Capture-Recapture Estimation of Local Food Farms in the United States

Michael Hyman*

Abstract

Over the past decade, the popularity of farmer's markets and locally produced agricultural products have grown rapidly. The U.S. Department of Agriculture's (USDA's) National Agricultural Statistics Service (NASS) conducts many surveys; however, no survey currently focuses directly on the marketing practices of operations that distribute products locally. In 2016, NASS initiated a project with the objective to learn more about this agricultural system. Samples were selected from (1) the NASS sampling frame of potential farms, and (2) a new frame built via advanced web-scraping technology. A survey questionnaire specific to local food marketing practices was distributed to each of the sampled operations. Dual system estimation methodology was used to estimate the total number and sales of local foods operations at the U.S. and state levels. In this study, the dual system estimator used to estimate the population total for agricultural operations that distribute food locally is described. National level results from this new study are discussed, and some of the statistical challenges addressed.

1. Introduction

In recent years, concerns over the sustainability of current agricultural systems, as well as the social, economical, and environmental benefits of locally produced agriculture have led to increased support and publicity of local farming. Local foods often costs less than the equivalent food bought on the international market or from a supermarket, due to reduced transportation costs and more direct distribution from farm to table [1, 2, 3, 4]. Local foods reduce the environmental burden of agricultural transportation by reducing the distance from food production to food consumption [1, 5]. As a result, the demand for local foods has grown. The increase in farmers' markets and community supported agriculture (food delivery subscription schemes) is a clear indication that consumer demand for local foods is continuing to grow rapidly [6, 7, 8, 9]. These marketing channels, which allow consumers to purchase agricultural products directly from the producer, often supply fresher products and a more transparent distribution path [6, 3]. As consumer demand has increased, many supermarkets and restaurants are beginning to promote themselves by offering

*National Agricultural Statistics Service

locally grown food options [10, 11, 12]. Other institutions, including schools, colleges, and hospitals have also shown interest in providing local foods [13, 2]. The benefits of locally produced agriculture have led some to believe local foods will play an important role in the development of sustainable agricultural systems at the community, national, and global levels. [10, 14].

As the demand for local foods increases, it becomes more important to learn about the operations that produce locally grown products and this unique agricultural system as a whole. This will allow policy amendment that encourages local food products and reduction of barriers that have prevented consumers and organizations from purchasing local foods directly from farmers. The first step in learning about the local food agricultural system is to quantify the prevalence of local farming and characteristics about the operations that participate.

The National Agricultural Statistics Service (NASS) of the U.S. Department of Agriculture (USDA) conducts hundreds of surveys annually regarding the nation's crop acreage, agricultural production, livestock inventory, commodity prices, and more. These surveys and successive reports enable the USDA to understand the quantity, production, and characteristics of agricultural businesses of different types in the U.S. Operations that produce and distribute local foods are included in current and past surveys administered by NASS, however, no survey currently focuses directly on these operations' marketing practices. The Census of Agriculture, the most comprehensive survey that NASS conducts on all agricultural businesses in the US, contains several questions regarding local food sales. This includes whether an operation has direct to consumer sales and the quantity of these sales. However, the primary focus of NASS is large production agriculture, and this information is inadequate to fully understand the dynamics of local food businesses or identify all businesses that sell products locally.

While large production agricultural operations can distribute products locally, most local food operations differ from large-scale farms, raising additional challenges in producing accurate estimates. The majority of the countries agriculture is located in rural areas, whereas local food farms are often found closer to urban areas to reduce transportation distance [14]. Because of their proximity to urban areas, local food operations tend to be much smaller in acreage than agricultural operations in the rural U.S [15]. Restrictions in physical space and reduced marketing from local food farms limit production and sales [16, 17]. Operations that produce livestock typically raise far fewer animals than large production livestock farms [9]. As a result of smaller scaled production, local food businesses go in and out of business more frequently than large-scale farms [10]. Limited space also results in local food operations being more dispersed than most farms [16]. Local food producers typically grow more diverse products than large-scale farms, often selling small quantities of numerous agricultural products at farmer's markets [7, 6]. Because of these factors, local food businesses are challenging to identify, and the estimates of local food distribution based on NASS's current data have not been as precise as other agricultural sectors.

In 2015, NASS received approval and funding to conduct the Local Food Marketing Practices (LFMP) survey. This is the first survey NASS conducted focused directly on local food businesses in the U.S. and the products that they produce. The objective was to estimate the quantity and characteristics of agricultural operations that distribute food locally at the U.S. and state level. In addition to conducting a new survey, novel methods of list building via web scraping have been employed to assess under-coverage of these businesses by the NASS's current list of agricultural operations.

To accurately estimate the number of agricultural businesses that sell products locally, the target population needs to be defined. Previous research of local agricultural systems defines local food in many different ways, often using a measure of distance or “food miles" [18, 19]. The mileage necessary for a product to be local varies greatly for different regions of the country and is not an ideal indication of locally sourced products at a national level. The USDA defines a farm as an operation from which $1,000 or more of agricultural goods were sold or normally would have been sold within a year. In addition to fulfilling the USDA criteria for being a farm, an operation is considered part of the target population of local food farms if it has sales in one of the following categories:

1. Direct to the consumer
2. Direct to a retail market
3. Direct to an institution
4. Direct to an intermediate market that mark the product as “local”.

The quantity of sales to these marketing channels is not a consideration for an operation to be considered part of the target population. If an operation meets the USDA's criteria of a farm and reports sales of any quantity to any of the above marketing channels, it is considered “in-scope” for the purposes of this study, that is, a member of the target population of local food farms. Sales information to each of these marketing channels is not known based on current NASS surveys, and thus, a new survey questionnaire, focused directly on the marketing practices of local food operations, is developed and distributed to potential farms.

NASS has extensive experience with estimating population totals, both through modeling and sampling designs. One of the most comprehensive examples of this is the Census of Agriculture, a nation wide survey conducted every five years to estimate the dynamics of all farm activity in the U.S. The Census of Agriculture estimates many population totals regarding the U.S. agricultural industry, including the total number of agricultural operations at the national and state levels, and for many categories of interest. For the 2012 Census of Agriculture, a variation of dual system estimation (DSE) was developed for these estimates. This required two different lists or surveys to adjust for under-coverage. In this case, these are the NASS List Frame (an up-to-date list of potential agricultural operations in the U.S.) and NASS's annual June Area Survey (JAS). This corrects for various sources of error that would be incorporated into estimates by simply using the information that NASS currently has available. Similar methodology was proposed to estimate the quantity and characteristics of local food farms.

