*Journal of Survey Statistics and Methodology (2022) **10**, 979–1004*

CAPTURE–RECAPTURE ESTIMATION OF CHARACTERISTICS OF U.S. LOCAL FOOD FARMS USING A WEB-SCRAPED LIST FRAME

MICHAEL HYMAN*
LUCA SARTORE
LINDA J. YOUNG

The emerging sectors of agriculture, such as organics, urban, and local food, tend to be dominated by farms that are smaller, more transient, more diverse, and more dispersed than the traditional farms in the rural areas of the United States. As a consequence, a list frame of all farms within one of these sectors is difficult to construct and, even with the best of efforts, is incomplete. The United States Department of Agriculture's (USDA's) National Agricultural Statistics Service (NASS) maintains a list frame of all known and potential U.S. farms and uses this list frame as the sampling frame for most of its surveys. Traditionally, NASS has used its area frame to assess undercoverage. However, getting a good measure of the incompleteness of the NASS list frame using an area frame is cost prohibitive for farms in these emerging sectors that tend to be located within and near urban areas. In 2016, NASS conducted the Local Food Marketing Practices (LFMP) survey. Independent samples were drawn from (1) the NASS list frame and (2) a web-scraped list of local food farms. Using these two samples and capture–recapture methods, the total number and sales of local food operations at the United States, regional, and state levels were estimated. To our knowledge, the LFMP survey is the first survey in which a web-scraped list frame has been used to assess undercoverage in a capture–recapture setting to produce official statistics. In this article, the methods are presented, and the challenges encountered are reviewed.

MICHAEL HYMAN is with Roundtable Analytics Inc., Raleigh, NC, USA. LUCA SARTORE is with USDA National Agricultural Statistics Service, Washington D.C., USA, National Institute of Statistical Sciences, Washington D.C., USA and LINDA J. YOUNG is with USDA National Agricultural Statistics Service, Washington D.C., USA
*Address correspondence to Michael Hyman, Roundtable Analytics Inc., Raleigh, NC, USA; E-mail: mhyman1@gmail.com

doi: 10.1093/jssam/smab008 Advance access publication 16 August 2021

Best practices and open research questions for conducting surveys using web-scraped list frames and capture–recapture methods are discussed.

KEYWORDS: Capture-recapture; Population estimation; Sampling; Coverage; Web scraping.

1. INTRODUCTION

An inherent concern associated with list frames, which often serve as the sampling frames for surveys, is list undercoverage and the resulting potential bias (Federal Committee on Statistical Methodology 2001). The increasing use of cell phones and the unwillingness to have landlines publicly listed have led to decreasing coverage of list-assisted landline frames (Blumberg and Luke 2019). For an election survey, Lipps, Pekari, and Roberts (2015) found the bias owing to landline undercoverage to be greater than that owing to noncontact. Sala and Lillini (2017) noted the low coverage of the Italian population when using the white pages for a telephone survey and found this undercoverage resulted in bias. Similarly, in assessing the effects of undercoverage when using a list of addresses derived from the U.S. Postal Service's Computerized Delivery Sequence as a sampling frame, Amaya et al. (2018) reported the risk of undercoverage bias ranging from low for some variables to high for others.

Multi-frame methods are often used to account for undercoverage (Wallgren and Wallgren, 2016; Chipperfield et al. 2017; Brown et al. 2019). The USDA's National Agricultural Statistics Service (NASS) has traditionally used its area frame to account for undercoverage on its list frame, which is composed of all known farms or potential farms in the United States. Wallgren and Wallgren (2016) also used an area frame to account for undercoverage of a register-based list frame. Surveys using multiple frames typically require the sampling population (i.e., the union of all frames) to cover the entire population of interest for resulting samples to be representative. Using an area frame ensures that complete coverage.

Using the NASS area frame to measure for undercoverage, in which areal land units are sampled and combined to identify all farms within, works well in the rural areas of the United States where production agriculture is centered. However, the emerging sectors of agriculture, such as urban agriculture, organics, and local food, tend to be dominated by farms that are smaller, more diverse in production, more transient, and more spatially dispersed than the traditional farms. Their often small sizes and sometimes nontraditional appearances (e.g., an abandoned warehouse) make them challenging to identify within an areal unit. Thus, sampling areal units to measure undercoverage of the NASS list frame for these populations is cost prohibitive and plagued by substantial misclassification (Abreu et al. 2010). Alternative, cost-effective, time-sensitive methods for measuring the undercoverage of a list frame are needed.

Although there are many examples of using web scraping to obtain information, its use for list building is in the early stages of exploration. Rhodes et al. (2015) used web scraping to create a list of all electronic nicotine delivery systems (ENDS) vape stores in the State of Florida and verified whether or not each record on the list was truly a vape store through crowdsourcing (the process of obtaining information from a large group of people, usually via the internet). In 2014 and 2015, NASS explored the use of web scraping to develop a list frame of urban farms that could serve as an alternative to the area frame for assessing list undercoverage (Young et al. 2018). From this early work, it was evident that, when considering the population of all U.S. farms, the union of the list frames would not provide full coverage.

Capture–recapture techniques can account for population units that are not on either list frame (Bird and King 2018) and have been used in the production of official statistics. The U.S. Census Bureau has used capture–recapture estimation, called dual-system estimation when applied to human populations, to evaluate the undercoverage of the Decennial Census of Population since 1980 (Hogan 1993; U.S. Census Bureau 2004, 2008; Mule 2012), and the Brazilian Institute of Geography and Statistics applied dual-system estimation to estimate net coverage, undercoverage, and overcoverage for its 2010 Brazilian Census (da Silva et al. 2015). NASS used capture–recapture techniques to adjust for undercoverage, nonresponse, and misclassification in its 2012 Census of Agriculture (Young et al. 2017).

In 2016, growing national interest in local food led NASS to implement the Local Food Marketing Practices (LFMP) survey to quantify and characterize operations that produce locally sourced agriculture. The objectives of this survey were to generate estimates for the total number of U.S. farms that distribute their products locally and to learn more about these local food farms (LF farms), including the amount of produce sold at the national, regional, and state levels. For the purposes of this study, local sales are defined as sales in one of the four distinct categories: direct to consumer, direct to retail, direct to institution, or direct to an intermediate source that labels a product as "local." This allows for a uniform definition of local sales, as the distance a product must travel to be deemed "local" differs regionally. In addition to having local sales, an operation must sell at least $1,000 of agricultural products during a typical year to qualify as a farm.

As with other emerging sectors of agriculture, the NASS list frame is known to be substantially incomplete for LF farms, and it is too costly to use the area frame to adjust for undercoverage. A web-scraped list frame was created from publicly accessible websites. However, the resulting list was also incomplete, leaving potential for farms that were not identified by either list. Thus, NASS employed capture–recapture methods to use the web-scraped list to adjust for undercoverage, nonresponse, and misclassification. When the official statistics from this study were released in December 2017, to our knowledge, the LFMP survey is the first *survey* in which a web-scraped list frame was used within a

capture–recapture framework to adjust for the undercoverage of a traditional list frame to produce official statistics.

In Section 2, web scraping is defined in more detail and the lists, samples, and modeling methodology are described. Results from the LFMP survey, with respect to the web-scraped list frame, the sample characteristics, and the estimation process are presented in Section 3. In Section 4, considerations for applying these methods more broadly in survey practice and open research questions are discussed.

2. METHODOLOGY

2.1 Development of the Web-Scraped List

Research, commercial, and government organizations can gather unstructured information through web scraping and convert that information into useful data sets. The ability to navigate through large quantities of web pages quickly and target specific information allows data collection on a larger scale than a single person is capable (Landers et al. 2016). MIT's Sloan School of Management's Billion Prices Project demonstrates the efficiency and utility of web scraping as a research tool, collecting pricing data for millions of products, from hundreds of online retailers weekly, to analyze short- and long-term trends (Cavallo 2018). Although each application requires customized programming, the speed and absence of human interaction make web scraping a generally inexpensive method of collecting large amounts of data (Cavallo and Rigobon 2016).

