Agricultural Research Using Social Media Data

Samuel C. Zipper*

ABSTRACT
The use of social media in scientific research is rapidly increasing, typically focusing on discrete events of interest to many people and/or spatially mapping a variable of interest. Relatively little research has been done on the utility of social media for monitoring the spatiotemporal patterns of day-to-day life, and none within the agricultural sciences. Here, I discuss the potential applications and limitations of social media data for agricultural research. As an example, I demonstrate the ability of Twitter to map state-level corn and soy planting progress in the conterminous United States. Results compare favorably to traditional survey-based crop progress monitoring, with mean absolute differences of <10% for most state-crop combinations. I also highlight the additional contextual information available from social media data including factors contributing to replanting decision-making and the evolution of farmer sentiment through time. Using analogs from other disciplines, I then discuss key opportunities and challenges for agricultural research using social media. Social media is particularly well-suited for identifying emerging agricultural issues (e.g., weather, crop pests) and guiding extension and outreach directly to affected areas. However, limited data and unknown representativeness of social media users relative to the overall agricultural population are challenges which must be addressed for social media-based agricultural research in the future.

Core Ideas
• Social media data provide quantitative and qualitative data on agricultural practices.
• Twitter data accurately captures timing of crop planting across the United States.
• Maximum positive sentiment in planting tweets aligned with optimal planting period.
• Opportunities for mapping emerging agricultural issues and targeted intervention.
• Challenges include data availability and representativeness of social media users.

Information about agricultural practices is difficult and costly to collect, yet is essential to diverse stakeholder groups including agronomists, farm advisers and extension agents, agricultural economists, and soil/water conservation groups, to name a few (Lehecka, 2014; Gao et al., 2017). Traditionally, collecting agricultural management information has required time- and labor-intensive practices such as farm visits, face-to-face meetings, or surveys conducted by telephone or mail (USDA NASS, 2017). Technological advances such as satellite and aerial remote sensing (e.g., Landsat) and sensor-equipped agricultural equipment are providing a new data source on agricultural management and health (Ozdogan and Gutman, 2008; Zipper and Loheide, 2014; Zipper et al., 2015; Azzari et al., 2017; Yang et al., 2017). However, in the case of sensor-equipped equipment these data remain primarily sparse and/or private, while the ability to translate remotely sensed changes in radiance to on-the-ground agricultural practices remains challenging (Kang et al., 2016). Thus, there is a need to identify complementary data sources to improve understanding of modern agriculture, particularly of the environmental, economic, and cultural context surrounding the biophysical data aggregated by such sensors.

One intriguing possibility for aggregating publicly available, user-volunteered data is the use of social media. Here, I use the term “social media” to refer to content from social network sites (SNS). The SNS are online platforms in which users can (i) create a user profile which can be either public or semi-public; (ii) connect with other users; and (iii) see lists of connections for both themselves and other users (Boyd and Ellison, 2007). There are estimated to be hundreds of SNS (Boyd and Ellison, 2007), with the most popular globally including Facebook, Twitter, weibo, and VK. Examples of social media use for environmental research include species distribution and biodiversity assessments (Stafford et al., 2010; Barve, 2014; Daume, 2016; ElQadi et al., 2017), studying animal behavior (Dylewski et al., 2017), mapping hazards such as wildfire (Wang et al., 2016) or earthquakes (Earle et al., 2012; Crooks et al., 2013); water resource monitoring (Michelsen et al., 2016; Giuliani et al., 2016); and studying public perception of environmental changes such as tree mortality (Fellenor et al., 2017), to name a few.

Among SNS, Twitter is disproportionately represented in the scientific literature, likely due to the ease of accessing and analyzing data; unlike Facebook, for example, the vast majority

*Corresponding author (samuelczipper@gmail.com).