The NASS list frame contains all types of agricultural operations in the U.S., including those that distribute food locally, and can be used as one source to estimate the number of local food farms with DSE. However, the JAS is an areal survey and more sparse in the number of operations that it collects. As local food farms are thought to consist of only a small proportion of all agricultural operations, the JAS likely does not contain enough operations that meet the criteria. To use DSE, an alternative list is required. To correct for under-coverage of the NASS list frame, web scraping was used to develop a new list of agricultural operations. Web scraping is automated data collection of publicly available information from websites. The list was built specifically for this project and comprised of U.S. agricultural operations that are likely to have local sales. While web scraping has been used to collect data and possibly to compose a list of records, a web-scraped list has not been used to produced official estimates of population totals using DSE. Operations from the web-scraped list need to be matched to operations on NASS list frame and the data must be applicable to the dual system estimation methodology. In addition, assumptions regarding the lists and their records must be met to produce accurate estimates. In this paper, we discuss these assumptions and whether web scraping is a viable method for list building to implement dual system estimation.

This paper describes and evaluates the survey process and statistical methods used to estimate the total number and characteristics of agricultural operations with local food sales at the national, regional, and state level, and the measurement of error associated with these estimates. Section

2 describes the methods used, including the web-scraping methods for building a second list of operations, the sample that was drawn, data collection, and estimation methodology developed for the total population size and sales of local food farms, and the uncertainty associated with these estimates. Section 3 describes some of the national results that were obtained from this study. Section 4 examines the assumptions involved in the statistical methodology and how closely they were met by employing web-scraping list building methods. In Section 5, conclusions are drawn from this pilot study and some of the statistical challenges that were encountered are discussed.

2. Estimation Methodology

The foundation for the estimation of local food agricultural operations is dual system estimation methodology. This work incorporates ideas from the 2012 U.S. Census of Agriculture, the U.S. Census Bureau, as well as traditional DSE methods developed for estimation of animal populations [20, 21, 22, 23]. Estimation of the total number of local food operations, as well as other characteristics about the population, must account for under-coverage, non-response, and misclassification. Two types of misclassification can affect estimates: incorrectly labeling an in-scope operation as out-of-scope (undercounting), and incorrectly classifying an out-of-scope operation as in-scope (overcounting).

Implementation of DSE requires at least two different lists or capture events used to assess under-coverage. This methodology relies on several assumptions regarding these lists [24]. Firstly, the lists are independent; the inclusion of an operation on one list does not impact the probability of the operation being on the other list. Secondly, the population of interest is closed; no individuals enter or leave the population during sampling. Thirdly, the lists and any samples from the lists must be a random sample of the population. Finally, all individuals captured by both lists can be matched. Fulfillment of these assumptions allows estimation of the population total as the number of records on the first list that are captured by the second list is approximately proportional to the total population that is captured by the first list. Violation of these assumption can result in biased estimates [25]. If violation of any of the assumptions is detected, adjustments can often be made to the estimators [26, 27, 28]. The two lists used in the local foods estimation methodology are the NASS list frame, and a list of potential local food operations built specifically for this project by the Multi-Agency Collaboration Environment (MACE). Because samples were drawn from each of these lists, they are also referred to as sampling frames. Each sample frame is described in further detail.

2.1 Sampling Frames

The primary sampling frame used to estimate local food operations is the NASS list frame. This is a continuously growing collection of potential farms in the U.S., composed by NASS. NASS updates the list frame continuously, adding new or omitted records discovered through other USDA lists, producer association lists, or other sources, to keep the list frame as complete as possible. If a new operation is found that cannot be linked to an operation on the NASS list frame, the record is added as a potential farm. Surveys are then conducted to determine whether the operation has farm status. Thus, records on the NASS list frame are operations that are likely to be farms; however, it is possible that they do not meet the farm criteria, or that they have gone out of business since being added to the list. The list frame contains information about each record including business name, address, person of contact, and any data that has been collected from the business during past surveys. The NASS list frame is not a complete list of all agricultural operations in the U.S. and thus, the quantity of operations that are unaccounted for needs to be assessed.

To adjust for under-coverage, a second list of agricultural operations was created by the Multi-Agency Collaboration Environment (MACE). The MACE sampling frame was built using web-scraping technology and consists of potential local food operations. Keywords specific to local foods were searched in various internet sources (e.g., Google Maps, YP.com, etc.), and search results were automatically scanned for indicators of local food sales. In addition, other web lists containing operations qualifying as local food businesses were scraped and added to the sampling frame. The final list was comprised of all search results with any evidence of agricultural sales direct to consumers or an intermediate source, thus, creating a list of records likely to be to meet the definition of local food farms. The list was delivered to NASS with no additional follow-up surveys or vetting of the operations for farm or local foods status.

To implement DSE, the NASS list frame records are first matched to the MACE frame. Records are linked based on an operation's name, address, phone number, person of contact or other criteria. Linkage software is used and returns a likelihood that records are a match. This helps account for errors or discrepancies between names, address or other inputs from the two lists. Records continued to be investigated throughout the project to ensure the most accurate record linkage between lists. Table 2.1 shows the number of agricultural operations on the NASS list frame, the MACE list frame, and the number of operations that were matched between the frames. In addition to record linkage, duplicate records are also identified and removed.

Table 1: The total number of records on each sampling frame and the number of records linked between the MACE and NASS list frames.

Frame	Total Number of Records on Frame
NASS List Frame	2,006,626
MACE Frame	33,262
Total Records on Both Frames	28,986

2.2 Sampling

Independent samples were drawn from the NASS list frame and the MACE list frame. The MACE list frame is composed of 33,262 records, identified by MACE to be potential local food agricultural businesses. These records were grouped by state, and a systematic sample was selected from the entire frame. This ensured that that sample size from each state was approximately proportional to the total number of operations from each state on the frame. The total sample sizes from both lists were determined based on anticipated response rates, in-scope rates, and desired coefficients of variation for the estimator. The sample from the MACE list frame contained 19,365 records.

Limited information regarding local food sales is available for some NASS list frame operations from the 2012 Census of Agriculture. This information included whether the operation reported any direct to consumer sales in 2012 and the quantity of these sales. Furthermore, information was obtained from the NASS regional field offices regarding operations they believed to be involved in the local foods market. Using the available information, the NASS list frame is parsed into four sampling groups:

Operations were further stratified into one of the following groups prior to sampling:

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

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

C) Farms that were identified by NASS regional field offices as having local food sales but which are not in groups A or B,

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

Sampling groups A and B have indications of being part of the target population. Operations in

Table 2: The total number of records in each sampling frame and sampled from each frame. The last column represents the number of potential farms sampled from that frame that were matched to a record on the opposite frame.