Applications of web scraping primarily focus on extracting specific details from web sources for each record in a prespecified list and compiling the information into a usable data set (Vargiu and Urru 2013). These processes can be replicated quickly to gather the desired information for extensive lists of records (Chow et al. 2011). Adapting these processes to identify and retrieve *new* records for a sampling frame introduces additional concerns about accuracy and coverage of web sources. Application programming interfaces (APIs), such as Google Maps and YellowPages.com, have made record retrieval for simple search queries trivial. Overly specific or ill-defined search criteria, including farm status of a location or local sales of products, can be difficult to mechanically verify using extracted information, presenting questions about whether the records obtained are in the target population. A member of the target population may be identified through web scraping if that member has a website or is present on another publicly available website, such as in an association's member list or in a list on a state, county, or local website. Assessing whether some segments of the target population cannot be identified from web sources is also challenging. The resulting undercoverage of the list could lead to biased estimates. Other questions can be raised regarding the

impact that the varying number of websites from which a member can be identified may have on the likelihood of finding new records.

Initial efforts to use web scraping to develop a list frame have been made. Zhang and Tang (2016) introduced and evaluated an algorithm that automatically builds place–name data sets of specific business types from the results of a web-search engine. The method was applied to 10 city blocks in a U.S. city and the results were compared with search results from Google Maps and OpenStreetMap. Site names and addresses were identified through Google Street View and used to ground truth results. Results showed that the authors' list building methodology generated place–name data sets as effectively as Google Maps and better than other open source web-search engines.

Rhodes et al. (2015) examined comprehensive list building via web scraping by creating a list of ENDS stores in Florida based on search results from Google Maps, Yelp, and YellowPages. Crowdsourcing was used to verify sales of ENDS products by the identified ENDS stores. ENDS stores in Florida are required to be registered as tobacco retailers (29,039 total tobacco retailers but not all sell ENDS products), providing a gold standard list for comparison. The authors found that 131 (32.5 percent) of the 403 stores on the web-scraped list were on Florida's tobacco licensure list, and the remaining 272 results were not.

Young et al. (2018) developed a list of urban agricultural sites within the City of Baltimore, MD, using a combination of web scraping and satellite imagery. The resolution of the available satellite imagery was too coarse to identify agricultural characteristics, but a list of 505 potential sites was obtained via scraping publicly available web sources. This list was sampled and 108 (57.4 percent) of 188 responding, sampled sites reported at least some type of agricultural activity. Interviewers were able to observe evidence of agriculture (visible vegetable gardens, animal enclosures, etc.) at an additional 31 (39.7 percent) of 78 nonresponding sampled sites when visiting in an effort to secure a response.

For this study, spatial scope and explicit search criteria make developing a comprehensive list of LF farms challenging. The process began with the development of a set of keywords that could be indicative of local food sales. The list consisted of 41 words and phrases, including "local food farm," "organics farms," "local food sales," and "farmer's market." A combination of existing APIs (e.g. the sets in Google Maps and YellowPages) and internet search queries were used to generate a list of potential LF farms. Application programming interfaces generally focus on small spatial regions, and searches had to be replicated for states or even smaller regions. To scrape the requisite information needed for list building, software programs, which require a substantial amount of development time, are created. These must be modified for each internet source (Chow et al. 2011). Relevant information from records showing evidence of local sales was drawn either directly from a web application (e.g., Google Maps in which LF farms are identified by place names and associated

information on a map), from websites with lists of potential LF farms, or from a website identified for each record. Acquired information consisted of site-specific details important for linking records to an existing list (e.g. business name, address, and phone number). Young et al. (2018) provided a more technical discussion of the web-scraping methods which were also used in this study.

The quality of a web-scraped list of records is constrained by the extent and depth of web sources scraped. The total time available for the LFMP survey, from initiation to the publication of results, was 14 months owing to the factors beyond the control of the Agency. As a consequence, the time allocated for developing a web-scraped list frame was December 14, 2015, to January 18, 2016, a period of about 1 month that included major holidays. In any scenario, identifying and using all available web sources is difficult, but an anticipated consequence of the short time frame was that there would be incomplete harvesting of collected sources and, hence, a more incomplete web-scraped list frame than otherwise possible. Every effort was made to develop a national web-scraped list frame of LF farms in the time available.

After the list building was concluded, the Matchware Technologies programs Automatch and Autostan (Day 1996) were used to conduct probabilistic record linkage between the NASS list frame and the web-scraped list. Uncertain matches were reviewed and contacted to verify whether a potential match was correct. Resulting matches were considered exact and without error. As a consequence, the records could be separated into three groups: those on only the NASS list frame, those on only the web-scraped list frame, and those on both frames.

Typically, NASS screens all new records identified as potential farms to determine whether they meet the definition of a farm. Only records confirmed to be farms during the screening process are eligible for participation in NASS surveys. Given the compressed time frame and the considerable amount of time necessary to screen records, the records identified on the web-scraped list frame were not put through the screening process. The farm status (confirmed or potential) of web-scraped records that were also on the NASS list frame could be determined. However, all records only on the web-scraped list were potential farms. As a result of not being able to confirm farm status for all web-scraped records (and to then remove the non-LF farms from that list), the web-scraped list was expected to have more non-farms than the NASS list frame. Sales to local marketing channels were unknown for all records on both lists, although information from the 2012 Census of Agriculture indicated past local sales for a subset of NASS list frame records.

2.2 Overview of Capture–Recapture Methodology

Capture–recapture is frequently used to estimate wildlife populations by resampling the population of interest and comparing individuals captured in

each sample (Seber 2002; Bird and King 2018). For these wildlife studies, no list frame is available and hence multiple samples are collected by traps, seines, or some other mechanism. For this study, two list frames were developed, and a probabilistic sample was drawn from each frame. The two samples (one from each list frame) are treated as independent samples from the target population (Hogan 2003).

Estimation of farm counts via traditional capture–recapture using two list frames requires five basic assumptions (Pollock et al. 1994): (1) The population must be closed; no farms enter or leave the population between samples. (2) The two lists are independent; that is, a record's inclusion on one list is independent of its being on the other list. (3) All farms are equally likely to be caught in each sample. (4) Capturing a farm in one sample does not affect its catchability in the other sample. (5) Finally, farms must be identifiable and linked between samples without error. These assumptions imply that the proportion of individuals captured in one sample that are also captured in a subsequent sample is approximately equals to the proportion of the entire population captured during the first sample, allowing estimation of population coverage by the first sample. Rarely, all of the assumptions are met. For example, the probability of capture may vary with sample, a farm respondent to one survey may avoid another, and the capture probabilities of farms may be heterogeneous (see Otis et al. 1978 for an ecological perspective and Wolter 1986 for coverage models relating to censuses and surveys). Pledger (2000) provides a framework to account for violations of one or more of the assumptions using logistic finite mixture models, which is the approach used here to model coverage. In Section 3, whether each of these assumptions is satisfied for this study is considered and, in Section 4, these are summarized and research needs highlighted.

Although the two used different approaches (Hogan 2003; Young et al. 2017), the coverage estimates of the 2010 U.S. Census (Mule 2012) (a population census) and the 2012 U.S. Census of Agriculture (an establishment census) estimates that accounted for undercoverage, nonresponse, and misclassification (Young et al. 2017) were based on capture–recapture methods. For both the U.S. Census and the U.S. Census of Agriculture, all records on the primary list frame were surveyed, and a sample was drawn from a second frame to assess the coverage of the primary frame. A full census of the NASS list frame to identify LF farms would be cost prohibitive. For this study, a sample drawn from the primary frame was used for estimation and the sample drawn from the secondary frame provided coverage evaluation.

