DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 34, No. 2, 2018, pp. 323–340, http://dx.doi.org/10.2478/JOS-2018-0015

Exploring a Big Data Approach to Building a List Frame for Urban Agriculture: A Pilot Study in the City of Baltimore

Linda J. Young[1], *Michael Hyman*[1], *and Barbara R. Rater*[2]

The United States Department of Agriculture's National Agricultural Statistics Service (NASS) has the responsibility of quantifying the nation's agricultural production. Historically, it has focused on large production agriculture. With interest and activity increasing in urban areas, NASS has begun exploring how to better quantify urban agriculture. This segment of agriculture is particularly challenging to enumerate because the agricultural holdings tend to be small, diverse, widely dispersed, and more transient than the predominantly large farms in rural areas. In collaboration with the Multi-Agency Collaboration Environment (MACE), a new approach to list building was explored in a pilot study conducted in the City of Baltimore, Maryland. Using a big data approach, areas of potential agricultural activity were identified by gathering information (state and local permits, Facebook and twitter feeds, interest groups, etc.) via the web. A sample was drawn from the list, and an in-person survey was conducted to assess whether or not the identified areas had agricultural activity. The results of the pilot study are presented. Lessons learned from the study and next steps are discussed.

Key words: Web scraping; urban agriculture; list building.

1. Introduction

In the United States, urban agriculture has been receiving increasing attention at the local, state, and national levels. Although urban agriculture has long been present, many cities began actively promoting it after the start of the Great Recession in an effort to ensure that the vacant lots, abandoned buildings, and other under-utilized land are used productively (Goldsmith 2014; Santo et al. 2016). Some cities, such as Detroit, Michigan, have goals of increasing their food resiliency by growing a large portion of the fruits and vegetables used by their citizens within the city limits (Colasanti et al. 2010). Urban agriculture includes not only urban farms, but also a diverse array of agriculture, including backyard gardens, school or community gardens, greenhouses, hoop houses, converted warehouses, vertical gardens, and aquaponics. Land use planners and landscape designers have the opportunity to be involved in the integration of urban agriculture, such as community farms, allotment gardens, edible landscaping, urban forests, and rooftop gardening, into a sustainable urban environment (Lovell 2010).

[1] USDA National Agricultural Statistics Service, Research and Development Division, 1400 Independence Avenue SW, Washington DC 20250, U.S.A. Emails: Linda.Young@nass.usda.gov and Michael.Hyman@nass.usda.gov
[2] USDA National Agricultural Statistics Service, Census and Survey Division, 1400 Independence Avenue SW, Washington DC 20250, U.S.A. Email: Barbara.Rater@nass.usda.gov

© Statistics Sweden

Local policies have facilitated access to land, enabled use of land for agricultural purposes through revised zoning rules, and provided initiatives and legislation supporting urban agriculture (PolicyLink 2012). Examples include the City of Seattle's Urban Garden Share program, which matches experienced gardeners who live in condos and apartments with local gardens with growing space to share, and the City of Austin's Sustainable Urban Agriculture and Community Garden Program, which provides a framework of guidelines for an established local food system (Popovitch 2014). Federal funding programs for urban agriculture, such as The New Farmers Initiative, have been established (USDA 2016). The Brownfield's Economic Development Initiative, a federal grant program assisting cities with redeveloping abandoned, idled, and under-used industrial and commercial facilities, was not designed specifically for urban agriculture, but can be used to benefit it (PolicyLink 2012). The U.S. Department of Agriculture (USDA) has developed a website with a wealth of information, including funding opportunities for urban farmers (USDA 2016). Community-based organizations focused on urban agriculture have emerged.

Although USDA's National Agricultural Statistics Service (NASS) has always included urban farms in its counts of farms and farm production, the estimates have not been as precise as those for other sectors of the agricultural economy. Because urban agriculture is responsible for a relatively small portion of the total farm production and the traditional approach to identifying agricultural activity in urban areas is costly, NASS has not devoted its scarce resources to providing more precise estimates. Policy-makers at all levels are increasingly interested in the efficacy of programs designed to promote and expand urban agriculture. Thus, in 2014, the Under Secretary for USDA's Research, Education and Economics asked NASS to explore cost-effective ways to better quantify the extent and food production of urban agriculture.

Because urban farms tend to be smaller, more diverse, more transient, and more widely dispersed than the more traditional farms in rural areas of the United States, these farms are challenging to identify and thus to quantify. The NASS list frame, which is a list of all known farms and potential farms in the United States, is anticipated to have substantial undercoverage of urban farms. Furthermore, because these farms tend to be dispersed and not concentrated within the urban areas, it is cost prohibitive to obtain sufficient numbers of urban farms in a sample drawn from the NASS area frame, which includes all states except Alaska. Thus, NASS began to investigate alternative ways to build a list of urban farms that would be independent of the NASS list frame, provide good coverage, and be relatively inexpensive, so that it could be used to assess undercoverage of the NASS list frame.

NASS began exploring the use of a "big data" approach to creating a list frame for urban agriculture. Although big data is an abstract construct and has no agreed-upon definition, the term generally is applied to data sources characterized by three Vs (Mayer-Schönberger 2014): (1) Volume: a massive number of observations or variables, (2) Velocity: data generated frequently or in real-time, and (3) Variety: a variety of data formats and structures, including no structure. Chen et al. (2014) add a fourth V, Value, because the data generally have high value but very low density. Here, two big data approaches are explored as a potential means to generate a list frame of urban agriculture sites: satellite imagery and web scraping. Earth observation satellites provide satellite images of the Earth. Some satellite imaging companies sell or license the satellite imagery

to governments or businesses, such as Apple Maps or Google Maps. Web scraping is an automated approach to extracting publically available data from websites. Using key words, webpages with potentially relevant information are identified, and the information is gathered. (See Krijnen et al. (2014) for an introductory discussion of web scraping.) Both satellite imagery and web scraping provide a variety of data being generated continuously (high velocity) in large volumes and with high value.

In 2014 and 2015, an urban agriculture pilot study was conducted in the City of Baltimore. The goal was to assess the viability of creating a list frame using satellite imagery and web scraping. Baltimore was selected because it has an active urban agriculture program, is close to the Washington DC area where the foundational research was conducted, and has available interviewers who can collect data from identified sites for assessing the efficacy of using these approaches to develop a list frame of urban farms. Recent Baltimore urban agricultural initiatives include the Baltimore Office of Sustainability's Land Leasing Initiative, the Urban Agriculture Tax Credit, and the Urban Agriculture Training Program (Baltimore Office of Sustainability 2016). With a plethora of vacant buildings and vacant lots, Baltimore City is focused on redevelopment strategies for the city's land-use policies and on greening initiatives, such as urban agriculture.

