liedtke, and L. Porter (2015). "Identifying e-Cigarette Vape Stores: Description of an Online Search Methodology", Tobacco Control, 25, 19-23. DOI: http://dx.doi.org/10.1136/tobaccocontrol-2015-052270.

References

- Rhodes, B. B., A. E. Kim, and B. R. Loomis (2016). "Vaping the Web: Crowdsourcing and Web Scraping for Establishment Survey Frame Generation", In Proceedings of the 2015 Federal Committee on Statistical Methodology Research Conference, available at: https://nces.ed.gov/fcsm/pdf/H3_Rhodes_2015FCSM.pdf.
- ten Bosch, O., D. Windmeijer, A. van Delden, and G. van den Heuvel (2018). "Web Scraping Meets Survey Design: Combining Forces", In Proceedings of the BIGSURV18 Conference, available at: https://www.europeansurveyresearch.org/bigsurv18/uploads/73/61/20180820_BigSurv_WebscrapingMeetsSurveyDesign.pdf
- Barcaroli, G., D. Fusco, P. Giordano, M. Greco, V. Moretti, P. Righi, and M. Scarno (2016). "ISTAT Farm Register: Data Collection by Using Web Scraping for Agritourism Farms", In Proceedings of the ICAS VII Seventh International Conference on Agricultural Statistics, 1017-1086. DOI: https://doi.org/10.1481/icasVII.2016.f29d.
- Arora, S. K., S. Kelley, and S. Madhavan (2021). "Building a Sample Frame of SMEs Using Patent, Search Engine, and Website Data", Journal of Official Statistics, 37(1), 1-30. DOI: http://dx.doi.org/10.2478/JOS-2021-0001.
- Johnson, P., and D. Williams (2010). "Comparing ABS vs. Landline RDD Sampling Frames on the Phone Mode", Survey Practice, 3(3), 1-10. DOI: https://doi.org/10.29115/SP-2010-0012.
- Fulton, B. R., and King, D. P. (2022). "Using Google Maps to Generate Organizational Sampling Frames", SocArXiv Pre-print. DOI: https://doi.org/10.31235/osf.io/qtu8n.

Thank you

Samuel Garber	samuel.garber@usda.gov	202-692-0284
Michael Gerling	michael.gerling@usda.gov	202-692-0277
Katherine Vande Pol	katherine.vandepol@usda.gov	217-493-2999
Tyler Wilson	tyler.wilson@usda.gov	202-692-0290