Abbreviations: API, Application Program Interface; SNS, social network sites; MAD, mean absolute difference.

doi:10.2134/agronj2017.08.0495

Supplemental material available online
Available freely online through the author-supported open access option

Copyright © 2018 American Society of Agronomy
5585 Guilford Road, Madison, WI 53711 USA
This is an open access article distributed under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

Published November 30, 2017
CROP ECONOMICS, PRODUCTION & MANAGEMENT

Agronomy Journal • Volume 110, Issue 1 • 2018
of postings on Twitter are publicly available (Tufekci, 2014). The instantaneous nature of Twitter makes it well suited to understanding the propagation of information about events experienced by many, such as flood or earthquake monitoring (Earle et al., 2012; Crooks et al., 2013; Zhou and Xu, 2017) or disease outbreaks (Schmidt, 2012; Broniatowski et al., 2013). However, relatively less analysis has focused on Twitter as a tool for monitoring the spatiotemporal evolution of events (Steiger et al., 2015; Zhou and Xu, 2017), particularly events which do not attract the interest of large swaths of the population (such as day-to-day aspects of agricultural management).

In this study, I explore the potential role of social media in providing both quantitative and qualitative information about agricultural practices, using Twitter as an example. First, I provide a brief overview of the structure and availability of data from Twitter. I then test the ability of Twitter data to provide useful information regarding planting progress for key U.S. commodity crops {corn [Zea mays L.] and soybean [Glycine max (L.) Merr.]} and highlight the additional contextual information gained from the use of social media. Finally, I use analogs from other disciplines to discuss challenges and opportunities for the future use of social media in agricultural research.

OVERVIEW OF TWITTER

The basic element of data on Twitter is referred to as a “tweet”, which is a short (≤280 character) status message posted by a user (Fig. 1). Within the tweet, users can embed images or links; mention other users (@UserName); or include a hashtag (#keyword) which is a keyword or phrase corresponding to a topic of conversation. Twitter is considered social media because users can “follow” other users, meaning that the follower will receive that user’s tweets on their home screen; and interact with either their own or other users’ tweets by “retweeting” (sending the tweet to their followers, with or without added commentary), “replying” (responding to the tweet and initiating a conversation with the initial poster and/or other respondents), or “liking” (signaling an appreciation for the content). Each user may also include metadata in their profile, including information such as a real name or pseudonym, geographic location, birthday, and/or description of themselves (Fig. 1). Users also have the option of embedding geographic information (latitude/longitude) within a tweet, though <0.5% of tweets contain this information (Cheng et al., 2010). While this description is specific to Twitter, many SNS share some or all of these characteristics.

Scientific research using Twitter typically uses the Twitter Application Program Interface (API), a set of tools and routines which allow access to Twitter’s raw data via a developer portal. A wide variety of functions are available within this API, including information-gathering functions such as searching for and downloading tweets and metadata for a specific topic or user; and interactive functions, such as following users or retweeting. Search capabilities are limited temporally to tweets occurring within the past week (approximately), and therefore longitudinal studies must also have a separate system for storing search results (Daume, 2016). A detailed overview and tutorials on the Twitter API can be found at https://dev.twitter.com/overview/api.

Fig. 1. Screenshot of (a) user profile and (b) tweet with key sources of information annotated with green numbers: (1) Display name; can be real or pseudonym. (2) User name. (3) User profile description including (4) mentions of other users. (5) Location provided by user. (6) User website. (7) User join date. (8) Tweet text, including (9) hashtags. (10) link to external website, and (11) attached image. (12) time and date of tweet. Number of (13) replies to tweet, (14) retweets by other users, and (15) likes by other users. All of this information and more is accessible via Twitter Application Program Interface (API). Figure content and design modeled after Darling et al. (2013). Image in tweet from Zipper et al. (2017).
MONITORING PLANTING PROGRESS USING SOCIAL MEDIA DATA

To demonstrate the capabilities of SNS data for research-
ing agricultural management, I asked the questions: (1) Can Twitter data be used to quantitatively estimate weekly planting progress (the proportion of the crop planted) for corn and soy in the United States?; and, (2) What additional contextual information can be obtained from tweets related to planting?

Methodology

To estimate planting dates, data went through three discrete steps: gathering, filtering, and model-fitting. A generic overview of the process of mining social media for data is provided by Daume (2016).