The results of the City of Baltimore pilot study are presented in this article. In Section 2, the target population for a national urban agricultural study is defined. The basic process used to develop an urban agriculture list frame from satellite imagery and web scraping is outlined in Section 3. A pilot study of a web-scraping approach to identifying urban agriculture in the City of Baltimore, with considerations for national implementation, is described in Section 4. The final section reviews the lessons learned and future directions.

2. Urban Agriculture – The Target Population

Historically, NASS has not reported urban agriculture separately; it has been combined with all other types of agriculture. To report on it separately, urban agriculture must be defined so that an operation can be unambiguously identified as either being or not being urban agriculture. This requires that both "urban" and "agriculture" be clearly defined in this context. With the 1950 US Census of population (referred to here as Census), a distinction was made between urban and rural areas. The definition of urban has evolved with each Census. With the advent of geographical information systems (GISs), which capture, store and display geographic and spatial data, the process of identifying urban areas took a major leap forward in the 2000 Census and was further refined for the 2010 Census. An urbanized area is identified through an iterative process. The primary criteria for delineating an urbanized area are based on Census tract and block population density, count, and size thresholds. Other criteria are thresholds for the area covered by impervious surfaces, such as roads, cement parking lots, and buildings. Initially, census tracts with area less than three square miles and a population density of at least 1,000 persons per square mile (ppsm) are identified, and contiguous census tracts meeting these thresholds are aggregated to form initial urban cores. A census tract that is contiguous to an urban core is added to the core if it has a land area less than three square miles and a population density of at least 500 ppsm. Then a contiguous census block is added if it has a population density of at least 1,000 ppsm. To account for a mix of residential and nonresidential urban uses, contiguous census blocks

with at least 500 ppsm or meeting thresholds based on the percent of land area covered by impervious surfaces are aggregated into the initial urban core. Complex rules are also in place for including noncontiguous areas into the initial urban core as well as for combining and splitting urban cores. Once the final urban cores have been identified, those with populations of at least 50,000 are the urbanized areas, and those with populations between 2,500 and 50,000 are the urbanized clusters (see U.S. Census Bureau 2011 for full details). For the purposes of defining urban agriculture in this pilot study, urbanized areas, but not urbanized clusters, are the areas defined as urban (see Figure 1).

The next step is to determine what should be included as agriculture within these urbanized areas. In the United States, the definition of a farm is any operation that produces and sells, or has the potential to sell, USD 1,000 or more of agricultural products in a year. Thus, an individual with a large backyard garden who sells the excess produce at a farmers' market is a farmer if the produce sells for at least USD 1,000 over the course of a year. If an operation typically has sales that qualify it as a farm, but weather or some other factor negatively affected production so that sales are less than USD 1,000 for a given year, then it is still considered to be a farm. If an operation has less than USD 1,000 in sales, its agricultural potential is assessed by assigning points to its agricultural activity. For example, each horse is given 200 points; a dairy cow is worth 2,800 points; an acre of dry beans that has been harvested is valued at 350 points; a harvested acre of vegetables is rated at 3,600 points; and an acre of fruit is assigned 3,100 points. If the points total at least 1,000, the operation is classified as a farm. If none of the agriculture products a farm produces are sold, as is the case when a family consumes the food grown in a backyard garden, then the operation is not classified as a farm.

Urban agriculture is quite diverse. Historically, personal gardens, either in the backyard or at some other location, have been tended by individuals or a single family. Community gardens (private or public land gardened by a group of people as either individual or shared plots), neighborhood gardens (private or public land gardened by a group of people where

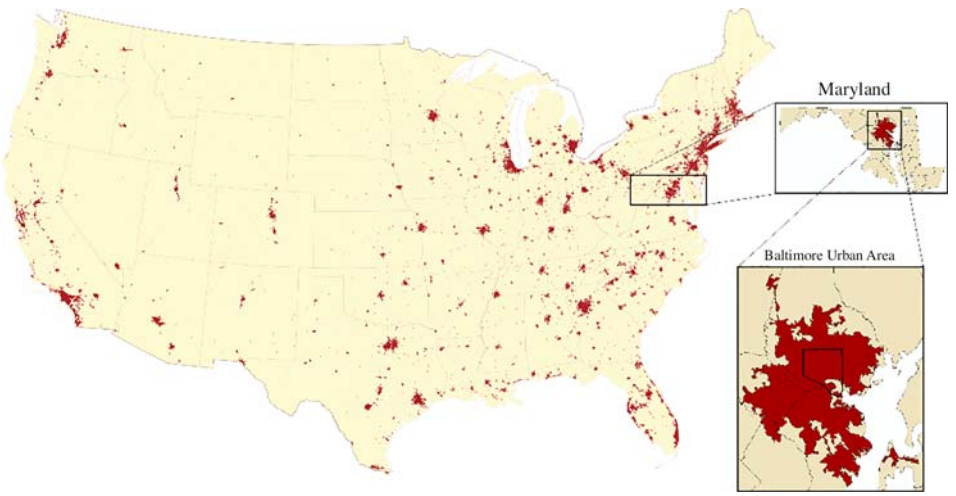

Fig. 1. Urbanized areas within the United States (left), area of pilot study within United States (upper right), and Baltimore City boundaries contrasted with Baltimore urbanized area (lower right).

each person or family rents an individual plot at a nominal fee), residential gardens (gardens shared among individuals in multiple housing units, such as apartment buildings, assisted living facilities, and low income housing), institutional gardens (gardens attached to public or private organizations, such as schools, rehabilitation centers, and prisons), and demonstration gardens (gardens used in educational and recreational settings) are common forms of urban agriculture (Marin Master Gardeners 2017). For gardens comprised of small plots, such as community or neighborhood gardens, NASS has traditionally considered each plot as an operation, which either qualifies as a farm and is counted or found to not qualify as a farm and is not counted. Often none of the plots produces enough to qualify as a farm even though the total production of the community garden could well be enough to be classified as a farm. This could contribute to the perception that NASS is not fully reporting agricultural activity in urban areas.