Sampling Frame	Total Number of Records on Frame	Sample Size	Records Linked to the other Frame
NASS List Frame	2,006,626	24,907	2,509
MACE Frame	33,262	19,365	15,669
Records on Both Frames	28,986	1,466	–

group C are thought to have a higher probability of being in the target population than a randomly chosen operation, although there is no objective evidence. Sampling group D is thought to be composed largely of non-target operations.

In addition to being stratified into the these groups, records in groups A and B are also stratified based on reported or estimated local food sales during the 2012 Census of Agriculture. Records in sampling group A are stratified into five additional sampling strata, with each strata corresponding to the set of records composing approximately one fifth of the total state level local food sales. Records in stratum 5 are operations reporting the largest local food sales and were automatically included in the survey (i.e., probability of being sampled equals 1). The average value of local food sales at the state level from sampling group A is used to infer an average value of local food sales for records in sampling groups B and C. The average value of local food sales in stratum 1 of sampling group A (records with the smallest local food sales) is used to infer an average value of direct food sales for records in sampling group D. Using these sampling strata and local food sales, sample sizes from each group and strata where determined to control the Type I and Type III errors, as well as the coefficient of variation of state local food sales. The sample from the NASS list frame consisted of 24,907 records. Of these records, 1,466 matched records that were also sampled from the MACE list frame. Table 2.2 shows the size of each sampling frame, the total samples sizes drawn from each frame, and the number of records from each sample that match records on the other list.

2.3 Survey and Data Collection

The Local Food Marketing Practices (LFMP) survey is a questionnaire designed specifically to inquire about sales and marketing practices of farms that distribute products locally. Questionnaire content and format were evaluated by NASS through a specifications process, where requests for changes were evaluated and approved or disapproved. NASS survey methodologists also conducted cognitive interviews before the questionnaire was finalized. All modes of data collection were tested prior to obtaining responses. The survey contained questions assessing specific information regarding operation characteristics, specific information regarding local food sales, and demographics of the operation. In addition to asking marketing practice questions, all survey instruments collected information to verify the sampled unit, determine any changes in the name or address, identify any partners to detect possible duplication, and verify that the operation is in fact in the local food farms population.

Data collection was performed through a mailed paper version of the questionnaire, as well as, web and Computer Assisted Telephone Interview (CATI) instruments, built to model the paper survey. To increase response rates and reduce response burdens, respondents received a pre-survey postcard in March 2016, and the questionnaire, along with a cover letter and instructions for EDR reporting were mailed in April 2016. Mail, web, telephone and face-to-face interview modes of data collection were utilized for the survey. Respondents who did not return their survey by the end of May 2016 were sent a follow-up mailing at that time. In June 2016, NASS began face-to-face enumeration for any remaining non-respondents. Data collection concluded in August 2016.

As survey data were collected and captured, records were edited for consistency and reasonableness using automated systems. The edit logic ensures administrative coding follows the methodological rules associated with the survey design. Relationships between data items on the current survey were verified. The edit determined the status of each record to be either “dirty” or “clean”, where dirty records indicate possible errors in the response. Dirty records were either updated or certified by an analyst to be accurate. Corrected data were reedited interactively.

Once data were cleaned, NASS determined the local food farm status of each record. The operation was marked as in-scope for this project if they qualified as a farm to the USDA and reported sales to any of the marketing channels described in Section 1, to be local. Qualification as a local food farm is based solely on responses to the LFMP survey, and does not consider other surveys conducted by NASS. To obtain estimates for local food businesses at the state and national levels, in-scope operations were weighted to account for under-coverage, non-response, and misclassification of the NASS list frame and survey responses.

2.4 Estimation

DSE weights are calculated for each NASS list frame record that is identified as a member of the target population, based on the probability that the operation is identified. The probability that a local food operation is identified by the NASS list frame, π_C, is unknown and must be estimated. For a local food farm to be identified by the NASS list frame, it must be on the NASS list frame, selected in the sample, respond to the LFMP survey, and be correctly identified as a member of the target local food farm based on the survey response. Let

- S = 1 if the operation is in the sample,
  = 0 if the operation is not sampled,

- R = 1 if the operation responded to the survey,
  = 0 if the operation did not respond to the survey,

- N = 1 if the operation is on the NASS list frame,
  = 0 if the operation is not on the NASS list frame,

- LF = 1 if the operation is classified as a local food farm based on survey response,
  = 0 if the operation is not classified as a local food farm based on survey response,

- IS = 1 if the operation is truely a local food farm (in-scope),
  = 0 if the operation is not a local food farm.

Then the capture probability, π_C is

$$\begin{aligned} \pi_C &= P\left(S, R, N, LF|IS\right) && (1) \\ &= P(LF|N, R, S, IS)P(N|R, S, IS)P(R|S, IS)P(S|IS) && (2) \end{aligned}$$

The term $P(S|IS)$ represents the conditional probability that a local food farm is selected in the sample. The term $P(R|S, IS)$ is the probability that a sampled local food operation responded to the survey. The probability that a respondent local food farm is a record on the NASS list frame is expressed as $P(N|R, S, IS)$. Finally, $P(LF|N, R, S, IS)$ represents the conditional probability that a local food farm on the NASS list frame is correctly identified as such from the survey response.

In addition to the capture probability, misclassification of records also presents challenges to estimation of population totals. Consider the two types of classification error that can occur. First, a local food farm can be misclassified as an out-of-scope operation with probability [1 −

$\pi(LF|N, R, S, IS)]$. This error results in under-estimating the population. Note that the probability of correctly classifying local food farms (not making this misclassification error) is encountered in determining the probability of capture π_C above. The second type of classification error results when a response to the survey is classified as a local food farm when the operation does not meet the criteria. Errors of this type result in over-estimating the population. To account for this error, the probability of an operation that classified as a local food farm being correctly classified as such must be estimated; that is, $\pi_{NLF} = P(IS|S, R, N, LF)$.

To adjust for under-coverage, non-response, and misclassification, each record sampled from the NASS list frame and classified as a local food farm based on its response to the survey, should be given a weight, w_i, equal to the ratio of the probability of correct classification of a local food farm and its probability of capture. Based on the assumptions, the dual system estimator of the number of local food operations is then

$$\hat{T}_{LF} = \sum_{i \in F} w_i = \sum_{i \in F} \frac{\pi_{NLF,i}}{\pi_{C,i}}. \tag{3}$$

where F represents operations sampled from the NASS list frame that responded to the survey and were classified as a local food farm. To estimate the number of operations by state, marketing channel, or any other category of interest, the set of operations F refers to all identified operations within the particular category of interest.