To estimate the number of LF farms, the 2012 Census of Agriculture's capture–recapture methodology was modified to account for a sample being drawn from the NASS list frame instead of a full census (Young et al. 2017). As a consequence, the probability of selection is a factor in a farm's capture probability that must be accounted for in the estimation process. To be captured, a LF farm must be on the NASS list frame, included in the sample, respond to

the LFMP questionnaire, and be classified as a farm with local sales based on its response to the questionnaire. Two types of misclassification could occur. Based on the response to the LF questionnaire, a LF farm may be classified as a non-LF farm or a non-LF farm may be classified as a LF farm. Both types of misclassification are considered. Denoting the probability of a farm being captured during the survey as $\pi_{C,i}$ and the probability of a record identified as a LF farm based on its survey response being correctly classified as $\pi_{M,i}$, the weight w_i associated with each respondent i classified as a farm on the LFMP survey is

$$w_i = \frac{\pi_{M,i}}{\pi_{C,i}} = \frac{P(F|M,\ C,\ R,\ S)}{P(M,\ C,\ R,\ S|F)}, \tag{1}$$

where F indicates that the record is part of the target population of LF farms, C denotes that the record is on the NASS list frame, S signifies that the record was included in the sample, R indicates the record responded to the survey, and M represents a record classified as a LF farm. It should be noticed that the potential misclassification of a LF farm as a non-LF farm is accounted for in the capture probability; however, the possible misclassification of a non-LF farm as a LF farm is not and must be addressed by the probability in the numerator of the weight in (1). Weights adjust each record for the number of uncollected LF farms resulting from undercoverage, sampling, nonresponse, or either type of misclassification. For example, a record with a weight equal to 3 accounts for two additional LF farms that were not included in estimation owing to sampling or one of the identified sources of error.

The capture probability of a LF farm, $\pi_{C,i} = P(M,\ C,\ R,\ S|F)$, reflects record being a LF farm, on the NASS list frame, included in the sample, responding to the survey, and classified as being a LF farm. The probability, $\pi_{M,i} = P(F|M,\ C,\ R,\ S)$, is the probability that a record identified as a LF farm by the survey is classified correctly.

The capture probability can be reformulated into four probabilities attributed to each capture requirement as follows:

$$P(M,\ C,\ R,\ S|F) = P(M|C,R,S,F)P(C|R,S,F)P(R|S,F)P(S|F). \tag{2}$$

When the sample is drawn, whether the record is associated with a LF farm is not known until a response to the questionnaire is received. The coverage of the NASS list frame can only be appropriately assessed after it has been determined whether a record is associated with a member of the population of interest (LF farms). Thus, the probability of coverage is conditional on response and, for estimation purposes, the capture probability $\pi_{C,i}$ may be expressed as

$$P(M,\ C,\ R,\ S|F) = P(M|C,R,S,F)P(C|R,S,F)P(R|S)P(S), \tag{3}$$

Equivalence of (2) and (3) requires the assumptions that a record's sampling probability and response probability are independent of the record's status as a LF farm:

$$P(S) = P(S|F) \tag{4}$$

and

$$P(R|S) = P(R|S, F). \tag{5}$$

That is, the probability of being included in the sample and the probability of obtaining a response from a sampled record is the same for LF farms and non-LF farms. These assumptions are discussed further in Section 3.

The probability of being selected for the sample, $P(S)$, is based on the sample design and is thus known. The $P(C|R, S, F)$, which assesses the coverage of the NASS list frame, is estimated using the sample drawn from the web-scraped list frame. The NASS list frame sample is used to estimate $P(R|S)$. A separate misclassification study is conducted to estimate both $P(M, C, R, S|F)$ and $P(M|C, R, S, F)$.

The capture–recapture estimator for a numeric characteristic of LF farms in the United States is the sum of each weight w_i multiplied by the corresponding variable of interest, y_i, for all LF farms identified from the list frame sample,

$$\sum_{i \in \mathcal{F}} y_i w_i = \sum_{i \in \mathcal{F}} \frac{y_i \pi_{M,i}}{\pi_{C,i}} \tag{6}$$

where $\mathcal{F}$ is the set of responding LF farms in the list frame sample (Young et al. 2017). For estimates of LF farm counts, y_i is equals to 1 for all responding LF farms and is 0 otherwise. The variance of this estimator was computed using bootstrapping methods to be described later.

2.3 Sampling

To implement the capture–recapture methodology, samples were drawn independently from the NASS list frame and the web-scraped list frame. Sample sizes accounted for the expected portion of LF farms on each list, anticipated response rates, and desired precision of the estimator.

Results from NASS's urban agriculture pilot study in Baltimore showed 57 percent of respondents reported having some type of agriculture (Young et al. 2018). Based on this fact, it was assumed that approximately half of the records on the web-scraped list frame would qualify as LF farms. This projected accuracy rate and an anticipated response rate of 70 percent were considered in establishing the sample size of 19,365 potential LF farms from the NASS list frame. The web-scraped list frame contained only contact information, such as business name and address, for each record. To achieve sample sizes proportional to the number of records identified within each state, the web-scraped

records were sorted by state and a systematic sample was selected from the entire frame.

When the sample was drawn, the NASS list frame contained 2,007,110 confirmed farms and potential LF farms. Only a small proportion of these farms were believed to be selling locally. To better target LF farms, information from the 2012 Census of Agriculture and from NASS field offices was used to stratify records. A stratified random sample was drawn from the list frame. The sample was stratified by state and, within each state, four strata were identified:

(A) Farms reporting local food sales on the 2012 Census of Agriculture and a value for local food sales,

(B) Farms reporting local food sales on the 2012 Census of Agriculture but not reporting a value for local food sales,

(C) Potential farms that were identified by NASS regional field offices as potentially having local food sales but not in groups A or B,

(D) All other farms not in groups A, B, or C.

For each state, records in sampling strata A and D were stratified further based on total value of sales (not only local sales) reported during the 2012 census. The sampling fraction varied across strata with strata A, B, and C having a higher probability of selection than stratum D and the records with higher total values of sales in strata A and D having a higher selection probability.

Based on the record linkage, which was conducted after the formation of the web-scraped list frame and before the samples were drawn, each record in both samples was identified as being in one of three groups: (1) in both the NASS list frame and the web-scraped list frame samples, (2) in the NASS list frame sample but not the web-scraped list frame sample, or (3) in the web-scraped list frame sample but not the NASS list frame sample.

The NASS list frame is designed for record linkage based on farm operators. The web-scraping process led to a list frame focused on the farm operation (not the operator). This made record linkage more challenging, and the links between the two frames were continually refined during the sampling and data collection process. Although all known linkage errors were corrected, some matching errors likely remain. The rate with which this occurred is unknown.

All sampled records were sent a LFMP questionnaire. Records included in both samples were sent one questionnaire to avoid the additional burden of asking for two responses to the same set of questions. Questionnaires from respondents were reviewed and each record's scope (whether or not the record is classified as a LF farm and thus part of the target population) was determined. Respondent records identified as LF farms were in-scope for the LFMP survey and used during estimation. Records indicating no local food sales or failing to meet the definition of a farm were out-of-scope and were removed.

2.4 Estimation

The NASS list frame sample was the primary sample for estimation. For in-scope records in this sample, the sampling weights were adjusted to account for undercoverage of the NASS list frame, survey nonresponse, and misclassification of LF farms as out-of-scope records and of non-LF farms as in-scope records.

The estimated probability of capture was obtained by estimating the individual components of undercoverage, nonresponse, both types of misclassification, and sampling (3). The probability of inclusion in the sample from the NASS list frame, P(S), was based on the stratified random sampling design and thus known for each record.