Increasingly, a variety of agricultural enterprises are being developed in urban areas (Specht et al. 2014). An abandoned warehouse may become the site for aquaculture (a process of raising animals, such as snails and fish, in tanks), hydroculture (a process of raising animals in water), or aquaponics, which is a combination of aquaculture and hydroculture. Traditional commercial greenhouses, perhaps located on the top of a commercial building, can produce fruits and vegetables throughout the year. Hoop houses, also called high tunnels, are inexpensive, passive solar greenhouses that are usually unheated, rounded structures. Hoop houses can be used to extend the growing season and hence the time urban farmers have to sell fresh, local produce each year (Conner et al. 2009). Cultivating plants or raising animals within skyscrapers or on vertically inclined surfaces is referred to as vertical farming (Despommier 2010). Although cultivation of plants is dominant, fish, shrimp, bees, chickens, goats, and other animals are part of today's urban agriculture.

For the purposes of the pilot study, all types of urban agriculture are defined to be in-scope with few exceptions. Some gardens, such as those in a park, around a person's residence, or on the top of a residential building, are for personal enjoyment and not designed for either food production or the sale of plants or animals grown in them; these were excluded from the target population. Based on the resolution of the satellite imagery, a minimum area of 36 square feet had to be devoted to the agricultural activity for it to be in-scope. This led to excluding plants grown in pots on the front porch or in a window box. In addition, mobile gardens, such as those in the back of a pickup or on the top of a bus, were excluded because they do not have a permanent geospatial location.

The effort to define urban agriculture more broadly than entities satisfying the definition of a farm reflects the USDA's interest in supporting all types of urban agriculture. Because NASS traditionally reports on only entities meeting the definition of a farm, those operations qualifying as a farm are to be reported separately from those that do not qualify as a farm, such as ones that produced food only for home consumption.

3. Developing a "Big Data" List Frame

Satellite imagery and web scraping are two big data approaches that have the potential to be used for list frame development. NASS currently uses satellite and aerial imagery to develop its area frame, which covers the United States, except for Alaska. It also uses

remotely sensed data to create the Cropland Data Layer (CDL), which identifies land devoted to crops throughout the contiguous 48 states. The CDL is published as CropScape at the end of the growing season. In developing the NASS list frame, staff often search the web to identify potential farms. However, to create a list frame using satellite imagery, web scraping, or both is a new endeavor for NASS. Here, previous work related to this effort is reviewed, and then the approach taken by NASS will be described.

3.1. Previous Efforts

In Google Earth, high-resolution satellite images, aerial photography (pictures taken from sub-orbital instruments, such as airplanes or drones), and GIS data are combined to produce 3D maps of Earth (Wikipedia 2017). Working with Google Earth images in conjunction with ArcGIS (GIS software for mapping and spatial analysis, see Law and Collins 2013), Taylor and Lovell (2014) used manual interpretation to identify potential areas of urban agricultural activity in Chicago, Illinois. They began by combining lists of community gardens from several non-governmental organizations to identify 1,236 potential community gardens in Chicago. The sites were visited in 2010, and 12.9% were found to be food gardens with the others being ornamental gardens/parks, streetscaping projects, or no garden. The confirmed food and non-food gardens were analyzed to develop a manual visual classification approach to classifying urban agriculture based on agricultural markers identified in the reference images. Then, in a labor-intensive effort requiring approximately 400 hours, the Google Earth images of Chicago were analyzed visually, and 4,493 potential areas of agricultural activity that were not included in the 1,236 areas on the lists were identified. In a follow-up survey of 194 of the sites, 166 (89.6%) were found to have agriculture. This approach was effective for identifying urban agriculture sites, but it would be cost prohibitive when scaled to the national level. Automating the process by combining satellite imagery and web scraping may be a more cost-effective way to create a list frame of areas of urban agricultural activity.

Forster et al. (2009) explored mapping urban agriculture in a peri-urban area of Hanoi, Vietnam, using high-resolution satellite imagery. The average farm size of 0.22 hectares (approximately 0.54 acres) and the diversity of agriculture made classification challenging. They used segmentation of shape and size in an object-oriented classification approach to assess the accuracy of classifying data from known classes (bare soil, fallow, maize, sweet potato, orchard, and tree/hedge). The overall classification accuracy was 67%, and the overall kappa coefficient was 0.61.

Dumbacher and Capps (2016) report on a study of the Quarterly Summary of State and Local Government Tax Revenue in which methods of data collection based on unstructured data, text analytics, and machine learning were explored. Polidoro et al. (2015) used web scraping techniques to collect data on consumer electronics and airfares. In these cases, the focus was on the collection of data from identified sources.

In an effort more closely related to that of NASS, 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 crowdsourcing to verify whether or not a store identified through web scraping was truly a vape store. In Florida, all ENDS vape stores are required to register with the state as tobacco retailers. The final list of 403 stores was compared to the

full state of Florida's tobacco licensure list. They found 131 of the 403 stores on the licensure list (32.5%); the remaining 272 stores were not found on the list.

3.2. Methods: Satellite Imagery

NASS contracted with the Multi-Agency Collaboration Environment (MACE) to build a list frame of urban agriculture in the City of Baltimore. MACE is a consortium of government agencies and contractors that solves complex data problems. In collaboration with the Air Force Research Laboratory, MACE explored two approaches to creating a list of urban farms within the City of Baltimore: (1) satellite imagery and (2) web scraping.

Imagery from three different satellites and different seasons were obtained. The acquisition dates and resolution for the satellites Pleiades, GeoEye, and Worldview-2 were, respectively, August 15, 2014 at 0.7 meter (m); May 17, 2012 and July 27, 2014 at 0.41 m; and September 15, 2012 and January 23, 2013 at 0.5 m (see Stoney 2006, for more information on satellite imagery). Data from these satellites are freely available for public use. In remote sensing, the reflected or emitted radiation from different bodies is measured (European Association of Remote Sensing Laboratories 2017). Objects with different surface features, such as color, structure, and texture, have different spectral reflectance patterns. Spectral reflectance is comprised of wavelengths of different colors. These include the red, green, and blue bands that humans can see as well as bands, such as near-infrared, that humans cannot see. These spectral reflectance patterns are analyzed to identify earth surface features or materials. For example, the spectral reflectance curve of healthy green vegetation has little reflectance in the visible portion of the spectrum, but increases dramatically in the near-infrared. The fraction of incoming solar radiation that is reflected from the Earth's surface is known as Top-of-Canopy (TOC) reflectance, which is retrieved from satellite images by correcting for atmospheric effects, illumination, angle, etc. TOC reflectance is the most basic remotely sensed surface parameter and provides the primary input for other geophysical parameters, such as vegetation indices and texture images (see Baret and Buis 2008, for a fuller discussion).