2.4.1 Probability of Sample

The probability of an operation being sampled is based on the sampling sizes determined for each frame. This is described in further detail in Section 2.2. The probability of being sampled from the NASS list is the sample size of a sampling strata and state divided by the total number of operations in the strata and state combination. That is,

$$\hat{P}(S) = \frac{S_J}{N_J}, \tag{4}$$

where N_J is the total number of records in group J (sampling strata and state), and S_J is the sample size from group J. The sample probability from the MACE list is approximately equal for all operations on the MACE sampling frame. Sampling probabilities are estimated to account for records that belonged to both the NASS and MACE sampling frames.

2.4.2 Probability of Response

The response probability in Equation 1 is conditional on the record being a local food farm. However, the local food status of the record cannot be established without a response to the survey. Thus, it is necessary to make the assumption that

$$P(R|S, IS) \approx P(R|S) \tag{5}$$

to estimate the response probability. This assumption and its implications are discussed in further detail in Section 4.

To estimate the probability of response conditional on a record being sampled, the entire sample from the NASS list frame was used. Only the 24,907 operations sampled from the NASS list frame were used in the estimators and thus, response probabilities for operations found only on the MACE list frame were disregarded. The estimate of the response probability is the ratio of survey respondents to total sample size, S_J, for some subset of records J:

$$\hat{P}(R|S) = \frac{\sum_{i=1}^{S_J} I(\text{record } i \text{ responded to survey})}{S_J}. \tag{6}$$

It is expected that the response probability varies based on different criteria and thus, the response probabilities are estimated for different subgroups of operations. To calculate the probability of response conditional on an operation being sampled, the operations were categorized in groups according to the sampling frame and sampling group they belonged to and the region (six total regions in the U.S.) the operation is in. Operations were first categorized into two groups: operations that are only found on the NASS list frame and operations that are found on both the NASS list frame and the MACE frame. Records from the NASS frame sample are further classified on the basis of sampling group and the region while records from the MACE frame are classified on the basis of region only. In the event that subgroups had fewer than 100 records, response rates from contiguous regions with similar response rates were aggregated. Prior to aggregating two contiguous regions, a chi-square test was used to determine whether aggregated regions had statistically different response rates.

2.4.3 Probability of Coverage

Once the response is obtained, the operation is classified as as either in-scope or out-of-scope as a local food farm. The probability of coverage, $P(C|R, S, IS)$, is estimated. This is the probability that an in-scope record is found on the NASS list frame. These probabilities are modeled using logistic regression. The 4,322 operations that were sampled from the MACE sampling frame, responded to the survey, and determined to be local food farms are used to fit a logistic model. The resulting model is then used to predict the coverage probabilities for 4,410 records on the NASS sampling frame that responded and were determined to be local food farms. Only 368 records sampled from the MACE frame were found to be both in-scope and not identified by the NASS list frame. Thus, model selection was limited and categorical variables from the questionnaire had to be collapsed to form groups with reasonable numbers of records.

The response variable in the logistic regression model has a value of 1 if the operation is on the NASS list frame and a value of 0 if it is a MACE record that is not on the NASS list frame. Variables considered in coverage modeling consist of total value of product sales sold to the four marketing channels of interest (direct to consumers, retail markets, institutions, or intermediate markets), indicators of whether the operation reported sales to each of the four marketing channels (i.e., four indicator variables indicating whether the operation sold direct to consumer, to a retail market, to institutions, or to intermediate markets), and farm type. Although the questionnaire response to farm-type includes 16 categories, the variable used in a model was collapsed to include only three categories: production of crop-based products, production of livestock-based products, and production of both crop- and livestock-based products. The survey response to total value of sales includes 13 different categories. Several of these categories were combined such that the modeling variable included only 9 categories for total value of sales. Collapsing of sales categories is performed by comparing estimated model coefficients using only the total value of sales as a predictor of the operation being contained by the NASS list frame. If subsequent categories (e.g., $10,000 - $24,999 and $25,000 - $49,999) have estimated coefficients that are not statistically different, these categories were combined. Two-way interactions between covariates are also considered in modeling.

Stepwise model selection was used to choose which interactions were included in the final model. Stepwise selection selected covariates that minimized the AIC of the model. The final model for coverage included total value of sales - marketing channel (sales to consumer, retail, institution, intermediate) - farm type (crop, livestock, both crop and livestock) - interaction between farm type and marketing channel - interaction between consumer and retail (i.e., an indicator of whether the operation reported both direct to consumer and retail sales).

After a model is fit using records sampled from the MACE list frame, the predicted coverage probabilities, $\hat{P}(C|R, S, IS)$, are estimated for all sampled, in-scope records on the NASS list frame.

2.4.4 Probability of Misclassification

A supplemental questionnaire was designed to investigate misclassification of operations. Systematic samples were selected from responses to the LFMP survey, including both in-scope and out-of-scope responses. The questionnaire was delivered to each of the sampled operations with screening questions used to determine if the operations meets the criteria of a local food farm. Based on the responses to this survey, misclassification rates are estimated.

As mentioned in Section 2, two types of misclassification are included in the weighting for the estimates. The first type of misclassification, that of falsely identifying a local food farm to be a non-target farm based on it's response to the LFMP survey, corrects for undercounts. This probability, $P(LF|C, R, S, IS)$, is included in the operation's probability of capture. The probability of correct classification is assumed to be equal for all operations and is estimated by

$$\hat{P}(LF|C, R, S, IS) = \frac{N_{LFIS}}{N_{R,LF}}, \tag{7}$$

where $N_{R,LF}$ represents the total number of operations that responded to the misclassification survey and were determined in-scope on the LFMP survey and N_{LFIS} is the total number of operations that responded to the misclassification survey and were determined to be in-scope on both the LFMP survey and the misclassification survey.

The second type of misclassification is incorrectly identifying an out-of-scope operation as in-scope. This adjusts for overcounting, as incorrectly labeling non-target operations as part of the target population will result in an estimate that is too large. The probability of interest is $\pi_{NLF} = P(IS|LF, C, R, S)$, that is, the probability that an operations classified as a local food farm based on it's survey response is in fact a local food farm. Recall that this is the numerator in Equation 3. The probability that an operation is a member of the target population given that it is labeled as such is assumed to be equal for all operations and is estimated by

$$\hat{P}(IS|LF, C, R, S) = \frac{N_{LFIS}}{N_{R,IS}}, \tag{8}$$

where $N_{R,IS}$ represents the total number of operations that responded to the misclassification survey and determined in-scope, and N_{LFIS} is the total number of operations that responded to the misclassification survey and were determined to be in-scope on both the LFMP survey and the misclassification survey.