Gathering Twitter Data

To gather data, I used a temporal analysis of semantically driven data, in which data is collected using specific keyword searches and analyzed for patterns through time (Brooker et al., 2016). First, I defined a Boolean search string which captured a broad swath of tweets relevant to agricultural plant-
ging: “[(corn OR soy OR wheat) AND (plant OR planting OR planted OR plants OR #plant17 OR #plant2017)] OR #corn17 OR #corn2017 OR #soy17 OR #soy2017 OR #wheat17 OR #wheat2017”. The hashtags used (#plant17, etc.) were identified based on exploratory analysis of the tweet content of agricultural Twitter users. This search string is deliberately broad (a tweet containing the text “Corn is a plant” would be returned) to be as inclusive as possible, so that the maximum number of tweets could be archived and filtered for relevance later.

Given that this study aims to track crop progress through time, I designed a system (referred to as AgroStream) which automatically searches, filters, aggregates metadata, and stores tweets. AgroStream uses the “twitteR” package for R (Gentry, 2015; R Core Team, 2017) to search the Twitter API daily for the string defined above. Search results are then filtered to eliminate users with a reported location outside of the conter-
minous United States, as well as users with overly broad location descriptors (e.g., “USA”).

The remaining tweets are then stored in a SQLite database. AgroStream is set to run daily on a standard desktop computer to collect the previous day’s tweets using Windows Task Scheduler. This system was initiated on 26 Feb. 2017 and data collection is ongoing; the period of interest for this study was defined as a 20-wk period containing planting dates typical of corn and soybeans for the Midwest (27 Feb.–16 July 2017).

Filtering Data

As the string used to gather tweets was deliberately broad, I filtered the data to eliminate tweets which were irrelevant and/ or did not include sufficient geospatial information for analysis. First, tweets were extracted for the 20-wk period of interest (n = 11,514). Second, tweets were filtered for relevance. Given that the focus of analysis was planting, only tweets containing one of the following widely used hashtags were retained: “#plant17”, “#corn”, or “#soy”. If the tweet contained only “#plant17”, it had to include either the word “corn” or “soy”. Matching at the subword level was allowed, so that using “soy” as a search term returned results including longer strings such as “soybean” or “soybeans”. These filtering criteria were identi-

Estimating Planting Dates

I estimated planting dates for each state and crop meet-
ing the criteria defined above. Only state-crop combinations meeting the following criteria were analyzed: (a) NASS crop progress data are available for comparison; (b) average of >1 tweet/week over the period of analysis; and (c) no single user contributed >25% of the tweets. Criterion (c) was necessary, as in some states with high tweet totals a single user was responsi-
ble for an inordinate proportion of the tweets. For example, in Kansas one user was responsible for 211 of 280 (75%) of tweets mentioning corn, and 67 of 74 (91%) of tweets mentioning soy. Criterion (c) was also exceeded for other states with high tweet totals, including Nebraska and Ohio for both corn and soy. The accounts contributing these disproportionately high number of tweets tended to be in agriculture-affiliated industries, such as commodity brokers or seed companies, and therefore do provide useful contextual information; however, they were eliminated from analysis to avoid having a single user influence state-level comparison with NASS. Texas met all of the above criteria but was also eliminated from analysis because a signifi-
cant portion of planting occurred before AgroStream began collecting data in late February (USDA NASS, 2017). In the end, I analyzed 10 state-crop combinations: corn in Colorado, Iowa, Illinois, Indiana, Michigan, Missouri, North Dakota, and South Dakota; and soy in Iowa and Illinois.

For each state-crop combination, tweets were aggregated to weekly totals to correspond with USDA crop progress reports (Fig. 2a), and fit with a nonparametric cubic smoothing spline (Hastie and Tibshirani, 1990) (Fig. 2b). I then defined the start and end of the planting period as the maximum and minimum of the second derivative of the fitted function, respectively, with an adjustable buffer period at both the start and end (Fig. 2c) (Zipper et al., 2016). All state-crop combinations had a positive maximum second derivative and negative minimum second derivative, indicating that curves had both convex and concave second derivatives.
I then calculated weekly crop progress as the proportion of the total tweets over the course of the planting period to have occurred by the end of each week (Fig. 2d–2e). Dates before and after the start/end of the planting period were estimated as 0 and 100% progress, respectively.