Here the purpose of obtaining the satellite imagery and associated products was to identify farms. Training data for classifying the imagery were obtained through two approaches. First, the analysts verified that farm locations obtained from webpages of local websites were farms, using visual inspection of satellite imagery and Google Earth. Then, polygons were drawn around farms as well as other land cover classes, such as grass, man-made surfaces, and water. Approximately 150 farms were identified using this approach. Half were used as training data, and the other half were retained for validation. The other approach used semi-automatic classification (conducted using the semi-automatic classification plugin for the GIS software QGIS) to identify nine classes (farm, lawn/park, car, building, tree, water body, asphalt, pool, and dirt/lot) (Congedo 2014). Each class was evaluated against the TOC mosaic.

In addition to the classifier assessments, texture metrics were employed in an effort to identify agricultural operations from the remotely sensed data. One method of examining texture is based on the gray-level co-occurrence matrix (GLCM). To create a GLCM, the frequencies with which pixel pairs with specific values and in a specified spatial

relationship occur in an image are calculated. Then statistical measures can be derived from the GLCM. As examples, the local variations in the GLCM provide a measure of contrast, and the joint probability of occurrence of the specified pixel pairs is a measure of correlation. Using training data, these can be combined with edge mapping algorithms to identify areas of interest, here farms. In this study, Google Earth Engine, Matlab, or the open source GRASS GIS were used to evaluate 21 different GLCMs and edge mapping algorithms with multiple radii and neighborhood sizes (Schowengerdt 2007; Zujovic et al. 2009; Neteler 2010). To help differentiate between nominal vegetation and garden features, the Gabor filter, which utilizes frequency and orientation representations similar to the human eye, was used (Manjunath and Ma 1996; Singh and Hemachandran 2012).

3.3. Methods: Web Scraping

In addition to satellite imagery, the potential of web scraping to build a list of agricultural operations within the City of Baltimore was explored. In preparation for web scraping, NASS developed a set of key words that are often associated with urban agriculture. The list included terms such as urban farm, organic, farm-to-table, community garden, school garden, and local foods. Using these key words, the initial focus was on acquiring the "right" data, that is, determining where to look on the web for urban agricultural activity. A few iterations were needed to determine where and what urban agricultural information for the City of Baltimore was present in open source data. For example, were people using social media (Twitter, Facebook, etc.) to share information about locations of urban agriculture or was Four-Square, a mobile application for location sharing, used more frequently? Analysts explored these and numerous other digital platforms. Social media posts, such as tweets, retweets, and Facebook posts, were monitored to discover blog posts and to capture hashtags or accounts that identified locations or activities relating to urban agriculture in Baltimore. Sites hosting lists of operations with state and local animal husbandry permits were identified. Using the Yahoo application programming interface (API), search queries were initiated with key words relating to urban agriculture and Baltimore, such as community garden, school garden, beekeeping, and so on, and the webpages with the largest number of hits were retrieved (Krijnen et al. 2014). The open source Heritrix web crawler (Mohr et al. 2004) was employed in a focused web crawl to identify and download millions of webpages (Kausar et al. 2013). These webpages were filtered further for pages likely to contain data on urban agriculture sites.

The webpages found using the Yahoo API and the web crawler were the foundation for developing an automated urban agriculture finder, which built upon the Fathom Natural Language Processing (NLP) features (Hirschberg and Manning 2015). The algorithm read the pages and attempted to extract a description of the agriculture at the identified location as well as point-of-contact information, including personal name, organization, e-mail address, and telephone number, from each identified agriculture site. The naïve Bayes algorithm, a machine-learning method, was used to score each location identified through web scraping, indicating the likelihood that it was associated with urban agriculture in Baltimore (Vidhya and Aghila 2010). The algorithm was trained on manually labeled examples.

The geospatial coordinates for each potential urban agriculture site were found by passing the final set of addresses through Google's geocoder. Duplicates of urban agriculture sites geocoded to the same location were removed. Efforts were made to validate agriculture sites using either Google Earth imagery or Digital Globe. If the site could be verified through visual inspection of the imagery, then it was rated as having a high level of confidence for urban agriculture; otherwise, it was rated as having a low level of confidence.

4. City of Baltimore Pilot Study

Given the available satellite imagery, traditional imagery analysis techniques were not found to be useful for automatically identifying the small areas of agriculture common in urban farms. The satellite imagery available had resolutions ranging from 0.41 to 0.7 m. An example of that imagery is provided on the left side of Figure 2. On the right is a 15-cm aerial imagery of the same area. (The small dots toward the center of each figure are at the same point in space on each.) The human eye is more effective than algorithms in detecting items within imagery. The agricultural activity evident in the 15-cm imagery is not visible in the 0.5 m satellite imagery. With the aerial imagery, automatic identification of agricultural sites is potentially possible. However, higher resolution satellite imagery and aerial imagery were prohibitively expensive, so NASS was unable to take advantage of either. Therefore, MACE had to rely on the web-scraping techniques associated with the text analytics to develop the urban agriculture list frame.

The use of text analytics was more effective than the satellite imagery, identifying more than 1,500 urban agriculture locations based on the multiple data sources. Because many websites were visited, the list had duplicate identifications of numerous agricultural sites. In all, 57% of the locations were duplicative and removed. Further, 9% were confirmed to have no agriculture through a manual review. These tended to be associated with operations with names or website information that included a key word, but were obviously not a farm, such as a hair salon that used organic products. The remaining 505 (34%) locations were determined to be unique and with potential for agricultural activity (see Figure 3).

Some potential agricultural sites were identified from multiple websites. However, of the 505 potential agricultural sites, 386 were extracted from only one website (a sole source). Understanding which of the sole source websites provided a number of potential agricultural sites may be useful in developing future list frames using web scraping.

Fig. 2. Comparison of visibility of agricultural area using 0.5-m satellite imagery (left) compared to 15 cm aerial imagery (right). The light dots are at the same point in space.