Response probabilities, which vary among records, were estimated from the primary sample (and did not include the web-scraped list frame sample records). To account for this variation when estimating P($R|S$), the NASS list frame sample records were grouped by state/region and whether or not the record linked to a record on the web-scraped list. States were grouped together to form regions if the number of respondent records was less than 80 records. The probability of response was estimated by the proportion of all responding records within a group.

To model the probability of a LF farm record being on the NASS list frame, the LF farms (as identified based on the responses to the LFMP questionnaire) in the web-scraped list frame sample were used. The response variable was whether or not the LF farm was on the NASS list frame. To reflect the differential catchability of LF farms, a logistic model of a list frame record's coverage probability, P($C|R, S, F$), that allowed the probability to vary based on the corresponding LF farm's characteristics was developed. Independent variables considered during modeling included a record's U.S. region, indicator variables of sales to each of the four local marketing channels, total sales to each marketing channel, total sales (including nonlocal food sales), farm type (crop farms, livestock farms, or both), and all two-way interactions. Variable selection was performed using stepwise regression with fivefold crossvalidation to avoid over fitting the model. The fitted model was used to predict the coverage probability for all in-scope records from the NASS list frame sample.

To estimate both misclassification probabilities, a misclassification sample was drawn from all respondents in the NASS list frame sample. The sample was stratified by whether or not the record was in-scope. The LF farm status of each record in the sample was independently determined through a telephone survey. Records with discrepancies in their LF farm status between the LFMP survey and the misclassification survey were further investigated to establish true LF farm status. Misclassification probabilities were estimated based on the proportion of records that changed scope from the original to the final determination of LF farm status.

Weights were calculated for each in-scope record from the NASS list frame sample under the assumptions identified in (3):

$$\begin{aligned}\tilde{w}_i &= \widehat{\pi}_{M,i}/\widehat{\pi}_{C,i} = \widehat{P}_i(F|M,C,R,S)/\widehat{P}_i(M,C,R,S|F) \\ &= \widehat{P}_i(F|M,C,R,S)/[\widehat{P}_i(M|C,R,S,F)\widehat{P}_i(C|R,S,F)\widehat{P}_i(R|S)\widehat{P}_i(S)]\end{aligned} \tag{7}$$

Prior to producing estimates and standard errors, weights were calibrated to reduce variability caused by large sampling weights. Ten national-level target estimates were calculated using the precalibration weights (7): (1) total farm count, (2) total amount of local sales, (3) total direct to consumer farm count, (4) total direct to consumer sales, (5–7) total farm counts for three categories of local sales amounts (<$10,000, $10,000–$99,999, ≥$100,000), (8–10) total farm counts for three categories of direct to consumer sales amounts (<$10,000, $10,000–$49,999, ≥$50,000). Integer calibration was used to round the weights $\tilde{w}_i$ to integers and to calibrate the integer weights to meet the national-level target estimates. Through the calibration process, the integer weights were restricted to a maximum of 550 while ensuring that the estimates remained within 1 percent of each of the targets (see Sartore et al. 2019 for full details on the method). The final calibrated weight of record i is denoted by $\widehat{w}_i$.

The national estimate of the number of LF farms was calculated as the sum of all calibrated weights. Other estimates, such as state or regional estimates of the number of LF farms, were calculated as the sum of all weights in the corresponding subset of farms. Local and total sales were estimated as the sum of the weighted sales for each farm.

Standard errors were computed using an approach based on a bootstrap methodology. The bootstrap weights approach (Rao et al. 1992) was conducted to draw bootstrap weights instead of selecting bootstrap samples from either a pseudo-population or through direct bootstrapping (Mashreghi et al. 2016). The bootstrap weights were generated using the generalized bootstrap method of Beaumont and Patak (2012). For each record i, the bootstrapped weight of that record can be written as $\widehat{w}_i^* = a_i^* \ \widehat{w}_i$, where $\widehat{w}_i$ is the integer-calibrated weight of record i and a_i^* represents the adjustment to $\widehat{w}_i$ owing to bootstrapping. The distribution of a_i^* is chosen so that the estimator of the variance of the population totals and subtotals is unbiased. That is, $a_i^* \sim \mathrm{N}(1, (\widehat{w}_i - 1)/\ \widehat{w}_i)$. To estimate the variance of an estimator $\widehat{T}$ of the population total T, the total T was estimated from calibrated weights $\widehat{w}_i$, and the variance was estimated using the bootstrapped adjusted weights $\widehat{w}_i^*$.

3. RESULTS

3.1 Development of the Web-Scraped List

The final web-scraped list frame contained 36,228 records, each of which was a potential farm with evidence of local sales and sufficient information to link

to the NASS list frame. For the assumption of independence between lists to be satisfied, a farm's probability of being captured during one sample cannot affect the probability that it is captured during another sample; that is,

$$P(Record\ is\ on\ List1|Record\ is\ on\ List2) = P(Record\ is\ on\ List\ 1). \quad (8)$$

Statistically verifying independence between sampling frames requires at least three frames (Lohr 2009). During this study, the web-scraped sampling frame was built independently from the NASS list frame. Information, such as producer association lists, is gathered from the web during the list building of the NASS list frame, leading some to question whether the two frames are independent. Having an overlap in records is necessary for capture–recapture estimation, but an overlap in sources is not required. Whether the overlap in sources or some other aspect of the frame development process leads to a lack of independence has yet to be evaluated.

The web-scraped records were linked to the NASS list frame based on business name, address, phone number, or other features. In total, 27,986 (77.2 percent) records were linked to a record on the NASS list frame. However, the linkage information was not used in the sample design for the samples from either the NASS list frame or the web-scraped list frame.

For the 2012 Census of Agriculture, 2 percent of the records from the area frame used to measure undercoverage were initially, but incorrectly, thought to not match a NASS list frame record; these matches were corrected (Young et al. 2017). The error in this study is likely to be larger, and the rate of these errors is unknown. Therefore, the assumption of identifying and linking LF farms between samples without error was likely not met.

3.2 The Samples and Their Characteristics

Of the 19,365 records in the systematic sample drawn from the web-scraped list frame, 15,543 (80.3 percent) were linked to a record on the NASS list frame (Table 1). In contrast, 2,513 (10.1 percent) of the 24,907 records sampled from the NASS list frame were linked to a record on the web-scraped frame. As matching was conducted prior to the selection of the samples, the portion of the sampled records common to both frames can be partitioned into those records common to both samples, those in the NASS list frame sample but not the web-scraped list frame sample, and those in the web-scraped list frame sample but not the NASS list frame sample. In total, 1,471 records were in both the NASS list frame and the web-scraped list frame samples; that is, 5.9 percent of the NASS list frame sample and 7.5 percent of the web-scraped list frame sample were in both samples. These records were included in response rates and in-scope rates for each list, but only a single response and farm status were considered for each respondent. Briefly, 14,072 records of the 19,365 (72.7 percent) sampled from the web-scraped list frame were not in the

Table 1. Frame sizes and sample sizes for NASS's list frame and a web-scraped list.

Sampling frame	Sampling stratum	Frame size	Sample size	Number sampled records linked	Percent sample linked
NASS list	All	2,007,110	24,907	2,513	10.09
	A	97,326	9,956	1,999	20.08
	B	23,776	3,599	162	4.50
	C	7,760	1,532	211	13.77
	D	1,878,248	9,820	141	1.44
Web-scraped list	All	36,228	19,365	15,543	80.26

NASS list frame sample but were on the NASS list frame, and 1,042 records of the 24,907 (4.2 percent) sampled from the NASS list frame linked to a record on the web-scraped list frame but were not in both samples. Only the records in the two samples are used in the capture–recapture analysis that follows.