The buffer period was intended to capture the baseline periods of low activity before and after the peak period of planting, and was used as a model calibration parameter. As the start of the planting period tends to exhibit a more abrupt change in week-to-week tweet volumes than the end of the planting period (Fig. 2a), the primary function of the buffer period is to define the total number of tweets which are used to convert from cumulative weekly tweets to percent progress. The 10 state-crop combinations were first split into calibration ($n = 5$) and validation ($n = 5$) samples. I then varied the start and end buffer period from 0 to 5 wk and selected the combination with the minimum mean absolute difference (MAD) between Twitter and NASS estimates (see Model Verification) within the calibration sample, and used the validation data to assess model fit. I repeated the calibration procedure for all possible 50/50 splits of the state-crops ($n = 252$) and selected the buffer period with the lowest average MAD for all calibration samples.

**Model Verification**

I developed the model and verified its performance via comparison to state-level weekly crop progress reports from the U.S. Department of Agriculture National Agricultural Statistics Service (USDA NASS, 2017). Crop progress reports are based on extensive surveys of local observers, and have been used in other studies examining changes in agricultural management (Kucharik, 2006, 2008; Sacks and Kucharik, 2011; Gao et al., 2017). More information regarding these data and survey methodology is available at the USDA NASS National Crop Progress website (https://www.nass.usda.gov/Publications/National_Crop_Progress/).

**Contextual Information**

I also explored the potential of using the tweets to provide spatiotemporal contextual information regarding agricultural management. To examine common topics and themes, I created a wordcloud using the text from all the tweets retained after filtering after removing common words (“the”, “and”, etc.). Wordclouds are a type of graphic that takes in text and displays commonly used words in sizes proportional to their usage (many options exist for making wordclouds; I used wordclouds.com). Words contained in the tweets and the wordclouds provided background information about what types of information could potentially be extracted from the tweets.

To demonstrate the potential contextual information present in the tweets, I then extracted all tweets including “replant” or derivative words (replanting, replanted, etc.). These tweets were used to calculate both the timing and magnitude of replanting (expressed as a proportion of the total planting-related tweets) and different decision-making factors related to replanting. I manually identified four categories to which replanting tweets could belong: expressions of uncertainty or decision-making considerations; guidance and advice from expert sources; information related to the causes of replanting; and indicators of crop progress and/or performance. Tweets could belong to multiple or zero categories.

I then examined the utility of Twitter data for quantitatively assessing farmer sentiment using the AFINN-111 dataset (Nielsen, 2011). This widely used dataset contains 2476 words with an integer sentiment score ranging from −5 (negative sentiment) to +5 (positive sentiment). To evaluate changes in farmer sentiment through time, I calculated the daily average sentiment score of all tweets, ignoring words not present in the AFINN dataset.
RESULTS AND DISCUSSION

Data Collection

Of the >11,000 tweets aggregated by AgroStream, 3307 were identified as relevant to the research question posed here with sufficient geolocation data (see Filtering Data). Of these, 2628 mentioned corn and 679 mentioned soy. The retained tweets represent 45 of the 48 states in the conterminous United States and the District of Columbia, with no tweets retained from Utah, West Virginia, or Rhode Island. Tweets were unevenly distributed geographically and varied over 2.5 orders of magnitude across states, with the highest tweet counts in U.S. Corn Belt states such as Illinois (n = 479), Nebraska (n = 463), and Iowa (n = 410) (Fig. 3a). I also identified several temporal patterns in tweet frequency. At a daily resolution, there is strong within-week variability, with lower tweet counts on Saturdays (Fig. 3b). At a weekly resolution, tweets are approximately normally distributed around Week 18, which ends 7 May (Fig. 3c).

From a sub-state perspective, I was able to geolocate 400 tweets to the county level for the state of Illinois. These tweets were unevenly distributed among Illinois’ 102 counties (Supplemental Fig. S1). At least one tweet was recorded from 50 counties, though only 11 of these counties had greater than five tweets. The highest concentrations of tweets were in Cook County (n = 100) and McLean County (n = 90), totaling a combined 47.5% of all Illinois tweets. These counties are the home of Illinois’ largest population (the city of Chicago) and Illinois State University, respectively.