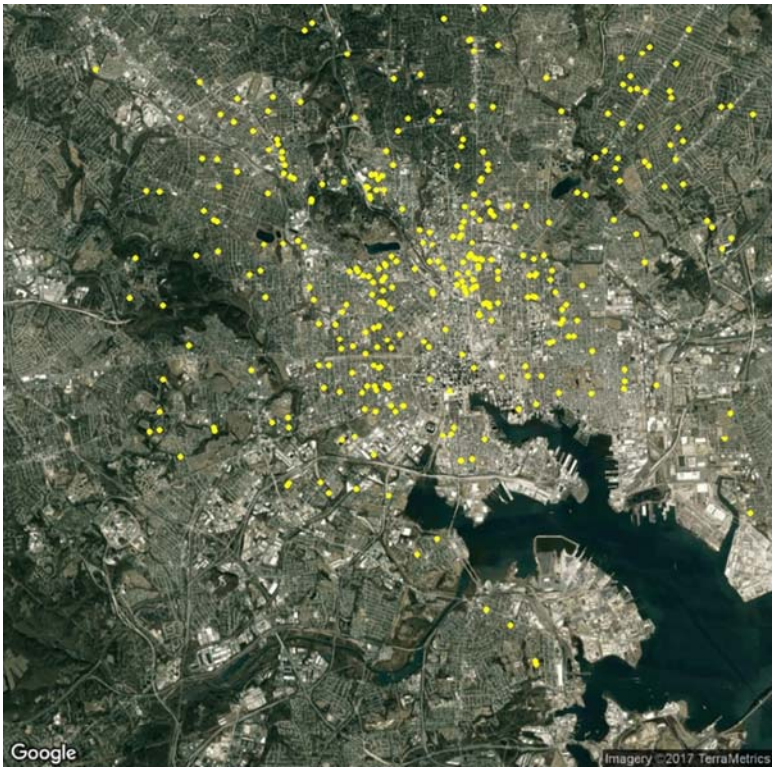

Fig. 3. Locations (light dots) of 505 potential areas of agricultural activity in the City of Baltimore.

The City of Baltimore maintains an Adopt-A-Lot list, which is made publicly available (see http://static.baltimorehousing.org/pdf/adopt_properties.pdf). This list accounted for 48% (185) of the sole source sites and thus, was extremely important to the completeness of the list. Community Greening Resource Network (https://www.climateinteractive.org/multisolving-in-action/multisolving-leaders/baltimores-community-greening-resource-network-supports-urban-gardeners/) and animal husbandry permit data each had 15% of the unique identifications from the web crawl. Open source research and the Yahoo API contributed the remainder (see Figure 4). This indicates that, at least for this study, identifying local internet sources was key to developing a full list.

The City of Baltimore had a 2013 list of areas of agricultural activity that was not available for the development of either the NASS list frame or the web-scraped list. Given the challenges of creating a complete list, this list is highly unlikely to have captured all of the agricultural sites in Baltimore. However, it is used here to provide insight into how well the web-scraped list performed and how that performance relates to the NASS list frame. When making these comparisons, it is important to remember that the NASS list frame has only farms and potential farms. In contrast, the web-scraped list includes agricultural activity of any size. The list from the City of Baltimore also includes agricultural sites that are not farms.

The Baltimore list had 159 agricultural operations, 70 school garden areas and 89 non-school agricultural sites. Twenty-four of the 70 school garden sites were on the

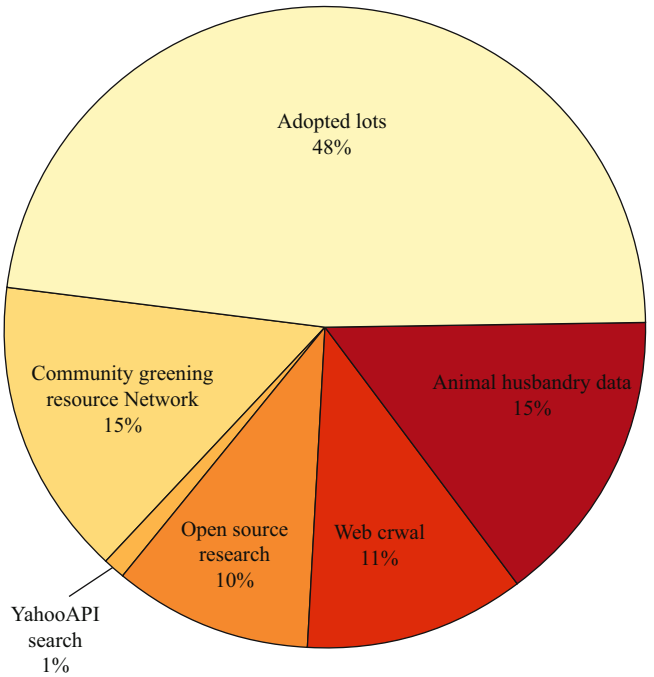

Fig. 4. Of the 386 agricultural sites identified through only one source, percentage found by each source.

web-scraped list of 505 operations, and 60 of the 89 non-school garden sites were on the web-scraped list. Baltimore's list included 13 urban farms. The web-scraped list included all 13 farms; only one of them was on the NASS list frame. Thus, although the web-scraped list did not include all of the agricultural sites identified on the City of Baltimore list, it provided good coverage of the Baltimore urban farms, and the NASS list frame was less complete for this target population.

To further evaluate the web-scraped list, an in-person survey was conducted to assess whether or not the identified areas had agricultural activity. Because the primary purpose was to evaluate the efficacy of building a web-scraped list frame and not to obtain estimates, a random sample of the potential urban agriculture sites was selected. Given the time that interviewers could devote to the study, the sample size was set at 266. For the community gardens, efforts were made to sample the plots within the community garden. Only twelve of the plot surveys were completed in six community gardens. It was extremely difficult to identify the plot operators for an interview; contacting the manager or coordinator of the community garden was much easier. Because finding the individual plot operators would be extremely time consuming, if even possible, it became clear that this approach would be cost prohibitive. Of the 266 sites, 71% (188) of the interviews were completed. In 21% (73) of the cases, the operator could not be found, and the nonresponse rate was 2% (5).

If the operator could not be contacted, the interviewer was asked to make an effort to observe whether or not agriculture was present. In five percent of the cases, the presence or absence of agriculture could not be ascertained from either an interview or by observation

Table 1. Presence of agriculture.

Survey status	Agricultural activity			Totals
	Yes	No	Unknown	
Completed	108	80	0	188
Not completed	31	34	13	78
Totals	139	114	13	266

(see Table 1). For approximately half (52%) of all sampled sites, agriculture was present. For some of the non-agricultural sites, it was evident that agriculture had been present previously. Vacant lots were also commonly found at the sites determined to be non-agricultural. During the 2012 US Census of Agriculture, half of the urban operations on the list frame were identified as farms; the other half were non-farms. Of course, some of those on the web-scraped list identified as having agricultural activity were not farms.