2.4.5 Weight Calculation

The final estimate for the number of local food farms in the U.S. is the count of in-scope, NASS list frame respondents adjusted for under-coverage, non-response, misclassification, and sampling. The dual system estimator uses a weight for each record to adjust for these components. Recall from Equation 3 that the estimator for the population total is

$$\begin{aligned} \hat{T}_{LF} &= \sum_{i \in F} \frac{\hat{\pi}_{NLF,i}}{\hat{\pi}_{C,i}} = \sum_{i \in F} \hat{w}_i & (9) \\ &= \sum_{i \in F} \frac{\hat{P}_i(IS|LF, C, R, S)}{\hat{P}_i(LF|C, R, S, IS)\hat{P}_i(C|R, S, IS)\hat{P}_i(R|S)P_i(S)}. & (10) \end{aligned}$$

where F represents the total number of local food operations sampled from the NASS list frame, and that responded to the LFMP survey. Estimators of other operation categories are summed over some subset $F_J \subset F$, where F_J represents all responding, sampled operations from the NASS list frame that belong to category of interest (e.g., state level population estimates, estimates of farms of a particular types, estimates of farms of a particular size, etc.). Total local food sales are estimated by weighting the sales from each operation in F by the appropriate weight,

$$\hat{T}_{sales} = \sum_{i \in F} \frac{\hat{\pi}_{NLF,i}}{\hat{\pi}_{C,i}} * \text{sales}_i. \qquad (11)$$

2.4.6 Calibration

The weights associated with each operation are calibrated. Calibration reduces state-level estimate variance resulting from weights that are largely inflated by small sample probabilities. The record weighting methods were applied to produce 10 adjusted, national level estimates: (1) national level farm number, (2) national level local food sales, (3) direct to consumer farm number, (4) direct to consumer sales, (5) three categories of total local food sales (<\$10,000, \$10,000 - \$99,999, ≥\$100,000), (6) three categories of direct to consumer sales. These 10 estimates provided targets for the calibration process.

An algorithm is used to restrict each weight to a maximum allowable limit, such that the calibrated weights sum to each target, within a specified error. This algorithm is repeated on a sequence of possible maximum values that weights can obtain. The value that minimized the summed absolute error between the calibrated weights and the 10 target values is determined to be the optimal maximum allowable weight. Based on this criteria, weights are restricted to 550;

records with a sample probability of 1, had a maximum allowable weight of 10. The calibrated weights are then integerized. The calibrated weights are used for national, regional, and state level estimates using equations 9 and 11.

2.4.7 Measures of Uncertainty

Nonparametric bootstrapping is used to quantify uncertainty associated with the local food farm estimates. Standard errors are estimated after calibration and integerizing the weights. Implementation of bootstrapping requires drawing B independent samples from the sample population of p weighted records. In each bootstrap sample S_b, for $b = 1, ..., B$, each weighted record i is sampled $M_i^{(b)}$, where $M_i^{(b)} \sim \text{Bin}(\hat{w}_i,\ \frac{1}{\hat{w}_i})$. Once sample S_b is drawn, the bootstrap population estimate is computed as

$$\hat{y}^{(b)} = \sum_{i=1}^{p} \hat{w}_i M_i^{(b)}. \tag{12}$$

If the estimate of interest is for a total quantity (e.g., sales, acreage, etc.), then the bootstrap estimate is

$$\hat{y}^{(b)} = \sum_{i=1}^{p} a_i \hat{w}_i M_i^{(b)}. \tag{13}$$

where a_i is the quantity of interest for operations i. The variance is estimated as

$$var(\hat{y}^{(b)}) = \frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{y}^{(b)} - \bar{y} \right) \tag{14}$$

and the coefficients of variation is calculated as

$$CV(\hat{y}^{(b)}) = \sqrt{\frac{1}{B-1} \sum_{b=1}^{B} \left(\frac{\hat{y}^{(b)}}{\bar{y}} - 1 \right)} \tag{15}$$

where $\bar{y} = \sum_{b=1}^{B} \hat{y}^{(b)}$. The variance estimator in Equation 14 is the Horvitz-Thompson estimator for population totals and under the proper conditions, is an unbiased estimator [29, 30].

3. Results

Results from the Local Food Marketing Practices Survey were released to the public on December 20th, 2016. Results were reported at the national, regional and state levels. Regional results were reported for 6 regions in the U.S., while state results were reported for 36 states and the remaining 14 states reported as 6 regions. These 14 unreported states have sample populations that are too small to produce estimates with the accuracy that NASS requires. In addition to total population of local food farm numbers, the proposed weighting method was used to determine many characteristics and marketing practices of the target population, including total local sales, farm counts for different categories of operations and types of products, and distance to consumers. 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.

The estimate of total local food farms in the U.S. is 167,009 operations, with a standard deviation of 5,845 operations. This resulted in a coefficient of variation of 3.5% for the national estimate. Reported state estimates for the total number of operations ranged from 1,269 operations in New Hampshire to 14,315 in California (smaller farm numbers were estimated in some unreported states). The estimated total local food sales of all farms is $8,747,222,197 with a standard deviation of $892,216,664. This resulted in a coefficient of variation of 10.2% for the total local food sales. State level estimates for total local food sales ranged from $28,235,565 in Utah to $2,869,192,534 in California (smaller sales were estimated in some unreported states).

The misclassification survey was sent to 682 respondent operations that were classified as in-scope and 712 classified as out-of-scope from the LFMP survey. Of the 682 in-scope operations sampled, 481 responded, and 405 responses were determined to be in-scope from the follow-up questionnaire. Thus probability of misclassifying an out-of-scope operation as a member of the target population is 0.158. Of the 714 out-of-scope operations sampled for the misclassification survey, 360 responded, and 60 responses were determined to be in-scope from the follow-up questionnaire. The probability of misclassifying an in-scope operation as out-of-scope is 0.167. The overall adjustment made to estimates from misclassification, the ratio of the probabilities in Equation 9, was 14.3%.

Of particular interest are the number and sales of farms that sold products directly to consumers. Estimates for the number of farms with direct to consumer sales were obtained from the 2012 Census of Agriculture, and this information was used to create the NASS list frame sampling groups. The 2012 Census of Agriculture reported 144,530 operations that reported direct to consumer sales. The number of these operations estimated during the LFMP survey was 114,801 with a standard error of 3,558. However, only 52.3% of responses from operations reporting sales directly to the consumer on the 2012 Census of Agriculture qualified as local food farms in the survey.