Recall that strata A, B, and D of the NASS list frame are composed of confirmed farms. Thus, sampling stratum C of the NASS list frame sample is the most comparable to the web-scraped list frame; both are composed of potential (not confirmed) farms, thought to sell locally (stratum C farms identified by NASS field offices and web-scraped farms identified through web sources). From stratum C, 211 records (13.8 percent) linked to the web-scraped list frame. Stratum A, which consisted of farms reporting a value of local food sales during the 2012 Census of Agriculture, had 1,999 (20.1 percent) records linked to the web-scraped frame. Only 4.5 percent (162) of records in stratum B matched to the web-scraped list. Based on linkage percentages from strata A and C, this was lower than expected because stratum B records had an indication of local sales, but no reported value of those sales, during the 2012 Census of Agriculture. The small percentage of matches from stratum D (141 or 1.4 percent of the records) was anticipated, given that these records had no indications of prior local sales and were likely to be out-of-scope for this study.

Local food farms tend to be more transient than more traditional types of agricultural operations, emphasizing the challenge of satisfying the assumption of population closure (Low et al. 2015). Any lag time between capture events can violate this assumption. For this study, the survey was conducted simultaneously for the two samples; thus, this assumption was met.

The overall response rate for the web-scraped list sample was 61.8 percent. Response rates from the NASS list frame sample varied among sampling strata but it was 72.3 percent for all records (Table 2). The response rate for stratum C, the portion of the NASS list frame sample most comparable to the web-scraped list, was 58.9 percent, the lowest among all sampling strata. The

Table 2. Responses and record scope for samples from the NASS list frame and the web-scraped list frame.

Sampling frame	Sampling stratum	In-scope	Out-of-scope	Out of business	Non-response	Total
NASS list	All	5,697 (22.87%)	10,632 (42.69%)	1,673 (6.72%)	6,905 (27.72%)	24,907
	A	4,218 (42.37%)	3,211 (32.25%)	542 (5.44%)	1,985 (19.94%)	9,956
	B	750 (20.84%)	1,656 (46.01%)	189 (5.25%)	1,004 (27.90%)	3,599
	C	339 (22.13%)	430 (28.07%)	134 (8.75%)	629 (41.06%)	1,532
	D	390 (3.97%)	5,335 (54.33%)	808 (8.23%)	3,287 (33.47%)	9,820
Web-scraped list	All	4,685 (24.19%)	5,775 (29.82%)	1,516 (7.83%)	7,389 (38.16%)	19,365

Percentages are based on the sample sizes from each sampling stratum (row totals).

confirmed farms comprising the other strata were more likely to respond, with the highest response rate (80.1 percent) being from those who reported values of local sales in the 2012 Census of Agriculture. For the web-scraped sample, the response rate for records linked to a NASS list frame record (60.1 percent) was significantly lower than the response rate for records that did not link (66.7 percent) (Pearson's $\chi^2 = 56.30$; $p < .0001$). A similar discrepancy was not found when comparing the NASS list frame sample records that were and were not linked to the web-scraped list.

The percentage of in-scope records (responding records that were identified as being LF farms based on the questionnaire responses) was surprisingly low. The in-scope rate from NASS list frame sample respondents was 31.6 percent $[= 100(5{,}697)/(24{,}907-6{,}905)]$, but varied based on sampling stratum (figure 1). The 39.1 percent in-scope rate for the web-scraped sample respondents was most comparable to the 37.5 percent in-scope rate from stratum C. As anticipated, sampling stratum A had the highest in-scope rate. Yet, although this stratum was composed of records that reported local sales during the 2012 Census of Agriculture and were considered highly likely to be LF farms, the in-scope rate was only 52.9 percent, an indication of the transient nature of these farms. Respondents from sampling stratum B, which was also considered likely to include in-scope records, produced only 28.9 percent in-scope records, which is lower than that of the web-scraped list sample. As anticipated, sampling stratum D had the lowest in-scope rate at 5.9 percent.

Figure 1. Response rates (left) and proportion of in-scope responses (right) for each sampling stratum.

Table 3. Number of in-scope records returned from NASS list frame and web-scraped list frame samples.

Sample frame	Linked to other frame	Not linked	Total
NASS list	528	4,373	4,901
Both lists	796	0	796
Web-scraped list	3,521	368	3,889
Total	4,845	4,741	9,586

The samples drawn from the NASS and web-scraped list frames produced, respectively, 5,697 and 4,685 LF farms. Of the 1,471 records selected in both samples, 796 were in-scope and included in both totals. Therefore, 9,586 unique LF farms were identified in the combined samples (Table 3). There were 4,845 in-scope records common to both frames, 528 of which were in the NASS list sample and not the web-scraped sample (although they were in the web-scraped frame) and 3,521 of which were in the web-scraped list frame sample and not the NASS list frame sample (although they were in the NASS list frame). The web-scraped sample returned 368 in-scope records not linked to the NASS list frame, which is 7.9 percent of all in-scope records from this sample.

Response rates were calculated using AAPOR RR1 from the American Association of Public Opinion Research Response Rate Standard Definitions manual (AAPOR 2016). Response rates accounted for whether or not the record was linked to the web-scraped list and were estimated for each U.S. state/region (states were combined to form regions if state totals were less than 80).

Varying coverage probabilities among frames can result in the types of farms collected differing between web scraping and traditional collection methods. The range of sales for records from both lists was examined, and as anticipated, the distributions of total sales differed between LF farms from the NASS list frame and farms from the web-scraped list. Web scraping identified a higher proportion of farms with total sales between $5,000 and $250,000 as compared to the NASS list frame sample, but fewer farms outside of this range (figure 2). These differences in farm size were accounted for during analysis. Other factors were also considered, and no gap in coverage was found in either list. However, unidentified coverage differences between the two lists could lead to bias in the estimates.

3.3 Estimation

Capture–recapture methods were used to estimate farm numbers, sales, and other characteristics of the local food industry. Weights were calculated for each in-scope record in the NASS list frame sample and were composed of the

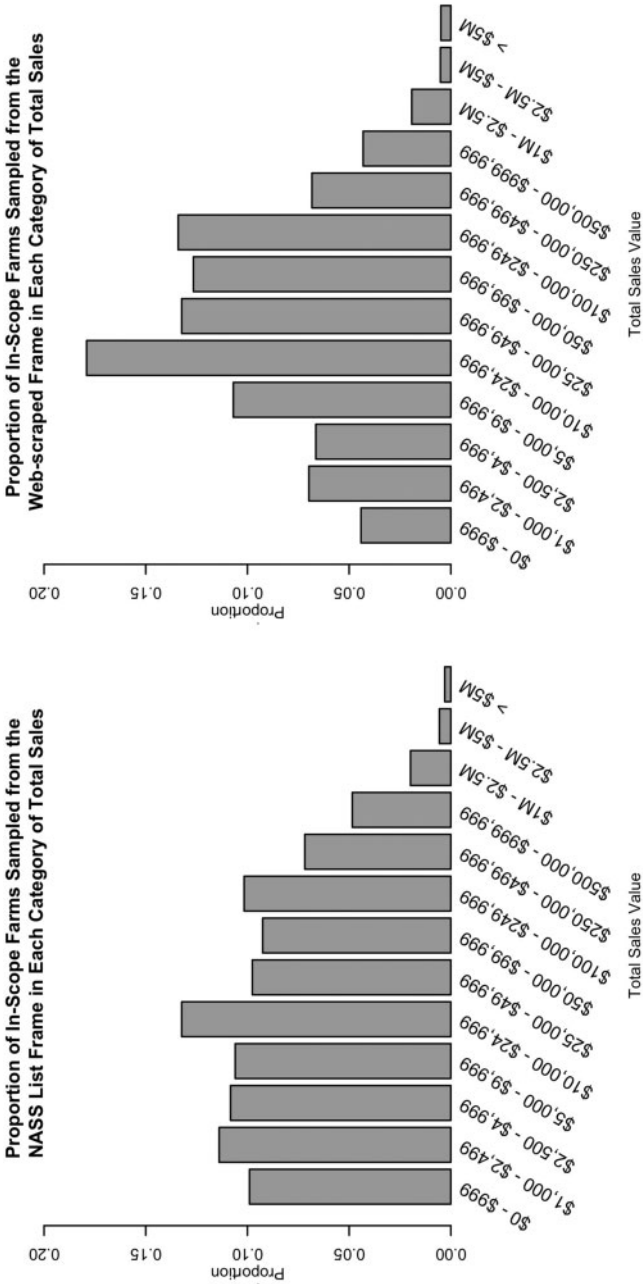

Figure 2. The proportions of in-scope records identified from the NASS list frame sample (left) and web-scraped list sample (right) in each category of total sales.

record's sample inclusion, response, coverage, and misclassification probabilities (3). Sample inclusion probabilities were known for all records.