Each of the potential agricultural sites on the web-scraped list frame was given a confidence rating as to whether or not the site was associated with a farm, based on visual inspection of satellite imagery. Of the 505 sites of potential agricultural activity, 159 were given a low confidence rating and 346 were given a high confidence rating of being agricultural. Based on the 253 sites in the sample for which the presence of agricultural activity could be determined, those that were initially rated as having a high level of confidence were significantly more likely to be agriculture sites, as compared to those initially rated as having low confidence based on a corrected Pearson's chi-squared test ($\chi^2 = 6.8371$, $p = 0.00893$). About two-thirds of the sites rated as having high confidence that agricultural activity was present actually had agriculture, whereas slightly under a half of those rated as having a low confidence regarding the presence of agricultural activity had agriculture (see Table 2).

From the 188 completed interviews, 108 unique operations were identified with agricultural activity. Personal gardens (34), school gardens (29), and community gardens (20) were the most common types of operations. Urban farms, vacant lot gardens, roof top gardens, aquaponics, hydroponics, and a commercial enterprise were also present (see Table 3).

Based on the 108 operations with agricultural activity, people grew a variety of produce on urban agriculture sites, with fruits and vegetables being most common, but a sizeable proportion raised or kept farm animals and produced animal products (see Table 4). As one would expect, the areas dedicated to agriculture tended to be small, with about 2/3 having an area of less than 1,000 square feet (see Figure 5).

Table 2. Relationship between high/low level of confidence in the presence of agricultural and the presence of agricultural activity.

	Agricultural	Non-agricultural	Totals
High confidence	61	31	92
Low confidence	78	83	161
Totals	139	114	253

Table 3. Types of urban agricultural operations observed in survey.

Operation type	Number	Operation type	Number
Personal gardens	34	Vacant lot gardens	5
School gardens	29	Rooftop gardens	4
Community gardens	20	Aquaponics	1
Demonstration gardens	9	Commercial enterprise	1
Urban farms	8	Other	7

Consider the 108 agricultural sites identified in the sample of 266 from the web-scraped list. Eleven of the 29 school gardens were on the Baltimore list; 18 of the 79 non-school agricultural sites were on the Baltimore list. Thus, the web-scraped list had some school gardens that were not on the Baltimore list, just as the Baltimore list had some school gardens not on the web-scraped list. This is a clear indication of how difficult it is to get a complete list of urban agricultural activity, especially when no produce is sold.

5. Discussion

The ideal list frame includes every unit in the target population and does not include any unit that is not part of the target population. However, most list frames are not perfect. In the development of the NASS list frame, all identified potential farms are included with the full knowledge that some do not satisfy the definition of a farm. Yet, even with this approach, some farms are not on the list frame. For farms as well as numerous other populations, there is a trade-off in developing a list that has a high percentage of units in the target population units and one that has as many target population units as possible at the expense of also including more units that are not in the target population. This trade-off also exists when web scraping is used to develop the list frame.

Of the potential urban agriculture sites in the sample, 52% had agricultural activity, which is higher than the 32.5% in-scope rate Rhodes et al. (2016) reported when using web scraping to create a list of ENDS vape stores in Florida. The substantial online information maintained by the City of Baltimore may be one reason that the success rate was higher for the current study. At the same time, the success in removing non-agricultural sites from the list could have also resulted in removing some agricultural sites. Determining the best balance between maximizing the number of agricultural sites and minimizing the number of non-farm units included in the web-scraped list is important. Furthermore, because a web-scraped list that has a large percentage of the target population will likely have a

Table 4. Types of products produced in urban agriculture sites in the sample.

Types of Products	Number
Vegetables, potatoes, or melons	83
Fruits, nuts, or berries	51
Greenhouse or nursery crops	27
Grains, hay, legumes, or field crops	5
Farm Animals	37
Farm animal products	32
Certified organic products	2

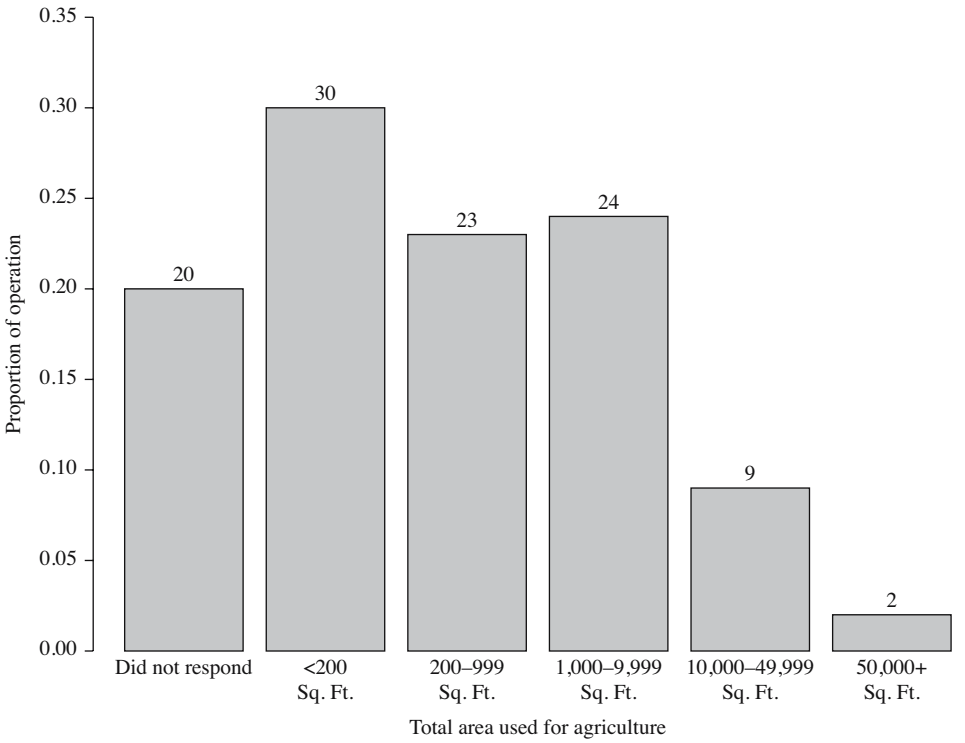

Fig. 5. Based on the Urban Agriculture Pilot Study sample, distribution of land by square footage, both in percentages and counts (on the top of each bar).

number of non-target units as well, it is important to do some type of screening of the list to remove the non-farms before using it as a list frame for a survey.