Tracking Planting Dates

Twitter data accurately captured planting progress across most of the state-crop combinations studied, with MAD between Twitter and NASS estimates less than 10% in 8 of 10 states (Fig. 4 and 5). The average MAD across our 252 validation samples was 8.20%, with a minimum of 3.52% and a maximum of 12.47%. The calibrated buffer periods were 0 wk at the start of the planting period and 4 wk at the end of the buffer period; this indicates that the second-derivative technique used to identify the start of the planting period accurately captures the onset of planting, but that identifying the end of the planting period is more challenging. This may be due to the slight positive skew to tweet counts through time (e.g., Fig. 2a), in which weekly totals change more gradually later in the growing season. The underlying reason for this distribution in time may be due to more concentrated planting activity at the beginning of the planting period, or a loss of interest in social media and corresponding decrease in tweets as the planting period progresses.

Twitter data tended to estimate slightly higher planting progress early in the growing season and slightly lower planting progress later in the growing season, with an additional 2 to 3 wk of Twitter-predicted planting period in most states (Fig. 4); however, note that NASS data do not start at 0% or go all the way to 100% in most state-crop combinations. The largest observed differences between Twitter and NASS were for corn in Indiana and soy in Illinois, where Twitter predicted...
that given the current agricultural Twitter activity, this is possible for corn in agriculturally dominated states, but not at the county level. However, one difference between our approach and NASS survey-based estimates is the temporal availability. Calculating the proportion of planting completed using cumulative tweets inherently requires the planting season to be complete to define the total tweets during the planting period, while NASS estimates are provided weekly throughout the growing season. Other approaches, such as the combination of short-term and long-term averaging used by Earle et al. (2012), may be better suited to identifying discrete events such as the onset of planting. Alternately, using multiple years’ worth of data to define total tweet estimates in advance may provide an opportunity to estimate planting progress in real time as data availability increases in the future.

Given these limitations, I do not intend to suggest that Twitter can be used as a replacement for traditional monitoring at this time. Rather, as noted by other studies, it provides additional information which may be valuable for better contextualizing existing, expert-collected data sources (Crooks et al., 2013; Daume et al., 2014; ElQadi et al., 2017). Future methodological improvements may improve the utility of these data to monitor planting progress in real time.

**Contextual Information**

In addition to tweet frequency, which was used to estimate planting dates, the content of the tweets contains contextual information about several themes (Fig. 6a). There was frequent repetition of words related to weather (rain, warm, wet, dry, frost, hail, freeze); crop condition and pests (stress, tall, short, borer, damage, struggling); soil or water management (soil, saline, clay, irrigation); crop choice and cover crops (alfalfa

---

Fig. 5. Comparison of estimated crop progress between Twitter and NASS for all states. Black line is linear best-fit to data and gray line is a 1:1 relationship.

Fig. 6. (a) Wordcloud showing frequently used words in tweets related to agricultural management. Word size corresponds to frequency of usage; word position is random. Not all words shown in cloud. Created at wordclouds.com. (b) Example of contextual information available regarding replanting decisions; plot shows proportion of weekly tweets mentioning replanting, and italicized text shows representative tweets regarding extension/outreach activities, factors contributing to replanting decision-making, and local heterogeneity in planting progress.
[Medicago sativa L.], rice [Oryza sativa L.], oat [Avena sativa L.], millet [Panicum miliaceum L.], cotton [Gossypium hirsutum L.], sorghum [Sorghum bicolor (L.) Moench]; management decisions (fertilizer, N, NH3, no-till, replanting); timing (slow, soon, late, late); mood or outlook (good, hope, hard, bad, poor, success); and economic factors (insurance, futures, markets, prices, ethanol). This variety of themes suggests that Twitter data is able to provide diverse contextual information regarding a wide spectrum of agricultural management practices.

To demonstrate the potential of contextual information present in social media data, I focused on tweets related to replanting. Replanting is a difficult and economically important decision for farmers, as it requires additional investment in seed costs and field labor, but may also yield a profitable crop (Lauer, 1997). Therefore, understanding the extent, causes, and decision-making process related to replanting decisions provides important insight into agricultural management (Benson, 1990). In the tweets analyzed here, tweets containing the word “replant” (including derivatives such as #replant, replanting, etc.) increase in frequency throughout the growing season to a maximum of ~8% of total tweets during Week 23 (Fig. 6b). Illustrative tweets highlighted on Fig. 6b demonstrate the contextual information present in these tweets. Contextual information observed includes farmer expressions of uncertainty and decision-making considerations (9.4% of replanting tweets); guidance and advice from expert sources (18.8%); information related to the causes of replanting, primarily excess moisture (29.7%); and indicators of crop progress and performance (42.2%).