4. Evaluation of Web Scraping to Build Lists

Dual system estimation requires assumptions regarding the lists and their associated records. Failure to meet these assumptions results in estimates that are biased, or standard errors that do not account for the true variability of the estimator [25, 27]. If violations of one or more assumptions are known, estimators can typically be developed that account for these violations, or adjustments can be made after estimation [28, 31, 32]. However, identifying violations to these assumptions can be difficult and often biases go undetected. In this section, the assumptions required for this estimator are stated and a list built by scraping the internet for data will be evaluated based on its ability to uphold these assumptions.

Four assumptions are required in fundamental DSE statistical methodology [24]. The population that is sampled must be closed between the two capture events. This implies that during the time that samples are drawn, no local food farms begin business or go out of business. Each capture-event must be a random sample of the population of interest. Capture events are required to be independent from one another. In this case, the two lists used to draw samples are assumed to be independent. Finally, the records on each list can be linked without error.

In general, local food farms are smaller than most large agricultural farms and as a result, more transient [15, 10]. It is expected that over some time, new farms will begin operation and some farms will go out of business. During this study, samples are drawn from the lists at the same time and a single survey is sent to each sampled operation. The classification of an operations as a local food farm is made directly from the survey response. Thus, operations do not have the opportunity to go in or out of business during the survey process.

While web scraping can identify potential records, it lacks additional data useful for estimating properties of the record. This includes estimation of the probability of a record being in-scope and the probability of the record being misclassified. To estimate misclassification in this study, a separate survey was conducted using a sample of the surveyed records. While there is no time lag between the two lists used in estimation, there was a short time lag between the LFMP survey and the misclassification survey, allowing a short opportunity for operations to change scopes between them. If this does happen, it is likely that these rates would be captured in the misclassification adjustment. Future surveys may better account for population dynamics by identifying records that changed scope between the two surveys.

The web-scraped list developed by MACE was built within a short time span compared to many of the agricultural lists used by NASS, and has the potential to be delivered at any time. For list building, this allows it to more adequately uphold the assumption of closure than other lists or surveys that NASS conducts. While a longer time line may allow extraction from more and less prevalent web sources and thus, more records to be identified, the list can be delivered in little time to prevent a time lag between a second capture event. For example, during the Census of Agriculture, almost six months elapses between the June Area Survey and the Census Survey, allowing operations to possibly change scope between them. Survey questions have been developed to account for scope changes during the Census of Agriculture. In other examples of DSE, it is often assumed that rates of records entering or leaving the target population are equal to account for time lags [27, 26]. However, these assumptions may not be necessary when utilizing a web-scraped list, as a list can be completed in a short time frame.

Simple DSE assumes that captured records are a simple random sample of the population of interest [24]. It is assumed that the probability of capture is equal for all members of the population. This assumption simplifies estimators in the case where little is known about individuals in the population, but it is not necessary for DSE [21]. In the case of this study, identification of local food farms can be dependent on farm size, commodities produced, marketing channel, as well as many other factors. The other components required for an operation to be captured (i.e. response and correct classification) also account for characteristics of the farm. To account for these, the probability of capture is adjusted by estimating coverage, sampling, response and misclassification probabilities using logistic regression or grouping similar operations based on these operations' characteristics. As a result, estimated capture probabilities varied greatly in this study, ranged from approximately 0.002 to approximately 1.

While equal capture probabilities are not necessary for unbiased estimates, the lists are assumed to be samples from the same populations. This ensures that all individuals of the population can be captured by each list. It is not necessary that individuals have equal probability of being added to each list, however, any discrepancies in the coverage of the lists must be accounted for based on the operation's characteristics. Similarly, it is assumed that the probabilities for response and misclassification are equal and independent after adjusting for the necessary operation characteristics. For example, if the coverage probability is a function of an operation's total sales, then two operations with the same sales have the same probability of being on the NASS list frame. This implies that all factors affecting a operation's probability of capture have been identified and correctly adjusted for.

One of the consequences on using web scraping to build a list is there is a fraction of the population that is missed. If an operation has no presence on the internet, the probability that it is identified via web scraping is zero. The proportion of operations that have no web presence is unknown and assumed to be small. Also, the more time and effort that is applied to web scraping results in more complete lists [33]. However, it is unknown whether the characteristics of the lists are uniform throughout the web-scraping process. For example, large farms might immediately be identified while small farms take longer and less traveled websites must be scraped. The duration and effort involved in the list building process may affect the type of operations that are captured.

It is assumed that the capture probabilities by each lists are independent [24, 21, 20]. Equivalent capture probabilities by the two lists are not required for accurate estimates, however, it is assumed that the probability of an individual being captured by one list has no affect on the individual being captured by the second event. In the case of this study,

P(Record i is on MACE list|Record i is on NASS list) =P(Record i is on MACE list) and

P(Record i is on NASS list|Record i is on MACE list) =P(Record i is on NASS list).

Independence between list frames is difficult to obtain. In this study, the web-scraped list frame was built independently from the NASS list frame. However, operations may have higher exposure due to farm size or type. Some farms may also have greater web presence due to these characteristics. If these events coincide, the likelihood that an operation is on one list may affect that of the other list. In addition, much of the NASS list frame has been created from other lists produced by NASS regional field offices, producer association lists, and other sources. If these lists are available on the internet, the same sources could be used to add records to both the NASS list frame and the MACE list frame, resulting in list dependency.

Independence between lists can also be difficult to verify. During the Census of Agriculture, the NASS list frame is paired with the June Area Survey to assess under-coverage. Because the JAS is an areal survey, in which parcels of land are fully sampled and all operations on these parcels are obtained, it is considered independent from the NASS list frame. During the 2012 Census of Agriculture, the average coverage probability (the total proportion of farms identified by the JAS that were also captured by the NASS list frame was 0.928. The proportion of local food farms from the MACE sampled operations that were matched to a NASS list frame record was 0.931. While the JAS is far more extensive than the MACE sample and the target population of farms is different, similar coverage proportions between these two studies provide some evidence that there is no obvious dependence between the web-scraped list frame and the NASS list frame.