Any unit in the target population with no online presence is excluded from a web-scraped list frame. The online presence does not need to be through a website or a webpage, such as a Facebook page, developed specifically for that population unit. In this study, it could have been through permits, Facebook or Twitter activity, or any other form that linked the operation to some type of agricultural activity. Nonetheless, some segment of the population is missed because those units do not have an online presence. The proportion missed is likely to depend upon the target population. For this study, agricultural operations that meet the definition of a farm may be more likely to have obtained a permit related to the farming activity, to have joined some type of farm organization, to have a website or a webpage for the operation, or to have had some other farm-related presence on the web than those that do not qualify as a farm, making them more likely to be on the web-scraped list. Determining how to best assess what proportion of the target population units (agricultural operations) do not have a web presence, how that proportion may vary with the type of operation (e.g., urban farm versus backyard garden), and how the characteristics of interest differ between those that have a web presence and those that do not are important steps toward broader use of web scraping to create list frames.

Several lessons were learned from the Baltimore pilot study. Only areas of at least 36 square feet were considered during the pilot study because it was anticipated that

satellite imagery would be used to identify agricultural sites in urban areas. Because satellite imagery did not prove to be useful at the resolution available, the minimum area requirement has been dropped for future work.

Community gardens are challenging. Whereas few, if any, plots qualify as a farm, the total output of the community garden may qualify as a farm, which could result in the perception that the contribution of urban agriculture is under-estimated.

In the future, alternative methods to naïve Bayes should be considered in the web-scraped list development process. One of the referees noted that for his/her data, naïve Bayes usually performs the worst.

When more than one incomplete frame is available, multiple frame methods have been used to provide improved estimates of population parameters (see Hartley 1962; Lohr and Rao 2006; and Lohr 2011). A primary assumption associated with these methods is that the union of all lists provide complete coverage of the population. For hard-to-survey populations, such as urban agriculture, this assumption is unlikely to hold, even approximately. Thus, capture-recapture and other methods that are able to account for population units not included on any list need to be considered. As an example, for the 2012 US Census of Agriculture, capture-recapture was used to adjust for undercoverage, nonresponse, and misclassification (Lohr 2010; and Young et al. 2012). Independent samples were drawn from the Census mail list (a subset of the NASS list frame) and the NASS area frame, and a 12.3% adjustment in the number of farms was made for undercoverage of the Census mail list. Could web scraping either be the foundation for a third sample or replace the area frame as the second sample?

With constant pressure to provide official statistics on emerging and rapidly changing issues in short time frames, it is important to identify new, cost-effective approaches to addressing the questions of interest to policy-makers and other stakeholders. Web scraping, technology, and secondary data sources may be tools that are used increasingly.

6. References

Baltimore Office of Sustainability. 2016. Urban Agriculture. Available at: http://www.baltimoresustainability.org/projects/baltimore-food-policy-initiative/homegrown-baltimore/urban-agriculture-2/ (accessed April 25, 2018).

Baret, F. and S. Buis. 2008. "Estimating Canopy Characteristics from Remote Sensing Observations: Review of Methods and Associated Problems." In *Advances in Land Remote Sensing*, edited by S. Liang, 173–201. Netherlands: Springer.

Chen, M., S. Mao, and Y. Liu. 2014. "Big Data: A Survey." *Mobile Networks and Applications* 19: 171–209.

Colasanti, K., C. Litjens, and M. Hamm. 2010. *Growing Food in the City: The Production Potential of Detroit's Vacant Land*. East Lansing, MI: The C.S. Mott Group for Sustainable Food Systems at Michigan State University.

Congedo, L. 2014. *Semi-Automatic Classification Plugin User Documentation*. Release 5.0.0.1. Technical Report. Available at: https://www.researchgate.net/profile/Luca_Congedo/publication/265031337_Semi-Automatic_Classification_Plugin_User_Manual/links/57cafe2d08ae59825183576d.pdf (accessed May 28, 2017).

Conner, D., A.D. Montri, D.N. Montri, and M.W. Hamm. 2009. "Consumer Demand for Local Produce at Extended Season Farmers' Markets: Guiding Farmer Marketing Strategies." *Renewable Agriculture and Food Systems* 24: 251–259.

Despommier, D. 2010. *The Vertical Farm: Feeding the World in the 21st Century*. New York: Dunne Books/St. Martin's Press.

Dumbacher, B., and C. Capps. 2016. "Big Data Methods for Scraping Government Tax Revenue from the Web." In *2016 Proceedings of the Joint Statistical Meetings, Section on Statistical Learning and Data Science*: 2940–2954.

European Association of Remote Sensing Laboratories. 2017. "Introduction to Remote Sensing." Available at: http://www.seos-project.eu/modules/remotesensing/remotesensing-c01-p05.html (accessed September 30, 2017).

Forster, D., Y. Buehler, and T.W. Kellenberger. 2009. "Mapping Urban and Peri-Urban Agriculture Using High Spatial Resolution Satellite Data." *Journal of Applied Remote Sensing* 3(1): 033523. Doi: https://dx.doi.org/10.1117/1.3122364

Goldsmith, S. 2014. "Milwaukee's Push to Turn Vacant Land into Urban Farms." *Governing the States and Localities*. April 16, 2014. Available at: http://www.governing.com/blogs/bfc/gov-milwaukee-mayor-tom-barrett-home-grown-vacant-lots-urban-agriculture.html (accessed April 25, 2018).

Hartley, H.O. 1962. "Multiple Frame Surveys." *Proceedings of the Social Statistics Section of the American Statistical Association*: 203–206.

Hirschberg, J. and C.D. Manning. 2015. "Advances in Natural Language Processing." *Science* 349: 261–266.

Kausar, M.A., V.S. Dhaka, and S.K. Singh. 2013. "Web Crawler: A Review." *International Journal of Computer Applications (0975-8887)* 63: 31–36. Available at: https://pdfs.semanticscholar.org/7086/cfbc441e1ae956e4600a115b45c8cc84e4a7.pdf (accessed May 29, 2017).

Krijnen, D., R. Bot, and G. Lampropoulos. 2014. "Automated Web Scraping APIs." Online: http://mediatechnology.leiden.edu/images/uploads/docs/wt2014_web_scraping.pdf (accessed April 25, 2018).

Law, M. and A. Collins. 2015. *Getting to Know ArcGIS*, 4th Ed. pp. 768: ESRI Press.

Lohr, S. 2011. "Alternative Survey Sample Designs: Sampling with Multiple Overlapping Frames." *Survey Methodology* 37: 197–213.