At a broader scale, Twitter data can also provide contextual information regarding farmer sentiment. There is a clear evolution of sentiment related to planting through time (Fig. 7), with mean sentiment scores peaking early in the planting season (~DOY 100 = 10 April) at approximately +1 and then declining toward the end of the growing season, with frequent negative sentiment scores occurring after ~DOY 150 = 30 May).

This seasonal pattern mirrors the response of yield to planting date, which peaks in April or early May, depending on location (Lauer, 1996; Klein, 2009; Nafziger, 2017). For example, in Illinois, peak yield for corn corresponds to planting during early to mid-April with a decline in yield for planting occurring in May (Nafziger, 2017). Thus, this sentiment data may reflect an awareness of the timing of planting relative to the optimum for a location, or a decrease in optimism as the season progresses.

However, the relatively small number of tweets retained in our study limits our ability to draw broader conclusions or identify spatiotemporal patterns using these data. Using “irrigate” as an example, this word (including other forms) was mentioned only 17 times (0.5% of tweets) by 10 different users—far too few to draw any conclusions about irrigation extent or management. Similarly, while sentiment data tracks temporal patterns when aggregated over all tweets retained, there is insufficient data to draw strong conclusions about the spatiotemporal evolution of farmer sentiment.

**SOCIAL MEDIA POTENTIAL FOR AGRICULTURAL RESEARCH**

The example above highlights both the potential utility and limitations of social media data for agricultural research. Here, I discuss several opportunities and associated challenges of using social media data for agricultural research based on disciplines in which the use of social media data is more mature.

**Opportunities**

The results from our case study demonstrate that social media data is a powerful tool for monitoring crop progress over large spatial and temporal domains, despite the relatively limited number of tweets available. While the crop progress analysis in this study employs a post facto approach, social media also provides an opportunity for monitoring agricultural issues in real time. Social media is particularly well suited for “horizon scanning” in which activity is used to identify emerging issues (Amanatidou et al., 2012). This makes social media a potentially useful tool for tracking emerging agricultural issues (e.g., drought, disease) at both local and national scales, much as it can be used to track disease outbreaks in the human population (Signorini et al., 2011; Schmidt, 2012; Bernardo et al., 2013; Broniatowski et al., 2013).

Integration of social media’s horizon-scanning capabilities with analysis of social media user data also provides an opportunity for targeted extension and outreach, as has been demonstrated in the public health sector. Monitoring social media for discussion of an emerging problem (e.g., a particular crop disease) may provide an indicator of the spatial areas most affected by that problem, allowing experts to then target users in affected areas with information and outreach via both social media and/or more traditional practices. This process is analogous, for example, to monitoring spatiotemporal dynamics of HIV infection (Young et al., 2014; Nielsen et al., 2017) and the use of social media to spread health information to affected populations (Bull et al., 2012; Pedrana et al., 2013; Fung et al., 2014).
In addition to real-time monitoring, social media data aggregated over long periods also has the potential to track shifting beliefs over time. In regions where there has been investment in promoting a particularly agricultural practice, social media offers a tool to assess the uptake of said practice; for example, Nielsen et al. (2017) note the promise of social media for monitoring uptake of HIV-prevention measures. This provides a tool to evaluate and refine outreach and extension campaigns to enhance their effectiveness at reaching target populations, such as farmers in a specific region.

While our study focuses on Twitter data, social network sites other than Twitter can provide potentially valuable information. The use of repeat photographic imagery in evaluating phenology of natural ecosystems (Sonnentag et al., 2012; Nijland et al., 2013; Klosterman et al., 2014; Keenan et al., 2014), and the photo-sharing website Flickr has been used for biodiversity monitoring (Stafford et al., 2010; Barve, 2014; ElQadi et al., 2017). Given these capabilities, geolocated images including crops from sources such as Flickr, Facebook, Twitter, or Instagram may be a useful tool to monitoring crop condition in real time over large spatial areas, for example by using photograph-derived vegetation indices to identify areas with crop stress.