The last assumption required for accurate DSE estimates requires individuals from multiple lists to be matched accurately [24]. Web scraping has the ability to retrieve all necessary information used to match operations across lists. This includes information such as operation name, address, multiple phone numbers, email address, person(s) of contact and fax numbers. All data collected via web scraping are cleaned and edited for errors. Matching is then performed and results are classified into three categories: matches, non-matches, and possible matches. Possible matches are scrubbed extensively to determine if a link can be made. Thus, web scraping is not expected to result in any additional false matches or missed matches as compared to current record linkage between other NASS lists. For the purposes of this project, all matches and non-matches were assumed to be perfect and thorough.

In addition to the basic criteria necessary for dual system estimation, assumptions must also be made due to the methods of data collection and the available data provided by web scraping. Data from the list is limited and records only consist of operation names, addresses and persons of contact. Both lists contain records that do not qualify as local food farms, so reducing the available records to include only these operations is necessary. The only method of determining a record's local sales status is based on responses to the Local Food Marketing Practice Survey. Thus, in order for an operation to be classified as a local food farm, it must be sampled from one of the lists and must respond to the survey. Conditioning response and sample probabilities on the local food farm status of an operation is not possible. It is necessary to assume that response and sample probability is independent of local food farm status. Specifically, the study required the assumption that

$$P(\text{Sampled}|\text{Local Food Farm}) = P(\text{Sampled})$$

and

$$P(\text{Response}|\text{Sampled, Local Food Farm}) = P(\text{Response}|\text{Sampled}).$$

While the web-scraped list is built to include operations that have potential to sell products locally, the true local food status is not known, and sampling is conducted independently of qualification as a local food farm. Sampling from all operations on the NASS list frame would have resulted in very few local food farms (less than 4% of all operations with no local sales indications were in-scope) in the samples. Independence of response probability and local food farm status is difficult to discern as all classification of operations as local food farms is conditional on a response. The assumption can be simplified to

$$\begin{aligned} P(\text{Response}|\text{Local Food Farm}) &= P(\text{Response}|\text{Not Local Food Farm}) \text{ or} \\ P(\text{Local Food Farm}|\text{Response}) &= P(\text{Local Food Farm}|\text{No Response}). \end{aligned}$$

Future studies may allow us to determine or estimate the local food status of operation's prior to a response. Providing further information about records on the lists may allow for estimation of the probability that an operation qualifies as a local food farm, prior to a survey response. This can ease the assumptions requiring conditioning probabilities on local food farm status. However, initial attempts at using web scraping to build a list has not produced operation data other than the business's name and contact information. Thus, more research is required to determine the extent of data than can be acquired through this process.

5. Discussion

The Local Food Marketing Practices Survey was the first survey NASS has conducted focused directly on the marketing practices of U.S. agricultural operations with local food sales. As such, the survey questionnaire, sampling methodology, estimation methodology, and even the definition of local food businesses were created for the purposes of this study. Web scraping provided a list necessary to implement the proposed DSE methodology and obtain estimates of local food farms at national and state levels. Estimates of total local food operations were lower than anticipated from the 2012 Census of Agriculture. The Census of Agriculture estimates of direct to consumer sales were larger than that from the LFMP survey. However, approximately 52.3% of the operations that reported direct to consumer sales during the 2012 Census of Agriculture were classified as local food farms in this study, possibly accounting for discrepancies between these estimates. Overall, results from using web scraping to adjust for under-coverage using dual system estimation were encouraging.

The samples used for current dual system estimation implemented by NASS require considerable time and manpower to obtain a second list of farms. The JAS, used in the DSE estimators during the Census of Agriculture, require many people to create the sampling frame, sample each land unit, and edit the data that is obtained. The list of farms require many months to complete and hundreds of people working over these months. Comparatively, web scraping is efficient in cost, time and manpower. The resulting list required only a fraction of the cost as compared to other surveys NASS conducts. The list was planned and built in approximately one month, reducing the time taken to produce other list frames. While some technical knowledge is required in web scraping, the total manpower needed to produce the frame was far less than other samples. Web scraping could be a cost-efficient alternative for future agricultural projects, as well as, other fields in which list building is necessary.

While building a web-scraped list farm is efficient, the total size of the list was smaller than that of other surveys. The list composed by MACE contained 33,394 total records. Based on the sample taken from this list, approximately 25.2% of these records returned local food farms. The total number of in-scope operations returned in this study was far fewer than other surveys. This limited the amount of data used to estimate capture probabilities and resulted in simplified models. A greater number of records or increased accuracy would allow these probabilities to be more accurately adjusted for operation characteristics. The MACE frame was composed in a short amount of time, and additional time and effort would allow further online sources to be scraped. This could result in a larger initial list and potentially more usable records. However, changes in accuracy during list building process may occur as deeper web sources are used.

Another challenge encountered was minimal data per record available from web scraping. The web-scraped list produced a business name and contact information. Further information about each operation may result in more accurate samples in the future. This may also allow the probability of an operation being in-scope to be estimated, easing assumptions regarding independence between response probabilities and a farm's local sales. Obtaining more data for each record would require increased programming abilities and data may not be consistent across records.

The ability of web scraping to maintain the assumptions necessary to implement DSE during the construction of the list frame are currently being researched further. In this study, there is no indication that using a web-scraped list violates any of these assumptions. However, it is difficult to either detect or reject the presence of any violations. The largest concern is that web scraping may not produce an independent list from the NASS list frame, resulting in biased estimates. This could arise from similar sources being used to build each of the lists. Matching both samples to a third list frame may help detect dependencies between the lists.

The number of local food farms with no online presence is another concern in the accuracy of resulting estimators. Any operation with no online presence is omitted from a list built by web scraping. Thus, some proportion of the population is missed during sampling. It is likely that this proportion differs based on the target population of the produced list. Similarly, the number of online sources connected to an operation may alter capture probabilities and is difficult to account for during development of estimators. Estimating the proportion of a population with no web presence or quantifying the web presence of a record may increase the accuracy of estimates, the completeness of the composed list frame, and extent of web scraping to produce list frames in other disciplines.

Even in the event that web scraping violates the assumptions of DSE or omits a proportion of the target population, the use of a web-scraped list may still be useful as a third source to implement triple system estimation. This would provide a step forward in quantifying the dependence between the lists used, and perhaps result in more accurate estimators and smaller measurements of error. Web scraping may also capture a segment of the population that is missed by one of the other lists involved. An ongoing study in WA is investigating web scraping to increase the number of small farms that are often missed by other NASS surveys.

As the amount of available information continues to increase, the task of filtering data to provide official statistics and answer questions becomes more difficult. The need for fast and efficient retrieval of relevant data becomes increasingly important. As more data becomes available on the internet, extraction of this data and transformation of unstructured data sources to usable data sets is necessary. Web scraping has potential to build new data sources or supplement existing data quickly and cost efficiently. However, the implications of using these sources must be assessed to ensure that extracted data and resulting statistics are accurate.