Lohr, S.L. 2010. *Sampling: Design and Analysis*, 2nd Ed. Brooks/Cole: Cengage Learning.

Lohr, S., and J.N.K. Rao. 2006. "Estimation in Multiple-Frame Surveys." *Journal of the American Statistical Association* 101: 1019–1030.

Lovell, S.T. 2010. "Multifunctional Urban Agriculture for Sustainable Land Use Planning in the United States." *Sustainability* 2: 2499–2522. Doi: http://dx.doi.org/10.3390/su2082499.

Manjunath, B.S. and W.Y. Ma. 1996. "Texture Features for Browing and Retrieval of Image Data." *IEEE Transactions on Pattern Analysis and Muchine Intelligence* 18: 837–842.

Marin Master Gardeners. 2017. *Community Gardens*. University of California, Division of Agriculture and Natural Resources. Available at: http://ucanr.edu/sites/MarinMG/Great_Gardening_Information/Marin_Community_Gardens/ (accessed May 26, 2017).

Mayer-Schönberger, V., and K. Cukier. 2014. *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Eamon Dolan/Mariner Books.

Mohr, G., M. Stack, I. Ranitovic, D. Avery, and M. Kimpton. 2004. "An Introduction to Heritrix: An Open Source Archival Quality Web Crawler." *4th International Web Archiving Workshop*. Available at: http://iwaw.europarchive.org/04/Mohr.pdf (accessed May 29, 2017).

Neteler, M., M.H. Bowman, M. Landa, and M. Metz. 2012. "GRASS GIS: A Multi-Purpose Open Source GIS." *Environmental Modelling & Software* 31: 124–130.

Policy Link. 2012. *Growing Urban Agriculture: Equitable Strategies and Policies for Improving Access to Healthy Food and Revitalizing Communities*. Available at: http://www.policylink.org/sites/default/files/URBAN_AG_FULLREPORT.PDF (accessed April 24, 2018).

Polidoro, F., R. Grannini, R.L. Conte, S. Mosca, and F. Rossetti. 2015. "Web Scraping Techniques to Collect Data on Consumer Electronics and Airfares for Italian HICP Compilation." *Statistical Journal of the IAOS* 31: 165–176.

Popovitch, T. 2014. "10 American Cities Lead the Way with Urban Agriculture Ordinances." *SeedStock Newsletter*. May 27, 2014. Available at: http://seedstock.com/2014/05/27/10-american-cities-lead-the-way-with-urban-agriculture-ordinances/ (accessed April 24, 2018).

Rhodes, B.B., A.F. Kim, and B.R. Loomis. 2015. "Vaping the Web: Crowdsourcing and Web Scraping for Establishment Survey Farm Generation." *Proceedings of the 2015 Federal Committee on Statistical Methodology Research Conference*. Available at: https://fcsm.sites.usa.gov/files/2016/03/H3_Rhodes_2015FCSM.pdf (accessed October 30, 2016).

Santo, R., A. Palmer, and B. Kim. 2016. *Vacant Lots to Vibrant Plots: A Review of the Benefits and Limitations of Urban Agriculture*. Johns Hopkins Center for a Livable Future. Available at: http://www.jhsph.edu/research/centers-and-institutes/johns-hopkins-center-for-a-livable-future/_pdf/research/clf_reports/urgan-ag-literature-review.pdf (August 19, 2016).

Schowengerdt, R.A. 2007. *Remote Sensing: Models and Methods for Image Processing*, 3rd Ed. Elsevier: San Diego, CA, USA.

Singh, S.M., and K. Hemachandran. 2012. "Content-Based Image Retrieval Using Color Moment and Gabor Based Image Retrieval Using Color Moment and Gabor Texture Feature." *UCSI International Journal of Computer Science Issues* 9: 1694-0814. Available at: http://s3.amazonaws.com/academia.edu.documents/33984493/IJCSI-9-5-1-299-309.pdf?AWSAccessKeyId=AKIAIWOWYYGZ2Y53UL3A&Expires=1495994274&Signature=XS%2F%2FueQk9Kg1xteMNIf%2BfswT4HI%3D&response-content-disposition=inline%3B%20filename%3DIJCSI-9-5-1-299-309.pdf (accessed May 30, 2017).

Specht, K., R. Siebert, I. Hartmann, U.B. Feisinger, M. Sawicka, A. Werner, S. Thomaier, D. Henckel, H. Walk, and A. Dierich. 2014. "Urban Agriculture of the Future: An Overview of Sustainability Aspects of Food Production in and on Buildings." *Agriculture and Human Values* 31: 33–51. Doi: http://dx.doi.org/10.1007/s10460-013-9448-4.

Stoney, W.E. 2008. *ASPRS Guide to Land Imaging Satellites*. Noblis Inc. Available at: http://www.asprs.org/a/news/satellites/ASPRS_DATABASE_021208.pdf (accessed September 30, 2017).

Taylor, J.R., and S.T. Lovell. 2014. "Mapping Public and Private Spaces of Urban Agriculture in Chicago Through the Analysis of High-Resolution Aerial Images in Google Earth." *Landscape and Urban Planning* 108(1) : 57–70.

United States Census Bureau. 2011. Urban Area Criteria for the 2010 Census: Notice. *Federal Register* 76. No 164.

United States Department of Agriculture (USDA). 2016. *Urban Agriculture*. https://newfarmers.usda.gov/ (accessed April 25, 2018).

Vidhya, K.A. and G. Aghila. 2010. "A Survey of Naīve Bayes Machine Learning Approach in Text Document Classification." *International Journal of Computer Science and Information Security* 7: 200–211. Available at: https://pdfs.semanticscholar.org/6861/d02328e18e84fe98b30658100b1c8e7d9891.pdf (accessed May 29, 2017).

Wikipedia. 2017. Google Earth. Online: https://en.wikipedia.org/wiki/Google_Earth (accessed September 30, 2017).

Young, L.J., A.C. Lamas, and D.A. Abreu. 2012. "The 2012 Census of Agriculture: A Capture-Recapture Analysis." *Journal of Agricultural Biological and Environmental Statistics* 22: 523–539. Doi: https://dx.doi.org/10.1007/s13253-017-0303-8 (accessed September 30, 2017).

Zujovic, J., T.N. Pappas, and D.L. Neuhoff. 2009. "Structural Similarity Metrics for Texture Analysis and Retrieval." *Proceedings of the International Conference on Image Processing*. Cairo, Egypt. 2225–2228.

Received October 2016
Revised November 2017
Accepted December 2017