**Challenges**

Challenges for the use of social media data for agricultural research can be split broadly into two categories: those inherent to research using social media data, and those specific to agricultural applications. For social media-based scientific research in general, semantically driven analyses such as the planting example explored here are impacted by the selection of keywords and hashtags used for searching (Tufekci, 2014; Zhou and Xu, 2017). Given that our analysis is based on the use of specific hashtags, our results will necessarily be skewed to users aware of these hashtags and participating in these discussions online, and will therefore exclude potentially relevant tweets from analysis. Text-based analyses may also have difficulty parsing typed language for subtleties of human intention such as sarcasm (Tufekci, 2014).

The representativeness of social media users relative to the population as a whole is a challenge for social media-based research in general (Ruths and Pfeffer, 2014). Work on the broader population has found that social media users tend to be younger, more educated, more politically attentive, and more liberal than the general population (Bogdanou et al., 2013; Mellon and Prosser, 2016), while users embedding geographic information in their tweets are more likely to have higher incomes, be younger, and live in coastal or urban settings (Malik et al., 2015). However, no studies have looked at social media use among agricultural users to evaluate whether these biases (or others) exist and, if so, how they may relate to agricultural practices—for example, people receptive to emerging social technologies (e.g., social media) may also be more receptive to the use of emerging agricultural technologies (e.g., variable-rate nutrient application systems). Given that younger generations are disproportionate users of social media relative to society at large (Mellon and Prosser, 2016), this challenge may also provide an opportunity to learn about the practices of the next generation of farmers.

Finally, the quantity of useful data is a key challenge for social media use in agricultural research. The public Twitter API limits the number of tweets provided to free users (Daume et al., 2014), though paid services exist which provide streaming access to 100% of tweets. Our retained sample of 3307 tweets is on the same order of magnitude as other studies in the environmental sciences (Barve, 2014; Daume, 2016); however, these studies do not attempt to track both spatial and temporal patterns simultaneously. Our sample size is far lower than event-based studies in which social media data is more commonly used, for example earthquake monitoring where samples of dozens to hundreds of tweets per minute are common (Earle et al., 2012; Crooks et al., 2013). Hope for improvement may lie in better techniques for geolocation and filtering relevant tweets from the larger database; research in the public health sector has shown that machine learning approaches may be better able to identify and select relevant tweets (Mytilinou et al., 2013).

Given that the majority of the world’s population is urban (United Nations, 2015) and internet access can be limited in rural areas (particularly in developing nations), this necessarily limits the degree to which agricultural information will be conveyed over social media. Therefore, the data availability problem may be particularly acute for agricultural research relative to other disciplines. However, social media usage is increasing rapidly worldwide—the number of global social media users has doubled over the past 6 yr—and is expected to exceed three billion users by 2021 (Statista, 2017). As overall social media usage continues to increase worldwide, it is likely that a corresponding increase in agriculturally relevant social media use as well.

**CONCLUSIONS**

In this study, I discuss the potential role of social media in agricultural research, using crop progress as an illustrative example. I find that Twitter is a useful data source for monitoring spatiotemporal dynamics of planting progress for major agricultural crops at the U.S. state level, with comparable results to traditional survey-based crop progress reports across a sample of 10 states and crops. Data from social media also provide additional contextual information, for example on the causes and status of replanting decisions and the evolution of farmer sentiment throughout the planting window. However, given the relatively limited number of agricultural tweets, contextual data remains too sparse for detailed monitoring and generalization of agricultural practices at fine spatial and temporal scales.

More broadly, social media have great potential for future agricultural research. In particular, social media are well suited to identify emerging agricultural issues, providing targeted extension and outreach, and mapping the extent of different agricultural practices. However, several key challenges remain, particularly in terms of limited data availability and determining the representativeness of agricultural social media users relative to the agricultural population as a whole. As social media usage continues to spread globally, there will likely be an increase in the utility of social media data for agricultural research as well.

**ACKNOWLEDGMENTS**

This manuscript has been improved substantially by the comments of Dr. Roger Lawes and three anonymous reviewers. Thanks to Kelly Helm Smith for discussion and feedback on a previous draft of this manuscript; and to Kaitlin Stack Whitney and Carly Ziter for convincing me that social media is not (always) a waste of time. All code and data used in this paper are available online under a GNU General Public License at https://github.com/szipper/AgroStream.
SUPPLEMENTAL MATERIAL
Fig. S1. Map of tweet frequency by county for the state of Illinois.