Because farms could not be screened prior to sampling, a survey response was necessary to determine whether a record was a LF farm. This required the record to be in the sample and to respond during the survey. Thus, two additional assumptions are required for estimation ((4) and (5) in Section 2). The first, that $P(S|F) = P(S)$, holds because all records on the NASS and web-scraped list frames within a sampling stratum had an equal probability of inclusion.

The second assumption is that the probability of response of a sampled record is independent of whether or not the record is a LF farm, $P(R|S, F) = P(R|S)$. The NASS list frame sample produced 12,305 responses that were either out-of-scope records or out-of-business (60.9 percent of respondents), indicating that many non-LF farms responded. The scope of non-respondents is unknown, and hence it is not known whether this assumption is satisfied. If LF farms are more likely to respond than non-LF farms, response probabilities will be underestimated and the resulting estimates will be upwardly biased. Response rates between in-scope and out-of-scope records were not expected to differ as most out-of-scope records were agricultural businesses with no local sales and likely to respond to an agricultural survey. Multiple surveys with independent scoping from NASS list frame records and web-scraped list records would help alleviate this assumption by allowing response probabilities to be estimated from only in-scope records.

Local food farms do not have equal capture probabilities within the same list (figure 2). For example, larger farms are often more likely to be identified when creating a sampling frame, leading to higher coverage percentages of large farms as compared to small farms. In addition to variation among farms, coverage probabilities between frames varied with farm characteristic. As an illustration, it is believed that web scraping is more likely to identify small farms than the traditional approach, and this is supported by the data shown in figure 2. Response rates can also vary with farm characteristics. To adjust for this differential catchability, the probability of coverage was estimated using logistic regression, and the probability of response was estimated by grouping records based on farm characteristics.

To estimate coverage probabilities, a logistic regression model was fit using the in-scope records from the web-scraped list sample to estimate model coefficients. The fitted model was then used to predict the coverage probabilities of in-scope NASS list frame records. Low in-scope rates and high coverage of the NASS list frame resulted in only 368 in-scope LF farms in the web-scraped sample that were not linked to a NASS list frame farm. This limited the number of variables considered during modeling. The fitted model included the following covariates: total value of local sales, indicator variables for each marketing channel (consumer, retail, and institution/intermediate), indicator variables for crop and livestock farm types, and an interaction between the

presence of sales to retail and consumer marketing channels. The model's dependent variable was an indicator variable of whether an in-scope record in the web-scraped sample was on the NASS list frame; that is, it had a value of 1 if a LF farm in the web-scraped sample was on the NASS list frame and a value of 0 otherwise. Based on the Hosmer and Lemeshow goodness-of-fit test (Hosmer et al. 2013), the model provided an adequate fit to the data ($\chi^2 = 4.9944$, $df = 8$, $p = .7582$).

Random samples of 682 in-scope records and 714 out-of-scope records were drawn from the NASS list frame sample respondents to estimate misclassification probabilities. A record's scope determined from the misclassification survey was considered its true scope. In total, 481 sampled in-scope records responded and 76 (15.8 percent) were determined out-of-scope from the misclassification questionnaire and subsequent review. Briefly, 360 of the out-of-scope records sampled responded to the misclassification survey, 60 (16.7 percent) of which were determined to be in-scope. The probabilities associated with misclassification were assumed to be constant for all responding records.

In addition to national estimates, regional estimates were produced for six regions, which covered the United States. Based on prior information, the sample was designed to produce state-level estimates for the 36 states with the greatest LF production, with the remaining 14 states represented at the regional level. In total, 393 estimates were reported at the national level, 33 estimates were reported at the regional level, and 15 estimates were reported at the state level. Low in-scope rates resulted in fewer fit-for-use estimates at state and regional levels than anticipated.

The national estimate for LF farms in the United States was 167,009 operations, with a standard error of 5,845 farms (CV = 3.5 percent). State estimates ranged from 1,269 farms in New Hampshire to 14,315 farms in California, with smaller farm numbers in some unreported states. National estimated local sales were $8,747,222,197 with a standard error of $892,216,664 (CV = 10.2 percent). State-level estimates for local sales ranged from $28,235,565 in Utah to $2,869,192,534 in California (USDA NASS 2015).

Local Food Marketing Practices estimates were reasonable compared to previous estimates for local food farming. In 2012, the USDA's Economic Research Service estimated 163,675 LF farms and $6.1 billion in sales through local marketing channels (Low et al. 2015). Although the farm numbers increased by only 2 percent, sales to local channels increased by 42 percent. The estimated number of LF farms based solely on the NASS list frame sample and disregarding coverage adjustments was 151,122 farms, indicating that the total coverage adjustment was 10.5 percent. The coverage adjustment for the number of U.S. farms from the 2012 Census of Agriculture was 15.1 percent, but varied with size of farm from 2.2 percent for farms with sales of $500,000–$999,999 to as high as 22.1 percent for farms with <$1,000 in sales (USDA NASS 2019).

4. DISCUSSION

Bird and King (2018) provide several examples of the use of capture–recapture methods to inform public policy, including estimating the size of the homeless population, the extent of human slavery, and the number of civilian deaths during war. In each case, multiple lists were available for these hard-to-identify populations and served as a foundation for analysis. Federal statistical agencies often have access to numerous sources for list frame development. As an example, as part of its list building efforts, NASS routinely acquires administrative data from the USDA Farm Services Agency and the Internal Revenue Service and a variety of lists, such as the membership lists of farm organizations. As a consequence, no alternative lists are known to be available, and the options for assessing undercoverage of the list frame become limited. Area frames provide complete coverage and can be used to assess undercoverage. However, when the population units are widely dispersed and transient, data collection costs can be prohibitive, and misclassification of sample units may be high (Abreu et al. 2010). In these cases, a web-scraped list frame and capture–recapture methods may be a viable approach. In this section, some best practices for conducting surveys when using web-scraped lists to assess undercoverage using capture–recapture methods and open research questions are discussed.

Web-scraped list frames may be especially effective for establishment surveys for which population units are likely to have a web presence either through individual websites, membership in an organization with a website, or through permit or other filings at the local or state levels.

Extensive thought and effort need to be devoted to developing a set of keywords that will effectively identify members of the population, including those with rare characteristics. Developing a set of keywords that will lead to a large and inclusive web-scraped list generally results in the inclusion of units that are not part of the population of interest. As an example, for the local food study, restaurants and hair salons using organic products were included in the initial web-scraped list. Thus, it is important to allocate additional time and effort to the development of a list of negative keywords that can be used to remove the units that should not be on the list frame prior to manual review. For new lists, the development of the negative keyword list could extend into the early stages of web scraping so that more relevant negative keywords can be identified. Text analysis could be a useful tool in determining additional positive and negative keywords and phrases. As an example, the corpus of web-scraped text data could be divided into two groups: in-scope LF farms and out-of-scope LF farms. Then one could observe which words or phrases tend to appear in each group. Alternatively, a text classification model that predicts whether a web search result is in-scope could be fit to the corpus of web-scraped text data and the most predictive words and phrases identified. Without a negative keyword list, extensive manual effort will be spent to

remove the nonpopulation units, or resources will be wasted in the survey process.