References

[1] V. Caputo, R. M. Nayga, and R. Scarpa, "Food miles or carbon emissions? exploring labelling preference for food transport footprint with a stated choice study," *Australian Journal of Agricultural and Resource Economics*, vol. 57, no. 4, pp. 465–482, 2013.

[2] B. Halweil, *Home grown: The case for local food in a global market*, vol. 163. Worldwatch Institute, 2002.

[3] S. D. Hardesty, "The growing role of local food markets," *American Journal of Agricultural Economics*, vol. 90, no. 5, pp. 1289–1295, 2008.

[4] R. P. King, *Comparing the structure, size, and performance of local and mainstream food supply chains*, vol. 99. Diane Publishing, 2010.

[5] G. W. Feenstra, "Local food systems and sustainable communities," *American journal of alternative agriculture*, vol. 12, no. 01, pp. 28–36, 1997.

[6] R. Govindasamy, M. Zurbriggen, J. Italia, A. O. Adelaja, P. Nitzsche, R. VanVranken, *et al.*, "Farmers markets: Consumer trends, preferences, and characteristics," tech. rep., Rutgers University, Department of Agricultural, Food and Resource Economics, 1998.

[7] R. Dodds, M. Holmes, V. Arunsopha, N. Chin, T. Le, S. Maung, and M. Shum, "Consumer choice and farmers' markets," *Journal of agricultural and environmental ethics*, vol. 27, no. 3, pp. 397–416, 2014.

[8] M. Coit, "Jumping on the next bandwagon: An overview of the policy and legal aspects of the local food movement," *J. Food L. & Pol'y*, vol. 4, p. 45, 2008.

[9] S. A. Low and S. J. Vogel, "Direct and intermediated marketing of local foods in the united states," *USDA-ERS Economic Research Report*, no. 128, 2011.

[10] S. Martinez, *Local food systems; concepts, impacts, and issues.* Diane Publishing, 2010.

[11] J. B. Dunne, K. J. Chambers, K. J. Giombolini, and S. A. Schlegel, "What does local mean in the grocery store? multiplicity in food retailers' perspectives on sourcing and marketing local foods," *Renewable Agriculture and Food Systems*, vol. 26, no. 01, pp. 46–59, 2011.

[12] S. M. Inwood, J. S. Sharp, R. H. Moore, and D. H. Stinner, "Restaurants, chefs and local foods: insights drawn from application of a diffusion of innovation framework," *Agriculture and Human Values*, vol. 26, no. 3, pp. 177–191, 2009.

[13] J. M. Bagdonis, C. C. Hinrichs, and K. A. Schafft, "The emergence and framing of farm-to-school initiatives: civic engagement, health and local agriculture," *Agriculture and Human Values*, vol. 26, no. 1-2, pp. 107–119, 2009.

[14] C. Brown and S. Miller, "The impacts of local markets: a review of research on farmers markets and community supported agriculture (csa)," *American Journal of Agricultural Economics*, vol. 90, no. 5, pp. 1298–1302, 2008.

[15] M. Mirosa and R. Lawson, "Revealing the lifestyles of local food consumers," *British Food Journal*, vol. 114, no. 6, pp. 816–825, 2012.

[16] P. Kremer and T. L. DeLiberty, "Local food practices and growing potential: Mapping the case of philadelphia," *Applied Geography*, vol. 31, no. 4, pp. 1252–1261, 2011.

[17] K. Darby, M. T. Batte, S. Ernst, and B. Roe, "Decomposing local: a conjoint analysis of locally produced foods," *American Journal of Agricultural Economics*, vol. 90, no. 2, pp. 476–486, 2008.

[18] J. N. Pretty, A. S. Ball, T. Lang, and J. I. Morison, "Farm costs and food miles: An assessment of the full cost of the uk weekly food basket," *Food policy*, vol. 30, no. 1, pp. 1–19, 2005.

[19] R. S. Pirog and A. Benjamin, “Checking the food odometer: Comparing food miles for local versus conventional produce sales to iowa institutions,” 2003.

[20] J. M. Alho, “Logistic regression in capture-recapture models,” *Biometrics*, pp. 623–635, 1990.

[21] J. M. Alho, “Analysis of sample based capture-recapture experiments,” *Journal of Official Statistics*, vol. 10, no. 3, p. 245, 1994.

[22] M. K. Soisalo and S. M. Cavalcanti, “Estimating the density of a jaguar population in the brazilian pantanal using camera-traps and capture–recapture sampling in combination with gps radio-telemetry,” *Biological conservation*, vol. 129, no. 4, pp. 487–496, 2006.

[23] K. Tilling and J. A. Sterne, “Capture-recapture models including covariate effects,” *American journal of epidemiology*, vol. 149, no. 4, pp. 392–400, 1999.

[24] S. Lohr, *Sampling: Design and Analysis*, vol. 1. Duxbury Press, 1999.

[25] K. H. Pollock, J. D. Nichols, C. Brownie, and J. E. Hines, “Statistical inference for capture-recapture experiments,” *Wildlife monographs*, pp. 3–97, 1990.

[26] L. J. Young and J. H. Young, “Capture recapture: Open populations,” in *Statistical Ecology*, pp. 357–389, Springer, 1998.

[27] C. J. Schwarz and A. N. Arnason, “A general methodology for the analysis of capture-recapture experiments in open populations,” *Biometrics*, pp. 860–873, 1996.

[28] A. Chao, “Estimating the population size for capture-recapture data with unequal catchability,” *Biometrics*, pp. 783–791, 1987.

[29] D. G. Horvitz and D. J. Thompson, “A generalization of sampling without replacement from a finite universe,” *Journal of the American statistical Association*, vol. 47, no. 260, pp. 663–685, 1952.

[30] L. Satore, K. Toppin, and C. Spiegelman, “Estimated covariance matrices associated with calibration,” 2017.

[31] G. M. Jolly, “Explicit estimates from capture-recapture data with both death and immigration-stochastic model,” *Biometrika*, vol. 52, no. 1/2, pp. 225–247, 1965.

[32] K. H. Pollock, “A capture-recapture design robust to unequal probability of capture,” *The Journal of Wildlife Management*, vol. 46, no. 3, pp. 752–757, 1982.

[33] E. Vargiu and M. Urru, “Exploiting web scraping in a collaborative filtering-based approach to web advertising,” *Artificial Intelligence Research*, vol. 2, no. 1, p. 44, 2012.