REFERENCES
www.sciencedirect.com/science/article/pii/S0034425717301645
(accepted 17 May 2017).
24:194–199. doi:10.1016/j.ecoinf.2014.08.008
Bernardo, T.M., A. Rajic, I. Young, K. Robiadek, M.T. Pham, and J.A.
Funk. 2013. Scoping review on search queries and social media for
disease surveillance: A chronology of innovation. J. Med. Internet
Bogdanou, T., C.B. Starr, A. Weatherall, and A.D. Leslie. 2013. Use
of the internet and social media in the forestry profession in the
230. doi:10.1111/j.1083-6101.2007.00393.x
Influenza Surveillance through Twitter: An analysis of the 2012–
2013 influenza epidemic. PLoS One 8(12):e83672. doi:10.1371/jour-
nal.pone.0083672
Brooker, P., J. Barnett, and T. Cribbin. 2016. Doing social media analyt-
Cheng, Z., J. Caverlee, and K. Lee. 2010. You are where you tweet: A
text-based approach to geo-locating Twitter users. p. 759–768.
In: Proceedings of the 19th ACM International Conference on
Information and Knowledge Management. CIKM ’10. Assoc. for
Computing Machinery, New York. doi:10.1145/1871437.1871535
Crooks, A., A. Croitoru, A. Stefanidis, and J. Radzikowski. 2013. #Earthquake: Twitter as a distributed sensor system. Trans. GIS
Darling, E., D. Shiffman, I. Côté, and J. Drew. 2013. The role of Twitter
in the life cycle of a scientific publication. Ideas in Ecology and
24:194–199. doi:10.1016/j.ecoinf.2014.08.008
54(6). http://www.annalsofgeophysics.eu/index.php/annals/arte-
2017. Mapping species distributions with social media geo-tagged
images: Case studies of bees and flowering plants in Australia. Ecol.
Fellenor, J., J. Barnett, C. Potter, J. Urquhart, J.D. Mumford, and C.P.
Quine. 2017. The social amplification of risk on Twitter: The case
of ash dieback disease in the United Kingdom. J. Risk Res. 0(0):1–
S0140-6736(14)62418-1
2017. Toward mapping crop progress at field scales through fusion
188:9–25. doi:10.1016/j.rse.2016.11.004
Gentry, J. 2015. twitterR: R Based Twitter Client.
crowdsourced web content for informing water systems operations
ed. Chapman and Hall/CRC, Boca Raton, FL.
Kahle, D., and H. Wickham. 2013. ggmap: Spatial visualization with
sensed vegetation indices and crop leaf area index? A global assess-
ment. Remote Senses. 8(7):597. doi:10.3390/rs8075979
Kecen, T.F., B. Darby, E. Felts, O. Sonnentag, M.A. Friedl, K. Hufkens et al. 2014. Tracking forest phenology and seasonal physiology
24(6):1478–1489. doi:10.1890/13-0652.1
Klein, B. 2009. How corn planting date can affect yield. Univ. of
Nebraska-Lincoln Ext. http://cropwatch.unl.edu/how-corn-plant-
Lavine et al. 2014. Evaluating remote sensing of deciduous forest
phenology at multiple spatial scales using PhenoCam imagery. Bio-
Kucharik, C.J. 2008. Contribution of planting date trends to increased
maize yields in the central United States. Agron. J. 100(2):328–
336. doi:10.2134/agronj2007.0145
Kucharik, C.J. 2008. Contribution of planting date trends to increased
maize yields in the central United States. Agron. J. 100(2):328–
336. doi:10.2134/agronj2007.0145
of Wisconsin-Ext., Madison, WI.
Lehecka, G.V. 2014. The value of USDA crop progress and condition
information: Reactions of corn and soybean futures markets. J.
in geotagged Tweets. In: Ninth International AAAI Conference
on Web and Social Media. AAAI Press, Palo Alto, CA.
Mellon, J., and C. Prosser. 2016. Twitter and Facebook are not representa-
tive of the general population: Political attitudes and demographics
2016. YouTube as a crowd-generated water level archive. Sci. Total
Environ. 568:189–195. doi:10.1016/j.scitotenv.2016.05.211