For the two list frames to be linked, each record on a web-scraped list must have sufficient information to determine whether or not it is also on the traditional list frame. For this study, the record had to have at least a name and city/state/zip code for it to link to a NASS list frame record; records that did not have sufficient contact information were dropped. Developing cost-effective methods for acquiring missing information needed for record linkage can lead to a more complete web-scraped list.

Although the use of negative keywords, manual review, and record linkage will lead to some nonpopulation units being identified and removed from the web-scraped list frame, others will remain. A survey that screens each remaining record on the list to determine whether or not it is in the population is a helpful tool. By eliminating records that are not in the population, the survey sample will have higher in-scope rates.

Unlike most capture–recapture studies, the number of records on the portion of the traditional list frame but not on the web-scraped list, the portion of the web-scraped list but not on the traditional list frame, and the portion common to both lists can be determined in advance of drawing the sample. Using both lists in the same sample areas has operational advantages. Furthermore, the sample can be designed to draw sufficient numbers of records from these three segments to provide a good foundation for obtaining modeled estimates of the capture–recapture weights. Determining the best sample design for this type of capture–recapture study is a topic for future research.

As noted in Section 2.2, capture–recapture estimation is based on five assumptions. For applications similar to this study, three of the five are fully addressed. Because the survey is sent to both samples at the same time, the population is closed; units do not leave or enter the population between the time of the two surveys. Furthermore, a population unit's propensity to respond is the same whether it is in the sample from the traditional or web-scraped list frame. Although it is unlikely that all population units are equally likely to be caught in each sample, adjustments for this differential catchability can be made in the modeling process. In this study, the probabilities of coverage and response were estimated using logistic regression with the independent variables representing population characteristics chosen to adjust for differential catchability and response, respectively. The assumption of record linkage without error underlies this as well as more traditional approaches and is unlikely to be met. The validity of the assumption of independence of the two list frames is an open question that should be explored. More robust capture–recapture estimation is not as sensitive to departures from these assumptions, and capture–recapture models can account for violations of one or more assumptions (for a review, see Chao and Huggins 2005). Acquiring a better understanding of the extent to which these assumptions are not met and

identifying and/or developing robust estimation methods are important areas of future research.

Although web-scraped list frames have been constructed before (Webb et al. 2015; Young et al. 2018), to our knowledge, the LFMP Survey is the first *survey* in which a web-scraped list frame was used to adjust for the undercoverage of a traditional list frame to produce official statistics. Although implementing new methodology in a production environment is always challenging, national, regional, and state-level estimates, providing valuable information about the U.S. local food industry were produced. The combined use of web-scraped list frames and capture–recapture methods has the potential to provide a cost-effective approach to precise estimates with valid measures of uncertainty.

ACKNOWLEDGEMENTS

The authors thank Christy Meyer and Sarah Goodale for providing production information and rerunning the local food analysis. The authors are also appreciative of the careful review of earlier versions of this article by an associate editor and four reviewers; their comments led to an improved article. The first author was at the USDA National Agricultural Statistics Service when this research was conducted.

REFERENCES

Abreu, D. A., J. S. McCarthy, and L. A. Colburn (2010), "Impact of the Screening Procedures of the June Area Survey on the Number of Farms Estimates," *Research and Development Division*. RDD Research Report Number RDD-1003. Washington, DC: USDA, National Agricultural Statistics Service. Available at https://ideas.repec.org/p/ags/unasrr/234374.html. Last Accessed May 8, 2020.

Amaya, A., S. Zimmer, M. Morton, and R. Harter (2018), "Does Undercoverage on the U.S. Address-based Sampling Frame Translate to Coverage Bias?" *Sociological Methods & Research*. DOI: https://doi.org/10.1177/0049124118782539. Last Accessed May 8, 2020.

American Association for Public Opinion Research (AAPOR (2016)), *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys* (9th ed.), AAPOR. Available at https://www.aapor.org/AAPOR_Main/media/publications/Standard-Definitions20169theditionfinal.pdf. Last Accessed May 8, 2020.

Beaumont, J. F., and Z. Patak (2012), "On the Generalized Bootstrap for Sample Surveys with Special Attention to Poisson Sampling," *International Statistical Review*, 80, 127–148.

Bird, S. M., and R. King (2018), "Multiple Systems Estimation (or Capture-Recapture Estimation) to Inform Public Policy," *Annual Review of Statistics and its Application*, 5, 95–118. DOI: https://doi.org/10.1146/annurev-statistics-031017-100641. Last Accessed August 12, 2020.

Blumberg, S. J., and J. V. Luke (2019), "Wireless Substitution: Early Release of Estimates from the National Health Interview Survey, July–December 2018," National Health Interview Survey Early Release Program, National Center for Health Statistics. Available at https://www.cdc.gov/nchs/data/nhis/earlyrelease/wireless201906.pdf. Last Accessed June 29, 2019.

Brown, J. J., C. Sexton, O. Abbott, and P. A. Smith (2019), "The Framework for Estimating Coverage in the 2011 Census of England and Wales: Combining Dual-system Estimation with Ratio Estimation," *Statistical Journal of the IAOS*, 35, 481–499. DOI: https://doi.org/10.3233/SJI-180426. Last Accessed August 12, 2020.

Cavallo, A. (2018), "Scraped Data and Sticky Prices," *The Review of Economics and Statistics*, 100, 105–119. Available at https://www.nber.org/system/files/working_papers/w21490/w21490.pdf. Last Accessed June 29, 2019.

Cavallo, A., and R. Rigobon (2016), "The Billion Prices Project: Using Online Research for Measurement or Research," *Journal of Economic Perspectives*, 30, 151–178. Available at https://pubs.aeaweb.org/doi/pdfplus/10.1257/jep.30.2.151. Last Accessed June 29, 2019.

Chao, A., and R. M. Huggins (2005), "Modern Closed Population Capture-Recapture Models." in *Handbook of Capture-Recapture Analysis*, eds. S. C., Amstrup, T. L. McDonald, and B. F. J. Manly, pp. 58–87, Princeton: Princeton University Press.

Chipperfield, J., J. Brown, and P. Bell (2017), "Estimating the Count Error in the Australian census," *Journal of Official Statistics*, 33, 43–59.DOI: https://doi.org/10.1515/jos-2017-0003. Last Accessed May 8, 2020.

Chow, T. E., Y. Lin, and W. D. Chan (2011), "The Development of a Web-Based Demographic Data Extraction Tool for Population Monitoring," *Transactions in GIS*, 15, 479–494. DOI: https://doi.org/10.1111/j.1467-9671.2011.01274.x. Last Accessed August 12, 2020.

da Silva, A. D., M. P. S. de Freitas, and D. G. C. Pessoa (2015), "Assessing Coverage of the 2010 Brazilian Census," *Statistical Journal of the IAOS*, 31, 215–225. DOI: 10.3233/SJI-150897. Last Accessed August 12, 2020.

Day, C. (1996), "Record Linkage II: Experience Using AUTOMATCH for Record Linkage in NASS," US Department of Agriculture, National Agricultural Statistics Service STB Research Report Number STB-96-01. Available at https://www.nass.usda.gov/Education_and_Outreach/Reports,_Presentations_and_Conferences/Survey_Reports/Record%20Linkage%20II%20Experience%20Using%20AUTOMATCH%20for%20Record%20Linkage%20in%20NASS.pdf. Last Accessed June 29, 2019.

Federal Committee on Statistical Methodology. (2001), "Statistical Working Paper 31: Measuring and Reporting Sources of Errors in Surveys." Available at https://nces.ed.gov/fcsm/pdf/spwp31.pdf. Last Accessed June 29, 2019.

Hogan, H. (1993), "The 1990 Post-Enumeration Survey: Operations and Results," *Journal of the American Statistical Association*, 88, 1047–1066.

———. (2003), "The Accuracy and Coverage Evaluation: Theory and Design," *Survey Methodology*, 29, 129–138.

Hosmer, D. W., Jr., S. A. Lemeshow, and R. X. Sturdivant (2013), *Applied Logistic Regression* (3rd ed.), Hoboken NJ: Wiley.

Landers, R. N., R. C. Brusso, K. J. Cavanaugh, and A. B. Collmus (2016), "A Primer on Theory-Driven Web Scraping: Automatic Extraction of Big Data from the Internet for Use in Psychological Research," *Psychological Methods*, 21, 475–492. DOI: 10.1037/met0000081.

Lipps, O., N. Pekari, and C. Roberts (2015), "Undercoverage and Nonresponse in a List-Sampled Telephone Election Survey," *Survey Research Methods*, 9, 71–82. DOI: https://doi.org/10.18148/srm/2015.v9i2.6139. Last Accessed February 20, 2021.

Lohr, S. (2009), *Sampling Design and Analysis* (2nd ed.), Boston MA: Cengage Learning.

Low, S., A. Adalja, E. Beaulieu, N. Key, S. Martinez, A. Melton, A. Perez, K. Ralston, et al. (2015), "Trends in U.S. Local and Regional Food Systems: A Report to Congress," (AP-068). U.S. Department of Agriculture, Economic Research Service. Available at http://scholarship.sha.cornell.edu/articles/1058. Last Accessed May 8, 2020.

Mashreghi, Z., D. Haziza, and C. Léger (2016), "A Survey of Bootstrap Methods in Finite Population Sampling," *Statistics Surveys*, 10, 1–52. DOI: 10.1214/16-SS113. Last Accessed May 8, 2020.

Mule, T. (2012), "2010 Census Coverage Measurement Estimation Report: Summary of Estimates of Coverage for Persons in the United States," DSSD 2010 Census Coverage Measurement Memorandum Series #2010-G-01. Washington DC: U.S. Census Bureau. Available at https://www.census.gov/coverage_measurement/pdfs/g01.pdf. Last Accessed May 8, 2020.

Otis, D. L., K. P. Burnham, G. C. White, and D. R. Anderson (1978), "Statistical Inference from Capture Data on Closed Animal Populations," *Wildlife Monographs*, 62, 3–135. Available at https://pubs.er.usgs.gov/publication/70119899. Last Accessed February 20, 2021.

Pledger, S. (2000), "Unified Maximum Likelihood Estimates for Closed Capture-Recapture Models using Mixtures," *Biometrics*, 56, 434–442. DOI: https://doi.org/10.1111/j.0006-341X.2000.00434.x. Last Accessed February 20, 2021.

Pollock, K. H., S. C. Turner, and C. A. Brown (1994), "Use of Capture-Recapture Techniques to Estimate Population Size and Population Totals when a Complete Frame is Unavailable," *Survey Methodology*, 20, 117–124.

Rao, J. N. K., W. C. F. J, and K. Yue (1992), "Some Recent Work on Resampling Methods for Complex Surveys," *Survey Methodology*, 18, 209–217. Available at https://www150.statcan.gc.ca/n1/pub/12-001-x/1992002/article/14486-eng.pdf. Last Accessed February 20, 2021.

Rhodes, B. B., A. F. Kim, and B. R. Loomis (2015), "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 http://sitesusa.s3.amazonaws.com/wp-content/uploads/sites/242/2016/03/H3_Rhodes_2015FCSM.pdf. Last Accessed February 20, 2021.

Sala, E., and R. Lillini (2017), "Undercoverage Bias in Telephone Surveys in Europe: The Italian Case," *International Journal of Public Opinion Research*, 29, 133–156. DOI: https://doi.org/10.1093/ijpor/edv033. Last Accessed August 12, 2020.

Sartore, L., K. Toppin, L. J. Young, and C. Spiegelman (2019), "Developing Integer Calibration Weights for Census of Agriculture," *Journal of Agricultural, Biological and Environmental Statistics*, 24, 26–48. DOI: https://doi.org/10.1007/s13253-018-00340-4. Last Accessed August 29, 2020.

Seber, G. A. F. (2002), *The Estimation of Animal Abundance and Related Parameters* (2nd ed.), Caldwell NJ: Blackburn Press.

U.S. Census Bureau (2004), "Accuracy and Coverage Evaluation of Census 2000: Design and Methodology." Available at http://www.Census.gov/prod/2004pubs/dssd03-dm.pdf. Last Accessed May 8, 2020.

U.S. Census Bureau (2008), "2010 Coverage Measurement Estimation Methodology," DSSD 2010 Census Coverage Measurement Memorandum Series (Report Number #2010-E-18). Available at https://www2.census.gov/programs-surveys/decennial/2010/technical-documentation/methodology/ccm-workshop/2010-e-18.pdf. Last Accessed May 8, 2020.

US Department of Agriculture National Agricultural Statistics Service (USDA NASS) (2015), Local Food Marketing Practices Survey. Available at https://www.nass.usda.gov/Surveys/Guide_to_NASS_Surveys/Local_Food/index.php. Last Accessed May 8, 2020.

US Department of Agriculture National Agricultural Statistics Service (USDA NASS) (2019), Appendix A: Census Agriculture Methodology. Available at https://www.nass.usda.gov/Publications/AgCensus/2017/Full_Report/Volume_1,_Chapter_1_US/usappxa.pdf. Last Accessed May 8, 2020.

Vargiu, E., and M. Urru (2013), "Exploiting Web Scraping in a Collaborative Filtering-based Approach to Web Advertising," *Artificial Intelligence Research*, 2, 44–54. DOI: 10.5430/air.v2n1p44. Last Accessed February 20, 2021.

Wallgren, A., and B. Wallgren (2016), "Frames and Populations in a Register-based National Statistical system," *Journal of Mathematics and Statistical Science*, 2, 208–216. DOI: http://www.ss-pub.org/wp-content/uploads/2016/04/JMSS15121601.pdf. Last Accessed February 20, 2021.

Webb, L. M., D. M. Gibson, Y. Wang, H. C. Chang, and M. Thompson-Hayes (2015), "Selecting, Scraping, and Sampling Big Data Sets from the Internet: Fan Blogs as Exemplar," in *SAGE Research Methods Case*, London: SAGE Publications, Ltd., Available at https://pdfs.semanticscholar.org/48f8/4535a0607c7e64f87c61faf176ef3c5161fc.pdf?_ga=2.244060101.1587555821.1595364779-1428202309.1595364779. Last Accessed August 12, 2020.

Wolter, K. M. (1986), "Some Coverage Error Models for census data," *Journal of the American Statistical Association*, 81, 338–346.

Young, L. J., M. Hyman, and R. 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, 323–340. DOI: 10.2478/JOS-2018-0015. Last Accessed February 20, 2021.

Young, L. J., A. Lamas, and D. Abreu (2017), "The 2012 Census of Agriculture: A Capture-recapture Analysis," *Journal of Agricultural Biological and Environmental Statistics*, 22, 523–539. DOI: 10.1007/s13253-017-0303-8. Last Accessed February 20, 2021.

Zhang, Z., and W. Tang (2016), "Analysis of Spatial Patterns of Public Attention on Housing Prices in Chinese Cities: A Web Search Engine Approach," *Applied Geography*, 70, 68–81. DOI: https://doi.org/10.1016/j.apgeog.2016.03.004. Last Accessed May 8, 2020.